申报	系列：教师系列教学科研并重型
	专业：计算机科学与技术
	职称：副教授

业绩成果材料

（申报人的业绩成果材料包括论文、科研项目、获奖以及其他成果等）

单　位（二级单位）数学与信息学院

姓　名　黄立峰

材料核对人：　　单位盖章：

核对时间：

华南农业大学制

目　录

一、教学研究业绩

二、科研项目

三、论文、著作等

四、科研成果

五、其他业绩

广 州 市 教 育 局

广州市教育局关于公示市教育局协同育人项目拟立项名单的通知

各有关高校、省市属普通高中，各区教育局：

为推进“新时代高等教育优化提升工程”（市教育事业发展“十四五”规划重点工程之一），联合广州地区高校与普通高中共建协同育人平台，我局组织开展了局协同育人项目申报、评审工作。

经市教育评估和教师继续教育指导中心组织专家评审及我局审定，共拟立项27项，其中重点项目9项、一般项目16项、专业指导项目2项。现对拟立项名单（见附件1）予以公示，并就有关事项通知如下。

一、拟资助经费

（一）根据市财政安排，项目经费额度下调。重点项目拟资助经费21万元，一般项目拟资助经费10万元，专业指导项目拟资助经费7万元，项目建设周期均为3年。

（二）省部属高校、民办高校项目经费于2024年一次性下达。

（三）广州大学、广州医科大学项目经费纳入学校部门预算，资助经费额度最终以年度财政批复为准。香港科技大学（广州）项目经费在学校财政补助经费中自筹。

二、公示期限

公示期为自本通知发布之日起5个工作日。公示期内，凡对拟立项项目持有异议的，可向市教育局（越秀区西湖路83号4楼高教处，邮编：510030）提出署名书面意见。来信来电必须提供真实姓名、工作单位和联系方式，以单位名义反映的需加盖单位公章，否则不予受理。

因项目资助经费额度下调提出放弃立项的，请向我局提出申请。需要对项目立项信息进行变更的，请填写广州市教育局协同育人项目拟立项项目变更申请表（见附件2）。

三、任务书签订工作

公示期满后，拟立项的项目需由项目承担单位与市教育局签订项目任务书（见附件3）并明确项目预期成果后，方可正式立项。

请各高校于2024年10月15日前将纸质版任务书（一式三份）送到市教育局高教处。

附件：1.广州市教育局协同育人项目拟立项名单

2.广州市教育局协同育人项目拟立项项目变更申请表

3-1.广州市教育局协同育人项目任务书

3-2.广州市教育局专业指导项目任务书

广州市教育局

2024年9月27日

（联系人：胡洋，联系电话：22083707）

附件1

广州市教育局协同育人重点项目拟立项名单

序号	项目类别	项目名称	项目承担高校	高校项目负责人	项目承担高中	所属区域	高中项目负责人	拟资助总经费（万元）	实际资助经费（万元）
1	协同育人重点项目	以学科竞赛为牵引的拔尖人才贯通培养体系建设——以依托“天河二号”的信息学特长生培养为例	中山大学	万海	广州大学附属中学	市属	欧卫国	21	/
2	协同育人重点项目	大中衔接基础学科创新人才贯通培养机制探索	中山大学	王猛	广州市执信中学	市属	林间开	21	/
3	协同育人重点项目	探索拔尖人才培养机制	华南理工大学	陈浩文	广东实验中学	省属	阳珂	21	/
4	协同育人重点项目	高校与高中贯通的拔尖创新人才培养机制探索	暨南大学	卢惠辉	广州市培正中学	越秀区	张志红	21	/
5	协同育人重点项目	高校-中学协同培养拔尖创新人才项目	华南师范大学	阳成伟	华南师范大学附属中学	省属	黄爱国	21	/
					广东广雅中学	市属	苏科庚		
					广州市执信中学	市属	林间开		
6	协同育人重点项目	高校与高中拔尖创新人才培养协同育人机制的探索与实践	广东工业大学	胡钦太	广东番禺中学	番禺区	胡展航	21	/
					广州大学附属中学	市属	姚海霞		
					广州市第十六中学	越秀区	王勇		
					广东广雅中学	市属	苏庚科		

7	协同育人重点项目	未来生物科学家培育项目	广东药科大学	吴凤麟	广州市第五中学	海珠区	汪爱华	21	/
					北京师范大学广州实验学校	黄埔区	梁肇栋		
8	协同育人重点项目	高中生网安意识及创新能力培养机制探索	广州大学	鲁辉	广州大学附属中学	市属	王晓鹏	21	以年度财政批复为准
9	协同育人重点项目	科技赋能优才计划	香港科技大学（广州）	伍楷舜	广州市第二中学	市属	何院琴	21	以年度财政批复为准
					广州市执信中学	市属	陈民		
					广州大学附属中学	市属	欧卫国		
					广州外国语学校	市属	李伟		
					广东广雅中学	市属	苏庚科		
					广州协和学校	市属	周寅博		

广州市教育局协同育人一般项目拟立项名单

序号	项目类别	项目名称	项目承担高校	高校项目负责人	项目承担高中	所属区域	高中项目负责人	拟资助总经费（万元）	实际资助经费（万元）
1	协同育人一般项目	提升高中生科学素养	华南农业大学	余明	广州市第三中学	越秀区	胡峥嵘	10	/
2	协同育人一般项目	基于气象卫星及数据分析技术的协同创新型人才培育模式探索与实践	华南农业大学	黄立峰	广州市育才中学	越秀区	张先锋	10	/
3	协同育人一般项目	学生人文艺术科学素养协同培养项目	华南师范大学	张学波	华南师范大学附属南沙中学	南沙区	林天伦	10	/
					广州中学	天河区	李强		
					广州市第三中学	越秀区	胡峥嵘		
					广州市第十七中学	越秀区	莫明珉		
4	协同育人一般项目	数字技术与非遗艺术教学融合创新	广东工业大学	张丽平	广州市花都区圆玄中学	花都区	杨美英	10	/
5	协同育人一般项目	基于科创模式迁移与资源共享的高中生科学素养协同培养机制探索与实践	广东财经大学	陈建超	广州市海珠外国语实验中学	海珠区	陈智	10	/
6	协同育人一般项目	医药科学素养提升项目	广东药科大学	廖丽贞	广东实验中学越秀学校	越秀区	罗伟雄	10	/
7	协同育人一般项目	高校与普通高中协同育人下的艺术素养贯通培养平台构建	星海音乐学院	卢格霖	广州大学附属中学	市属	张蕾	10	/
8	协同育人一般项目	基于多学科融合的高中生科学素养提升教育的创新与实践	广东技术师范大学	杜灿谊	广州市第六十五中学	白云区	罗锐杰	10	/
					广州市花都区新华中学	花都区	陈彩红		

9	协同育人一般项目	“文艺领航 筑梦未来”UGS协同育人平台	广东第二师范学院	钟香炜	广州市花都区邝维煜纪念中学	花都区	高秀丽	10	/
					广州市花都区邝维煜纪念中学附属雅正学校	花都区	黄家和		
					广州市花都区第一中学	花都区	田长鹏		
					广州市花都区第二中学	花都区	吕律彬		
10	协同育人一般项目	科技筑梦，立德树人—广州医科大学呼吸疾病全国重点实验室联合广州市第三中学科技研学项目	广州医科大学	关伟杰	广州市第三中学	越秀区	胡峥嵘	10	以年度财政批复为准
11	协同育人一般项目	科技赋能育才计划	香港科技大学（广州）	毛进宇	广州市执信中学	市属	陈民	10	以年度财政批复为准
					广州大学附属中学	市属	欧卫国		
					广州市第二中学	市属	何院琴		
					广州外国语学校	市属	李伟		
					广东广雅中学	市属	苏庚科		
					广州协和学校	市属	周寅博		
12	协同育人一般项目	双高联动·美育未来——艺科融合赋能普通高中美育发展实践	广州商学院	赖伟成	广州市增城区第一中学	增城区	张 薇	10	/
					广州市增城区派潭中学	增城区	张健文		
					广州市增城区永和中学	增城区	赵伟坤		
13	协同育人一般项目	空天科技人才培养特色研修班	广州理工学院	王瑛	广州市育才中学	越秀区	杨世鹏	10	/

14	协同育人一般项目	广城理与秀中开展机器人与人工智能协同育人	广州城市理工学院	阮安正	广州市花都区秀全中学	花都区	卢桂湟	10	/
15	协同育人一般项目	“双高联动”育英苗——广州南方学院-广州市第一中学、广州南方学院附属番禺中学创新及职业能力培养项目	广州南方学院	王玉平	广州市第一中学	荔湾区	郑妍	10	/
					广州南方学院番禺附属中学	番禺区	阳小平		
16	协同育人一般项目	华南农业大学珠江学院与广州市从化区第四中学协同育人培养路径研究	华南农业大学珠江学院	陈军	广州市从化区第四中学	从化区	谭克斌	10	/

广州市教育局协同育人专业指导项目拟立项名单

序号	项目类别	项目名称	项目承担高校	项目负责人	拟资助总经费（万元）	实际资助经费（万元）
1	专业指导项目	在穗高校与普通高中协同育人项目质量管理与绩效评价	广州大学	聂衍刚	7万	以年度财政批复为准
2	专业指导项目	在穗高校与高中协同育人项目跟踪调查与成效评估	华南理工大学	项　聪	7万	/

花瓣云科技有限公司

2025年第二批产学合作协同育人-新工科建设项目

（HarmonyOS）立项协议书

甲方：花瓣云科技有限公司 乙方：华南农业大学

为了推进教育部产学合作协同育人项目的顺利实施，保证新工科建设项目达到预期目标，依据教育部规范产学合作协同育人项目管理要求，甲、乙双方根据各自的权利和所承担的义务，特签订本协议如下：

甲方：

1. 向乙方提供HarmonyOS开发技术文档、学习资料，用于指导老师进行项目开发；
2. 于协议签订后，乙方提交开课计划，并经甲方验收确认后，由甲方向乙方支付25000元人民币（含税，大写：贰万伍仟圆整人民币），其中不含税金额为24271.84元人民币，于收到乙方开具的合格发票后30个自然日内支付；乙方完成建设任务并通过甲方的结题验收后，由甲方向乙方支付25000元人民币（含税，大写：贰万伍仟圆整人民币），其中不含税金额为24271.84元人民币；
3. 检查、监督项目建设进展情况，组织项目验收、评价和优秀成果交流活动，负责项目中止、撤销等工作。

乙方：

校方收款账户信息：

开户行支行名称：中国工商银行

开户账号：3602002609000310520

开户学校名称：华南农业大学

对接人姓名：黄立峰

对接人联系方式：13929500478

学校地址：广东省广州市天河区五山路483号华南农业大学

按照甲方项目规定和申报指南要求，制定 基于 HarmonyOS 的《操作系统》课程建设与新农科融合创新教学实践 项目的整体规划和实施计划，并将成果提交甲方评审，并于 2026年10月1日 前完成项目建设；

1. 项目负责人基于已有的HarmonyOS课程资源，将HarmonyOS技术融入本校相关专业课程，包括但不限于《移动应用开发》、《操作系统》、《物联网应用与开发》等课程；
2. 教学负责人牵头成立教学工作组、并拉通学院内几位任课老师，共同确定每学期的开课内容和计划，包括课程授课PPT编写及授课等；
3. 项目建设期内至少完成1轮校内授课，其中HarmonyOS技术课程不低于16学时，引导不少100位学生通过HarmonyOS应用开发者基础认证；
4. 交付件：与HarmonyOS相关的教学课件内容，包括但不限于盖学院章的教学大纲、课件PPT/视频、实验手册/案例、人才培养解决方案、学生通过HarmonyOS应用开发者基础认证的证书/考试记录等；
5. 按项目申报书承诺，完成建设任务，接受甲方对项目完成情况的结题验收；如乙方未能通过甲方验收，则乙方应重做相关建设和交付件，直至通过甲方验收；或乙方应按甲方要求退还本项目下甲方支付的所有经费。
6. 甲乙双方共享本合作项下交付件的版权，将项目建设成果放在华为开发者联盟平台无偿开放和共享，甲方及其关联公司有权在全球范围内免费分发、复制、展示、编辑、修改合作项涉及的交付件并在前述范围内进行转授权。
7. 合理安排使用项目资源和建设经费，做到专款专用，提高资源使用效率。
8. 保密：鉴于双方在项目中的合作关系，乙方在合作过程中将了解到甲方不对外披露的“保密信息”，双方同意：

   8.1 保密信息的定义。 保密信息是指甲方（“披露方”）以口头或书面向乙方（“接收方”）披露的、被指定为具保密性的或鉴于信息的性质和有关信息披露的情况应合理视具保密性的所有信息。保密信息不包括(1)乙方无论作为或者不作为都是现在或将来逐渐为公众所知的；(2)在披露前为乙方合法占有且未违反对披露方的保密义务；(3)依法向乙方披露而不受限制的；(4)由乙方独立开发的信息。

   8.2 保密信息的保护。乙方同意针对保密信息的保密的义务与责任应持续有效，

直至该等保密信息根据8.1条款被认定为非保密信息而不需保密为止。乙方对甲方的保密信息具有永久保密义务。乙方同意采取适当措施保护另一方的保密信息，且在任何情况下，其保护另一方的保密信息的谨慎程度不得低于保护自己保密信息的程度。乙方仅可将保密信息披露给其需要了解且其负有不低于本条款规定的保密义务的老师、学生或相关员工。乙方应事先与其老师、学生或相关员工签署不低于本协议保密义务持续有效的保密协议，保证前述人员遵守本协议中约定的义务。若乙方老师、学生或相关员工违反保密义务，乙方同意就前述人员违反保密义务的行为向甲方承担连带责任。乙方老师、学生或相关员工保密义务不因离职或与乙方的合同解除、终止而终止 。除甲方另有授权外，乙方仅可将乙方的保密信息用于履行本协议的目的，不得将保密信息向第三方披露或公开。尽管有相反规定，乙方可在法律诉讼中或按法律要求向政府披露甲方的保密信息。未经甲方书面许可，乙方不得以明示或暗示的任何方式、或以任何媒体、宣传渠道发布与甲方的任何合作信息，包括但不限于官方网站、报纸、宣传材料、广播、电视、杂志等。合作信息包括但不限于双方的合作关系、合作领域、合作金额、当前合作项目、客户信息，双方正在或即将进行某种磋商、或缔结某种合作关系的可能性；或双方即将缔结、已缔结、或已终止某种合作关系的事实等。

8.3 乙方保证拥有其所提供的交付件的知识产权并对其负责，若任何第三方指控乙方交付件包含的任何内容侵犯其知识产权，乙方应当承担全部知识产权侵权赔偿责任并保护甲方（包括甲方关联公司）及其客户免受任何损失。

8.4 违约责任。如因乙方违反本协议而对甲方导致任何损失或损害的，乙方应采取补救措施以弥补另一方遭受的全部损失和支出的费用，并使其免于陷入任何的诉求、请求或诉讼；但该等补救措施不影响另一方行使任何其他权利或采取任何其他的补救措施包括但不限于终止合作项目，要求接收方立即停止违约行为，并采取补救措施等。

9. 双方均为中华人民共和国大陆境内主体，应按如下约定执行：

（1） 双方按照中华人民共和国法律各自承担税款。

（2） 开票总金额应与结算金额（含增值税）一致。

（3） 发票主体必须是本合同的签约方。

（4） 在本协议执行过程中，如果税种、税率等要素因税法修改而发生变动，双方

可直接依据不含税金额及变动后的税率调整合同金额后继续履行合同并按变动后的新税率开具发票。

（5） 如果乙方属于增值税一般纳税人或小规模纳税人，提供的业务不属于免税业务的，应开具增值税专用发票；否则，乙方应开具增值税普通发票。

（6） 如因发票开具方提供的发票不合规，造成发票接收方的一切经济损失（如发票不能抵扣，不能作为相关支出费用列支凭证），由发票提供方全额赔偿。

10. 其它约定事项

10.1 甲、乙方是本协议的独立合作方，不得因本协议规定而被解释为法律上的代理、合伙、合资、聘用或任何种类的正式商业组织。

10.2 任何情况下，任何一方或其关联方均不对另一方因本协议或与本协议相关而产生的间接性或结果性的损失或损害，或任何信誉、数据或数据使用的损害负责。即使某一方或其关联公司已被告知或已经意识到这一损害或可能由另一方造成损失的可能性。双方确认在法律允许的最大范围内，甲方对乙方的赔偿责任不应超过甲方自违约之日起前12个月因本协议所获得利益金额的总额。

10.3 本协议中的附件（如有）为本协议不可分割的部分，与本协议具有相同的法律效力。

本协议经甲、乙双方签字并盖章后立即生效，双方各尽其职，协议中未尽事宜应由双方协商解决。本协议的成立、生效、履行、解释以及纠纷解决，适用中华人民共和国法律（不包括冲突法）。若甲乙双方发生任何纠纷或争议，协商不成的，双方均有权向本协议签订地，深圳市龙岗区有管辖权的法院通过诉讼的方式解决。本协议一式两份，甲、乙双方各执一份。

甲方：花瓣云科技有限公司

（盖章）

日期：2025.10.9

乙方：华南农业大学

项目负责人（签字）：

（盖章）

日期：2025.10.21

华南农业大学文件

华南农教〔2025〕7 号

关于公布华南农业大学 2024 年度课程思政示范项目立项名单的通知

各学院、部处、各单位:

根据《关于开展 2024 年度校级课程思政示范建设项目申报工作的通知》，学校组织开展了 2024 年度课程思政示范项目评选工作。经项目负责人申请、所在单位遴选推荐、学校组织专家评审、校内公示等程序，决定立项建设华南农业大学 2024 年度课程思政示范项目 91 项，包括课程思政试点学院 4 个、课程思政示范团队 6 个、课程思政示范课程 14 门、课程思政示范课堂 25 个、课程思政典型案例 42 个（名单详见附件），现予以公布。

本次立项的课程思政示范项目建设期至 2026 年 12 月，建设期内，示范项目原则上不允许更换负责人或变更项目团队成员。各项目负责人要严格按照 2024 年度申报通知要求，及时开展工作，加

快推进课程思政改革，确保高质量完成建设目标和任务。

请各学院充分认识课程思政改革的重要意义，认真贯彻《华南农业大学课程思政实施方案》，加强对教师的相关培训、指导、引领和支持，带动教师全员积极参与课程思政教学改革，持续深入抓典型、树标杆、推经验，全面提升本科人才培养质量。

附件：华南农业大学2024年度课程思政示范项目立项名单

华南农业大学

2025年1月24日

公开方式：主动公开

华南农业大学党政办公室2025年1月24日印发

附件

华南农业大学2024年度课程思政示范项目立项名单

序号	类别	项目名称	所属单位	项目负责人	团队成员（不含负责人）
kcsz2024001	试点学院	课程思政试点学院	动物科学学院	王文策	
kcsz2024002	试点学院	课程思政试点学院	艺术学院	张艳河	
kcsz2024003	试点学院	课程思政试点学院	数学与信息学院、软件学院	陈文艺	
kcsz2024004	试点学院	课程思政试点学院	公共管理学院	蔡茂华	
kcsz2024005	示范团队	《公共管理学》课程思政示范团队	公共管理学院	史传林	唐斌、姜国兵、方敏、曾小龙、吴彦
kcsz2024006	示范团队	藏粮于“技”，“数”说农业：数字化转型课程思政教学团队	数学与信息学院、软件学院	熊俊涛	黄文玲、韩方珍、古万荣、邓金、张建桃、宋歌、韦婷婷、邓成剑、黄丽清
kcsz2024007	示范团队	农业昆虫学课程群思政示范团队	植物保护学院	陆永跃	王磊、齐易香、岑伊静、黄振、程代凤、桑文、冼继东、吴建辉、潘慧鹏、何晓芳、汪荤荤、何娜芬
kcsz2024008	示范团队	大数据与机器学习类课程思政	经济管理学院	陈有华	何勤英、伍敬文、文乐、

		示范团队			李景荣
kcsz2024009	示范团队	水利工程类专业课程群课程思政示范团队	水利与土木工程学院	齐龙	韦未、王慧琳、杨海燕、卢玉华、余长洪、张巍
kcsz2024010	示范团队	光电信息类课程群思政团队	电子工程学院（人工智能学院）	林芳	刘金龙、欧阳强强、胡旭波、刘建斌、翁嘉文、徐初东、杨初平、林上港、曾应新
kcsz2024011	示范课程	水污染控制工程	资源环境学院	余光伟	梁瑜海、种云霄、陈澄宇、仲海涛、黄柱坚
kcsz2024012	示范课程	行政法与行政诉讼法	人文与法学学院	李燕	杨正喜、欧仁山、左卫霞、孙梦
kcsz2024013	示范课程	森林培育学	林学与风景园林学院	何茜	邱权、陈祖静、刘效东、潘澜、苏艳
kcsz2024014	示范课程	植物学实验	生命科学学院	梁祥修	白玫、李雁群、张荣京、何韩军
kcsz2024015	示范课程	瑜伽	体育教学研究部	何灵捷	田甜、周文英、王顺熙、姚叶戴、戴金明
kcsz2024016	示范课程	乒乓球（国球）	体育教学研究部	陈华东	钞飞侠、张波、李嘉鹏、赵东升、王一
kcsz2024017	示范课程	数据挖掘与大数据分析	经济管理学院	文乐	李琴、陈有华、何勤英、李景荣
kcsz2024018	示范课程	翻译理论与实践	外国语学院	陈喜华	王之杰、张欢、杨敏、严晓蓉，周彧丰

kcsz2024019	示范课程	咨询心理学	公共管理学院	钟向阳	何小芹、唐晓容、卓彩琴、吕玲玲、李锦顺、韩丽
kcsz2024020	示范课程	大学英语 IV（学术英语）	外国语学院	刘玉花	钟志英、柳青、李莉
kcsz2024021	示范课程	西方园林史	林学与风景园林学院	潘建非	夏宇、陈意微、李晓雪
kcsz2024022	示范课程	金融工程	经济管理学院	周丽云	陈标金、王雄志、林伟芬、陈晓洁
kcsz2024023	示范课程	民事诉讼法	人文与法学学院	刘万洪	张艳琼、王琳、杜国明
kcsz2024024	示范课程	农业技术经济学	经济管理学院	蔡键	薛春玲、周伟、苏柳方、易智敏，蔡轶
kcsz2024025	示范课堂	《物理化学》第四章第 4 节（单组元系统两相平衡）	材料与能源学院	刘维	陈明洁、胡航
kcsz2024026	示范课堂	《生态工程学》第五章第 1 节（湿地保护与生态修复）	资源环境学院	田纪辉	蔡昆争
kcsz2024027	示范课堂	《食品分析》第八章第 2 节（蛋白质的测定）	食品学院	徐振林	温棚、罗林
kcsz2024028	示范课堂	《生态农业工程》第七章 第 1 节（碳中和与农业绿色低碳）	资源环境学院	秦俊豪	赵本良
kcsz2024029	示范课堂	《篮球》--传接球与挡拆技术专题	体育教学研究部	郭城	张波、于江杨、黄燕玲
kcsz2024030	示范课堂	《大学生心理健康教育》第七	经济管理学院	何凯	刘思蓓、崔翱鸽

		章第 3 节（情绪与情感调节的方法）			
kcsz2024031	示范课堂	《金工实习》--电火花线切割加工技术专题	基础实验与实践训练中心	徐相华	陈海波、温威
kcsz2024032	示范课堂	《环境毒理学》第五章第 1 节（试验基础）	资源环境学院	银仁莉	
kcsz2024033	示范课堂	《空间分析原理与应用》第四章第 4 节(全局空间自相关分析)	资源环境学院	陈永康	钟晓兰、赵寒冰
kcsz2024034	示范课堂	《金工实习》--注射成型及零件检测专题	基础实验与实践训练中心	张殿武	温威、陈海波
kcsz2024035	示范课堂	《水球》--踩水与水球游戏专题	体育教学研究部	张俊龙	李梅、卢三妹
kcsz2024036	示范课堂	《大学英语III》 unit 5 Gender Equality	外国语学院	陈国华	苏君、李志英
kcsz2024037	示范课堂	《国际政治学》第二章 第 2 节（世界格局）	公共管理学院	欧阳晓东	黄剑飞
kcsz2024038	示范课堂	《C 语言程序设计》第五章--循环结构程序设计之 while 语句专题	数学与信息学院、软件学院	郭艾侠	王栋、邢仲璟
kcsz2024039	示范课堂	《乒乓球》--正手攻球技术专题	体育教学研究部	黄燕玲	吕立、张晓萍、郭城
kcsz2024040	示范课堂	《移动应用开发》第九章--丰富你的程序，运用程序多媒体	数学与信息学院、软件学院	杨春	

		专题			
kcsz2024041	示范课堂	《程序设计与算法语言》第三章第 1 部分--数据表示与数据类型专题	数学与信息学院、软件学院	廖彬	蔡贤资、张连宽
kcsz2024042	示范课堂	《社会保障基金管理》第四章（社会保障基金管理的对象专题）	公共管理学院	杨亚丽	
kcsz2024043	示范课堂	《流体力学》第七章第 2 节（黏性流体的两种流态）	水利与土木工程学院	黄俐	韦未
kcsz2024044	示范课堂	《种子的力量》第十章 第 2 节（植物生理学课堂）	生命科学学院	罗娜	刘亚林、刘慧丽
kcsz2024045	示范课堂	《英语国家概况》下册第四单元（美国政治制度专题）	外国语学院	张雅娜	侯金萍，张国俊
kcsz2024046	示范课堂	《森林培育学》第六章第 5 节（无性繁殖苗培育-扦插育苗）	林学与风景园林学院	陈祖静	邱权
kcsz2024047	示范课堂	《云计算与大数据》第 4 章（数据分析专题）	数学与信息学院、软件学院	古万荣	
kcsz2024048	示范课堂	《MEMS 及其应用》第四章 第 3 节（牺牲层技术）	电子工程学院（人工智能学院）	梁亨茂	
kcsz2024049	示范课堂	《人才测评、培训与开发》第十讲（雇佣与选拔中的测评 2 专题）	经济管理学院	陈灿	

kcsz2024050	典型案例	《新文科创新创业思维与就业力提升》第三讲 “谈谈读书：为什么和怎么做”	公共管理学院	周毅	申佐佐
kcsz2024051	典型案例	《审计学》第二章 第 3 节（职业道德概念框架）	经济管理学院	易智敏	
kcsz2024052	典型案例	《科技前沿与行业发展（植保）》微生物的语言系统之群体感应-Quorum Sensing	植物保护学院	崔紫宁	
kcsz2024053	典型案例	《大学生创新创业基础》--法学创新创业与国家发展、个人成长的关系	人文与法学学院	李玮舜	王琳
kcsz2024054	典型案例	《比较思想政治教育》--“颜色革命”与意识形态风险防范	马克思主义学院	韩谦	
kcsz2024055	典型案例	《生物技术与人类》--CRISPR 基因编辑技术	林学与风景园林学院	张俊杰	周玮
kcsz2024056	典型案例	《Java 与面向对象程序设计》第四章 第 2 节（类的分析）	数学与信息学院、软件学院	梁茹冰	黄小虎
kcsz2024057	典型案例	《跨文化交际》--“空间语言”	外国语学院	钟建玲	赵勇
kcsz2024058	典型案例	《兽医临床诊断学》兽医临床基本检查法之问诊	兽医学院	廖建昭	
kcsz2024059	典型案例	《机械设计基础》第一章 第 3 节（平面机构的自由度）	工程学院	卿艳梅	

kcsz2024060	典型案例	《大学英语 I》Unit 6 Text A Door closer, are you?	外国语学院	黄净	王莹
kcsz2024061	典型案例	《基础化学实验》--茶叶中茶多酚的提取与测定	基础实验与实践训练中心	王鹏程	郑明轩
kcsz2024062	典型案例	《算法分析与设计》--动态规划中的最长公共子序列	数学与信息学院、软件学院	张慧玲	司国东
kcsz2024063	典型案例	《遗传学实验课》--植物减数分裂过程染色体制片及观察	基础实验与实践训练中心	李亚娟	郭海滨
kcsz2024064	典型案例	《文学概论》--文学的超越性	人文与法学学院	姚海燕	王瑛
kcsz2024065	典型案例	《汇编语言程序设计》--循环程序设计	数学与信息学院、软件学院	吴理华	
kcsz2024066	典型案例	《生物多样性》--大陆漂移和动物地理分布	植物保护学院	刘卫欣	
kcsz2024067	典型案例	《农业英语阅读课程》--生物质能概况	外国语学院	郭敏	袁玮
kcsz2024068	典型案例	《瑜伽》--肩颈主题练习	体育教学研究部	田甜	何灵捷
kcsz2024069	典型案例	《Linux 系统及程序设计》Linux 发展史及国产化	数学与信息学院、软件学院	钟德祥	李舜鹏
kcsz2024070	典型案例	《复变函数与积分变换》--柯西积分公式	数学与信息学院、软件学院	刘曼莉	
kcsz2024071	典型案例	《生活营造》—中国茶的起源	艺术学院	王小玉	

kcsz2024072	典型案例	《工程热力学与传热学》第三章第 4 节（孤立系统熵增原理）	工程学院	李成杰	
kcsz2024073	典型案例	《数据挖掘与大数据分析》第三章第 2 节 （Seaborn 绘图基础）	经济管理学院	伍敬文	
kcsz2024074	典型案例	《高等数学》第五章第 1 节（定积分的概念）	数学与信息学院、软件学院	郑佳悦	
kcsz2024075	典型案例	《计算机科学绪论》第十三章信息保障与安全	数学与信息学院、软件学院	罗志坚	黄小虎
kcsz2024076	典型案例	《Linux 系统及程序设计》--Linux 程序设计原理	数学与信息学院、软件学院	黄立峰	罗浩宇
kcsz2024077	典型案例	《软件工程导论》第七章 第 7 节（黑盒测试之等价类划分与边界值分析）	数学与信息学院、软件学院	聂笃宪	刘鹏飞
kcsz2024078	典型案例	《美国简史》--美国进步主义运动	人文与法学学院	王跻崭	
kcsz2024079	典型案例	《通识管理训练--项目管理模块》--项目管理的涵义	基础实验与实践训练中心	谢虎	陈建军
kcsz2024080	典型案例	《PLC 原理与应用》--梯形图基本电路之兼顾公平与效率	工程学院	张闻宇	
kcsz2024081	典型案例	《数学分析》第九章第 1 节（定积分的概念）	数学与信息学院、软件学院	雷春林	王雪琴
kcsz2024082	典型案例	《数学建模》第三章--以微分方程、差分方程建模为例	数学与信息学院 软件学院	方平	张昕

kcsz2024083	典型案例	《物流与供应链管理》第八章第 1 节（供应链失调与牛鞭效应）	数学与信息学院、软件学院	陈炜颖	刘伟章、毛小娟
kcsz2024084	典型案例	《大学英语 I》第二册 Unit 2 Humanities：out of date?	外国语学院	荣莉	苏君
kcsz2024085	典型案例	《大学生心理健康教育》--与压力共处	植物保护学院	曹珂	梁丽梅
kcsz2024086	典型案例	《大学数学》第六章第 5 节（二重积分的概念）	数学与信息学院、软件学院	姚焕城	
kcsz2024087	典型案例	《数据库应用》第七章第 3 节（SQL 分组统计查询）	数学与信息学院、软件学院	张春玲	涂淑琴
kcsz2024088	典型案例	《兽医传染病学》第三章第 1 节（流行性感冒中禽流感防控）	兽医学院	代曼曼	
kcsz2024089	典型案例	《流体力学及其工程应用》第二章第 2 节（流体静平衡微分方程及其积分）	工程学院	罗远强	钟南
kcsz2024090	典型案例	《水产品加工技术》--水产加工原材料的特点	海洋学院	杨敏	李雪竹
kcsz2024091	典型案例	《食品机械与设备》--碟片式离心分离机械	食品学院	关甜	司徒文贝、宋贤良

华南农业大学校内公文

研究生院〔2025〕1号

关于公布2025年华南农业大学研究生教育创新计划立项项目的通知

各学院、部处、各单位:

为深入实施研究生教育创新计划,进一步提高人才培养质量,结合高水平大学建设的有关工作，学校开展了2025华南农业大学研究生教育创新计划项目的申报工作。

经个人申报、单位推荐、形式审查、专家评审和校内公示等程序,确定广东美涂士建材股份有限公司等2个联合培养研究生示范基地,《Plant Plasticity and Adaptation植物的可塑性和适应性》等4个一般性全英文课程建设项目,《现代水产动物医学》等3个课程思政建设项目,《植物生殖生物学》等4个示范课程理论课建设项目,《实验数据分析与处理》1个研究生在线开放课程建设项目,《农业生态与可持续耕作制度》等8个高水平研究生教材建设项目,《“双一流”学科建设与专业学位研究“双向融合”培养模式探索》等16个专业学位研究生实践教学资源建设与培养模式改革研究项目以及《植物病害教学案例库建设》等5个专业学位研究生课程案例库建设项目，共43个项

目为 2025 年华南农业大学研究生教育创新计划立项项目（详见附件），现予公布。

各项目负责人应按照项目既定研究周期，严格执行研究计划，扎实推进研究工作，确保按时完成研究任务，实现预期目标。

附件： 2025 年华南农业大学研究生教育创新计划项目立项名单

研究生院

2025 年 12 月 9 日

（联系人：黄超 ；电话：020-85280856）

附件：2025 年华南农业大学研究生教育创新计划项目立项名单

项目类别	序号	合作单位	学科领域	校内负责人	依托单位
（一）联合培养研究生示范基地	1	广东美涂士建材股份有限公司	材料与化工、农业工程、林业工程等	袁腾	材料与能源学院
	2	福州译国译民集团有限公司	翻译	林绿	外国语学院
项目类别	**序号**	**课程名称**	**课程类型**	**负责人**	**所在单位**
（二）一般性全英文课程建设项目	1	Plant Plasticity and Adaptation 植物的可塑性和适应性	专业选修课	张艳霞	农学院
	2	英文科技论文写作与交流	专业学位课	黄晓	生命科学学院
	3	世界农产品加工技术发展及其创新	专业选修课	吴绍宗	食品学院
	4	Agricultural UAV Technology for Crop Protection	专业选修课	陈鹏超	电子工程学院（人工智能学院）
（三）课程思政建设项目	1	现代水产动物医学	专业学位课	黄友华	海洋学院
	2	算法分析与设计	专业学位课	黄立峰	数学与信息学院、软件学院
	3	人工智能	专业学位课	涂淑琴	数学与信息学院、软件学院
（四）示范课程理论课建设项目	1	植物生殖生物学	专业选修课	吴锦文	农学院
	2	生态系统生态学	专业学位课	蔡昆争	资源环境学院
	3	昆虫免疫学	专业选修课	邓小娟	动物科学学院
	4	稀土化学	专业选修课	张学杰	材料与能源学院
（五）研究生在线开放课程建设项目	1	实验数据分析与处理	专业选修课	祁剑英	农学院

项目类别	序号	教材名称	学科领域	负责人	所在单位
（六）高水平研究生教材建设项目	1	农业生态与可持续耕作制度	作物学	王小龙	农学院
	2	家禽免疫学理论前沿与技术	预防兽医学	代曼曼	兽医学院
	3	植物细胞工程原理与实验	生物学	李静	生命科学学院
	4	光学农业	化学工程与技术	李唯	材料与能源学院
	5	人工智能应用案例	管理科学与工程	熊俊涛	数学与信息学院、软件学院
	6	深度学习的数学视角：从基础到前沿	人工智能	刘景锋	电子工程学院（人工智能学院）
	7	食品安全与食品健康法	法学	杜国明	人文与法学学院
	8	AI Python 翻译应用教程	文学	吕靖	外国语学院
项目类别	**序号**	**项目名称**	**学科领域**	**负责人**	**所在单位**
（七）专业学位研究生实践教学资源建设与培养模式改革研究项目	1	“双一流”学科建设与专业学位研究“双向融合”培养模式探索	教育学	张凯旋	农学院
	2	基于产业体系的农业硕士实践教学资源建设与培养模式创新研究	畜牧	刘吉平	动物科学学院
	3	服务于国家重大战略背景下农业专业学位研究生培养模式创新——以“低碳农业绿色发展”专项硕士班为例	低碳农业	蔡泽瀛	资源环境学院
	4	基于“双创”理念的饲料配制与投饲技术课程体系构建	渔业发展	徐超	海洋学院
	5	数智化赋能食品类专业学位研究生食品包装实践能力培养的模式	食品科学与工程	范小平	食品学院
	6	新文科视域下农业院校法律硕士“深协同”培养模式创新研究	法学	李燕	人文与法学学院
	7	新文科背景下法硕课程智慧教学探索研究----以华南农业大学法硕课程教学为例	法学	王琳	人文与法学学院
	8	法学专业研究生参与社区治理实践教学模式改革	法学	王慧	人文与法学学院
	9	AI 时代涉农翻译人才“产教-科教”双融合培养模式研究	翻译	邓飞	外国语学院
	10	农业科技术语英汉双语平行语料库建设与教学应用研究	翻译	李飞武	外国语学院
	11	DeepSeek 在农业院校 MTI 教学中的创新应用	翻译	李舸	外国语学院
	12	“双一流”建设背景下设计专业学位研究生培养模式创新研究——基于“科教+产教”双融合驱动视角	设计	陈薇薇	艺术学院

	13	面向乡村振兴“数智叙事”视角的新媒体艺术课程产教融合实践教学模式研究	戏剧与影视	李雷鸣	艺术学院
	14	乡村振兴背景下科教产融合创新模式研究与实践—以《融媒体实务》课程为例	戏剧与影视	赵娜	艺术学院
	15	数智时代涉农高校专业学位研究生“大思政课”实践教学创新模式研究	马克思主义理论	林晓燕	马克思主义学院
	16	服务“双碳”战略的农林废弃物资源化专业硕士产教融合实践教学改革研究	林业工程	姚业成	生物质工程研究院
（八）专业学位研究生课程案例库建设项目	1	植物病害教学案例库建设	植物保护	司徒俊键	植物保护学院
	2	小动物疾病诊断案例库	兽医	李英	兽医学院
	3	渔业案例分析与研讨	渔业发展	甘炼	海洋学院
	4	现代农业创新与乡村振兴案例库	农业	罗明忠	经济管理学院
	5	《审计理论与务实》教学案例库	会计	易智敏	经济管理学院

公开方式：主动公开

华南农业大学研究生院　　2025 年 12 月 12 日印发

国家自然科学基金资助项目批准通知

（包干制项目）

黄立峰　先生/女士：

根据《国家自然科学基金条例》、相关项目管理办法规定和专家评审意见，国家自然科学基金委员会（以下简称自然科学基金委）决定资助您申请的项目。项目批准号：62502165，项目名称：智能驾驶多模态感知模型的黑盒对抗防御方法研究，资助经费：30.00万元，项目起止年月：2026年01月至 2028年 12月，有关项目的评审意见及修改意见附后。

请您尽快登录科学基金网络信息系统（https://grants.nsfc.gov.cn），**认真阅读《国家自然科学基金资助项目计划书填报说明》并按要求填写《国家自然科学基金资助项目计划书》（以下简称计划书）**。对于有修改意见的项目，请您按修改意见及时调整计划书相关内容；如您对修改意见有异议，须在电子版计划书报送截止日期前向相关科学处提出。

请您将电子版计划书通过科学基金网络信息系统（https://grants.nsfc.gov.cn）提交，由依托单位审核后提交至自然科学基金委。自然科学基金委审核未通过者，将退回的电子版计划书修改后再行提交；审核通过者，打印纸质版计划书（一式两份，双面打印）并在项目负责人承诺栏签字，由依托单位在承诺栏加盖依托单位公章，且将申请书纸质签字盖章页订在其中一份计划书之后，一并报送至自然科学基金委项目材料接收工作组。纸质版计划书应当保证与审核通过的电子版计划书内容一致。**自然科学基金委将对申请书纸质签字盖章页进行审核，对存在问题的，允许依托单位进行一次修改或补齐。**

向自然科学基金委提交电子版计划书、报送纸质版计划书并补交申请书纸质签字盖章页截止时间节点如下：

1. **2025年9月5日16点：**提交电子版计划书的截止时间；

2. **2025年9月12日16点：**提交修改后电子版计划书的截止时间；

3. **2025年9月23日：**报送纸质版计划书（一式两份，其中一份包含申请书纸质签字盖章页）的截止时间。

4. **2025年10月9日：**报送修改后的申请书纸质签字盖章页的截止时间。

请按照以上规定及时提交电子版计划书，并报送纸质版计划书和申请书纸质签字盖章页，逾期不报计划书或申请书纸质签字盖章页且未说明理由的，视为自动放弃接受资助；未按要求修改或逾期提交申请书纸质签字盖章页者，将视情况给予暂缓拨付经费等处理。

附件：项目评审意见及修改意见表

国家自然科学基金委员会

2025年8月27日

项目批准号	62502165
申请代码	F0205
归口管理部门	
依托单位代码	51064208A0499-0932

国家自然科学基金

资助项目计划书

（包干制项目）

资助类别： 青年科学基金项目（C类）[原青年科学基金项目]

亚类说明：

附注说明：

项目名称： 智能驾驶多模态感知模型的黑盒对抗防御方法研究

资助经费： 30万元　　**执行年限：** 2026.01-2028.12

负 责 人： 黄立峰　　**BRID：** 01318.00.30603

通讯地址： 广东省 广州市 华南农业大学 数学与信息(软件)学院 514

邮政编码： 510000　　**电　　话：** 020-85285383

电子邮件： huanglf6@scau.edu.cn

依托单位： 华南农业大学

联 系 人： 郑雪宜　　**电　　话：** 020-85280070

填表日期： 2025年08月29日

国家自然科学基金委员会制

Version：1.004.680

国家自然科学基金资助项目计划书填报说明

（包干制项目）

一、项目负责人收到《国家自然科学基金资助项目批准通知》（以下简称《批准通知》）后，请认真阅读本填报说明，参照国家自然科学基金相关项目管理办法和《国家自然科学基金资助项目资金管理办法》（以下简称《资金管理办法》，请查阅国家自然科学基金委员会门户网站首页“政策法规”栏目），按《批准通知》的要求认真填写和提交《国家自然科学基金资助项目计划书》（以下简称《计划书》）。

二、填写《计划书》时要科学严谨、实事求是、表述清晰、准确。《计划书》经国家自然科学基金委员会相关项目管理部门审核批准后，将作为项目研究计划执行、检查和验收的依据。

三、《计划书》各部分填写要求如下：

（一）简表：由系统自动生成。

（二）摘要及关键词：各类获资助项目都应当填写中、英文摘要及关键词。

（三）正文：

1. 青年科学基金项目（C类）、青年学生基础研究项目：如果《批准通知》所附“项目评审意见及修改意见表”中“修改意见”栏目没有修改要求的，只需选择“研究内容和研究目标按照申请书执行”即可；如果《批准通知》中上述栏目明确要求调整研究期限或研究内容等的，须选择“根据研究方案修改意见更改”并填报相关修改内容。

2. 青年科学基金项目（A类）和青年科学基金项目（B类）按下列提纲撰写：

（1）研究方向；

（2）结合国内外研究现状，说明研究工作的学术思想和科学意义（限两个页面）；

（3）研究内容、研究方案及预期目标（限两个页面）；

（4）年度研究计划；

四、资助经费相关要求：

1. 资助经费批准时不再区分直接费用和间接费用。

2. 项目负责人在提交计划书时需签署承诺书，承诺尊重科研规律，弘扬科学家精神，遵守科研伦理道德和作风学风诚信要求，认真开展科学研究工作；承诺项目经费全部用于与本项目研究工作相关的支出，不得用于与本项目研究无关的支出。

3. 项目负责人提交计划书时，无需编制项目预算。项目资金由项目负责人自主决定使用，按照《资金管理办法》第九条规定的开支范围列支。有关管理费用的补助支出，由依托单位根据实际管理需要，在充分征求项目负责人意见基础上合理确定。绩效支出由项目负责人根据实际科研需要和相关薪酬标准自主确定，依托单位按照工资制度进行管理。对于青年学生基础研究项目，支付给项目负责人本人的劳务费用，应符合相关比例要求。其余用途经费无额度限制，由项目负责人根据实际需要自主决定使用。

4. 项目结题时，项目负责人根据实际使用情况编制项目经费决算，经依托单位财务、科研管理部门审核后，报自然科学基金委。依托单位应当在单位内部公开非涉密项目立项、主要研究人员、资金使用（重点是间接费用、外拨资金、结余资金使用等）、决算、大型仪器设备购置以及项目研究成果等情况，接受内部监督。

5. 自然科学基金委结合项目管理，对经费使用情况和依托单位管理情况定期开展抽查。

五、其他事项

（一）根据有关要求，国家自然科学基金资助项目研究形成的代表性论文中发表在我国科技期刊上的应占20%以上。

（二）国家自然科学基金资助项目研究形成的专利申请应按照《建立财政资助科研项目形成专利的声明制度实施方案》要求进行声明。

简表

项目负责人信息	姓　　名	黄立峰	性 别	男	出生年月	1990年02月	民 族	侗族
	学　　位	博士			职称	讲师		
	是否在站博士后	否		电子邮件		huanglf6@scau.edu.cn		
	电　　话	020-85285383		个人网页				
	工 作 单 位	华南农业大学						
	所 在 院 系 所	数学与信息(软件)学院						
依托单位信息	名　　称	华南农业大学					代码	51064208A0499
	联 系 人	郑雪宜		电子邮件		kycjhk@scau.edu.cn		
	电　　话	020-85280070		网站地址		http://kjc.scau.edu.cn/		
合作单位信息	单 位 名 称							
项目基本信息	项目名称	智能驾驶多模态感知模型的黑盒对抗防御方法研究						
	资助类别	青年科学基金项目（C类）[原青年科学基金项目]			亚类说明			
	附注说明							
	申请代码	F0205:网络与系统安全						
	执行年限	2026.01-2028.12						
	资助经费	30万元						

项目摘要

中文摘要：

智能驾驶系统的多模态感知模型面临严峻的对抗噪声攻击威胁，易引发交通环境误判等重大安全风险。实际对抗环境中攻防双方信息互不可知，构成黑盒场景，现有防御因学习数据质量低、鲁棒模型偏差大、黑盒场景评估难等问题，导致防护失效，呈显著脆弱性。为突破黑盒防御技术瓶颈，项目围绕“数据-模型-评估”展开：①数据层面，研究面向黑盒迁移的可泛化点云对抗学习数据生成方法，通过建模对抗脆弱性来解耦优化对抗噪声，以构造高质量点云对抗数据，支撑多模态对抗学习；②模型层面，研究基于归因理解的去偏差鲁棒模型优化方法，通过运用多模态对抗数据来解析模型偏好，以降低对抗偏差，消除鲁棒模型安全隐患；③评估层面，研究复杂对抗环境下安全评估驱动的模型优选方法，通过仿真多样化、跨模态攻击来建立黑盒场景安全基准，以实现资源约束下模型动态选择，提升系统稳定性。项目成果将形成安全可信的感知模型黑盒防御机制，为智能驾驶实现稳定可靠的环境感知。

Abstract：

Multimodal perception models in intelligent driving systems face severe threats from adversarial noise attacks, which can lead to critical safety risks such as erroneous environmental assessments. In real-world adversarial scenarios, both attackers and defenders have no knowledge of each other, forming black-box scenarios. Traditional defense mechanisms often fail due to low-quality learning data, biased robust models, and challenges in black-box evaluation, exposing significant vulnerabilities. To overcome the bottlenecks in black-box defense for intelligent driving, this project explores three key aspects. (1) Data Level: Investigating generalizable point cloud adversarial learning data generation for black-box transfer. By modeling adversarial vulnerability, the project aims to decouple and optimize adversarial noise, enabling the construction of high-quality point cloud adversarial data to support multimodal adversarial training. (2) Model Level: Developing debiasing robust model optimization based on attribution analysis. This involves leveraging multimodal adversarial data to interpret robust model preferences, reducing adversarial bias and eliminating potential security risks. (3) Evaluation Level: Establishing a security-driven model selection framework for complex adversarial environments. By simulating diverse cross-modal attacks, the project seeks to build a black-box security benchmark, enabling dynamic model selection under resource constraints and improving system stability. The project outcomes will contribute to a secure and trustworthy black-box defense mechanism for perception models, ensuring stable and reliable environmental perception in intelligent driving systems.

关键词(用分号分开)：模型安全；对抗防御；对抗样本；对抗噪声；黑盒场景

Keywords(用分号分开)： model security; adversarial defense; adversarial example; adversarial noise; black-box scenarios

报告正文

研究内容和研究目标按照申请书执行。

本项目研究形成的代表性论文中发表在我国科技期刊上的将占 20%以上。

本项目研究形成的专利申请将按照《建立财政资助科研项目形成专利的声明制度实施方案》要求进行声明。

国家自然科学基金项目负责人、依托单位承诺书

国家自然科学基金项目负责人承诺书

本人郑重承诺：我接受国家自然科学基金的资助，严格遵守中共中央办公厅、国务院办公厅《关于进一步加强科研诚信建设的若干意见》《关于进一步弘扬科学家精神加强作风和学风建设的意见》《关于加强科技伦理治理的意见》《科技伦理审查办法（试行）》等规定，和国家自然科学基金委员会关于资助项目管理、项目资金管理等各项规章，在《计划书》填写及项目执行过程中：

（一）按照《批准通知》《国家自然科学基金资助项目计划书填报说明》的要求填写《计划书》，未自行降低、更改目标任务或约定要求，或缩减研究（研制）内容；

（二）树立“红线”意识，严格履行科研合同义务，按照《计划书》负责实施本项目（批准号：62502165），切实保证研究工作时间，按时报送有关材料，及时报告重大情况变动，不违规将科研任务转包、分包他人，不以项目实施周期外或不相关成果充抵交差；

（三）遵守科研诚信、科技伦理规范和学术道德，认真开展研究工作，对资助项目发表的论著和取得的研究成果按规定进行标注，不在非本项目资助的成果或其他无关成果上标注本项目批准号，反对无实质学术贡献者“挂名”，不在成果署名、知识产权归属等方面侵占他人合法权益，并如实报告本人及项目组成员发生的违背科研诚信要求的任何行为；

（四）尊重科研规律，弘扬科学家精神，严谨求实，追求卓越，反对浮夸浮躁、投机取巧，不人为夸大学术或技术价值，不传播未经科学验证的现象和观点；

（五）将项目资金全部用于与本项目研究工作相关的支出，并结合科研活动需要，科学合理安排项目资金支出进度；

（六）做好项目组成员的教育和管理，确保遵守以上相关要求。

如违背上述承诺，本人愿接受国家自然科学基金委员会和相关部门做出的各项处理决定。

项目负责人（签字）：

年　　月　　日

国家自然科学基金项目依托单位承诺书

我单位同意承担上述国家自然科学基金项目，将保证项目负责人及其研究队伍的稳定和研究项目实施所需的条件，严格遵守中共中央办公厅、国务院办公厅《关于进一步加强科研诚信建设的若干意见》《关于进一步弘扬科学家精神加强作风和学风建设的意见》《关于加强科技伦理治理的意见》《科技伦理审查办法（试行）》等规定，和国家自然科学基金委员会有关资助项目管理、项目资金管理、科研诚信管理和科技伦理管理等各项规定，并督促实施。

依托单位（公章）

年　　月　　日

国家自然科学基金资助项目签批审核表

本栏目由自然科学基金委填写	科学处审查意见： 负责人（签章）： 年　　月　　日
	科学部审查意见： 负责人（签章）： 年　　月　　日

受理编号：c25140500001891

项目编号：2025A1515010030

文件编号：粤基金字〔2025〕10号

广东省基础与应用基础研究基金项目

任务书

项目名称： 黑箱对抗场景下自动驾驶视觉模型的鲁棒优化与集成研究

项目类别： 广东省自然科学基金-面上项目

项目起止时间： 2025-01-01　至　2027-12-31

管理单位（甲方）： 广东省基础与应用基础研究基金委员会

依托单位（乙方）： 华南农业大学

通讯地址： 广东省广州市天河区五山路483号

邮政编码： 510642　　**单位电话：** 020-85283435

项目负责人： 黄立峰　　**联系电话：** 13929500478

（广东科技微信公众号）　　（查看任务书信息）　　（受理纸质材料二维码）

广东省基础与应用基础研究
基金委员会
二〇二〇年制

填写说明

一、项目任务书内容原则上要求与申报书相关内容保持一致，不得无故修改。

二、项目承担单位通过广东省科技业务管理阳光政务平台下载项目任务书，按要求完成签名盖章后扫描上传到广东省科技业务管理阳光政务平台。

三、签名盖章说明。请分别在单位工作分工及经费分配情况页、人员信息页、签约各方页等地方按要求签字或盖章，签章不合规或错漏将不予受理。其中，人员信息页要求所有参与人员本人亲笔签名，代签或印章无效，漏签将不予受理。

四、本任务书自签字并加盖公章之日起生效，各方均应负本任务书的法律责任，不应受机构、人事变动影响。

五、根据《广东省科学技术厅广东省财政厅关于深入推进省基础与应用基础研究基金项目经费使用“负面清单+包干制”改革试点工作的通知》(粤科规范字〔2022〕2号)，2022年度及以后立项资助的全部省基金项目(包括省自然科学基金、省市联合基金、省企联合基金项目等)均适用“负面清单+包干制”，项目提交申请书和任务书时无需编制费用明细科目预算。

一、主要研究内容和要达到的目标

项目围绕当前自动驾驶视觉模型在黑箱对抗场景面临的安全隐患问题展开研究。针对现有方案面临的对抗噪音数据质量低、模型优化对抗偏差大和模型集成方案选择难等问题，项目拟从对抗数据生成、鲁棒模型优化和集成方案选择三方面展开研究，采用数据驱动和模型驱动的方式提升自动驾驶视觉模型在黑箱对抗场景下的鲁棒性，以防范恶意对抗噪音干扰。具体而言，项目拟深入研究点云通用对抗噪音生成、鲁棒模型去偏差优化和模型集成方案安全评估与选择等关键技术：

（1）研究黑箱场景可迁移的点云通用对抗噪音生成方法，从点云数据自身特性出发，在不依赖替代模型的前提下，设计并度量在黑箱场景下更加通用和泛化的空间重要性测度，旨在构造更高质量的点云对抗数据，为模型鲁棒优化（关键技术二）和模型集成方案评估（关键技术三）提供数据基础；

（2）研究基于归因理解的鲁棒模型去偏差优化方法，以高质量的通用对抗噪音数据（关键技术一）为基础，深入理解鲁棒模型的推理偏好并将其纳入优化过程，旨在降低对抗偏差，并进一步提升鲁棒模型在黑箱对抗场景的鲁棒性，为模型集成方案评估与选择（关键技术三）提供模型支撑。

（3）研究模型集成方案多样化的安全评估与选择方法，通过缓解现有方法在评估视觉模型安全性时面临的差异敏感性问题，结合黑箱场景点云对抗噪音数据（关键技术一），为不同的多模态鲁棒模型组合（关键技术二）建立黑箱场景安全性基准，进而在有限资源约束下选择安全性高的模型集成部署方案，为自动驾驶系统提供安全支撑。

本项目整体研究目标是实现黑箱场景下对抗鲁棒的自动驾驶视觉模型，为实现安全可信的自动驾驶提供理论和实践支撑。具体研究目标包括：

（1）设计黑箱场景可迁移的点云通用对抗噪音生成方法，构造高质量的点云对抗噪音数据，作为模型优化的学习数据和模型安全评估的测试数据；

（2）研究基于归因理解的鲁棒模型去偏差优化方法，提高视觉模型在黑箱场景下的鲁棒性，为模型集成方案提供鲁棒模型集合。

（3）提出模型集成方案多样化的安全评估与选择方案，在资源约束下从多种鲁棒模型组合中选择黑箱场景下安全性高的集成部署方案，为自动驾驶系统提供安全支撑。

二、项目预期获得的研究成果及形式

论文及专著情况	国家统计源刊物以上刊物发表论文（篇）		3		科技报告（篇）		1	
	其中被SCI/EI/ISTP收录论文数（篇）		2		培养人才（人）		3	
	专著（册）				引进人才（人）			
专利情况(项)	发明专利		实用新型专利		外观设计专利		国外专利	
	申请	授权	申请	授权	申请	授权	申请	授权
	2							

三、项目进度和阶段目标

（一）项目起止时间： 2025-01-01 至 2027-12-31		
（二）项目实施进度及阶段主要目标:		
开始日期	结束日期	主要工作内容
2025-01-01	2025-12-31	1、研究黑箱场景可迁移的点云通用对抗噪音生成方法； 2、构建高质量的点云对抗数据集； 3、搭建视觉模型对抗攻防仿真平台，并完成验证测试； 4、发表1篇学术论文，申请发明专利1件； 5、参加1-2次国内外知名学术会议或者相关学术活动。
2026-01-01	2026-12-31	1、研究基于归因理解的鲁棒模型去偏差优化方法； 2、构建黑箱场景下高鲁棒性的视觉模型集合，完成验证测试； 3、发表1篇学术论文，申请发明专利1件； 4、参加1-2次国内外知名学术会议或者相关学术活动。
2027-01-01	2027-12-31	1、研究集成部署方案多样化的安全评估与选择方案； 2、建立鲁棒模型组合在黑箱场景下的安全性基准，选择安全性高的集成部署方案，完成验证测试； 3、发表1-2篇学术论文，申请发明专利1件； 4、参加1-2次国内外知名学术会议或者相关学术活动； 5、准备结题工作，进行技术和研究成果汇总，完成结题报告。

四、项目总经费及省基金委经费预算

1.省基金委经费下达总额： （大写）壹拾万圆整；（小写 ）10万元；					
2.省基金委经费年度下达计划：					
年度	2025 年	年	年	年	年
经费(万元)	10.00				

五、人员信息

项目负责人								
姓名	证件号码	年龄	性别	职称	学历	在项目中承担的任务	所在单位	签名
黄立峰	431202199002010815	35	男	讲师	博士研究生	项目负责人	华南农业大学	黄立峰

项目组主要成员								
姓名	证件号码	年龄	性别	职称	学历	在项目中承担的任务	所在单位	签名
周子涵	342601199602060028	29	女	讲师	博士研究生	算法设计、架构设计	华南农业大学	周子涵
罗浩宇	360302198907132536	36	男	副教授	博士研究生	算法设计、架构设计	华南农业大学	罗浩宇
王涵	440105200008055432	25	男	未取得	本科	系统设计、软件开发	华南农业大学	王涵
刘名	370921200009074235	25	男	未取得	本科	系统设计、软件开发	华南农业大学	刘名

六、工作分工及财政经费分配

承担/参与单位名称（盖章）	工作分工	省级财政科技资金分配（万元）
华南农业大学	本项目由华南农业大学研究团队负责项目的整体组织、制定总体方案和研发规划，具体包括：1. 项目整体规划、研究方案制定和项目实施规划，确保项目按时按质地有序推进；2. 针对自动驾驶系统视觉模型，开展黑箱场景下的点云通用对抗噪音生成关键技术公关，从数据生成方面为模型安全提供基础。 聚焦于黑箱场景下的视觉模型鲁棒优化关键技术公关，研究基于归因理解的鲁棒模型去偏差优化方法，从模型构建方面提供支撑。研究自动驾驶系统中高鲁棒性视觉模型集成部署的关键技术公关，研究模型集成方案多样化的安全评估与选择方法。公开相关技术和模型。3. 实现自动驾驶视觉攻防原型系统，主要包含架构扩展模块、数据生成模块和攻防评估模块，涵盖本申请中研究方法以及已有的主流攻防算法，将重点面向互联网科技公司、创新创业企业和研究机构等，为自动驾驶系统相关开发者、使用者和研究者提供自动驾驶汽车视觉模型鲁棒性评估服务。4. 针对项目的研究成果进行总结，充分准备并配合项目的验收工作。	10.00
	合计	10.00

七、任务书条款

第一条 甲方与乙方根据《中华人民共和国民法典》及国家有关法规和规定，按照《广东省自然科学基金及联合基金项目管理实施细则》（粤科规范字〔2024〕5号）《省级科技计划项目任务书管理细则》（粤科规范字〔2022〕8号）等规定，为顺利完成（2025）年黑箱对抗场景下自动驾驶视觉模型的鲁棒优化与集成研究专项项目（项目编号：2025A1515010030）经协商一致，特订立本任务书，作为甲乙双方在项目实施管理过程中共同遵守的依据。

第二条 甲方的权利义务：1.按任务书规定进行经费核拨的有关工作协调。2.根据甲方需要，在不影响乙方工作的前提下，定期或不定期对乙方项目的实施情况和经费使用情况进行检查或抽查。3.根据《广东省科学技术厅科技计划项目科研诚信管理办法》（粤科规范字〔2024〕2号）《广东省基础与应用基础研究基金项目科研不端行为调查处理实施细则（试行）》（粤科规范字〔2023〕1号）等规定对乙方进行科技计划信用管理。

第三条 乙方的权利义务：1.确保落实自筹经费及有关保障条件。2.按任务书规定，对甲方核拨的经费实行专款专用，单独列账，并随时配合甲方进行监督检查。3.经费使用按照广东省级财政科研项目经费使用及省基金项目经费使用“负面清单+包干制”等有关规定进行管理。4.项目依托单位应制定经费使用“负面清单+包干制”内部管理制度并报甲方备案。5.使用财政资金采购设备、原材料等，按照《广东省实施〈中华人民共和国招标投标法〉办法》有关规定，符合招标条件的须进行招标。6.项目任务书任务完成后，或任务书规定的任务、指标及经费投入等提前完成的，乙方可提出验收结题申请，并按甲方要求做好项目验收结题工作。7.若项目发生需要终止结题的情况，乙方须提出终止结题申请，并按甲方要求做好项目终止结题工作。8.在每年规定时间内向甲方如实提交上年度工作情况报告，报告内容包含上年度项目进展情况、经费决算和取得的成果等。9.按照国家和省有关规定，提交科技报告及其他材料。10.利用甲方的经费获得的研究成果，项目负责人和参与者应当注明获得“广东省基础与应用基础研究基金（英文：Guangdong Basic and Applied Basic Research Foundation）（项目编号）”资助或作有关说明。11.乙方要恪守科学道德准则，遵守科研活动规范，践行科研诚信要求，不得抄袭、剽窃他人科研成果或者伪造、篡改研究数据、研究结论；不得购买、代写、代投论文，虚构同行评议专家及评议意见；不得违反论文署名规范，擅自标注或虚假标注获得科技计划（专项、基金等）等资助；不得弄虚作假，骗取科技计划（专项、基金等）项目、科研经费以及奖励、荣誉等；不得有其他违背科研诚信要求的行为。12.确保本项目开展的研究工作符合我国科技伦理管理相关规定。

第四条 在履行本任务书的过程中，如出现广东省相关政策法规重大改变等不可抗力情况，甲方有权对所核拨经费的数量和时间进行相应调整。

第五条 在履行本任务书的过程中，当事人一方发现可能导致项目整体或部分失败的情形时，应及时通知另一方，并采取适当措施减少损失，没有及时通知并采取适当措施，致使损失扩大的，应当就扩大的损失承担责任。

第六条 本项目技术成果的归属、转让和实施技术成果所产生的经济利益的分享，除双方另有约定外，按国家和广东省有关法规执行。

第七条 根据项目具体情况，经双方另行协商订立的附加条款，作为本任务书正式内容的一部分，与本任务书具有同等效力。

第八条 本任务书一式三份，各份具有同等效力。甲、乙方及项目负责人各执一份，三方签字、盖章后即生效，有效期至项目结题后一年内。各方均应负任务书的法律责任，不应受机构、人事变动的影响。

第九条 乙方必须接受甲方聘请的本项目任务书监理单位的监督和管理。监理单位按照甲方赋予的权利对本项目任务书的履行进行审核、进度调查，对项目任务书变更、经费使用情况进行监督管理及组织项目验收。

说明：1.本任务书中，凡是当事人约定无需填写的内容，应在空白处划（/）。

2.委托代理人签订本任务书的，应出具合法、有效的委托书。

八、本任务书签约各方

管理单位（甲方）：　　广东省基础与应用基础研究基金委员会　（盖章）

法定代表人（或法人代理）：　曾战　（签章）

2025 年　03 月　21 日

依托单位（乙方）：　华南农业大学　（盖章）

法定代表人（或法人代理）：　薛红卫　（签章）

联系人（项目主管）姓名：　夏杰　（签章）

Email：　kjcgxk@scau.edu.cn

电话：　020-85283435　/　13711345768

开户单位名称：　华南农业大学

开户银行名称：　广东广州工行五山支行

开户银行账号：　3602002609000310520

2025 年 4 月 9 日

联系人（项目负责人）姓名：　黄立峰　（签名）

Email：　huanglf6@scau.edu.cn

电话：　13929500478

2025 年 4 月 5 日

受理编号：c232019102400000026

项目编号：2023A1515110075

文件编号：粤基金字〔2024〕4号

广东省基础与应用基础研究基金项目

任务书

项目名称：面向跨域数据和异构模型的迁移场景对抗攻防方法研究

项目类别：区域联合基金-青年基金项目

项目起止时间：2023-11-01　至　2026-10-31

管理单位（甲方）：广东省基础与应用基础研究基金委员会

依托单位（乙方）：华南农业大学

通讯地址：广东省广州市天河区五山路483号

邮政编码：510642　　单位电话：020-85283435

项目负责人：黄立峰　　联系电话：13929500478

（广东科技微信公众号）　　（查看任务书信息）　　（受理纸质材料二维码）

广东省基础与应用基础研究
基金委员会
二〇二〇年制

填写说明

一、项目任务书内容原则上要求与申报书相关内容保持一致，不得无故修改。

二、项目承担单位通过广东省科技业务管理阳光政务平台下载项目任务书，按要求完成签名盖章后扫描上传到广东省科技业务管理阳光政务平台。

三、签名盖章说明。请分别在单位工作分工及经费分配情况页、人员信息页、签约各方页等地方按要求签字或盖章，签章不合规或错漏将不予受理。其中，人员信息页要求所有参与人员本人亲笔签名，代签或印章无效，漏签将不予受理。

四、本任务书自签字并加盖公章之日起生效，各方均应负本任务书的法律责任，不应受机构、人事变动影响。

五、根据《广东省科学技术厅广东省财政厅关于深入推进省基础与应用基础研究基金项目经费使用“负面清单+包干制”改革试点工作的通知》(粤科规范字〔2022〕2号)，2022年度及以后立项资助的全部省基金项目(包括省自然科学基金、省市联合基金、省企联合基金项目等)均适用“负面清单+包干制”，项目提交申请书和任务书时无需编制费用明细科目预算。

一、主要研究内容和要达到的目标

本项目旨在利用对抗攻击技术挖掘深度学习模型的安全漏洞，并基于此构建高鲁棒性的防御机制，为发展安全可信的人工智能提供基础和保障。针对实际迁移对抗场景中攻防技术存在的跨域迁移性低、逃逸能力弱和资源消耗多等挑战，分别设计面向跨域数据和异构防御的对抗攻击方法，以此探索深度学习技术的安全隐患，最终构建低成本高鲁棒性的防御机制。具体而言，本项目拟展开三方面研究。

研究内容一拟提出面向跨域数据的对抗样本生成方法，聚焦于如何提升对抗样本在不同数据域模型之间的迁移攻击能力，实现研究神经网络特征空间的脆弱机理的目标。首先，设计多类型合成域模型，即生成多种类型的合成增广数据，并根据深层特征相似性聚类多粒度伪标签集合，构造合成域模型集合，模拟神经网络在跨域数据中学习差异化的特征空间。然后，优化扰动生成模型，其优化目标是搜索通用的特征级对抗扰动，能对基于合成域模型建模的域泛化特征向量产生攻击效果，当对抗样本有能力误导多个特征空间差异较大的合成域模型时，即能以较高概率跨域迁移攻击未知目标域的黑盒模型。

研究内容二拟提出面向异构模型的迁移对抗攻击方法，主要考虑如何增强对抗攻击在迁移场景中面对差异化防御模型的逃逸能力，达到探索神经网络结构与鲁棒性关联关系的目标。首先，构建轻量级的结构增广神经网络，即将设计多分支神经网络，包含预测分布多样化的特征残差预测模块，并且嵌入轻量级随机化网络模块，避免对抗样本将过度依赖神经网络结构，降低对模型结构差异的敏感性。然后，增强对抗样本逃逸能力，即通过周期性修正动量来优化对抗样本，避免陷入局部最优解，同时随机选择不同的模型分支进行迭代计算，提高对抗样本多样性，使其以高概率规避异构防御模型的检测与识别。

研究内容三拟提出低成本的集成防御模型构建方法，重点探索如何快速高效地优化集成模型，以低成本建立高鲁棒性防御机制，最终实现兼顾人工智能安全性与算力成本的目标。首先，以训练友好的方式生成高质量增广数据，即以研究任务一生成的扰动生成模型为基础，采取单步特征蒸馏技术初始化数据，并结合历史知识提升增广质量，减少计算复杂度，大幅度节省算力成本。然后，以资源集约的形式的学习多样化特征，即将研究任务二的多分支神经网络作为基础结构，增加特征的内部相似性和外部差异性，与传统方法相比，不需要额外神经网络传播计算，显著降低优化开销，以此提升防御模型的构建效率和安全性。

二、项目预期获得的研究成果及形式

论文及专著情况	国家统计源刊物以上刊物发表论文（篇）		4		科技报告（篇）		0	
	其中被SCI/EI/ISTP收录论文数（篇）		4		培养人才（人）		0	
	专著（册）		0		引进人才（人）		0	
专利情况(项)	发明专利		实用新型专利		外观设计专利		国外专利	
	申请	授权	申请	授权	申请	授权	申请	授权
	2	0	0	0	0	0	0	0
其他								

三、项目进度和阶段目标

（一）项目起止时间： 2023-11-01 至 2026-10-31		
（二）项目实施进度及阶段主要目标:		
开始日期	结束日期	主要工作内容
2023-11-01	2024-10-31	1. 研究面向跨域数据的对抗样本生成方法，提升对抗样本针对不同数据域模型的跨域迁移性，初步搭建迁移场景对抗攻防仿真平台，并完成验证测试。 2. 发表1篇学术论文；申请发明专利1项；参加1-2次国内外知名学术会议或者相关学术活动。
2024-11-01	2025-10-31	1. 研究面向异构模型的迁移对抗攻击方法，增强对抗攻击面对不同神经网络结构黑盒防御模型的逃逸能力，完成迁移场景对抗攻防仿真平台搭建工作。 2. 发表1-2篇学术论文；参加1-2次国内外知名学术会议或者相关学术活动。
2025-11-01	2026-10-31	1. 研究低成本的集成防御模型构建方法，降低集成防御模型在数据增广阶段和正则项优化阶段的算力资源开销，并在仿真平台完成对抗鲁棒性验证测试。 2. 发表1-2篇学术论文；申请发明专利1项；参加1-2次国内外知名学术会议或者相关学术活动；准备项目结题工作，汇总项目技术和相关研究成果并完成结题报告。

四、项目总经费及省基金委经费预算

（一）省基金委经费下达总额： （大写）壹拾万圆整；（小写 ）10万元;					
（二）省基金委经费年度下达计划：					
年度	2023 年	年	年	年	年
经费(万元)	10.00				

五、人员信息

项目负责人								
姓名	证件号码	年龄	性别	职称	学历	在项目中承担的任务	所在单位	签名
黄立峰	431202199002010815	34	男	讲师	博士研究生	项目负责人	华南农业大学	黄立峰

六、工作分工及财政经费分配

承担/参与单位名称（盖章）	工作分工	省级财政科技资金分配（万元）
华南农业大学	申请人独立完成。	10
	合计	10

七、任务书条款

第一条 甲方与乙方根据《中华人民共和国民法典》及国家有关法规和规定，按照《广东省科学技术厅关于广东省基础与应用基础研究基金（省自然科学基金、联合基金等）项目管理的实施细则（试行）》《省级科技计划项目任务书管理细则》《广东省省级科技计划项目验收结题工作规程（试行）》等规定，为顺利完成（2023）年面向跨域数据和异构模型的迁移场景对抗攻防方法研究专项项目（项目编号：2023A1515110075）经协商一致，特订立本任务书，作为甲乙双方在项目实施管理过程中共同遵守的依据。

第二条 甲方的权利义务：

1.按任务书规定进行经费核拨的有关工作协调。

2.根据甲方需要，在不影响乙方工作的前提下，定期或不定期对乙方项目的实施情况和经费使用情况进行检查或抽查。

3.根据《广东省科研诚信管理办法(试行)》等规定对乙方进行科技计划信用管理。

第三条 乙方的权利义务：

1.确保落实自筹经费及有关保障条件。

2.按任务书规定，对甲方核拨的经费实行专款专用，单独列账，并随时配合甲方进行监督检查。

3.经费使用按照广东省级财政科研项目经费使用等有关规定进行管理。

4.项目依托单位应制定经费使用“负面清单+包干制”内部管理制度并报甲方备案。

5.使用财政资金采购设备、原材料等，按照《广东省实施〈中华人民共和国招标投标法〉办法》有关规定，符合招标条件的须进行招标。

6.项目任务书任务完成后，或任务书规定的任务、指标及经费投入等提前完成的，乙方可提出验收结题申请，并按甲方要求做好项目验收结题工作。

7.若项目发生需要终止结题的情况，乙方须提出终止结题申请，并按甲方要求做好项目终止结题工作。

8.在每年规定时间内向甲方如实提交上年度工作情况报告，报告内容包含上年度项目进展情况、经费决算和取得的成果等。

9.按照国家和省有关规定，提交科技报告及其他材料。

10.利用甲方的经费获得的研究成果，项目负责人和参与者应当注明获得“广东省基础与应用基础研究基金（英文：Guangdong Basic and Applied Basic Research Foundation）（项目编号）”资助或作有关说明。

11.乙方要恪守科学道德准则，遵守科研活动规范，践行科研诚信要求，不得抄袭、剽窃他人科研成果或者伪造、篡改研究数据、研究结论；不得购买、代写、代投论文，虚构同行评议专家及评议意见；不得违反论文署名规范，擅自标注或虚假标注获得科技计划（专项、基金等）等资助；不得弄虚作假，骗取科技计划（专项、基金等）项目、科研经费以及奖励、荣誉等；不得有其他违背科研诚信要求的行为。

12.确保本项目开展的研究工作符合我国科研伦理管理相关规定。

第四条 在履行本任务书的过程中，如出现广东省相关政策法规重大改变等不可抗力情况，甲方有权对所核拨经费的数量和时间进行相应调整。

第五条 在履行本任务书的过程中，当事人一方发现可能导致项目整体或部分失败的情形时，应及时通知另一方，并采取适当措施减少损失，没有及时通知并采取适当措施，致使损失扩大的，应当就扩大的损失承担责任。

第六条 本项目技术成果的归属、转让和实施技术成果所产生的经济利益的分享，除双方另有约定外，按国家和广东省有关法规执行。

第七条 根据项目具体情况，经双方另行协商订立的附加条款，作为本任务书正式内容的一部分，与本任务书具有同等效力。

第八条 本任务书一式三份，各份具有同等效力。甲、乙方及项目负责人各执一份，三方签字、盖章后即生效，有效期至项目结题后一年内。各方均应负任务书的法律责任，不应受机构、人事变动的影响。

第九条 乙方必须接受甲方聘请的本项目任务书监理单位的监督和管理。监理单位按照甲方赋予的权利对本项目任务书的履行进行审核、进度调查，对项目任务书变更、经费使用情况进行监督管理及组织项目验收。

说明：1.本任务书中，凡是当事人约定无需填写的内容，应在空白处划（/）。

2.委托代理人签订本任务书的，应出具合法、有效的委托书。

八、本任务书签约各方

管理单位（甲方）：　　广东省基础与应用基础研究基金委员会　（盖章）

法定代表人（或法人代理）：　曾路　（签章）

2024 年　04 月　08 日

依托单位（乙方）：　华南农业大学　（盖章）

法定代表人（或法人代理）：　薛红卫　（签章）

联系人（项目主管）姓名：　倪慧群　（签章）

Email：　kjcgxk@scau.edu.cn

电话：　020-85283435　/　15920301530

开户单位名称：　华南农业大学

开户银行名称：　广东广州工行五山支行

开户银行账号：　3602002609000310520

2024 年 4 月 10 日

联系人（项目负责人）姓名：　黄立峰　（签名）

Email：　huanglf6@scau.edu.cn

电话：　13929500478

2024 年 4 月 10 日

任务书编号：2024A04J4382

广州市科技计划项目

任务书

项目名称：	面向自动驾驶视觉感知模型的可迁移对抗攻防方法研究
承担单位：	华南农业大学
项目负责人：	黄立峰
计划类别：	基础研究计划
专题名称：	2024年度基础与应用基础研究专题
支持方向：	青年博士“启航”项目
组织单位：	华南农业大学
起止时间：	2024-01-01 至 2025-12-31
主管处室：	基础研究处

广州市科学技术局制

二〇二四年

填写说明

1.任务书甲方为广州市科学技术局；乙方为项目承担单位；丙方为项目组织单位。

2.任务书基于项目申报书转换而成，请按照“广州科技大脑”提示在线填写核实，若存在不填写内容的栏目，请用“无”表示；任务书中的单位名称应为规范全称，并与单位公章一致。

3.乙方与合作单位的合作协议自动从项目申报书中读取，如需变化调整，须待任务书签订后，按要求及时办理重大变更。

4.乙方完成项目任务书在线填写，依次提交丙方和甲方审核确认后，按要求登录“穗好办”APP完成电子签章。不具备电子签章条件的单位，经与业务主管处室沟通对接后，可下载电子版项目任务书用A4纸双面打印装订签章；一式六份报甲方和丙方签章，其中甲方两份丙方两份，项目承担单位和项目负责人各一份。

5.涉密项目请在“广州科技大脑”下载项目任务书模板，按保密要求离线填写报送。

6.项目申报书是项目任务书填报的重要依据，未经甲方许可，乙方不得修改考核指标，调整主要研究内容。项目任务书将作为项目实施管理、验收结题和监督评估的重要依据。

7.项目任务书中的“备注”，包括重要的必须补充的内容。

8.“广州科技大脑”是项目管理过程中重要通知和文书的电子送达平台。为确保电子送达渠道畅通，乙方和项目负责人应及时更新维护“广州科技大脑”的单位和个人信息。

9.根据相关要求，项目涉及人体临床研究的，项目需经医学伦理委员会审查通过并在任务书附件栏上传相关佐证材料。

一、项目基本信息

项目基本信息	项目名称	面向自动驾驶视觉感知模型的可迁移对抗攻防方法研究
	申请市财政科技经费	5(万元)
	研究期限	2（年）
项目摘要	保障自动驾驶可靠性的核心在于利用对抗攻防来研究视觉感知模型的脆弱机理并基于此提高安全性。然而，对抗攻击在跨领域和跨结构时难以发现安全漏洞，而高鲁棒性模型算力成本高，阻碍其产业化落地。针对于此，本项目拟提升点云对抗样本跨领域迁移性，分析特征空间弱点；增强图像对抗攻击跨结构迁移能力，探索模型结构隐患；低成本构建高鲁棒性模型，突破资源消耗多的限制。本项目成果能助力自动驾驶发展与应用，具有重要的科学意义。	

二、项目单位情况

项目承担单位	单位名称	华南农业大学	统一社会信用代码	12440000455416563 4
	注册时间	1952-01-01	单位类型	高等院校
	注册地址	广东省广州市天河区五山路483号		
	办公地址	广东省广州市天河区五山路483号		
	联系人	姓名	倪慧群	
		手机号码	13711345768	
		电子邮箱	kjcgxk@scau.edu.cn	
	开户银行	广东广州工行五山支行		
	开户户名	华南农业大学		
	银行账号	3602002609000310520		

三、项目负责人信息

姓名	黄立峰	证件类型	身份证
证件号码	431202199002010815	性别	男
出生日期	1990-02-01	民族	侗族
国籍	中国	学历	博士研究生
学位	博士	学位授予国家（或地区）	中国
职务	教师	职称	无
所学专业	网络空间安全	手机号码	13929500478
办公电话	020-85285393	电子邮箱	huanglf6@scau.edu.cn

四、项目经费信息

本项目总投入：¥（5）万元，其中，市财政科技经费：¥（5）万元，自筹经费：¥（0）万元。

经费下达计划			
资金来源	小计	市财政科技经费	自筹经费
2024	5	5	0
总计	5	5	0

（单位：万元）

注：本专题纳入“包干制”，市财政科技经费按市科技计划项目经费“包干制”相关规定执行。

五、预期代表性成果

项目负责人在项目实施期内，以该项目作为资助项目获得以下5种情形之一且经费使用符合规定的，由组织单位审核后通过验收。

（一）项目实施期内，以第一作者/通讯作者发表论文1篇或以上（须标注资助项目编号）；

（二）项目实施期内，以第一完成人申请或授权专利、软件著作1项或以上；

（三）项目实施期内，获省级以上科技计划项目或人才项目支持1项或以上；

（四）项目实施期内，获省级以上科技奖励（含列入获奖团队成员名单）1项或以上；

（五）项目实施期内，获得职称晋升。

六、备注

专题补充约定条款：

甲方对未履行勤勉尽责义务的相关责任主体，自作出处理结论之日起，依照法律法规规定或任务书约定实施惩戒5年，取消相关责任主体申报市科技计划项目、申领市科技计划项目经费的资格。

预期代表性成果需在实施期内获得。

项目承担单位（乙方）及项目负责人承诺书

承诺书

本单位/本人作为广州市科技计划项目承担单位/项目负责人，将严格遵守广州市科技计划管理相关规定，严格履行自身责任，加强对项目组人员及合作单位的管理，在此郑重承诺：

（一）确保与本项目有关的全部材料真实、合法、有效，未侵犯其他方知识产权等权利，不存在多头申报、重复申报行为；

（二）严格遵守《广州市科技创新条例》《广州市科技计划项目管理办法》《广州市科技计划项目经费管理办法》《广州市科技计划科技报告管理办法》等相关规定，实施项目和经费管理；

（三）严格遵守国家、省、市关于科研诚信和科技伦理的有关法律、法规，相关政策以及各项规定，加强项目实施过程中的科研诚信及科技伦理管理，恪守科研道德准则。

如有违反，本单位/本人愿意接受相关部门做出的各项处理决定，包括但不限于终止项目、停拨经费、核减经费、追回经费，取消一定期限广州市科技计划项目申报资格，记入科研失信行为数据库，将不良行为向社会公开等。

项目承担单位：华南农业大学

日期：2023年12月18日

项目负责人：黄立峰

日期：2023年12月15日

任务书签署

甲乙丙三方根据《广州市科技计划项目管理办法》《广州市科技计划项目经费管理办法》《广州市科技计划科技报告管理办法》等有关文件规定，以及有关法律、政策和管理要求，签署本任务书。

签订地点：广州市越秀区

广州市科学技术局（甲方）：广州市科学技术局
局项目经办人：蒋韬略　　联系电话：83124150
责任处室负责人：麦胜文

2024年01月17日

项目承担单位（乙方）：华南农业大学
二级部门：华南农业大学数学与信息学院
项目负责人：黄立峰
项目经费汇入账号
账户名：华南农业大学　　账号：3602002609000310520
开户银行：广东广州工行五山支行
财务负责人：肖斐

2023年12月18日

组织单位（丙方）：华南农业大学
项目经办人：倪慧群

2023年12月18日

技术开发（委托）合同

项目名称：猪业二部药物智能仓储数字化项目

委托方（甲方）：温氏食品集团股份有限公司

受托方（乙方）：华南农业大学

签订时间：2025 年 11 月 30 日

签订地点：广东省云浮市

有效期限：2025 年 11 月 30 日至 2026 年 6 月 30 日

中华人民共和国科学技术部印制

填 写 说 明

一、本合同为中华人民共和国科学技术部印制的技术开发（委托）合同示范文本，各技术合同认定登记机构可推介技术合同当事人参照使用。

二、本合同书适用于一方当事人委托另一方当事人进行新技术、新产品、新工艺或者新材料及其系统的研究开发所订立的技术开发合同。

三、签约一方为多个当事人的，可按各自在合同关系中的作用等，在“委托方”、“受托方”项下（增页）分别排列为共同委托人或共同受托人。

四、本合同书未尽事项，可由当事人附页另行约定，并可作为本合同的组成部分。

五、当事人使用本合同书时约定无需填写的条款，应在该条款处注明“无”等字样。

技术开发（委托）合同

委托方（甲方）：温氏食品集团股份有限公司

住　所　地：广东云浮新兴

法定代表人：温志芬

项目联系人：谢启钊

联系方式 ：13826862926

通讯地址：广东省云浮市新兴县新城镇东堤北路 9 号

受托方（乙方）：华南农业大学

住　所　地：广州市天河区五山路 486 号

法定代表人：薛红卫

项目联系人：黄立峰

联系方式 ：13929500478

通讯地址：广州市天河区五山路 486 号 数学与信息学院 514

电话：13929500478

电子信箱：huanglf6@scau.edu.cn

本合同甲方委托乙方研究开发 猪业二部药物智能仓储数字化项目 项目，并支付研究开发经费和报酬，乙方接受委托并进行此项研究开发工作。双方经过平等协商，在真实、充分地表达各自意愿的基础上，根据《中华人民共和国民法典》的规定，达成如下协议，并由双方共同恪守。

第一条 本合同研究开发项目的要求如下：

1．技术目标：智能仓储推荐算法通过“数据驱动 + 智能优化”的方式，实现仓储作业的精细化和智能决策，从而显著提升仓储体系的空间利用率与准确性，降低执行时间、缺货率与积压率。

2．技术内容：（1）方案设计，包括需求分析与方案设计；（2）算法实现，包括同批次合并优先模块、历史高度偏好分析（人因工程优化）、距离最短优先（动线优化）、空货位分配策略、多货位拆分与结果输出；（3）系统联调测试；（4）上线准备与相关操作。

3．技术方法和路线：（1）同批次合并优先模块。实现先进先出（FIFO），防止混批，便于质量追溯。通过查找当前仓库中是否存在与本次入库物料编码相同、生产日期相同、保质期相同的已有存储单元；筛选出其中可用容量 ≥ 本次入库数量的候选货位。（2）历史高度偏好分析。多个候选货位均满足“同批次+容量足够”条件。统计该物料在历史入库记录中，出现在高区、中区、低区三个垂直层级的频次；构建频率分布向量，如：[低区: 60%, 中区: 30%, 高区: 10%]；优先推荐位于历史出现频率最高的高度层级的货位；若多个货位处于同一高频层级，则选择物理高度更低者（如低区优于中区）。（3）距离最短优先。从多个等效候选货位中推荐距离最短者。（4）空货位分配策略。系统通过抓取该物料过去一段时间的出入库记录，综合考虑它被入库和出库的频繁程度、每次操作的数量大小等因素，计算出一个“热度得分”，通过“热度”定义“黄金分区”、“冷区”。系统不仅关注单个物料的热度，还会分析哪些物料经常一起出现，从而让“好伙伴”尽量住得近一些。这里同时考虑两种协同关系，即通过分析历史入库单据与历史出库单据，识别出哪些物料经常在同一张入库单和出库单中出现。例如，某批原材料 A 和 B 总是由一同送达，原材料 B 和 C 一同出库，说明它们在供应端高度关联。这样在库位分配时，生成日期不同的 A/B/C 物料都相近分配。最终形成一个统一的“物料亲和

关系网络”。（5） 多货位拆分与结果输出。如果单个空货位无法容纳本次全部入库数量，系统会循环执行上述推荐逻辑，依次为剩余数量分配次优、再次优的货位，直到全部入库量分配完毕，并输出最终的货位分配清单。

第二条 乙方应在本合同生效后 30 日内向甲方提交研究开发计划。研究开发计划应包括以下主要内容：

1. 总体设计方案、开发计划、详细原型设计；
2. 项目开发过程、分阶段开发清单和进度；
3. /。

第三条 乙方应按下列进度完成研究开发工作：

1. 2025.12-2026.02 算法实现，包括同批次合并优先模块、历史高度偏好分析（人因工程优化）、距离最短优先（动线优化）、空货位分配策略、多货位拆分与结果输出；

2. 2026.02-2026.03 系统联调测试、上线准备。

第四条 甲方应向乙方提供的技术资料及协作事项如下：

1. 技术资料清单：整体方案、详细原型设计、技术开发接口文档。

2. 提供时间和方式：双方协商。

3. 其他协作事项：无。

本合同履行完毕后，上述技术资料按以下方式处理：交还甲方。

第五条 甲方应按以下方式支付研究开发经费和报酬：

1. 研究开发经费和报酬（含税）总额为 216500 元（大写：人民币贰拾壹万陆仟伍佰元整）。

2. 研究开发经费由甲方 分期 （一次、分期或提成）支付乙方。具体支付方式和时间如下：

（1） 合同签订后 15 个工作日内，出具相应项目方案设计通过甲

方确认后，由甲方向乙方一次性支付人民币壹拾万捌仟元整（￥108000）

（2） 项目结题验收通过后15个工作日内，由甲方向乙方一次性支付人民币壹拾万捌仟伍佰元整（￥108500）

乙方开户银行名称、地址和帐号为：

开户银行：中国工商银行广州五山支行

地址：广州市天河区五山路483号

帐号：3602002609000310520

3．双方确定，甲方以实施研究开发成果所产生的利益提成支付乙方的研究开发经费和报酬的，乙方有权以双方协商确认的方式查阅甲方有关的会计帐目。

第六条 本合同的研究开发经费由乙方以双方协商确认的方式使用。甲方有权以双方协商确认的方式检查乙方进行研究开发工作和使用研究开发经费的情况，但不得妨碍乙方的正常工作。

第七条 本合同的变更必须由双方协商一致，并以书面形式确定。但有下列情形之一的，一方可以向另一方提出变更合同权利与义务的请求，另一方应当在15日内予以答复；逾期未予答复的，视为同意：

1．无；

2．无。

第八条 未经甲方同意，乙方不得将本合同项目部分或全部研究开发工作转让第三人承担。但有下列情况之一的，乙方可以不经甲方同意，将本合同项目部分或全部研究开发工作转让第三人承担：

1．无；

2．无。

乙方可以转让研究开发工作的具体内容包括：无。

第九条 在本合同履行中，因出现在现有技术水平和条件下难以克服的技术困难，导致研究开发失败或部分失败，并造成一方或双方损失的，

双方按如下约定承担风险损失：双方协商确认。

双方确定，本合同项目的技术风险按双方协商确认的方式认定。认定技术风险的基本内容应当包括技术风险的存在、范围、程度及损失大小等。认定技术风险的基本条件是：

1. 本合同项目在现有技术水平条件下具有足够的难度；

2. 乙方在主观上无过错且经认定研究开发失败为合理的失败。

一方发现技术风险存在并有可能致使研究开发失败或部分失败的情形时，应当在15日内通知另一方并采取适当措施减少损失。逾期未通知并未采取适当措施而致使损失扩大的，应当就扩大的损失承担赔偿责任。

第十条 在本合同履行中，因作为研究开发标的的技术已经由他人公开（包括以专利权方式公开），一方应在15日内通知另一方解除合同。逾期未通知并致使另一方产生损失的，另一方有权要求予以赔偿。

第十一条 双方确定因履行本合同应遵守的保密义务如下：

甲方：

1. 保密内容（包括技术信息和经营信息）：涉及本合同的技术文件、资料和商业秘密。

2. 涉密人员范围：项目组成员。

3. 保密期限：本合同项下乙方的保密义务在保密信息合法公开前持续有效。未征得甲方事先的书面同意，乙方不得以任何形式向任何第三方披露全部或部分保密信息。

4. 泄密责任：双方协商解决。

乙方：

1. 保密内容（包括技术信息和经营信息）：涉及本合同的技术文件、资料和商业秘密。

2. 涉密人员范围：项目组成员。

3. 保密期限：合同期内。

4. 泄密责任：双方协商解决。

第十二条 乙方应当按以下方式向甲方交付研究开发成果：

1．研究开发成果交付的形式及数量：（1）设计方案1份；（2）入库推荐算法1个；（3）配合进行系统联调测试；（4）配合上线准备。

2．研究开发成果交付的时间及地点：按照甲方确定的项目详细技术开发规定履行。

第十三条 双方确定，按以下标准及方法对乙方完成的研究开发成果进行验收：乙方按照合同约定提交成果且甲方验收，即视为验收通过。

第十四条 乙方应当保证其交付给甲方的研究开发成果不侵犯任何第三人的合法权益。如发生第三人指控甲方实施的技术侵权，乙方应当退还研发经费。

第十五条 双方确定，因履行本合同所产生的研究开发成果及其相关知识产权权利归属，按下列第1种方式处理：

1．双（甲、乙、双）方享有申请专利的权利。

专利权取得后的使用和有关利益分配方式如下：双方协商约定。

2．按技术秘密方式处理。有关使用和转让的权利归属及由此产生的利益按以下约定处理：

（1）技术秘密的使用权：甲方所有

（2）技术秘密的转让权：甲方所有

（3）相关利益的分配办法：双方协商约定

。

双方对本合同有关的知识产权权利归属特别约定如下：双方所有

。

第十六条 乙方不得在向甲方交付研究开发成果之前，自行将研究开发成果转让给第三人。

第十七条 乙方完成本合同项目的研究开发人员享有在有关技术成果文件上写明技术成果完成者的权利和取得有关荣誉证书、奖励的权利。

第十八条 乙方利用研究开发经费所购置与研究开发工作有关的

设备、器材、资料等财产，归 乙 （甲、乙、双）方所有。

第十九条 双方确定，乙方应在向甲方交付研究开发成果后，根据甲方的请求，为甲方指定的人员提供技术指导和培训，或提供与使用该研究开发成果相关的技术服务。

1．技术服务和指导内容： 项目相关算法技术 。

2：地点和方式： 远程指导或现场指导 。

3．费用及支付方式： 免费 。

第二十条 双方确定：任何一方违反本合同约定，造成研究开发工作停滞、延误或失败的，按以下约定承担违约责任：

1． 甲 方违反本合同第 四 条约定，应当 尽快支付研发经费。如因此导致乙方交付时间延迟，乙方不承担责任 （支付违约金或损失赔偿额的计算方法）。

2． 甲 方违反本合同第 五 条约定，应当 尽快支付研发经费。如因此导致乙方交付时间延迟，乙方不承担责任 （支付违约金或损失赔偿额的计算方法）。

3． 乙 方违反本合同第 二 条约定，应当 尽快采取补救措施，向甲方支付因此造成的损失 （支付违约金或损失赔偿额的计算方法）。

4． 乙 方违反本合同第 三 条约定，应当 尽快采取补救措施，向甲方支付因此造成的损失 （支付违约金或损失赔偿额的计算方法）。

第二十一条 双方确定，甲方有权利用乙方按照本合同约定提供的研究开发成果，进行后续改进。由此产生的具有实质性或创造性技术进步特征的新的技术成果及其权属，由 甲 （甲、乙、双）方享有。具体相关利益的分配办如下： 甲方享有 100% 。

乙方有权在完成本合同约定的研究开发工作后，利用该项研究开发成果进行后续改进。由此产生的具有实质性或创造性技术进步特征的新的技术成果，归 乙 （甲、乙、双）方所有。具体相关利益的分配办法如下：

乙方享有 100%。

第二十二条 双方确定，在本合同有效期内，甲方指定 谢启钊 为甲方项目联系人，乙方指定 黄立峰 为乙方项目联系人。项目联系人承担以下责任：

1. 对项目实施过程有关疑难问题进行沟通协调；

2. 对相关技术范围内难以解决的事宜，进行汇报并督促落实。

一方变更项目联系人的，应当及时以书面形式通知另一方。未及时通知并影响本合同履行或造成损失的，应承担相应的责任。

第二十三条 双方确定，出现下列情形，致使本合同的履行成为不必要或不可能的，一方可以通知另一方解除本合同；

1. 因发生不可抗力或技术风险；

2. _____

第二十四条：双方因履行本合同而发生的争议，应协商、调解解决。协商、调解不成的，确定按以下第 1 种方式处理：

1. 提交 中国云浮新兴 仲裁委员会仲裁；

2. 依法向人民法院起诉。

第二十五条 双方确定：本合同及相关附件中所涉及的有关名词和技术术语，其定义和解释如下：

1. _____

第二十六条 与履行本合同有关的下列技术文件，经双方确认后，无 为本合同的组成部分：

1. 技术背景资料：无；

2. 可行性论证报告：无；

3. 技术评价报告：无；

4. 技术标准和规范：无；

5. 原始设计和工艺文件：无；

6. 其他：无。

第二十七条 双方约定本合同其他相关事项为：无。

第二十八条 本合同一式肆份，甲乙双方各执贰份，具有同等法律效力。

第二十九条 本合同经双方签字盖章后生效。

甲方：温氏食品集团股份有限公司（盖章）

法定代表人/委托代理人：（签名）

年 月 日

乙方：华南农业大学（盖章）

法定代表人/委托代理人：薛红卫（签名）

年 月 日

印花税票粘贴处：

HXKJHT20231336

合同编号：

技术开发（委托）合同

项目名称：人工智能系统模型应用部署安全性评估与分析

委托方（甲方）：广东一码通信科技有限公司

受托方（乙方）：华南农业大学

签订时间：2023 年 3 月 29 日

签订地点：广州市

有效期限：2023 年 4 月 10 日至 2023 年 10 月 10 日

中华人民共和国科学技术部印制

填 写 说 明

一、本合同为中华人民共和国科学技术部印制的技术开发（委托）合同示范文本，各技术合同认定登记机构可推介技术合同当事人参照使用。

二、本合同书适用于一方当事人委托另一方当事人进行新技术、新产品、新工艺或者新材料及其系统的研究开发所订立的技术开发合同。

三、签约一方为多个当事人的，可按各自在合同关系中的作用等，在“委托方”、“受托方”项下（增页）分别排列为共同委托人或共同受托人。

四、本合同书未尽事项，可由当事人附页另行约定，并可作为本合同的组成部分。

五、当事人使用本合同书时约定无需填写的条款，应在该条款处注明“无”等字样。

技术开发（委托）合同

委托方（甲方）：广州一码通信科技有限公司

住　所　地：广州市天河区华强路 9 号 1601 房

法定代表人：阳程

项目联系人：阳程

联系方式 ：13922101617

通讯地址：广州市天河区华强路 9 号 1601 房

电话：020-87778346　传真：

电子信箱：

受托方（乙方）：华南农业大学

住　所　地：广州市天河区五山路 486 号

法定代表人：刘雅红

项目联系人：黄立峰

联系方式 ：13929500478

通讯地址：广州市天河区五山路 486 号 数学与信息学院 514

电话：13929500478　传真：

电子信箱：huanglf6@scau.edu.cn

本合同甲方委托乙方研究开发人工智能系统模型应用部署安全性评估与分析项目，并支付研究开发经费和报酬，乙方接受委托并进行此项研究开发工作。双方经过平等协商，在真实、充分地表达各自意愿的基础上，根据《中华人民共和国合同法》的规定，达成如下协议，并由双方共同恪守。

第一条　本合同研究开发项目的要求如下：

1．技术目标：随着人工智能技术的快速发展，多种人工智能模型已经集

成和部署在了实际的应用软件中。但是，人工智能模型的安全性存在较大的隐患，容易受到恶意攻击的干扰，导致最终做出错误的推理和预测。因此，项目通过对人工智能视觉系统模型（深度神经网络）的预测推理过程进行鲁棒性评估，分析模型的安全漏洞，以此提升人工智能模型安全性。

2．技术内容：

（1）设计日常应用场景下的人工智能视觉系统模型的鲁棒性分析方案；

（2）评估模型在面对对抗样本攻击时的鲁棒性和稳定性，并归纳测试基准。

3．技术方法和路线：

（1）设计对抗攻击方法以生成对抗样本数据，测试人工智能视觉系统模型预测推理的结果正确性，搜索模型的脆弱性漏洞；

（2）针对发现的安全漏洞，提出高效的对抗训练方法；

（3）提升人工智能视觉系统在面对恶意攻击时的健壮性和稳定性。

第二条 乙方应在本合同生效后 30 日内向甲方提交研究开发计划。研究开发计划应包括以下主要内容：

1. 总体设计方案、开发计划、各阶段研发文档；
2. 项目开发过程、分阶段开发清单和进度；
3. _____。

第三条 乙方应按下列进度完成研究开发工作：

1. 2023.04-2023.07 完成人工智能模型的安全性测试方案；
2. 2023.07-2023.10 设计人工智能模型的鲁棒性提升方案；
3. _____。

第四条 甲方应向乙方提供的技术资料及协作事项如下：

1．技术资料清单：人工智能系统模型说明文档。

2．提供时间和方式：项目开始两周内，通过电子邮件发送文档 。

3．其他协作事项：无。

本合同履行完毕后，上述技术资料按以下方式处理：乙方自行保管 。

第五条 甲方应按以下方式支付研究开发经费和报酬：

1．研究开发经费和报酬总额为 60000 元（大写：人民币陆万元整）。

2．研究开发经费由甲方 一次 （一次、分期或提成）支付乙方。具体支付方式和时间如下：

（1） 合同签订后 15 个工作日内，由甲方向乙方一次性支付人民币陆万元整（￥60000）

（2） 付款单位：广东一码通信科技有限公司

乙方开户银行名称、地址和帐号为：

开户银行： 中国工商银行广州五山支行

地址： 广州市天河区五山路 483 号

帐号： 3602002609000310520

3．双方确定，甲方以实施研究开发成果所产生的利益提成支付乙方的研究开发经费和报酬的，乙方有权以 双方协商确认 的方式查阅甲方有关的会计帐目。

第六条 本合同的研究开发经费由乙方以 双方协商确认 的方式使用。甲方有权以 双方协商确认 的方式检查乙方进行研究开发工作和使用研究开发经费的情况，但不得妨碍乙方的正常工作。

第七条 本合同的变更必须由双方协商一致，并以书面形式确定。但有下列情形之一的，一方可以向另一方提出变更合同权利与义务的请求，另一方应当在 15 日内予以答复；逾期未予答复的，视为同意：

1． ；

2． 。

第八条 未经甲方同意，乙方不得将本合同项目部分或全部研究开发工作转让第三人承担。但有下列情况之一的，乙方可以不经甲方同意，将本合同项目部分或全部研究开发工作转让第三人承担：

1． ；

2． 。

乙方可以转让研究开发工作的具体内容包括： 无

_____

_____。

第九条 在本合同履行中，因出现在现有技术水平和条件下难以克服的技术困难，导致研究开发失败或部分失败，并造成一方或双方损失的，双方按如下约定承担风险损失：互相承担风险，互不追究责任。

双方确定，本合同项目的技术风险按双方协商确认的方式认定。认定技术风险的基本内容应当包括技术风险的存在、范围、程度及损失大小等。认定技术风险的基本条件是：

1. 本合同项目在现有技术水平条件下具有足够的难度；

2. 乙方在主观上无过错且经认定研究开发失败为合理的失败。

一方发现技术风险存在并有可能致使研究开发失败或部分失败的情形时，应当在15日内通知另一方并采取适当措施减少损失。逾期未通知并未采取适当措施而致使损失扩大的，应当就扩大的损失承担赔偿责任。

第十条 在本合同履行中，因作为研究开发标的技术已经由他人公开（包括以专利权方式公开），一方应在15日内通知另一方解除合同。逾期未通知并致使另一方产生损失的，另一方有权要求予以赔偿。

第十一条 双方确定因履行本合同应遵守的保密义务如下：

甲方：

1. 保密内容（包括技术信息和经营信息）：涉及本合同的技术文件、资料和商业秘密。

2. 涉密人员范围：项目组成员。

3. 保密期限：合同期内。

4. 泄密责任：双方协商解决。

乙方：

1. 保密内容（包括技术信息和经营信息）：涉及本合同的技术文件、资料和商业秘密。

2. 涉密人员范围：项目组成员。

3. 保密期限：合同期内。

4．泄密责任：双方协商解决。

第十二条 乙方应当按以下方式向甲方交付研究开发成果：

1．研究开发成果交付的形式及数量：人工智能模型系统安全性测试报告一份。

2．研究开发成果交付的时间及地点：按照甲方确定的项目详细技术开发规定履行。

第十三条 双方确定，按以下标准及方法对乙方完成的研究开发成果进行验收：乙方按照合同约定提交成果即视为验收通过。

第十四条 乙方应当保证其交付给甲方的研究开发成果不侵犯任何第三人的合法权益。如发生第三人指控甲方实施的技术侵权，乙方应当退还研发经费。

第十五条 双方确定，因履行本合同所产生的研究开发成果及其相关知识产权权利归属，按下列第1种方式处理：

1．乙（甲、乙、双）方享有申请专利的权利。

专利权取得后的使用和有关利益分配方式如下：乙方享有100%。

2．按技术秘密方式处理。有关使用和转让的权利归属及由此产生的利益按以下约定处理：

（1）技术秘密的使用权：双方协商约定

（2）技术秘密的转让权：双方协商约定

（3）相关利益的分配办法：双方协商约定

双方对本合同有关的知识产权权利归属特别约定如下：双方所有。

第十六条 乙方不得在向甲方交付研究开发成果之前，自行将研究开发成果转让给第三人。

第十七条 乙方完成本合同项目的研究开发人员享有在有关技术成果文件上写明技术成果完成者的权利和取得有关荣誉证书、奖励的权利。

第十八条 乙方利用研究开发经费所购置与研究开发工作有关的设备、器材、资料等财产，归乙（甲、乙、双）方所有。

第十九条 双方确定，乙方应在向甲方交付研究开发成果后，根据甲方的

请求，为甲方指定的人员提供技术指导和培训，或提供与使用该研究开发成果相关的技术服务。

1．技术服务和指导内容：项目相关技术。

2：地点和方式：远程指导。

3．费用及支付方式：免费。

第二十条 双方确定：任何一方违反本合同约定，造成研究开发工作停滞、延误或失败的，按以下约定承担违约责任：

1．甲方违反本合同第四条约定，应当尽快提交。如因此导致乙方交付时间延迟，乙方不承担责任（支付违约金或损失赔偿额的计算方法）。

2．甲方违反本合同第五条约定，应当尽快支付研发经费。如因此导致乙方交付时间延迟，乙方不承担责任（支付违约金或损失赔偿额的计算方法）。

3．乙方违反本合同第二条约定，应当尽快采取补救措施，向甲方支付因此造成的损失（支付违约金或损失赔偿额的计算方法）。

4．乙方违反本合同第三条约定，应当尽快采取补救措施，向甲方支付因此造成的损失（支付违约金或损失赔偿额的计算方法）。

第二十一条 双方确定，甲方有权利用乙方按照本合同约定提供的研究开发成果，进行后续改进。由此产生的具有实质性或创造性技术进步特征的新的技术成果及其权属，由乙（甲、乙、双）方享有。具体相关利益的分配办如下：乙方享有100%。

乙方有权在完成本合同约定的研究开发工作后，利用该项研究开发成果进行后续改进。由此产生的具有实质性或创造性技术进步特征的新的技术成果，归乙（甲、乙、双）方所有。具体相关利益的分配办法如下：乙方享有100%。

第二十二条 双方确定，在本合同有效期内，甲方指定阳程为甲方项目联系人，乙方指定黄立峰为乙方项目联系人。项目联系人承担以下责任：

1．对项目实施过程有关疑难问题进行沟通协调；

2．对相关技术范围内难以解决的事宜，进行汇报并督促落实。

一方变更项目联系人的，应当及时以书面形式通知另一方。未及时通知并影响本合同履行或造成损失的，应承担相应的责任。

第二十三条 双方确定，出现下列情形，致使本合同的履行成为不必要或不可能的，一方可以通知另一方解除本合同：

1．因发生不可抗力或技术风险；

第二十四条：双方因履行本合同而发生的争议，应协商、调解解决。协商、调解不成的，确定按以下第 1 种方式处理：

1．提交 中国广州 仲裁委员会仲裁；

2．依法向人民法院起诉。

第二十五条 双方确定：本合同及相关附件中所涉及的有关名词和技术术语，其定义和解释如下：

1．无

第二十六条 与履行本合同有关的下列技术文件，经双方确认后，无 为本合同的组成部分：

1．技术背景资料：无；

2．可行性论证报告：无；

3．技术评价报告：无；

4．技术标准和规范：无；

5．原始设计和工艺文件：无；

6．其他：无；

。

第二十七条 双方约定本合同其他相关事项为：无。

第二十八条 本合同一式肆份，具有同等法律效力。

第二十九条 本合同经双方签字盖章后生效。

甲方：广东一码通信科技有限公司（盖章）

法定代表人/委托代理人：阳程（签名）

年　　月　　日

乙方：华南农业大学（盖章）

法定代表人/委托代理人：刘雅红（签名）

年　　月　　日

印花税票粘贴处：

（以下由技术合同登记机构填写）

合同登记编号：

1．申请登记人：_____

2．登记材料：（1）_____

（2）_____

（3）_____

3．合同类型：_____

4．合同交易额：_____

5．技术交易额：_____

技术合同登记机构（印章）

经办人：

年　　月　　日

Collaboration Agreement
合作协议

This Collaboration Agreement (the "Agreement") is entered into by and between South China Agricultural University, a P.R.C. registered university ("UNIVERSITY"), and Google Information Technology (China) Co., Ltd. ("Google"). This Agreement will be effective as of the date signed by Google below (the "Effective Date").
本合作协议（"协议"）由以下双方签订：华南农业大学，一所在中华人民共和国注册的大学（"大学"）；以及谷歌信息技术（中国）有限公司（"谷歌"）。本协议将从谷歌在下面签署的日期起开始生效（"生效日期"）。

背景信息
Background

The history of the UNIVERSITY can be traced back to the Guangdong Provincial Agricultural Experimental Field and Affiliated Agricultural Training Center, which was founded in 1909. In 1952, during the restructuring of colleges and universities across the country, South China Agricultural College was established by merging parts of the College of Agriculture of Sun Yat-sen University, the College of Agriculture of Lingnan University, and the Department of Animal Husbandry and Veterinary Medicine and the Department of Pest and Diseases of the College of Agriculture of Guangxi University, under the supervision of the Ministry of Agriculture, in 1984, it was renamed South China Agricultural University.
大学办学历史可追溯至始创于1909年的广东全省农事试验场暨附设农业讲习所。1952年，在全国高校院系调整时，由中山大学农学院、岭南大学农学院和广西大学农学院畜牧兽医系及病虫害系的一部分合并成立华南农学院，隶属农业部主管；1984年，更名为华南农业大学。

Google's mission is to organize the world's information and make it universally accessible and useful. Google seeks to support the development of computer science education in universities.
谷歌的使命是组织全球的信息，并使这些信息能够为公众所访问和使用。谷歌致力于支持高等院校计算机科学教育的发展。

Collaborating with IT companies (like Google) will be beneficial for the UNIVERSITY's computer science education. Collaboration with IT companies brings the latest technology into UNIVERSITY's curriculum and helps UNIVERSITY develop the skills that meet industry demand.
与 IT 公司（例如谷歌）合作将有利于大学的计算机科学教育，可将最新产业技术纳入课程教学、并帮助大学培养满足行业需求的技术技能。

With the appropriate coordination efforts, Google wants to work with the UNIVERSITY to implement the project: Lightweight and low-cost robust pest and disease detection technology and Android application development.
谷歌期望通过适当的合作，与大学开展项目：轻量级和低成本的鲁棒病虫害检测技术与Android应用开发。

Details of the project objectives and contents are in Attachment B: Project Proposal.
有关项目的目标和具体内容，请参阅附件B：项目申报书。

Agreement
协议

1. Google Commitments
谷歌的承诺

   a. Subject to the terms of this Agreement, Google agrees to provide UNIVERSITY an amount not to exceed RMB20,000. Such funds will be disbursed in accordance with the provisions of Attachment A and upon UNIVERSITY's demonstrating, to Google's reasonable satisfaction, that it has achieved each of the stipulated milestones ("Milestones").
   在受本协议约束的前提下，谷歌同意向大学提供不超过人民币20,000元的金额。谷歌将依照附件 A 的规定，并在大学证明已令谷歌适当满意地实现了所规定的各个里程碑（"里程碑"）后，支付该等资金。

   b. Google does not commit to any expenditure, technical assistance or branding beyond what is laid out in this Agreement.
   谷歌不承诺负责除本协议规定之外的任何开支、技术援助或品牌宣传。

2. UNIVERSITY Commitments
大学的承诺

   a. Upon the Effective Date, UNIVERSITY will appoint a full-time faculty member whose sole responsibility will be to coordinate the project. UNIVERSITY agrees to maintain such a role during the Term of the Agreement.
   自本协议生效日起，大学将委派一名全职教职人员专门负责本项目的协调工作。大学同意在协议期限内保留这一职务。

   b. UNIVERSITY will execute the project based on the plan in Attachment A.
   大学将根据附件 A 中的方案执行此项目。

   c. UNIVERSITY will report to Google about the funding usage status on a half year basis.
   大学将每半年向谷歌汇报一次资金使用情况。

   d. UNIVERSITY will provide Google with project status as per Google's request, especially the mid-year and year-end execution report.
   大学将根据谷歌提出的要求向谷歌提供项目状态，尤其是年中与年尾的执行报告。

3. Confidentiality
保密

   a. "Confidential Information" is information disclosed by one party to the other party under this Agreement that is marked as confidential or would normally under the circumstances be considered confidential information of the disclosing party. Confidential Information does not include information that the recipient already knew, that becomes public through no fault of the recipient, that was independently developed by the recipient, or that was rightfully given to the recipient by another

party.
“保密信息”是指一方根据本协议向另一方披露的、标记为保密信息或在相应情况下通常会被视作披露方保密信息的信息。保密信息不包括接收方已知的信息、非接收方错误而公开的信息、接收方独立开发的信息，或其他方通过合法途径提供给接收方的信息。

b. The recipient will not disclose the Confidential Information, except to affiliates, employees, and agents who need to know it and who have agreed in writing to keep it confidential. The recipient, its affiliates, employees, and agents may use Confidential Information only to exercise rights and fulfill obligations under this Agreement, while using reasonable care to protect it. The recipient may also disclose Confidential Information when required by law after giving reasonable notice to discloser.
除了需要知晓该信息且已书面同意对该信息保密的关联公司、员工和代理，接受方不得向其他人披露保密信息。接收方及其关联公司、员工和代理只能出于根据本协议行使权利和履行义务的目的而使用保密信息，同时须采取合理的谨慎态度来保护这些信息。接收方还可在法律要求时披露保密信息，但须先向披露方提供合理的通知。

4. Publicity. Neither party may make any public statement regarding the relationship contemplated by this Agreement without the other’s prior written approval, except when required by law after giving reasonable notice to the other.
公开。任何一方均不得在未经另一方事先书面同意的情况下就本协议所赋予的关系发表任何公开声明，除非法律有此要求，并且已向另一方提供了合理通知。

5. Warranties and Indemnities.
保证与赔偿。

a. Warranties. Each party represents and warrants that it has full power and authority to enter into and perform its obligations under this agreement.
保证。各方声明并保证其有充分的权利和权威来签订本协议并履行其相关义务。

b. Indemnities. To the fullest extent permitted by law, each party will indemnify and defend the other and their directors, officers, employees, and agents from third party claims arising from or related to a breach of such party’s representations and warranties.
赔偿。在法律允许的最大范围内，对于因一方违反其声明与保证而引起或与之相关的第三方索赔，该方须对另一方及其董事、高管、员工和代理进行赔偿，并保护其免受损害。

6. DISCLAIMERS
免责声明

UNIVERSITY EXPRESSLY UNDERSTANDS AND AGREES THAT:
大学明确理解并同意：

EXCEPT FOR THE EXPRESS WARRANTIES MADE BY THE PARTIES IN SECTION 5, GOOGLE EXPRESSLY DISCLAIMS ALL WARRANTIES OF ANY KIND, WHETHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NON-INFRINGEMENT.

除双方在第 5 条所作的明示保证外，谷歌明确拒绝任何类型的所有明示或默示保证，包括但不限于对适销性、针对某特定用途的适用性和非侵权性的默示保证。

7. REMEDIES, AND LIMITATION OF LIABILITY. TO THE MAXIMUM EXTENT PERMITTED BY LAW, EACH PARTY'S EXCLUSIVE REMEDY FOR BREACHES OF THIS AGREEMENT WILL BE MONETARY DAMAGES. EXCEPT FOR THE INDEMNITIES UNDER SECTION 5 AND BREACHES OF A PARTY'S INTELLECTUAL PROPERTY RIGHTS (INCLUDING LICENSE BREACHES), (A) NEITHER PARTY WILL BE LIABLE FOR LOST REVENUES OR INDIRECT, SPECIAL, INCIDENTAL, CONSEQUENTIAL, EXEMPLARY OR PUNITIVE DAMAGES, AND (B) NEITHER PARTY'S AGGREGATE LIABILITY FOR ANY CLAIM ARISING OUT OF OR RELATED TO THIS AGREEMENT WILL EXCEED $10,000.
救济与责任限制。在法律允许的最大限度内，各方针对本协议的违反可获得的唯一救济是金钱赔偿。除第 5 条规定的赔偿与侵犯一方知识产权（包括许可方面的违约）的情况外，(A) 任何一方均不对收入损失或者间接、特殊、偶发、继发、惩戒性或惩罚性损害负责；以及 (B) 任何一方对因本协议而引起或与之相关的任何索赔的总计责任均不超过 10,000 美元。

8. Entry Into Force, Duration and Termination, Miscellaneous Provisions
生效、期限和终止，其他规定

   a. Term: The effective date of this Agreement will be the date of Google's signature and the Agreement will be valid for a period of one year (the "Term").
   期限：本协议的生效日期为谷歌签名的日期，本协议的有效期为 1 年（"期限"）。

   b. Termination
   终止

      (i) Either party may terminate this Agreement immediately upon written notice to the other party if the other party is in material breach of this Agreement and has failed to cure such breach within 30 days after receiving notice from the first party identifying the breach.
      在下列情况下，任何一方均可在书面通知另一方后立即终止本协议：如果另一方实质性违反了本协议，并在收到首先发现其违约的一方的通知后的 30 天内未对此类违约进行补救。

      (ii) Either party may terminate this Agreement immediately upon written notice to the other party if the other party is unable to meet its obligations under this Agreement for more than 30 days due to force majeure.
      在下列情况下，任何一方均可在书面通知另一方后立即终止本协议：如果另一方因不可抗力而无法履行本协议所规定义务的时间超过 30 天。

      (iii) Google may terminate this Agreement immediately upon written notice to UNIVERSITY if UNIVERSITY breaches Section 5 (Representations and Warranties) or Section Section 8d (Compliance with Anti-Bribery Laws).
      如果大学违反第 5 条（声明与保证）或第8条第d款（遵守反贿赂法律）的规定，那么谷歌可以在书面通知大学后立即终止本协议。

      (iv) Survival. The following provisions will survive any termination or expiration of this Agreement: Sections 3 through 8.
      继续有效。以下条款在本协议终止或期满后将继续有效：第 3 条到第 8 条。

c. Notices. All notices of termination or breach must be in writing and addressed to the other party's Legal Department. The email address for notices being sent to Google's Legal Department is legal-notices@google.com. Notice will be treated as given on receipt, as verified by written or automated receipt or by electronic log (as applicable). All other notices must be in English, in writing and addressed to the other party's primary contact.
通知。有关协议终止或违约的所有通知均必须采取书面形式，并发送至另一方的法务部。谷歌法务部接收通知的电子邮件地址为：legal-notices@google.com。当书面或自动回执或者是电子记录予以确认之时，视作通知已送达。其他所有通知均必须采取英文书面形式,并发送给另一方的主要联系人。

d. Compliance with Anti-Bribery Laws. UNIVERSITY will comply with all applicable campaign finance and gift laws and anti-bribery laws, including the U.S. Foreign Corrupt Practices Act of 1977 and the UK Bribery Act of 2010, which prohibit corrupt direct or indirect offers of anything of value to anyone (including government officials), to obtain or keep business or to secure any other improper commercial advantage. UNIVERSITY will not (i) make any facilitation payments to induce government officials to perform otherwise required functions or (ii) directly or indirectly pay, offer, or agree to give any campaign contributions or gifts to government officials in connection with the program to implement. "**Government officials**" include any government employee; candidate for public office; and employee of government-owned or government-controlled companies, public international organizations, and political parties.
遵守反贿赂法律。反贿赂：大学将遵守所有适用的竞选资金和礼品法及反贿赂法律，包括1977年美国《反海外腐败法》和2010年英国《反贿赂法》。该等法律禁止直接或间接以贿赂手段向包括政府官员在内的任何人提供任何有价物，以获取或保留业务或取得任何其他不正当的商业好处。签约方将(i)不会支付任何疏通费诱使政府官员履行他们本应履行的职能；或 (ii) 也不会因要开展的项目而直接或间接向政府官员提供、支付或同意给予任何竞选献金或礼物。"**政府官员**"包括政府雇员、公共职位候选人、政府拥有或控制的公司和国际公共组织的雇员和政党。

e. Assignment. UNIVERSITY may not assign any part of this Agreement without the prior written consent of Google.
转让。大学不得在未经谷歌事先书面同意的情况下转让本协议的任何部分。

f. Force Majeure. Neither party will be liable for failure or delay in performance to the extent caused by circumstances beyond its reasonable control.
不可抗力。任何一方均无需对因超出其合理控制范围的情况而导致的未能履约或延迟履约行为负责。

g. No Waiver. Neither party will be treated as having waived any rights by not exercising (or delaying the exercise of) any rights under this Agreement.
非弃权。不得因一方未行使（或延迟行使）本协议所规定的任何权利而将其视作已放弃这些权利。

h. No Agency. This Agreement does not create any agency, partnership or joint venture between the parties.
非代理关系。本协议不在双方之间形成任何代理、合作或合资关系。

i. No Third Party Beneficiaries. This Agreement does not confer any benefits on any third party unless it expressly states that it does.

无第三方受益人。除非明确说明，否则本协议不向任何第三方授予任何利益。

j. Counterparts. The parties may execute this Agreement in counterparts, including facsimile, PDF, and other electronic copies, which taken together will constitute one instrument.
副本。本协议一式数份，包括传真件、PDF 版本和其他电子版本，各种文本共同构成一份协议。

k. Amendments. Any amendment must be in writing and expressly state that it is amending this Agreement.
修订内容。任何修改都必须以书面形式作出，并且明确说明修改了本协议之内容。

l. Entire Agreement. This Agreement sets out all terms agreed between the parties and supersedes all previous or contemporaneous agreements between the parties relating to its subject matter.
全部协议。本协议阐明了双方商定的所有条款，并将取代与本协议标的事宜相关的双方之间的所有先前或同期协议。

m. Severability. If any term (or part of a term) of this Agreement is invalid, illegal or unenforceable, the rest of the Agreement will continue in force unaffected.
可分割性。如果本协议的任何条款（或某条款的一部分）无效、不合法或无法执行，那么本协议的其余部分不受影响，将继续有效。

n. Governing Law.
适用法律

(i) Governing law. THIS AGREEMENT WILL BE GOVERNED BY THE LAWS OF THE PEOPLE'S REPUBLIC OF CHINA ("**PRC**"), EXCLUDING ITS CONFLICTS OF LAWS RULES.
适用法律。本协议将由中华人民共和国（"PRC"）法律管辖，但其冲突法规则除外。

(ii) Arbitration.
仲裁。

(1) Definitions. "**Dispute**" means any contractual or non-contractual dispute regarding this Agreement, including its formation, validity, subject matter, interpretation, performance, or termination.
定义。"争议"指关于本协议的任何合约性或非合约性争议，包括其构成、有效性、主题、解释、执行或终止。

(2) Settlement. The parties will try in good faith to settle any Dispute within 30 days after a party receives the first notice regarding the Dispute in accordance with Section 8.c (Notices). If the parties are unable to resolve the Dispute within this 30-day period, either party may refer the Dispute to arbitration in accordance with Section 8n(ii)(3) (Arbitration) below.
和解。双方将在收到根据第 8.c 条（通知）发出的有关争议的第一份通知后的 30 天内尽可能友善地解决任何争议。 如果双方无法在此30天内解决争议，任何一方均可根据下文第 8n(ii)(3) 条（仲裁）规定将争议提交仲裁。

(3) Arbitration. The parties will refer all Disputes to final, binding arbitration administered by the China International Economic and Trade Arbitration Commission ("**CIETAC**") in accordance with the CIETAC's Arbitration Rules in force as of this Agreement's Effective Date ("**Rules**"). The arbitration will be conducted in English by three arbitrators who will be appointed as follows: each party will appoint an arbitrator, and the party-appointed arbitrators will nominate a chairperson within 30 days after the confirmation of the last party-appointed arbitrator. If the party-appointed arbitrators fail to nominate a chairperson within 30 days after the confirmation of the last party-appointed arbitrator, CIETAC will nominate a chairperson. The chairperson may be appointed from outside CIETAC's panel of arbitrators in accordance with the Rules. The arbitration will be conducted in Beijing, PRC, which will be the seat of arbitration.
仲裁。双方将所有争议提交中国国际经济贸易仲裁委员会（**CIETAC**）根据其在本协议生效之日有效的仲裁规则（"**仲裁规则**"）进行仲裁，且裁决具有终局性，对双方均有约束力。仲裁将由根据以下规定指定的三位仲裁员用英语进行。协议各方各指定一位仲裁员，在较晚指定的仲裁员确认后30日内，双方指定的仲裁员将共同制定第三位仲裁员担任仲裁庭主席。若双方指定的仲裁员未能在该30日期间内指定仲裁庭主席，该主席将由中国国际经济贸易仲裁委员会指定。根据仲裁规则的规定，指定的该主席可能由中国国际经济贸易仲裁委员会仲裁员小组以外的人士担任。

(4) Confidentiality. The arbitration is Confidential Information (including the arbitration's existence and any oral or written information related to it). However, the parties may disclose to a competent court information necessary to execute any arbitral decision, but only if the confidentiality of those materials is maintained in those judicial proceedings.
保密。进行的仲裁（包括仲裁的存在和与仲裁有关的任何口头或书面信息）均属机密信息。不过，双方可在为执行仲裁裁决所需的限度内，且在在司法程序中须保持这些材料的机密性前提下，向有管辖权法院披露相关的保密信息。

(5) Non-Monetary Relief. The arbitrator(s) may only issue its award based on law, not in equity.
非金钱救济。仲裁员可能仅依据法律而非衡平法做出裁决。

(6) Fees and Expenses. Each party will bear its own lawyers' and experts' fees and expenses, regardless of the arbitrator's final decision regarding the Dispute.
费用与支出。本协议各方承担各自的律师费和专家费，无论仲裁庭对争议作出何种最终裁决。

The parties have executed this Agreement by persons duly authorized as of the Effective Date.
双方已于生效日期由合法授权代表签署本协议。

谷歌信息技术（中国）有限公司
GOOGLE INFORMATION TECHNOLOGY (CHINA) COMPANY LIMITED

Name: _____

Title: _____

Date: 2024.5.24

华南农业大学
South China Agricultural University

Name: _____

Title: _____

Date: _____

Attachment A
附件 A

Milestones
里程碑

Milestones **里程碑**	**Milestone Due Date** **里程碑到期日期**	**Payment Amount and Payment Date** **付款金额与付款日期**
1st Milestone: 第 1 个里程碑: ● Execution of Collaboration Agreement 签署合作协议。	May 31, 2024 2024年5月31日	N/A 不适用
2nd Milestone: 第 2 个里程碑 ● Kickoff the project development work. 立项启动项目开发任务。	Jun 1, 2024 2024年6月1日	N/A 不适用
3rd Milestone: 第 3 个里程碑 ● Funding: After UNIVERSITY appoints a full-time faculty member as the project coordinator, and submits invoice to Google, Google will pay the funding to UNIVERSITY. 资助款：在大学指定一名全职教职人员作为项目协调员及向谷歌提交发票之后，谷歌将先向大学支付资助款。	Jun 30, 2024 2024年6月30日	RMB20,000, subject to UNIVERSITY demonstrating to Google's reasonable satisfaction, that it has achieved the applicable Milestones 人民币20,000 元，前提是大学证明其已令谷歌适当满意地实现了相应的里程碑
4th Milestone: 第 4 个里程碑 ● Project checkpoint; 项目验收;	Oct. 31, 2025 2025年10月31日	N/A 不适用

项目批准号	62472182
申请代码	F0208
归口管理部门	
依托单位代码	51064208A0499-0932

国家自然科学基金

资助项目计划书

（预算制项目）

资助类别：面上项目

亚类说明：

附注说明：

项目名称：交通无线磁阻传感器网络深度学习去噪方法研究

直接费用：50万元　　执行年限：2025.01-2028.12

负 责 人：张足生　　BRID：03682.00.07960

通讯地址：广东省广州市天河区华南农业大学数学与信息(软件)学院510室

邮政编码：510642　　电　　话：020-85285383

电子邮件：zushengzhang@163.com

依托单位：华南农业大学

联 系 人：唐家林　　电　　话：020-85280070

填表日期：2024年08月26日

国家自然科学基金委员会制

Version：1.004.149

国家自然科学基金资助项目计划书填报说明
（预算制项目）

一、项目负责人收到《国家自然科学基金资助项目批准通知》（以下简称《批准通知》）后，请认真阅读本填报说明，参照国家自然科学基金相关项目管理办法和新修订的《国家自然科学基金资助项目资金管理办法》（以下简称《资金管理办法》，请查阅国家自然科学基金委员会官方网站首页“政策法规”栏目），按《批准通知》的要求认真填写和提交《国家自然科学基金资助项目计划书》（以下简称《计划书》）。

二、填写《计划书》时要科学严谨、实事求是、表述清晰、准确。《计划书》经国家自然科学基金委员会相关项目管理部门审核批准后，将作为项目研究计划执行、检查和验收的依据。

三、《计划书》各部分填写要求如下：

（一）简表：由系统自动生成。

（二）摘要及关键词：各类获资助项目都应当填写中、英文摘要及关键词。

（三）项目组主要成员：计划书中列出姓名的项目组主要成员由系统自动生成，与申请书原成员保持一致，不可随意调整。如果《批准通知》所附“项目评审意见及修改意见表”中“修改意见”栏目有调整项目组成员相关要求的，待项目开始执行后，按照项目成员变更程序另行办理。

（四）资金预算表：根据批准的项目资助额度，按规定调整项目预算，并按照《国家自然科学基金项目计划书预算表编制说明》填报资金预算表和预算说明书。

（五）正文：

1. 面上项目、地区科学基金项目：如果《批准通知》所附“项目评审意见及修改意见表”中“修改意见”栏目没有修改要求的，只需选择“研究内容和研究目标按照申请书执行”即可；如果《批准通知》中上述栏目明确要求调整研究期限或研究内容等的，须选择“根据研究方案修改意见更改”并填报相关修改内容。

2. 重点项目、重点国际（地区）合作研究项目、重大项目、重大研究计划重点支持项目、重大研究计划集成项目、国家重大科研仪器研制项目、联合基金项目、原创探索计划项目：须选择“根据研究方案修改意见更改”，根据《批准通知》的要求填写研究（研制）内容，不得自行降低、更改研究目标（或仪器研制的技术性能与主要技术指标、验收技术指标等）或缩减研究（研制）内容。此外，还要突出以下几点：

（1）研究的难点和在实施过程中可能遇到的问题（或仪器研制风险），拟采用的研究（研制）方案和技术路线；

（2）项目主要参与者分工，合作研究单位（如有）之间的关系与分工，重大项目还需说明课题之间的关联；

（3）详细的年度研究（研制）计划。

3. 创新研究群体项目：须选择“根据研究方案修改意见更改”，按下列提纲撰写：

（1）研究方向；

（2）结合国内外研究现状，说明研究工作的学术思想和科学意义（限两个页面）；

（3）研究内容、研究方案及预期目标（限两个页面）；

（4）年度研究计划；

（5）研究队伍的组成情况。

4. 基础科学中心项目：须选择“根据研究方案修改意见更改”，根据《批准通知》的要求和现场考察专家组的意见和建议，进一步完善并细化研究计划，按下列提纲撰写：

（1）五年拟开展的研究工作（包括主要研究方向、关键科学问题与研究内容）；

（2）研究方案（包括骨干成员之间的分工及合作方式、学科交叉融合研究计划等）；

（3）年度研究计划；

（4）五年预期目标和可能取得的重大突破等；

（5）研究队伍的组成情况。

5. 数学天元基金项目：天元前沿重点专项项目和数学与其他学科交叉联合资助项目，参照重点项目的方式进行选择和填写；其他类型项目，参照面上项目的方式进行选择和填写。

6. 对于其他类型项目，参照面上项目的方式进行选择和填写。

简表

项目负责人信息	姓　　名	张足生	性 别	男	出生年月	1980年09月	民 族	汉族
	学　　位	博士			职称	教授		
	是否在站博士后	否		电子邮件		zushengzhang@163.com		
	电　　话	020-85285383		个人网页				
	工 作 单 位	华南农业大学						
	所 在 院 系 所	数学与信息(软件)学院						
依托单位信息	名　　称	华南农业大学					代码	51064208A0499
	联 系 人	唐家林		电子邮件		kycjhk@scau.edu.cn		
	电　　话	020-85280070		网站地址		http://kjc.scau.edu.cn/		
合作单位信息	单 位 名 称							
项目基本信息	项 目 名 称	交通无线磁阻传感器网络深度学习去噪方法研究						
	资 助 类 别	面上项目			亚 类 说 明			
	附 注 说 明							
	申 请 代 码	F0208:物联网及其他新型网络			F0306:自动化检测技术与装置			
	基 地 类 别							
	执 行 年 限	2025.01-2028.12						
	直 接 费 用	50万元						

项目摘要

中文摘要：

无线磁阻传感器网络具有低功耗、低成本、微型化等特点，已广泛用于智能交通信息采集领域。磁阻传感器检测的“弱磁信号”容易受到多源干扰，目前的信噪分离方法无法提取高保真的车辆磁信号，导致交通检测精度急剧降低。如何在多源干扰的情况下实现车辆磁信号的高保真提取是一个亟待解决的关键问题，本项目将聚焦这一关键问题开展研究工作：首先对磁干扰信号和交通车辆磁信号分别进行数学建模，实现复杂环境下磁信号的生成；接着分析磁信号时间序列的特征规律，提出信噪辨识方法，将受噪声干扰的车辆磁信号识别出来；然后引入生成对抗网络框架，研究低信噪比条件下的深度学习去噪方法，实现高保真车辆磁信号的提取；并提出深度学习去噪模型压缩及分割部署方法，建立抗干扰的交通信息采集系统对上述理论研究的有效性进行验证。本项目将有效提高无线磁阻传感器网络在强干扰环境下检测精度，为低成本实时交通信息采集的大规模应用提供理论基础和技术支撑。

Abstract：

Wireless magnetic sensor networks, with their advantages of low power consumption, low cost and miniaturization, have been widely used in intelligent transportation information collection systems. "Weak magnetic signals" detected by magnetic sensors are susceptible to interference from multiple sources. Existing signal-to-noise separation methods are unable to extract high-fidelity vehicle magnetic signals, leading to a sharp decrease in traffic detection accuracy. How to achieve high-fidelity extraction of vehicle magnetic signals under multi-source interference is a key issue that urgently needs to be addressed. This project will focus on this key issue and carry out the following research work: Firstly, this project conducts separate mathematical modeling for magnetic interference signals and traffic vehicle magnetic signals, enabling the generation of magnetic signals in complex environments; Subsequently, we analyze the characteristic patterns of magnetic signal time series, propose methods for signal-to-noise identification, and distinguish vehicle magnetic signals affected by noise interference; Then, we introduce a generative adversarial network framework to investigate deep learning denoising methods under low signal-to-noise ratio conditions, achieving high-fidelity extraction of vehicle magnetic signals; Finally, a deep learning denoising model compression and segmentation deployment method is proposed, and an anti-interference traffic information collection system is established to verify the effectiveness of the above theoretical research. This project will effectively enhance the detection accuracy of wireless magnetic sensor networks in interference environments, providing both a theoretical foundation and technical support for the large-scale application of low-cost, real-time traffic information collection.

关键词(用分号分开)：传感器网络系统；物联网；磁干扰；交通检测；磁阻传感器

Keywords(用分号分开)： Sensor network systems; Internet of Things; Magnetic interference; Traffic detection; Magnetic sensor

项目组主要成员

编号	姓名	出生年月	性别	职称	学位	单位名称	电话	证件号码	项目分工	每年工作时间（月）
1	张足生	1980.09	男	教授	博士	华南农业大学	020-85285383	430422198009189038	项目负责人	10
2	黄立峰	1990.02	男	讲师	博士	华南农业大学	020-85285383	431202199002010815	信噪辨识算法设计	8
3	古万荣	1982.01	男	讲师	博士	华南农业大学	13719398640	440229198201270018	深度学习信号去噪	8
4	毛宜军	1979.12	男	讲师	博士	华南农业大学	020-85280320-511	41010519791206285X	深度学习模型压缩	8
5	肖克辉	1981.08	男	高级实验师	博士	华南农业大学	020-85285383	411502198108093039	磁信号建模与实验测试	8

总人数	高级	中级	初级	博士后	博士生	硕士生
9	2	3	0	0	0	4

国家自然科学基金预算制项目预算表

项目批准号：62472182　　项目负责人：张足生　　金额单位：万元

序号	科目名称	金额
1	一、科学基金资助项目直接费用合计	50.0000
2	1、 设备费	4.0000
3	其中：设备购置费	4.0000
4	2、 业务费	27.0000
5	3、 劳务费	19.0000
6	二、 其他来源资金	0.0000
7	三、 合计	50.0000

注：请按照项目研究实际需要合理填写各科目预算金额。

预算说明书

一、科学基金资助项目直接费用

请按照《国家自然科学基金项目计划书预算编制说明》等有关要求，按照政策相符性、目标相关性和经济合理性原则，实事求是编制项目预算。填报时，直接费用应按设备费、业务费、劳务费三个科目填报，每个科目结合科研任务按支出用途进行说明。

1. **设备费**（是指在项目实施过程中购置或试制专用仪器设备，对现有仪器设备进行升级改造，以及租赁外单位仪器设备而发生的费用。计算类仪器设备和软件工具可在设备费科目列支。填报时，应对设备费支出的必要性和测算的合理性等内容进行说明。单价大于50万元（含50万元）的设备需补充说明设备的主要性能指标、主要技术参数等内容;单价小于50万元的设备仅需按照设备购置费、试制改造费和租赁使用费分类进行说明即可。）

设备费4万元，说明如下：

项目购置一套交通磁阻传感器网络测试平台，用于动态和静态交通测试系统，每套平台单价3万元，共计3万元；购置2套太阳能充电装置（为网关及边缘节点提供电源），每套0.5万元，共计2*0.5=1万元整。

2. **业务费**（是指项目实施过程中消耗的各种材料、辅助材料等低值易耗品的采购、运输、装卸、整理等费用，发生的测试化验加工、燃料动力、出版/文献/信息传播/知识产权事务、会议/差旅/国际合作交流等费用，以及其他相关支出。）

业务费27万元，说明如下：

1）材料费11万元：主要用于购置传感器节点模具、磁阻传感器模块、无线通信模块、微处理器、聚合物锂电池、通信天线、电子标签、电子元器件等材料用于磁阻传感器节点、边缘节点、网关节点、手持移动节点等交通磁阻传感器网络所需节点的试制与研发，安排预算10万元，另外对数据存储硬盘、电脑配件等需预算1万元，共计11万元。

2）测试化验加工费2万元：委托相关企业对磁阻传感器节点、边缘节点、网关节点、手持移动节点等多类节点的模具、电路板、天线进行优化设计、测试与加工，需要预算2万元。

3）差旅费6万元：计划安排10人次参加国内外的学术或相关行业交流会议，每人次费用为0.5万元，共计5.0万元；项目组成员到相关的科研单位、企业交流所需交通费预算为1.0万元。

4）国际合作与交流费2万元：为了及时掌握相关领域的最新研究成果，项目实施期内拟邀请2位国际专家来华进行学术交流，来华专家每人预算1.0万元，安排预算2万元。

5）出版/文献/信息传播/知识产权事务费6万元：用于国际会议论文出版费2.0万元，国内外期刊论文出版费1.5万元，文献检索、资料印刷费及信息传播费预算0.5万元，发明专利申请费预算2.0万元，共计6万元。

3. **劳务费**（是指在项目实施过程中支付给参与项目研究的研究生、博士后、访问学者以及项目聘用的研究人员、科研辅助人员等的劳务性费用，以及支付给临时聘请的咨询专家的费用等。填报时，应综合考量劳务费支出对象所承担研究任务的必要性、投入本项目的工作时长、费用标准的合理性等因素。）

劳务费19万元，说明如下：

1）项目参与人员费16万元：参与项目研究及开发工作的研究生4人，每人10月/年，4年共计160人月，按平均每人月800元补助标准，预算13万元；项目需要聘用研究人员进行软/硬件开发，需要3人次，每人费用为0.5万/月，每人工作2个月，预算3万元。

2）专家咨询费3万元：拟邀请专家指导3人次/年，共12人次，每人次费用2500元，共计3万元。

二、其他来源资金

对其他来源资金的经费来源、主要用途、支出预算做简要说明。

无。

三、合作研究外拨资金

对合作研究单位承担研究任务及资金外拨情况等进行必要说明。如存在多个合作研究单位，请逐一说明。如存在资金外拨的情况，还需对外拨资金的金额进行重点说明。

无。

报告正文

研究内容和研究目标按照申请书执行。

国家自然科学基金项目负责人、依托单位承诺书

国家自然科学基金项目负责人承诺书

本人郑重承诺：我接受国家自然科学基金的资助，严格遵守中共中央办公厅、国务院办公厅《关于进一步加强科研诚信建设的若干意见》《关于进一步弘扬科学家精神加强作风和学风建设的意见》《关于加强科技伦理治理的意见》《科技伦理审查办法（试行）》等规定，和国家自然科学基金委员会关于资助项目管理、项目资金管理等各项规章，在《计划书》填写及项目执行过程中：

（一）按照《批准通知》《国家自然科学基金资助项目计划书填报说明》的要求填写《计划书》，未自行降低、更改目标任务或约定要求，或缩减研究（研制）内容；

（二）树立“红线”意识，严格履行科研合同义务，按照《计划书》负责实施本项目（批准号：62472182），切实保证研究工作时间，按时报送有关材料，及时报告重大情况变动，不违规将科研任务转包、分包他人，不以项目实施周期外或不相关成果充抵交差；

（三）遵守科研诚信、科技伦理规范和学术道德，认真开展研究工作，对资助项目发表的论著和取得的研究成果按规定进行标注，不在非本项目资助的成果或其他无关成果上标注本项目批准号，反对无实质学术贡献者“挂名”，不在成果署名、知识产权归属等方面侵占他人合法权益，并如实报告本人及项目组成员发生的违背科研诚信要求的任何行为；

（四）尊重科研规律，弘扬科学家精神，严谨求实，追求卓越，反对浮夸浮躁、投机取巧，不人为夸大学术或技术价值，不传播未经科学验证的现象和观点；

（五）将项目资金全部用于与本项目研究工作相关的支出，并结合科研活动需要，科学合理安排项目资金支出进度；

（六）做好项目组成员的教育和管理，确保遵守以上相关要求。

如违背上述承诺，本人愿接受国家自然科学基金委员会和相关部门做出的各项处理决定。

项目负责人（签字）：

年　　月　　日

依托单位科研管理部门：	依托单位财务管理部门：
负责人（签章）： 年　　月　　日	负责人（签章）： 年　　月　　日

国家自然科学基金项目依托单位承诺书

我单位同意承担上述国家自然科学基金项目，将保证项目负责人及其研究队伍的稳定和研究项目实施所需的条件，严格遵守中共中央办公厅、国务院办公厅《关于进一步加强科研诚信建设的若干意见》《关于进一步弘扬科学家精神加强作风和学风建设的意见》《关于加强科技伦理治理的意见》《科技伦理审查办法（试行）》等规定，和国家自然科学基金委员会有关资助项目管理、项目资金管理、科研诚信管理和科技伦理管理等各项规定，并督促实施。

依托单位（公章）

年　　月　　日

国家自然科学基金资助项目签批审核表

本栏目由自然科学基金委填写	科学处审查意见： 负责人（签章）： 年　　月　　日
	科学部审查意见： 负责人（签章）： 年　　月　　日

受理编号：c25140500000666

项目编号：2025A1515011539

文件编号：粤基金字〔2025〕10号

广东省基础与应用基础研究基金项目

任务书

项目名称： 基于语言协同式特征解耦的无参考图像质量评价方法研究

项目类别： 广东省自然科学基金-面上项目

项目起止时间： 2025-01-01　至　2027-12-31

管理单位（甲方）： 广东省基础与应用基础研究基金委员会

依托单位（乙方）： 华南农业大学

通讯地址： 广东省广州市天河区五山路483号

邮政编码： 510642　　**单位电话：** 020-85283435

项目负责人： 周子涵　　**联系电话：** 18819472687

（广东科技微信公众号）　　（查看任务书信息）　　（受理纸质材料二维码）

广东省基础与应用基础研究
基金委员会
二〇二〇年制

填写说明

一、项目任务书内容原则上要求与申报书相关内容保持一致，不得无故修改。

二、项目承担单位通过广东省科技业务管理阳光政务平台下载项目任务书，按要求完成签名盖章后扫描上传到广东省科技业务管理阳光政务平台。

三、签名盖章说明。请分别在单位工作分工及经费分配情况页、人员信息页、签约各方页等地方按要求签字或盖章，签章不合规或错漏将不予受理。其中，人员信息页要求所有参与人员本人亲笔签名，代签或印章无效，漏签将不予受理。

四、本任务书自签字并加盖公章之日起生效，各方均应负本任务书的法律责任，不应受机构、人事变动影响。

五、根据《广东省科学技术厅广东省财政厅关于深入推进省基础与应用基础研究基金项目经费使用“负面清单+包干制”改革试点工作的通知》(粤科规范字〔2022〕2号)，2022年度及以后立项资助的全部省基金项目(包括省自然科学基金、省市联合基金、省企联合基金项目等)均适用“负面清单+包干制”，项目提交申请书和任务书时无需编制费用明细科目预算。

一、主要研究内容和要达到的目标

主要研究内容：

本项目从人类视觉感知特性和无参考质量评价需求出发，通过挖掘面向图像质量的语言先验，从跨模态质量评价数据的分析和生成、解耦模型的构建及其在图像质量评价中的应用三个层面进系统性研究，以提高无参考评价模型在复杂环境下的准确性和泛化性，为图像质量评价领域开辟新思路。本项目拟从以下三个方面展开研究：

（1）针对面向图像质量的内容-退化的文本描述，研究自动生成策略，构建质量描述文本生成新范式。通过设计多维度图像质量描述文本的生成模型，为内容-退化解耦和图像质量评价提供更为丰富和准确的数据支持。

（2）针对内容-退化特征分解，研究语言协同引导的图像内容-退化解耦模型。通过设计语言信息协同约束的对比重构解耦框架，并结合视觉-语言模型，实现有效的内容-退化特征解耦。

（3）针对基于内容-退化分解的图像质量评价，研究多模态内容-退化特征的动态交互聚合策略，通过建立内容-退化驱动的动态交互机制,以显式建模两者对质量的交互影响。通过建立跨模态多任务学习框架，并设计自适应机制，以增强评价模型在复杂真实场景下的准确性和泛化能力。

研究目标：

本项目针对真实复杂场景下的无参考图像质量评价问题，对语言协同式内容-退化解耦机制及其在图像质量评价中的应用展开深入研究，旨在推动图像质量评价技术的发展及其在计算机视觉领域的广泛应用，并丰富图像表达，从而提升图像恢复等任务性能。具体目标包括：

（1）建立质量描述文本生成新范式，并提出一个多维度图像质量描述文本的生成模型，为内容-退化解耦和图像质量评价任务提供数据支持。

（2）提出视觉语言模型引导的图像内容-退化解耦模型，设计语言信息协同约束的对比重构解耦框架，以实现内容和退化的有效解耦。

（3）结合迁移学习技术，提出基于内容-退化特征动态交互的图像质量评价方法。通过设计自适应策略提高图像质量评价模型在真实场景的准确性和泛化能力。

（4） 整合以上成果，提出一套完整有效的基于特征解耦的无参考质量评价方法或系统。

基于上述研究成果，本项目旨在形成科技报告一篇，申请专利1项，发表 JCR 二区及以上学术期刊和 CCF-A 类会议论文共4篇。协助培养硕士3名。

二、项目预期获得的研究成果及形式

<table>
<tr><td rowspan="3">论文及专著情况</td><td colspan="2">国家统计源刊物以上刊物发表论文（篇）</td><td colspan="2">3</td><td colspan="2">科技报告（篇）</td><td colspan="2">1</td></tr>
<tr><td colspan="2">其中被SCI/EI/ISTP收录论文数（篇）</td><td colspan="2">0</td><td colspan="2">培养人才（人）</td><td colspan="2">3</td></tr>
<tr><td colspan="2">专著（册）</td><td colspan="2">0</td><td colspan="2">引进人才（人）</td><td colspan="2">0</td></tr>
<tr><td rowspan="3">专利情况(项)</td><td colspan="2">发明专利</td><td colspan="2">实用新型专利</td><td colspan="2">外观设计专利</td><td colspan="2">国外专利</td></tr>
<tr><td>申请</td><td>授权</td><td>申请</td><td>授权</td><td>申请</td><td>授权</td><td>申请</td><td>授权</td></tr>
<tr><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
</table>

三、项目进度和阶段目标

（一）项目起止时间： 2025-01-01 至 2027-12-31

（二）项目实施进度及阶段主要目标:

开始日期	结束日期	主要工作内容
2025-01-01	2025-12-31	收集和整理相关资料，调研最新研究进展，完善实验方案。完善复合退化函数模型，收集和准备无主观标记的合成图像数据集和有主观分数的图像质量评价数据集。针对上述视觉图像数据，细化结构化思考链和提示文本内容，微调大语言模型以构建多维度质量描述语料库，重点包括内容和退化描述。基于该方法开发质量描述文本生成软件原型系统，发表高水平学术论文1-2篇，并申请专利或软件著作权。
2026-01-01	2026-12-31	完善自监督学习框架的设计，包括细化内容编/解码器和退化编/解码器的结构。细化面向语言模型和图像质量的对比损失函数和训练方案，进行视觉内容-退化解耦网络的训练和验证。汇总本阶段和上阶段成果，发表高水平学术论文1-2篇，并申请专利。
2027-01-01	2027-12-31	进一步对相关领域进行调研，完善视觉-语言多模态特征融合机制、内容和退化特征动态交互机制和动态质量分数回归机制，并验证有效性。结合之前成果，将语言先验驱动的内容-退化解耦模型应用在无参考图像质量评价中，完善针对多模态的多任务框架，细化损失函数及多阶段训练策略，并验证有效性。汇总本阶段研究成果，发表高水平学术论文2-3篇，并申请专利。总结该项目在理论方法研究，模型设计和实际应用效果等方面所取得的成果，并提出与此项目相关的后续研究内容和发展方向。

四、项目总经费及省基金委经费预算

1.省基金委经费下达总额： （大写）壹拾万圆整；（小写 ）10万元；					
2.省基金委经费年度下达计划：					
年度	2025 年	年	年	年	年
经费(万元)	10.00				

六、工作分工及财政经费分配

承担/参与单位名称（盖章）	工作分工	省级财政科技资金分配（万元）
华南农业大学	本项目团队将全面负责项目的开展和进行，对各个研究内容进行统筹规划，对各项任务和其科研成果进展进行全面的把控。本团队还将负责科研任务的相关数据收集和处理、模型设计、算法设计以及程序设计。	10.00
	合计	10.00

五、人员信息

项目负责人								
姓名	证件号码	年龄	性别	职称	学历	在项目中承担的任务	所在单位	签名
周子涵	342601199602060028	29	女	讲师	博士研究生	项目负责人	华南农业大学	

项目组主要成员								
姓名	证件号码	年龄	性别	职称	学历	在项目中承担的任务	所在单位	签名
黄立峰	431202199002010815	35	男	讲师	博士研究生	模型设计	华南农业大学	
钟灿琨	445202199502277730	30	男	讲师	博士研究生	算法设计	华南农业大学	
肖新杰	460006200005106817	25	男	未取得	本科	实验测试	华南农业大学	
李亮辉	440902199711303230	28	男	未取得	本科	实验测试	华南农业大学	
肖雨婷	420112200011282740	25	女	未取得	本科	数据管理和清洗	华南农业大学	

七、任务书条款

第一条 甲方与乙方根据《中华人民共和国民法典》及国家有关法规和规定，按照《广东省自然科学基金及联合基金项目管理实施细则》（粤科规范字〔2024〕5号）《省级科技计划项目任务书管理细则》（粤科规范字〔2022〕8号）等规定，为顺利完成（2025）年基于语言协同式特征解耦的无参考图像质量评价方法研究专项项目（项目编号：2025A1515011539）经协商一致，特订立本任务书，作为甲乙双方在项目实施管理过程中共同遵守的依据。

第二条 甲方的权利义务：1.按任务书规定进行经费核拨的有关工作协调。2.根据甲方需要，在不影响乙方工作的前提下，定期或不定期对乙方项目的实施情况和经费使用情况进行检查或抽查。3.根据《广东省科学技术厅科技计划项目科研诚信管理办法》（粤科规范字〔2024〕2号）《广东省基础与应用基础研究基金项目科研不端行为调查处理实施细则（试行）》（粤科规范字〔2023〕1号）等规定对乙方进行科技计划信用管理。

第三条 乙方的权利义务：1.确保落实自筹经费及有关保障条件。2.按任务书规定，对甲方核拨的经费实行专款专用，单独列账，并随时配合甲方进行监督检查。3.经费使用按照广东省级财政科研项目经费使用及省基金项目经费使用“负面清单+包干制”等有关规定进行管理。4.项目依托单位应制定经费使用“负面清单+包干制”内部管理制度并报甲方备案。5.使用财政资金采购设备、原材料等，按照《广东省实施〈中华人民共和国招标投标法〉办法》有关规定，符合招标条件的须进行招标。6.项目任务书任务完成后，或任务书规定的任务、指标及经费投入等提前完成的，乙方可提出验收结题申请，并按甲方要求做好项目验收结题工作。7.若项目发生需要终止结题的情况，乙方须提出终止结题申请，并按甲方要求做好项目终止结题工作。8.在每年规定时间内向甲方如实提交上年度工作情况报告，报告内容包含上年度项目进展情况、经费决算和取得的成果等。9.按照国家和省有关规定，提交科技报告及其他材料。10.利用甲方的经费获得的研究成果，项目负责人和参与者应当注明获得“广东省基础与应用基础研究基金（英文：Guangdong Basic and Applied Basic Research Foundation）（项目编号）”资助或作有关说明。11.乙方要恪守科学道德准则，遵守科研活动规范，践行科研诚信要求，不得抄袭、剽窃他人科研成果或者伪造、篡改研究数据、研究结论；不得购买、代写、代投论文，虚构同行评议专家及评议意见；不得违反论文署名规范，擅自标注或虚假标注获得科技计划（专项、基金等）等资助；不得弄虚作假，骗取科技计划（专项、基金等）项目、科研经费以及奖励、荣誉等；不得有其他违背科研诚信要求的行为。12.确保本项目开展的研究工作符合我国科技伦理管理相关规定。

第四条 在履行本任务书的过程中，如出现广东省相关政策法规重大改变等不可抗力情况，甲方有权对所核拨经费的数量和时间进行相应调整。

第五条 在履行本任务书的过程中，当事人一方发现可能导致项目整体或部分失败的情形时，应及时通知另一方，并采取适当措施减少损失，没有及时通知并采取适当措施，致使损失扩大的，应当就扩大的损失承担责任。

第六条 本项目技术成果的归属、转让和实施技术成果所产生的经济利益的分享，除双方另有约定外，按国家和广东省有关法规执行。

第七条 根据项目具体情况，经双方另行协商订立的附加条款，作为本任务书正式内容的一部分，与本任务书具有同等效力。

第八条 本任务书一式三份，各份具有同等效力。甲、乙方及项目负责人各执一份，三方签字、盖章后即生效，有效期至项目结题后一年内。各方均应负任务书的法律责任，不应受机构、人事变动的影响。

第九条 乙方必须接受甲方聘请的本项目任务书监理单位的监督和管理。监理单位按照甲方赋予的权利对本项目任务书的履行进行审核、进度调查，对项目任务书变更、经费使用情况进行监督管理及组织项目验收。

说明：1.本任务书中，凡是当事人约定无需填写的内容，应在空白处划（/）。

2.委托代理人签订本任务书的，应出具合法、有效的委托书。

八、本任务书签约各方

管理单位（甲方）：　　广东省基础与应用基础研究基金委员会　（盖章）

法定代表人（或法人代理）：　曾路　（签章）

2025 年　03 月　21 日

依托单位（乙方）：　华南农业大学　（盖章）

法定代表人（或法人代理）：　薛红卫　（签章）

联系人（项目主管）姓名：　夏杰　（签章）

Email：　kjcgxk@scau.edu.cn

电话：　020-85283435　/　13711345768

开户单位名称：　华南农业大学

开户银行名称：　广东广州工行五山支行

开户银行账号：　3602002609000310520

2025 年 4 月 9 日

联系人（项目负责人）姓名：　周子涵　（签名）

Email：　joannezzh@outlook.com

电话：　18819472687

25 年 4 月 3 日

受理编号：c24140500001452

项目编号：2024A1515010950

文件编号：粤基金字〔2024〕7号

广东省基础与应用基础研究基金项目

任务书

项目名称： 物联网场景中面向无服务器边缘计算架构的工作流调度与资源优化

项目类别： 广东省自然科学基金-面上项目

项目起止时间： 2024-01-01 至 2026-12-31

管理单位（甲方）： 广东省基础与应用基础研究基金委员会

依托单位（乙方）： 华南农业大学

通讯地址： 广东省广州市天河区五山路483号

邮政编码： 510642 **单位电话：** 020-85283435

项目负责人： 罗浩宇 **联系电话：** 13022092227

（广东科技微信公众号） （查看任务书信息） （受理纸质材料二维码）

广东省基础与应用基础研究
基金委员会
二〇二〇年制

填写说明

一、项目任务书内容原则上要求与申报书相关内容保持一致，不得无故修改。

二、项目承担单位通过广东省科技业务管理阳光政务平台下载项目任务书，按要求完成签名盖章后扫描上传到广东省科技业务管理阳光政务平台。

三、签名盖章说明。请分别在单位工作分工及经费分配情况页、人员信息页、签约各方页等地方按要求签字或盖章，签章不合规或错漏将不予受理。其中，人员信息页要求所有参与人员本人亲笔签名，代签或印章无效，漏签将不予受理。

四、本任务书自签字并加盖公章之日起生效，各方均应负本任务书的法律责任，不应受机构、人事变动影响。

五、根据《广东省科学技术厅广东省财政厅关于深入推进省基础与应用基础研究基金项目经费使用“负面清单+包干制”改革试点工作的通知》(粤科规范字〔2022〕2号)，2022年度及以后立项资助的全部省基金项目(包括省自然科学基金、省市联合基金、省企联合基金项目等)均适用“负面清单+包干制”，项目提交申请书和任务书时无需编制费用明细科目预算。

一、主要研究内容和要达到的目标

本项目将无服务器技术与边缘计算框架结合，利用无服务器技术的事件驱动、资源按需加载和自动扩容等特征，为物联网应用的服务质量保障提供新的解决思路。考虑到无服务器计算架构中容器实例冷启动带来的影响以及物联网场景中业务请求调用频率差异大且具有短时突发性等问题，本项目研究边缘计算框架下的无服务器工作流执行优化方法，通过对容器化服务全生命周期的管理和工作流任务的优化调度，在物联网场景中有限计算资源的前提下提高对工作流请求的响应能力。研究内容由以下四个部分组成：

（1）冷启动优化的FaaS函数放置方法，包括无服务器工作流调用模式感知与预测、生命周期感知的自适应FaaS容器预热和重用混合策略。

（2）面向突发负载的区域化资源自动伸缩和放置策略，包括基于统计学习的区域突发负载预测方法和基于排队理论的资源自动伸缩和放置。

（3）面向分布式排队网络的工作流调度算法。

（4）面向无服务器边缘计算环境的工作流执行优化框架设计及验证。

本项目拟达到的研究目标包括方法创新、关键技术突破和方法有效性评估三个方面，具体如下：

（1）边缘计算架构下无服务器工作流执行优化方法创新。针对传统边缘计算架构在处理物联网场景中的工作流应用请求时资源利用率较低和服务质量难以保障的困境，以无服务器计算模式在基础设施管理方面独特的优势为契机，以FaaS容器冷启动优化和并发工作流任务调度及资源管理为基本手段，形成无服务器边缘计算架构下工作流执行优化新方法，有效缓解当前物联网场景中有限的计算资源及能量供应与物联网应用的服务质量高要求之间的矛盾。

（2）关键技术突破。突破面向FaaS函数链的容器冷启动优化技术，在减少冷启动所带来的响应时延的同时，提高系统资源利用率；突破面向突发负载的区域化资源管理技术，实现局部计算资源的弹性伸缩以应对物联网场景中局部短时突发的处理需求；突破分布式排队网络中的并发工作流调度方法，在尽可能少的资源消耗的前提下提高物联网平台的响应能力，保障服务质量。

（3）原型系统开发与方法有效性评估验证。以上述关键技术为核心，设计面向无服务器边缘计算环境的工作流执行优化框架并开发系统原型。确立多维综合的无服务器工作流执行优化方法和关键技术的评估验证体系，能够客观说明本项研究的方法和关键技术对“物联网场景中工作流应用的服务质量保障”任务的促进作用。

二、项目预期获得的研究成果及形式

<table>
<tr><td rowspan="3">论文及专著情况</td><td colspan="2">国家统计源刊物以上刊物发表论文（篇）</td><td colspan="2">5</td><td colspan="2">科技报告（篇）</td><td colspan="2">0</td></tr>
<tr><td colspan="2">其中被SCI/EI/ISTP收录论文数（篇）</td><td colspan="2">4</td><td colspan="2">培养人才（人）</td><td colspan="2">3</td></tr>
<tr><td colspan="2">专著（册）</td><td colspan="2">0</td><td colspan="2">引进人才（人）</td><td colspan="2">0</td></tr>
<tr><td rowspan="3">专利情况(项)</td><td colspan="2">发明专利</td><td colspan="2">实用新型专利</td><td colspan="2">外观设计专利</td><td colspan="2">国外专利</td></tr>
<tr><td>申请</td><td>授权</td><td>申请</td><td>授权</td><td>申请</td><td>授权</td><td>申请</td><td>授权</td></tr>
<tr><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
</table>

三、项目进度和阶段目标

（一）项目起止时间： 2024-01-01 至 2026-12-31		
（二）项目实施进度及阶段主要目标:		
开始日期	结束日期	主要工作内容
2024-01-01	2024-12-31	研究冷启动优化的FaaS函数放置方法，具体包括： ①优化研究方案并对其进行细化； ②研究无服务器工作流调用模式感知与预测； ③研究生命周期感知的自适应FaaS容器预热与重用混合策略。
2025-01-01	2025-12-31	研究面向突发负载的区域化资源自动伸缩和放置策略，具体包括： ①研究基于统计学习的区域突发负载预测方法； ②研究基于排队理论的资源自动伸缩和放置； ③并发工作流调度算法预研。
2026-01-01	2026-12-31	研究面向分布式排队网络的并发工作流调度算法，以及面向无服务器边缘计算环境的工作流执行优化框架设计及验证，具体包括： ①研究面向分布式排队网络的并发工作流调度算法； ②工作流执行优化框架设计。 ③原型系统开发

四、项目总经费及省基金委经费预算

1.省基金委经费下达总额： （大写）壹拾伍万圆整；（小写 ）15万元；					
2.省基金委经费年度下达计划：					
年度	2024 年	年	年	年	年
经费(万元)	15.00				

五、人员信息

项目负责人								
姓名	证件号码	年龄	性别	职称	学历	在项目中承担的任务	所在单位	签名
罗浩宇	360302198907132536	35	男	副教授	博士研究生	项目负责人	华南农业大学	

项目组主要成员								
姓名	证件号码	年龄	性别	职称	学历	在项目中承担的任务	所在单位	签名
彭超达	445221199009201918	34	男	副教授	博士研究生	面向分布式排队网络的并发工作流调度算法	华南农业大学	
黄立峰	431202199002010815	34	男	讲师	博士研究生	面向突发负载的区域化资源自动伸缩和放置	华南农业大学	
严思源	445302199412010333	30	男	未取得	本科	算法验证、原型系统开发	华南农业大学	
张金鹏	130427199809041934	26	男	未取得	本科	算法验证、原型系统开发	华南农业大学	

六、工作分工及财政经费分配

承担/参与单位名称（盖章）	工作分工	省级财政科技资金分配（万元）
华南农业大学	无	15.00
	合计	15.00

七、任务书条款

第一条 甲方与乙方根据《中华人民共和国民法典》及国家有关法规和规定，按照《广东省科学技术厅关于广东省基础与应用基础研究基金（省自然科学基金、联合基金等）项目管理的实施细则（试行）》《省级科技计划项目任务书管理细则》《广东省省级科技计划项目验收结题工作规程（试行）》等规定，为顺利完成（2024）年物联网场景中面向无服务器边缘计算架构的工作流调度与资源优化专项项目（项目编号：2024A1515010950）经协商一致，特订立本任务书，作为甲乙双方在项目实施管理过程中共同遵守的依据。

第二条 甲方的权利义务：

1.按任务书规定进行经费核拨的有关工作协调。

2.根据甲方需要，在不影响乙方工作的前提下，定期或不定期对乙方项目的实施情况和经费使用情况进行检查或抽查。

3.根据《广东省科研诚信管理办法(试行)》等规定对乙方进行科技计划信用管理。

第三条 乙方的权利义务：

1.确保落实自筹经费及有关保障条件。

2.按任务书规定，对甲方核拨的经费实行专款专用，单独列账，并随时配合甲方进行监督检查。

3.经费使用按照广东省级财政科研项目经费使用等有关规定进行管理。

4.项目依托单位应制定经费使用“负面清单+包干制”内部管理制度并报甲方备案。

5.使用财政资金采购设备、原材料等，按照《广东省实施〈中华人民共和国招标投标法〉办法》有关规定，符合招标条件的须进行招标。

6.项目任务书任务完成后，或任务书规定的任务、指标及经费投入等提前完成的，乙方可提出验收结题申请，并按甲方要求做好项目验收结题工作。

7.若项目发生需要终止结题的情况，乙方须提出终止结题申请，并按甲方要求做好项目终止结题工作。

8.在每年规定时间内向甲方如实提交上年度工作情况报告，报告内容包含上年度项目进展情况、经费决算和取得的成果等。

9.按照国家和省有关规定，提交科技报告及其他材料。

10.利用甲方的经费获得的研究成果，项目负责人和参与者应当注明获得“广东省基础与应用基础研究基金（英文：Guangdong Basic and Applied Basic Research Foundation）（项目编号）”资助或作有关说明。

11.乙方要恪守科学道德准则，遵守科研活动规范，践行科研诚信要求，不得抄袭、剽窃他人科研成果或者伪造、篡改研究数据、研究结论；不得购买、代写、代投论文，虚构同行评议专家及评议意见；不得违反论文署名规范，擅自标注或虚假标注获得科技计划（专项、基金等）等资助；不得弄虚作假，骗取科技计划（专项、基金等）项目、科研经费以及奖励、荣誉等；不得有其他违背科研诚信要求的行为。

12.确保本项目开展的研究工作符合我国科技伦理管理相关规定。

第四条 在履行本任务书的过程中，如出现广东省相关政策法规重大改变等不可抗力情况，甲方有权对所核拨经费的数量和时间进行相应调整。

第五条 在履行本任务书的过程中，当事人一方发现可能导致项目整体或部分失败的情形时，应及时通知另一方，并采取适当措施减少损失，没有及时通知并采取适当措施，致使损失扩大的，应当就扩大的损失承担责任。

第六条 本项目技术成果的归属、转让和实施技术成果所产生的经济利益的分享，除双方另有约定外，按国家和广东省有关法规执行。

第七条 根据项目具体情况，经双方另行协商订立的附加条款，作为本任务书正式内容的一部分，与本任务书具有同等效力。

第八条 本任务书一式三份，各份具有同等效力。甲、乙方及项目负责人各执一份，三方签字、盖章后即生效，有效期至项目结题后一年内。各方均应负任务书的法律责任，不应受机构、人事变动的影响。

第九条 乙方必须接受甲方聘请的本项目任务书监理单位的监督和管理。监理单位按照甲方赋予的权利对本项目任务书的履行进行审核、进度调查，对项目任务书变更、经费使用情况进行监督管理及组织项目验收。

说明：1.本任务书中，凡是当事人约定无需填写的内容，应在空白处划（/）。

2.委托代理人签订本任务书的，应出具合法、有效的委托书。

八、本任务书签约各方

管理单位（甲方）：　　广东省基础与应用基础研究基金委员会 （盖章）

法定代表人（或法人代理）：　　曾路　　（签章）

2024 年　05 月　22 日

依托单位（乙方）：　华南农业大学　　（盖章）

法定代表人（或法人代理）：　薛红卫　　（签章）

联系人（项目主管）姓名：　倪慧群　　（签章）

Email：kjcgxk@scau.edu.cn

电话：　020-85283435　/　15920301530

开户单位名称：　华南农业大学

开户银行名称：　广东广州工行五山支行

开户银行账号：　3602002609000310520

年　月　日

联系人（项目负责人）姓名：　罗浩宇　　（签名）

Email：haoyuluo@scau.edu.cn

电话：13022092227

年　月　日

中国南方电网
CHINA SOUTHERN POWER GRID

基于代码多模态分析与高阶关联特征的电力系统软件漏洞检测研究技术开发合同

CHINA SOUTHERN POWER GRID

合同编号：150000220240301O3XA00132
甲方：南方电网科学研究院有限责任公司
乙方：华南农业大学
签订地点：广州

委托方（甲方）：南方电网科学研究院有限责任公司

住 所 地：广东省广州市萝岗区科学城科翔路11号J1栋3、4、5楼及J3栋3楼

法定代表人（负责人）：吴宝英

开户行：中国建设银行广州南网中心支行

账　号：44001403304059888168

项目联系人：洪超

通讯地址：广东省广州市萝岗区科学城科翔路 11 号 J1 栋 3、4、5 楼及 J3 栋 3 楼

手　　机：18027371114

电　　话：18027371114

电子信箱：hongchao2@csg.cn

受托方（乙方）：华南农业大学

住 所 地：广东省广州市天河区五山路483号

法定代表人（负责人）：薛红卫

开户行：中国工商银行广州五山支行

账　号：3602002609000310520

项目联系人：邱少健

通讯地址：广东省广州市天河区五山路 483 号数学与信息学院

手　　机：15817190456

电　　话：15817190456

电子信箱：qiushaojian@scau.edu.cn

甲方委托乙方研究开发 基于代码多模态分析与高阶关联特征的电力系统软件漏洞检测研究 项目，并支付研究开发费用，乙方接受委托并进行此项研究开发工作。根据《中华人民共和国民法典》及相关法律法规的规定，双方经过平等协商，达成本合同，共同遵守。

1. 项目名称： 基于代码多模态分析与高阶关联特征的电力系统软件漏洞检测研究 。

2. 项目的目标、内容、指标要求和完成时间

2.1 目标： 详见附件《技术协议书》 。

2.2 内容： 详见附件《技术协议书》 。

2.3 指标要求： 详见附件《技术协议书》。

2.4 完成时间： 详见附件《技术协议书》。

3. 研究开发人员组成

乙方指派技术人员 邱少健、罗浩宇、黄立峰、程嘉濠、黄梦阳 组成研究开发团队，同时指派 邱少健 作为本项目的负责人。未经甲方同意，乙方不得随意更换研究开发人员。如果甲方认为乙方指派的研究开发人员不能胜任的，乙方应当及时更换。乙方指派的项目负责人和研究开发人员应实际参与本合同的技术开发工作。

4. 研究开发计划书

关于研究开发计划书，双方同意按以下第/款约定执行：

4.1 本合同不提交研究开发计划书。

4.2 乙方应在本合同生效之日向甲方提交研究开发计划书，研究开发计划书经甲方同意后，作为本合同附件。研究开发计划书的内容

应包括：/

5. 技术资料

5.1 乙方提出所需技术背景材料、技术资料及数据清单，甲方审核确认后提供。

5.2 乙方根据需要，可要求甲方补充必要的技术背景材料、技术资料及数据，甲方审核确认后提供。

5.3 双方联系人应在确认的资料移交清单上签字，清单应注明移交时间和方式。

5.4 乙方对于甲方提供的技术背景材料、技术资料、数据应当妥善保管，并只能用于本合同项下的技术研发工作。本合同履行完毕后，乙方应当将上述技术背景材料、技术资料、数据全部退还给甲方，并不得以复制、扫描等任何方式保存甲方的资料信息。

6. 甲方协作事项

6.1 调研：/。

6.2 办公条件：/。

6.3 其他：/。

7. 合同价款与支付

7.1 本合同价款，包括了乙方为完成本合同项下全部工作所需的全部费用，按以下第（1）种方式结算确定：

（1）固定价。本合同价款为人民币含税价小写70,000元（大写：柒万元整），税率为6%，其中，不含税价为小写66037.74元（大写：陆万陆仟零叁拾柒元柒角肆分）。

（2）暂定价。本合同价款暂定为人民币含税价小写/元（大写：/），税率为/）。最终合同价款按以下标准计算：/

（3）其他：/

7.2 双方同意本合同价款的支付按以下第（2）项约定执行：

（1）一次性支付

具体支付时间和方式为：/

（2）分期支付

具体支付时间和方式为：

分期	支付时间	支付条件	支付比例	支付金额
首付款	满足支付条件后 40 个工作日内	合同签订后 15 天内乙方向甲方提交《研究开发计划书》（即按合同附件二模板要求编制的《开放基金项目计划任务书》）。提供技术合同认定登记证明，并通过甲方的审查确认。	合同价款 33 %	23100 元（大写：贰万叁仟壹佰元）
进度款	满足支付条件后 40 个工作日内	乙方应按甲方要求时间（2025 年 6 月-9 月）向甲方提交中期检查交付物（包括：申请发明专利 1 项、收录或发表 SCI/EI/中文核心等级论文	合同价款 33 %	23100 元（大写：贰万叁仟壹佰元）

		1篇、中期技术报告1份），并通过甲方组织的中期检查。		
尾款	满足支付条件后40个工作日内	2026年7月31日前乙方完成合同项下所有工作，向甲方提交全部研究成果，并通过甲方组织的项目验收。	合同价款 34 %	23800 元 （大写：贰万叁仟捌佰元）

如本合同价款为暂定价款的，则上述 / 按最终结算价款计算支付，其他批次款则按本合同约定的暂定价款计算支付。

（3）其他：/

7.3 合同价款结算按第1种方式（1.转账/2.汇票/3.支票/4.其他：/）。如需使用商业汇票进行支付的，由款项支付方承担资金成本（买方付息贴现）。

乙方汇票开立信息如下：

汇票类型：☐银行承兑汇票 ☐商业承兑汇票

收款人全称：/

银行账户：/

开户银行：/

开户行行号：/

联系人：/

联系电话：/

7.4 乙方应在甲方付预付款前 40个工作 日开具收据等带有法

律效力的原始凭据，否则甲方有权迟延支付相应价款，乙方因此造成的损失由其自行承担。

乙方应在甲方支付首付款、进度款及尾款前 40 个工作 日开具等额的增值税专用发票，增值税专用发票上注明的银行账户应与本合同提供的相同，否则甲方有权迟延支付相应价款，乙方因此造成的损失由其自行承担。

如乙方不能开具增值税专用发票的，应向甲方提交相关证明材料并取得甲方书面同意后，可开具符合甲方要求的其他发票。

7.5 乙方收款账户信息如下：

账号：3602002609000310520

户名：华南农业大学

开户行：中国工商银行广州五山支行

7.6 乙方数字人民币对公钱包信息如下：

数字人民币钱包开户行：/

钱包名称：/

钱包编号：/

7.7 本合同项下乙方收款账户信息、乙方数字人民币对公钱包信息，以本合同第 7.5 条、7.6 条约定为准。乙方在签署本合同时，应对前述信息进行确认，本合同一经双方签署即视为乙方确认前述信息无误，甲方应按前述信息进行付款。乙方如需变更收款账户或数字人民币钱包信息的，应提前向甲方提出，并由双方签署变更协议后方可变更；否则，甲方仍应按本条约定的收款信息进行付款。甲方按本条

约定的收款信息付款后，视为已完成付款义务，因收款信息错误造成的全部损失由乙方自行承担，甲方不承担任何责任。

7.8 甲方增值税开票信息如下：

单位名称：南方电网科学研究院有限责任公司

税务识别号：914400005645564342

开户行：中国建设银行广州南网中心支行

银行账号：44001403304059888168

地址：广东省广州市萝岗区科学城科翔路 11 号 J1 栋 3、4、5 楼及 J3 栋 3 楼

联系电话：020-36625064

7.9 若乙方未能按上述要求开具增值税专用发票，或实际开票税率与上述要求不符，甲方有权根据实际收取的发票类型及税率从合同价款中直接扣除相应金额的合同款进行结算。甲方有权在任一批次的合同付款中直接扣减税款差额。若未支付合同款不足以弥补税款差额，乙方应将差额退还甲方。税款差额计算方式如下：

税款差额=承诺税率对应的增值税进项税额 - 实际开具发票的进项税额（若计算得出的扣款小于 0，则取 0）

其中：增值税进项税额=∑开票金额÷（1+税率）×税率

7.10 本合同项下不含税价格不因国家税率变化而变化，若在本合同履行期间遇国家税率调整，则价税合计相应调整，以开具发票的时间为准。

8. 履约担保与质量担保

8.1 履约担保

8.1.1 双方同意履约担保按以下第（1）种方式执行，如本合同价款为暂定价的，下列履约担保的数额按暂定价计算：

（1）本合同不提供履约担保。

（2）乙方于本合同生效之日起/日内，按本合同价款/%向甲方提供履约保函（☐ 银行保函 ☐ 保险保函），即人民币/元（大写/）。

（3）乙方于本合同生效之日起/日内，向甲方提供金额为人民币/元（大写：/）的履约保函（☐ 银行保函 ☐ 保险保函）。

8.1.2 乙方发生违约行为时，甲方有权按合同约定的违约金额在履约保函中兑付。

8.1.3 本合同履行完毕且乙方履行义务符合合同约定或甲方向乙方提交质量担保时，甲方应将保函退回乙方。

8.2 质量担保

8.2.1 双方同意质量担保按以下第（1）种方式执行，如本合同价款为暂定价的，下列质量担保的数额按最终结算价计算：

（1）本合同不提供质量担保。

（2）乙方向甲方申请结算时，按合同价款的/%向甲方提供质量保函（☐ 银行保函 ☐ 保险保函），即人民币/元（大写：/）。

（3）甲方在本合同约定的合同价款尾款支付中，扣留合同价款的/%作为质保金。

8.2.2 本合同质保期为/年，自/开始计算。如本合同质保

期内未发生质量问题或乙方提供的质保服务符合本合同约定质量标准的，甲方于质保期届满之日起 / 日内，向乙方无息退还质保金或质量保函。

如乙方在质保期内怠于履行质保义务的，甲方有权委托第三方代为履行，并有权用质保金或兑付质量保函予以支付所发生的费用，如有剩余，不再退还；如不足支付第三方代履行费用的，不足部分由乙方另行支付。

9. 研究开发成果的交付

乙方应按本合同约定向甲方交付最终的研究开发成果，交付方式如下：

9.1 交付时间：详见附件《技术协议书》。

9.2 交付形式：详见附件《技术协议书》。

9.3 交付数量：详见附件《技术协议书》。

9.4 交付地点：广州或甲方指定。

9.5 其他：/。

10. 验收

乙方在提交研究开发成果时应同时提出书面的验收申请，甲方应按照本条约定进行验收。

10.1 甲方应当在接到乙方书面验收申请后 30 天内组织验收。

10.2 验收标准：

10.2.1 按照本合同约定的要求。

10.2.2 其他：/。

10.3 验收费用的承担：乙方承担。

10.4 验收后的处理：

10.4.1 验收通过，甲方应出具验收合格的书面意见给乙方。

10.4.2 验收未通过，甲方应出具验收不合格的书面意见给乙方。乙方应当在 30 天内根据甲方的验收意见对研究开发成果进行修改和完善，并再次向甲方申请组织验收；若第二次验收仍未通过，甲方有权解除本合同，乙方应当按照本合同的约定承担违约责任。

10.5 甲方出具的验收合格书面意见，不能视为免除乙方对研究开发成果存在缺陷所应负的责任，如存在缺陷，乙方应免费予以解决。乙方不予解决的，甲方有权委托第三方解决，乙方应赔偿由此给甲方造成的一切损失。

11. 知识产权条款

11.1 本合同涉及的知识产权包括：在研发、采购、生产、销售、对外贸易等环节中所产生的作品（含计算机软件）、专利（含发明、实用新型、外观设计）、商标（含商业标识）、商业秘密（含技术秘密）、集成电路布图设计以及法律规定的其他客体。

11.2 知识产权归属

11.2.1 本合同项下研究开发成果知识产权的归属，按第（3）种约定执行：

（1）归甲方所有。

（2）归甲乙双方共同共有。

（3）其他：_

本合同项下研究开发成果知识产权（包括专利、专著等知识产权）归【甲乙双方共同共有】，不应出现第三方单位：

① 共有知识产权完成登记或注册并且取得相应证书之后，双方收益权比例如下：甲方【50%】、乙方【50%】。

② 双方均及时履行知识产权维护费缴纳义务时，双方的知识产权收益权分配比例保持不变。如一方未及时履行专利维护费缴纳的义务，另一方有权催告提醒未交方，如未交方明确答复不承担专利维护费，或在另一方催告提醒后25个工作日内仍未完成专利维护费缴纳的，视为未交方自愿放弃共有知识产权，且未交方应配合另一方做好专利权转让变更手续。

③ 如一方放弃共有知识产权并停止知识产权维护，应提前30个工作日通知对方，并配合另一方做好专利权转让变更手续。。

11.2.2 如本合同项下研究开发成果知识产权归甲乙双方共同共有的，使用规则如下：

（1）对本合同项目产生的技术成果，一方不同意申请知识产权的，另一方方不得申请知识产权。

（2）一方转让其共有知识产权申请权的，另一方享有以同等条件优先受让的权利；一方书面声明放弃其共有的知识产权申请权的，由未放弃一方单独申请。

（3）未经另一方书面同意，一方不能将项目产生的共有知识产权擅自转让、排他许可、独占许可、与第三人合作实施或者质押融资等，但可以自行实施或普通许可、开放许可。其中，普通许可、开放

许可净收益由甲乙双方另行协商分配，协商不成的按产权权益比例分配。

（4）未经另一方书面同意，一方不得赋予技术成果完成人或团队项目产生知识产权所有权或长期使用权。若另一方书面同意赋予技术成果完成人或团队项目产生知识产权所有权或长期使用权，甲乙双方应与技术成果完成人或团队签署三方书面协议，合理约定转化技术成果收益分配比例、转化决策机制、转化费用分担以及知识产权维持费用等。

（5）无论是否存在知识产权的许可和转让，甲乙双方及各自员工或学生拥有不可撤销的出于学术研究目的使用技术成果知识产权的权利。但涉及他方商业秘密的，须事先取得权利人的书面同意，否则不得实施。

（6）本合同期满或终止后，任何一方均有对本项目产生的知识产权进行改进或二次开发的权利，一方单独改进或二次开发的成果属于改进一方或二次开发方所有；另一方有同等条件下优先被许可或受让的权利。甲乙双方共同合作改进或二次开发的成果属于甲乙双方共有。

（7）任一方可以不定期（不限于项目期间）对任何涉嫌或实际侵犯本合同项目产生知识产权的行为采取一切合理必要措施，其他方应当为其采取上述措施提供合理的帮助。

（8）一方（及被赋所有权的技术成果完成人或团队）经同意后转让其共有部分知识产权的，同等条件下其他方享有优先受让该技术

成果知识产权的权利。任何一方有权以书面方式要求其他方（及被赋所有权的技术成果完成人或团队）不得就相同类似项目、相同类似技术和知识产权的许可或转让，与其竞争对手进行磋商或交易，但要求方须有合理证据证明其竞争关系。

11.3 关于本合同项下研究开发成果的学术发表，按第（2）种约定执行：

（1）未经甲方书面同意，乙方及其员工或学生等均不得使用本合同项下研究开发成果进行学术发表。

（2）在不违反本合同关于保密和知识产权规定，不影响知识产权有效性的情况下，甲乙双方均有权发表本项目的研究开发成果。任何一方在其发表的论文或出版物中应该明确标明另一方对本研究开发成果的贡献，以及其对本研究开发成果的资助情况。

11.4 关于本合同项下研究开发成果知识产权维护成本的承担，按以下第（4）种约定执行，包含了对研究开发成果知识产权的申请、复审、诉讼、年费等与维护知识产权有效性相关的费用成本：

（1）甲方承担。

（2）乙方承担。

（3）甲方双方共同承担。具体分配规则为：/

（4）其他：_

① 如本合同项下研究开发成果知识产权归【甲方所有】的，则上述知识产权维护成本由【甲方承担】。

② 如本合同项下研究开发成果知识产权归【甲乙双方共同共有】

的，则：共有知识产权完成登记或注册并且取得相应证书之前，因登记或注册而产生的所有费用均由【乙方承担】；共有知识产权完成登记或注册并且取得相应证书之后，双方均对共有知识产权的有效性维护承担义务，具体按以下约定执行：

共有知识产权完成登记或注册并且取得相应证书之后，年费由双方按年度轮流缴纳，第一年的年费由甲方缴纳，第二年的年费由乙方缴纳，以此类推，直至有效期满。

11.5 本合同研究开发需使用甲方或乙方已有知识产权的，甲方或乙方应当向对方提供该知识产权不侵犯第三方任何权利的担保。不能提供担保的，应提供该知识产权合法来源证明材料。如因此发生该知识产权侵权行为的，由该知识产权提供方承担所有责任。

12. 培训和服务

乙方向甲方交付研究开发成果后，免费为甲方相关人员提供 2 人次技术指导和培训，免费提供与使用该研究开发成果相关的技术服务，帮助甲方掌握研究开发成果。

13. 甲方权利和义务

13.1 甲方有权对乙方的研究开发工作和费用使用情况进行检查监督。

13.2 甲方协助乙方开展研究开发工作，及时向乙方提供有关资料数据，并及时解决乙方提出的需要甲方配合的相关事宜。

13.3 甲方应按合同约定及时、足额向乙方支付研究开发费用。

13.4 甲方应按合同约定及时对乙方提交的研究开发成果进行验

收。

13.5 其他：/。

14. 乙方权利和义务

14.1 乙方有权要求甲方按约支付研究开发费用，甲方拒不支付时，乙方有权中止研究开发工作。

14.2 乙方应按双方约定的开发进度完成开发工作，每一个阶段的工作完成后应报甲方认可方可进行下一阶段的工作。

14.3 乙方应当保证其交付给甲方的研究开发成果不侵犯任何第三人的合法权益，如第三方对研究开发成果主张权利的，由此产生的一切法律后果由乙方承担。

14.4 乙方应妥善保管甲方提供的技术资料、样品、设备等；在合同履行过程中，如发现继续工作对材料、样品或设备等有损坏危险时，应立即中止工作，并及时通知甲方。

14.5 乙方应支持甲方开展税前加计扣除工作，按甲方要求提供技术合同认定结果等相关材料。

14.6 乙方不得将本合同项目进行转包。未经甲方同意，乙方不得将本合同项目进行分包。

14.7 乙方应遵守《中华人民共和国网络安全法》《中华人民共和国密码法》以及信息安全技术网络安全等级保护（GBT 22239-2019）等法律、行政法规及规章的要求。如信息系统属于电力监控系统范畴，同时应遵守《电力监控系统安全防护规定》《国家能源局关于印发电力监控系统安全防护总体方案等安全防护方案和评估规范的通

知》（【2015】36号）、南方电网相关电力监控系统网络安全有关要求。如前述法律法规、规范性文件、南网公司制度有更新或修订的，以最新规定为准。

乙方选购的电力监控系统产品，应在合同中要求产品供应商提供产品符合《中华人民共和国网络安全法》《中华人民共和国密码法》以及信息安全技术网络安全等级保护（GBT 22239-2019）等法律法规要求，并在设备及系统全生命周期内对其负责。

列入国家网信部门制定、公布的网络关键设备和网络安全专用产品目录的产品，应按照《中华人民共和国网络安全法》和《中华人民共和国密码法》要求，经具备资格的机构检测认证合格，并提供相关证书。

可能影响国家安全的网络产品和服务，应按照《网络安全法》和《密码法》要求，通过国家网信部门会同国家密码管理部门等有关部门组织的国家安全审查。

14.8 其他：无

15. 保密义务

甲乙双方履行本合同应遵守如下保密义务，如甲乙双方签署了《保密协议》的，则保密义务按《保密协议》约定执行：

15.1 保密内容：包括但不限于因履行本合同而知悉的甲乙双方商业秘密、工作秘密、敏感信息及其他非公开的技术和经营信息、与本合同项下项目相关的知识产权信息、技术秘密等。

商业秘密是指在生产和经营活动中产生的不为公众知悉，影响

公司安全、经济利益，并经公司采取保密措施的经营信息和技术信息。

工作秘密是指泄露后会对甲乙双方工作带来被动和损害的内部敏感信息，包括但不限于有关工作内部方案、讨论记录、过程稿、征求意见等。

敏感信息内容包括但不限于：甲乙双方员工个人信息、公司运行管理数据、业务生产敏感数据、公司重要工作文件等。

15.2 涉密人员范围：参与实施本合同的甲乙双方全体人员。

15.3 保密期限：合同签订后至甲乙双方书面声明放弃该保密权利之日止。

15.4 对于由部分或全部保密信息构成的技术成果，任一方不得在任何国家或地区申请专利或其他知识产权，但原保密信息权利人书面同意的除外。

15.5 除本合同另有约定外，任何一方未经对方同意，不得在任何新闻、广告或其他宣传推广活动中使用对方的名称、标识或对方提供的主要人员的姓名。

15.6 任一方向国外申请专利（包括 PCT 专利申请）的，应当遵守我国相关法律法规的保密规定，接受保密审查。

15.7 泄密责任：乙方应于本合同项下项目结束后或合同解除后 5 日内向甲方退还甲方提供的相关资料，或经甲方同意后将相关资料全部销毁。如乙方未退还相关资料或违反本合同项下保密义务

的，应向甲方支付合同价款 20 %的违约金，还应赔偿甲方因此造成的损失。

15.8 任何一方披露以下信息，不构成对本合同项下保密义务的违反：

（1）该信息在不违反本合同保密义务和其他保密承诺的情况下已经被公开或为公众所知；

（2）该信息是本合同一方通过合法手段从第三方独立获得的，但该方明知第三方以违反保密义务的方式披露给自己的信息除外；

（3）能书面证明本合同一方从对方处获得保密信息之前已经熟知该信息，且知悉时尚未对对方承担任何保密义务；

（4）法律或者相关监管机构以及上级主管部门要求披露的信息。

16. 合同的变更和解除

16.1 在本合同履行过程中，经甲乙双方协商一致，可以对本合同的条款进行变更，不能就变更达成一致意见的，应当按照原合同条款执行。

16.2 乙方发生以下情形之一的，甲方有权解除本合同：

16.2.1 乙方丧失履约能力的。

16.2.2 乙方明确表示不能完成本合同约定的研究开发工作的。

16.2.3 乙方未按本合同的约定提交研究开发计划书超过 30 天的。

16.2.4 乙方未按合同约定提交研究开发成果超过 30 天的。

16.2.5 乙方将本合同项目转包的，或擅自进行分包的。

16.2.6 若乙方指派的负责人或技术人员未实际参与本合同开发工作或者乙方擅自更换上述人员，经甲方通知后 30 天内仍未纠正或已严重影响本合同履行的。

16.2.7 乙方研究开发进度不符合本合同约定，经甲方通知后 30 天内仍未纠正的。

16.2.8 乙方提交的研究开发成果经两次验收后仍不能通过验收的。

16.2.9 乙方将研究开发成果转让给第三方的。

16.2.10 其他可以解除合同的情形：/。

16.3 本合同甲方发生以下情形之一的，乙方有权解除本合同：

16.3.1 甲方迟延支付研究开发费用超过 90 天的。

16.3.2 甲方不审核确认或不提供乙方所需技术资料超过 45 天的。

16.3.3 其他可以解除合同的情形：/。

16.4 合同解除后之处理

16.4.1 因甲方原因解除本合同的，乙方已完成的开发成果应向甲方交付，乙方应退还甲方提交的基础资料，未支付的费用甲方不再支付，双方据实结算。甲方应同时按合同约定承担违约责任。

16.4.2 因乙方原因解除本合同的，乙方已完成的开发成果应向甲方交付，乙方应退还甲方提交的基础资料，退回甲方已支付而未履行部分的合同价款，并按照银行同期贷款利率向甲方支付该退回款项利息，未支付的费用甲方不再支付。乙方应同时按本合同约定向甲方

承担违约责任。

16.5 在合同履行过程中，合同项下的技术已经通过其它方式公开，甲方有权解除合同，合同未履行部分不再履行，其它事项由双方协商解决。

17. 违约责任

17.1 本合同生效后，甲乙双方均应当全面履行合同义务。任何一方违约，均应当承担违约责任，并赔偿对方由此受到的损失。

17.2 因甲方违约导致乙方解除合同的，甲方除按乙方已开展工作支付相应费用外，还应按照研究开发费用总额的10%向乙方支付违约金，违约金不足以弥补乙方损失的，甲方应赔偿乙方未能弥补的损失。

17.3 甲方迟延支付研究开发费用或未按照本合同的约定提供技术资料或没有完成协作事项而造成乙方研究开发工作停滞、延误的，乙方交付研究开发成果的时间相应顺延。

17.4 甲方如未按照本合同的约定支付研究开发费用的，每迟延一日，按照迟延支付研究开发费用的0.5‰向乙方支付违约金。

17.5 因乙方违约导致甲方解除合同的，乙方除应退还甲方已支付的费用、并按照银行同期贷款利率向甲方支付利息外，还应按照研究开发费用总额的10%向甲方支付违约金，违约金不足以弥补甲方损失的，乙方应赔偿甲方未能弥补的损失。

17.6 乙方迟延交付研究开发成果的，每迟延一日，应按照研究开发费用总额的0.5‰向甲方支付违约金。

17.7 乙方指派的负责人或技术人员未实际参与本合同研究开发工作或者擅自更换的，乙方应向甲方支付研究开发费用10%的违约金。乙方拒不按甲方要求改正这一违约行为的，甲方有权解除合同。

17.8 若由于乙方原因导致信息系统（包括电力监控系统）含有预置的安全漏洞、恶意代码，或导致网络故障、病毒感染、网络安全事件发生，由乙方全责承担，乙方应当立即进行整改，赔偿甲方因此遭受的损失，并向甲方支付合同价款10%的违约金。

17.9 乙方在质保期内不履行质保义务或履行质保义务不符合合同约定的，乙方应向甲方支付合同价款10%的违约金，该违约金不足以弥补甲方损失的，乙方还须赔偿甲方未得到弥补的损失。

18. 通讯与联络

18.1 为方便开展工作，提高双方的工作效率，甲方安排洪超负责与乙方保持日常联系，乙方安排邱少健负责与甲方保持日常联系。如双方确有必要更换联系人员时，应以书面形式提前通知另一方。甲方工作人员的联系方式是18027371114；乙方工作人员的联系方式是15817190456。

18.2 双方履行合同的有关事项，按照上述约定通知到对方联系人的，视为完成通知送达。

18.3 双方的通讯地址或者联系方式如发生变动，应书面通知对方，因未及时通知而造成的损失由其自行承担。

19. 风险责任的承担

19.1 在履行本合同的过程中，确因在现有水平和条件下难以克

服的技术困难，导致研究开发工作部分或全部失败，风险责任由甲乙双方平均分担。

19.2 本项目技术风险确认的方式为：通过专家评审确定。

20. 不可抗力

20.1 不可抗力事件是指合同各方在签署本合同时指不能预见、不能避免并不能克服的客观情况。包括但不限于：火山爆发、龙卷风、海啸、暴风雪、泥石流、山体滑坡、水灾、火灾、地震、台风、雷电等，以及政府行为、核辐射、战争、瘟疫、骚乱等。

20.2 如任何一方因不可抗力事件不能履行本合同，应立即通知对方，并在不可抗力事件发生后 10 日内向对方提供有关该事件的公证文书或书面说明。

20.3 受不可抗力影响的一方应尽最大努力履行合同并减少因不可抗力给对方造成的损失。如因其未尽最大努力而给对方造成损失的，应承担相应赔偿责任。

20.4 如果发生影响本合同履行的不可抗力事件，则双方应及时协商制定并实施补救计划和合理的替代措施，减少或消除不可抗力事件的影响。

20.5 不可抗力影响合同履行超过 30 日的，双方均有权解除合同，由此产生的损失由双方各自承担。

21. 廉洁条款

21.1 合同双方应严格遵守国家关于市场准入、招标采购、工程建设等市场经济活动的法律法规、政策及廉洁规定，不得为获取不正

当利益，损害国家、集体和合同双方权益。

21.2 甲方（包括甲方工作人员及其特定关系人，下同）应遵守廉洁规定，不得利用职权或者职务上的影响谋取不正当利益，包括但不限于不得索取或收受乙方（包括乙方及其委托人、代理人、中间人等相关单位，以及上述单位的工作人员及其特定关系人，下同）的礼品、礼金、回扣、有价证券等财物，以及其他非财产性利益；不得借用乙方的钱款、住房、车辆等；不得参加乙方安排的可能影响公正执行公务的宴请、旅游、健身、娱乐等活动；不得要求或接受乙方为个人装修住房、婚丧嫁娶及亲属工作安排等提供便利；不得向乙方介绍亲属或其他特定关系人参与可能获取不正当利益的经济活动；不得向乙方泄漏涉及有关业务活动的秘密。

21.3 乙方应遵守廉洁规定，不得利用本合同项下业务合作便利谋取不正当利益，包括但不限于不得向甲方提供或赠送礼品、礼金、回扣、有价证券等财物，以及其他非财产性利益；不得向甲方借出钱款、住房、车辆等；不得为甲方提供宴请、旅游、健身、娱乐等活动安排；不得为甲方装修住房、婚丧嫁娶及亲属工作安排等提供便利；不得为甲方参与可能获取不正当利益的经济活动提供便利；不得以谋取非正当利益为目的，与甲方就业务问题进行私下商谈或者达成利益默契。

发现甲方有违反廉洁规定的，应及时向甲方反映或举报。受理部门：南方电网科学研究院有限责任公司监督部；举报地址：广东省广州市黄埔区科学城科翔路 11 号南网综合基地 3 号楼南网科研院

监督部收；举报邮箱（网站）：kyyjdb@csg.cn；举报电话：020-38300042。

21.4 甲方违反国家及本合同有关廉洁规定的，由相关部门（机构）依法依规给予纪律处分或处理；涉嫌职务犯罪的，移交监察机构办理。

21.5 乙方违反国家及本合同有关廉洁规定的，甲方有权根据中国南方电网有限责任公司供应商失信处理有关规定，在南方电网公司范围内对乙方进行一定期限的不接受投标、市场禁入（包括暂停投标资格、取消中标或成交资格、不签订新的订单合同或发出新的订单）等不与其发生新的采购活动的处理，并有权单方解除合同，因此造成的损失由乙方承担；涉嫌犯罪的，移送司法机关依法追究刑事责任。

22. 争议解决方式

本合同在履行过程中发生争议的，应友好协商解决；不能协商解决的，向甲方所在地人民法院起诉。

在诉讼期间，本合同不涉及争议部分的条款仍须履行。

23. 其他约定

23.1 甲乙双方确定的其他约定为：_

23.1.1 双方确认本合同项下的研究内容 【属于】 基础研究领域。

23.1.2 双方同意本合同的技术合同认定登记工作按以下第【 2 】种方式执行：

（1）本合同不需要开展技术合同认定登记。

（2）乙方应在本合同签订后及时在其所在地技术合同认定登记机构

完成登记备案，并将登记证明文件报送甲方（合同采用一次性支付的，应于款项支付前提交；合同采用分期支付的，应于第一笔款项支付前提交），否则甲方有权迟延支付相应价款，乙方因此造成的损失由其自行承担。

23.1.3　乙方应在本合同签订后15天内向甲方提交研究开发计划书，研究开发计划书经甲方同意后，作为本合同附件。研究开发计划书的内容应包括：　1、研究目标；2、国内外研究现状；3、研究内容；4、技术路线、创新点与预期成果；5、组织架构与任务分工；6、时间进度安排；7、研究可能存在的问题及解决措施；8、质量保证措施。具体详见附件二《计划任务书》（格式）。

23.2 本条约定与本合同其他条款内容不一致的，以本条约定为准。

24. 合同签署与生效

24.1 本合同经双方法定代表人（负责人）或授权代表签字并加盖公章或合同专用章之日起生效，未尽事宜双方可协商并签署补充协议做出约定。

24.2 本合同附件包括 附件一《技术协议书》、附件二《计划任务书》（格式），均为合同组成部分，与本合同具有同等法律效力。

24.3 本合同文本一式 肆 份，甲方执 贰 份，乙方执 贰 份，具有同等法律效力。

（以下无正文）

【本页为基于代码多模态分析与高阶关联特征的电力系统软件漏洞检测研究技术开发合同（合同编号：1500002024030103XA00132）签署页】

甲方（盖章）：南方电网科学研究院有限责任公司

法定代表人（负责人）或授权代表（签名）：柳勇军

签订日期：2024 年 9 月 2 日

乙方（盖章）：华南农业大学

法定代表人（负责人）或授权代表（签名）：薛红卫

签订日期：2024 年 9 月 2 日

课题编号：YNSC24114

云南省服务计算重点实验室
开放课题合同书

课题名称：高可靠密文数据细粒度授权检索研究

委托方（甲方）：云南省服务计算重点实验室

受托方（乙方）：华南农业大学

甲方依托单位（丙方）：云南财经大学

课题负责人：肖媚燕　　手机：18675871082

课题起止时间：2025 年 1 月 1 至 2025 年 12 月 31 日

云南省服务计算重点实验室

2024 年

合 同 条 款

本合同甲方委托乙方就云南省服务计算重点实验室开放课题“高可靠密文数据细粒度授权检索研究”进行科学研究，并支付相应的开放课题经费。三方经过友好协商，在真实、充分地表达各自意愿的基础上，根据相关规定，达成如下合同条款，三方共同恪守。

第一条：甲方委托乙方进行科学研究的内容如下：

乙方研究内容及科技成果应覆盖申请书主要部分或与申请书主要内容一致。

第二条：丙方向乙方转账划拨开放课题经费金额及方式如下：

一、开放课题总额为：￥10000.00 元（壹万圆整）人民币。

二、开放课题经费由丙方 一次性 转账划拨给乙方。

三、需“先票后款”，即乙方先开票据之后，丙方再拨付款项。

四、时间要求：合同签订后 5 个工作日内，乙方向甲方提供事业单位往来收据或增值税普通发票（需盖章）；甲方收到乙方票据后 15 个工作日内，丙方以银行转账的方式向乙方拨付开放课题经费。

五、乙方开户银行名称、户名和账号为：

1、开户银行：广州工行五山支行

2、户　　名：华南农业大学

3、帐　　号：3602002609000310520

六、丙方开户银行名称、户名和账号为：

1、开户银行：中国农业银行股份有限公司昆明龙泉路支行

2、户　　名：云南财经大学

3、帐　　号：24013601040000374

七、乙方应当确保上述账户信息真实、合法、有效，因乙方提供的账户信息错误，导致的一切后果和责任，由乙方自行承担。

第三条：课题成果要求：

1、开放课题成果主要包括论文、发明专利、专著等，结项验收时须至少提

交1份研究报告及1篇以上SCI/SSCI期刊论文或CCF C类及以上会议论文或CSCD或北大中文核心期刊论文（不含预警期刊、负面清单期刊、增刊等）。

2、开放课题取得的研究成果和知识产权，由重点实验室与申请人共同所有和共享。鼓励与重点实验室专职人员开展合作研究。同时，申请人须在所取得的研究成果和知识产权中对重点实验室进行署名，否则不计入结题成果。署名方式如下：

（1）把重点实验室作为工作单位之一标注，中文单位信息："云南财经大学云南省服务计算重点实验室"，英文单位信息："Yunnan Key Laboratory of Service Computing, Yunnan University of Finance and Economics"。

（2）对于论文、专著等研究成果，除了署名工作单位信息外，还应在适当位置标明"云南省服务计算重点实验室开放课题"和资助编号。（中文为：云南省服务计算重点实验室开放课题（编号：YNSC24114），英文为：Supported by the Foundation of Yunnan Key Laboratory of Service Computing (No. YNSC24114)。

注：（1）（2）须同时满足。

第四条：课题验收：

一、乙方按本合同第一、三条要求完成开放课题的成果，论文须发表见刊（包括online）。

二、验收时间于2025年12月31日前，乙方应至少提前2周向甲方提交相关材料。

第五条：乙方若未能在合同期内取得本合同要求的成果，甲方有权作撤项处理，丙方有权追回课题经费。

第六条：本合同未尽事宜，按《云南省重点实验室建设与运行管理办法》《云南省服务计算重点实验室开放课题管理办法》《2024年度云南省服务计算重点实验室开放课题申报通知》等有关规定执行。

第七条：合同书一式3份，甲、乙、丙方各1份，本合同经三方签字盖章后生效。

附录 1 课题经费预算表

金额单位：万元（保留两位小数）

预算科目	财政经费	用途说明
合计	1.0	
1.材料费	0.1	采购实验材料用于方案测试
2.差旅费/会议费/国际合作交流费	0.2	用于项目调研、学术交流等
3.出版/文献/信息传播/知识产权事务费	0.14	用于论文和知识产权等费用
4.劳务费	0.4	用于支付研究生劳务费
5.专家咨询费	0.1	用于邀请专家进行项目咨询
6.管理费	0.06	用于学校项目管理费

合同书各责任方签字签章

委托方（甲方）	云南省服务计算重点实验室
经办人（签字）： 实验室主任（签章）： 单位公章 年　月　日	
受托单位（乙方）	华南农业大学
课题负责人（签字）： 法定代表人（签章）： 单位公章 2024 年 11 月 26 日	
甲方依托单位（丙方）	云南财经大学
经办人（签字）： 法定代表人（签章）： 单位公章 2024 年 11 月 26 日	

技术开发合同

项目名称：基于大模型的会议中控数据挖掘与分析模型研发

委托方（甲方）：广州市天誉创高电子科技有限公司

受托方（乙方）：华南农业大学

签订日期：2024年6月1日

签订地点：广州

技术服务合同

委托方（甲方）：广州市天誉创高电子科技有限公司

住　所　地：广州经济技术开发区明珠路 16 号三层、 四层

法定代表人：梁柱浓

项目联系人：李玉琪

联系方式 ：13798126568

通讯地址：广州市黄埔区经济开发区明珠路 16 号三楼

电子信箱 liyuqi@creator.com.cn

受托方（乙方）：华南农业大学

住　所　地：广州市天河区五山路 483 号

法定代表人：薛红卫

项目联系人：邱少健

联系方式 ：15817190456

通讯地址：广州市天河区五山路 483 号数学与信息学院 514 室

电子信箱：qiushaojian@scau.edu.cn

本合同为甲方委托乙方研发 基于大模型的会议中控数据挖掘与分析模型 ，并支付研究开发经费和报酬。乙方接受委托并进行此项研究开发工作。双方经过平等协商，在真实、充分地表达各自意愿的基础上，根据《中华人民共和国合同法》的规定，达成如下协议，并由双方共同恪守。

第一条 本合同研究开发项目的要求如下：

1．技术目标：研发基于大模型的会议中控数据挖掘与分析模型

2．技术内容：以广州市天誉创高电子科技有限公司技术要求为准

3．技术方法和路线：以甲乙双方具体实现技术方法为准

4．工作地点：广州

第二条 研究开发计划及进度主要内容如下：

内容	起止日期
模型需求确认	2023/06/01-2023/06/11
数据采集、分析和处理	2023/06/14-2023/12/31
模型构建和研发	2024/01/01-2024/05/31
模型上线和试用	2024/06/01-2024/12/30
项目验收	2024/12/31

第三条 甲方应向乙方提供的技术资料如下：

1．技术资料清单：基于大模型的会议中控数据挖掘与分析模型需求文档

2. 本合同履行完毕后，技术资料各自存档保管。

第四条 研究开发经费和报酬金额及支付方式：

模型上线（2024 年 6 月 1 日）后，甲方应在十个工作日内，支付乙方 50%研发经费，经费总额为 ¥10,000 元（人民币壹万元整）。

项目验收（2024 年 12 月 31 日）后，甲方应在十个工作日内，支付乙方 50%研发经费，经费总额为 ¥10,000 元（人民币壹万元整）。

未经甲方同意，乙方不得将本合同项目部分或全部研究开发工作转让第三方承担。

乙方开户银行信息为:

账户户名：华南农业大学

银行账号：3602-0026-0900-0310-520

开户银行：中国工商银行广州五山支行

单位地址：广州市天河区五山路 483 号

电话号码　020-85285508

第五条 乙方应派出专业技术人员组成技术研发组，按照约定的技术研发计划开展工作。乙方可以根据工作需要变更技术人员，但乙方项目负责人的变更需得到甲方同意。

第六条 乙方应当按以下方式向甲方交付研究开发成果：

研究开发交付的内容：研发各阶段文件、模型部署文档及模型文件。

第七条 验收标准：

从功能测试、安装测试、文档交付物等方面进行验收 。

乙方不得在向甲方交付研究开发成果之前，自行将研究开发成果转让给第三方。

第八条 乙方有责任在项目验收合格完成之后，向甲方提供 3 个月的免费维护服务（质保期），此维护仅指软件 bug 的修改以及小范围的功能性改动。

第九条 双方同意本项目应用软件的所有权益，包括但不限于所有权及知识产权，归甲方所有。乙方所购置与研究开发工作有关的设备、器材、资料等财产，归乙方所有。

第十条 乙方有责任对本合同的内容进行保密。

第十一条 双方确定，在本合同有效期内，甲方指定 李玉琪 为甲方项目联系人，乙方指定 邱少健 为乙方项目联系人。项目联系人承担以下责任：

1. 甲乙双方联系人需及时进行需求沟通；

2. 在项目进行期间，如若遇到疑问，甲乙双方联系人需及时沟通反馈；

3. 甲乙双方联系人需积极促进项目进行，遵守合约精神；

一方变更项目联系人的，应当及时以书面形式通知另一方。未及时通知并影响本合同履行或造成损失的一方，应承担相应的责任。

第十二条 双方确定，因发生不可抗力或技术风险，致使本合同的履行成为不必要或不可能的，一方可以通知另一方解除本合同，双方均不对该合同解除承担违约责任。

第十三条 本合同的成立、有效性、解释、签署、修订和终止以及争议的解决均应适用中华人民共和国法律；双方因履行本合同而发生的争议，应协商、调解解决。协商、调解不成的，应向乙方所在地人民法院提起诉讼。

第十四条 本合同的附件作为本合同的组成部分，与本合同具有同等法律效力。

第十五条 本合同一式 肆 份，具有同等法律效力。甲方保留贰份，乙方保留贰份。

第十六条 本合同自双方授权代表签字盖章之日起生效。

甲方：广州市天誉创高电子科技有限公司

签名（盖章）：

2024 年 6 月 1 日

乙方：华南农业大学

签名（盖章）：

2024 年 6 月 1 日

SCAULIB202625311

检索证明

根据委托人提供的论文材料，委托人华南农业大学数学与信息学院 黄立峰(学科类型:自然科学) 4 篇论文收录情况如下表：

序号	论文名称	发表刊物及发表的年月卷期/页码等	作者排名	论文等级	作者文中单位	收录情况	影响因子	中科院大类分区
1	FASTEN: Fast Ensemble Learning for Improved Adversarial Robustness	IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY 出版年：2024 卷期：19 页码：2565-2580 文献类型：Article	第一作者	T2 类	华南农业大学数学与信息学院	SCI	IF2-year=8.0 IF5-year=8.5 （2024）	计算机科学 1 区 Top 期刊：是 OA 期刊：否 （2023）
2	Erosion Attack: Harnessing Corruption To Improve Adversarial Examples	IEEE TRANSACTIONS ON IMAGE PROCESSING 出版年：2023 卷期：32 页码：4828-4841 文献类型：Article	第一作者	T2 类	华南农业大学数学与信息学院	SCI	IF2-year=10.8 IF5-year=12.1 （2023）	计算机科学 1 区 Top 期刊：是 OA 期刊：否 （2023）
3	LAFED: Towards robust ensemble models via Latent Feature Diversification	PATTERN RECOGNITION 出版年：2024 出版日期：JUN 卷期：150 页码：- 文献号：110225	并列第一作者	T2 类	华南农业大学数学与信息学院	SCI	IF2-year=7.6 IF5-year=7.9 （2024）	计算机科学 1 区 Top 期刊：是 OA 期刊：否 （2023）

		文献类型：Article						
4	DEFEAT: Decoupled feature attack across deep neural networks	NEURAL NETWORKS 出版年：2022 出版日期：DEC 卷期：156 页码：13-28 文献类型：Article	第一作者	T2 类	华南农业大学数学与信息学院	SCI	IF2-year=7.8 IF5-year=8.4 （2022）	计算机科学 1区 Top 期刊：是 OA 期刊：否 （2022）

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明:如未盖章,报告无效

检索员：尹银怀

华南农业大学图书馆

2026-03-04

中山大学图书馆

检索结果证明

编号：2026D0102

委托检索人（单位）： 黄立峰（华南农业大学数学与信息学院）

检索课题名称：检索黄立峰 (Huang Lifeng) 发表的 2 篇论文被 EI 数据库收录情况，以及黄立峰在论文作者中的署名情况。

检索数据库：

Engineering Village 2（EI）

检索结果：

经检索，黄立峰发表的 2 篇论文，被 EI 数据库收录 2 篇。

论文被收录情况，以及作者署名详细情况见附件。

声明：本证明检索的文献信息（题名、作者、刊名、卷期、页码等）均由委托人提供并承担真实性责任。

检索人：

中山大学图书馆

2026 年 03 月 05 日

编号：2026D0102（附件）

黄立峰论文的收录情况

序号	作者、题名、刊名、年卷期、页码	数据库收录	作者署名情况
1	Huang Lifeng, Su Tian, Gao Chengying, Liu Ning, Huang Qiong. (2025). AUTE: Peer-Alignment and Self-Unlearning Boost Adversarial Robustness for Training Ensemble Models. PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE. 39. (4): 3671-3679. DOI:10.1609/aaai.v39i4.32382	EI	第一作者
2	Qiu Shaojian, You Xiaokang, Rong Wei, Huang Lifeng, Liang Yun. (2024). Boosting Imperceptibility of Adversarial Attacks for Environmental Sound Classification. PROCEEDINGS - INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE, ICTAI: 790-797. DOI:10.1109/ICTAI62512.2024.00116	EI	通讯作者
合计		---	----

注：他引是指文献被除作者及合作者以外其他人的引用。

声明：本证明检索的文献信息（题名、作者、刊名、卷期、页码等）均由委托人提供并承担真实性责任。

FASTEN: Fast Ensemble Learning for Improved Adversarial Robustness

Lifeng Huang, Qiong Huang, Peichao Qiu, Shuxin Wei, and Chengying Gao

***Abstract*— Recent works show that adversarial attacks threaten the security of deep neural networks (DNNs). To tackle this issue, ensemble learning methods have been proposed to train multiple sub-models and improve adversarial resistance without compromising accuracy. However, these methods often come with high computational costs, including multi-step optimization to generate high-quality augmentation data and additional network passes to optimize complicated regularization. In this paper, we present the FAST ENsemble learning method (FASTEN) to significantly reduce training costs in terms of data and optimization. Firstly, FASTEN employs a single-step technique to initialize poor augmentation data and recycles optimization knowledge to enhance data quality, which considerably reduces the data generation budget. Secondly, FASTEN introduces a low-cost regularizer to increase intra-model similarity and inter-model diversity, with most of the regularization components computed without network passes, further decreasing training costs. Empirical results on various datasets and networks demonstrate that FASTEN achieves higher robustness while requiring significantly fewer resources than current methods. For example, a 5-member FASTEN speeds up the optimization process by 7× and 28× compared to state-of-the-art DVERGE and TRS, respectively. Moreover, FASTEN outperforms the stronger of the two methods by 26.3% and 6.1% under black-box and white-box attacks, respectively. FASTEN is also compatible with existing fast adversarial training techniques, making it an advantageous choice for enhancing robustness without incurring excessive costs. The source code is publicly available at https://github.com/mesunhlf/FASTEN.**

***Index Terms*— Adversarial attack, adversarial defense, ensemble learning, robustness.**

I. Introduction

DEEP neural networks (DNNs) have demonstrated excellent performance in various tasks and are widely adopted in many advanced AI-controlled systems [1], [2]. Nonetheless, they often show potential vulnerabilities to adversarial attacks, which artificially modify the clean data to create adversarial example can fool DNNs into making incorrect predictions in digital software [3], [4] and real-world applications [5], [6].

The threat of adversarial attacks has prompted researchers to develop a variety of defense mechanisms. In general, input transformation methods have several competitive advantages such as well accuracy and plug-and-play [7], [8], but their robustness is often overestimated [9], [10]. Adversarial training, on the other hand, is currently the most effective defense method, but at the expense of computational resources and clean accuracy [11], [12]. As a result, these two approaches are often unsuitable for practical scenarios. To address this problem, an alternative line of algorithms has been proposed in the literature — ensemble learning [13], [14], [15], [16], [17]. The solutions, based on the existing training paradigms and network structures, are conceptually simple: they optimize multiple diverse DNNs for joint predictions. The aggregated ensemble becomes naturally robust and difficult to break since attackers must fool most members to produce the same wrong results.

The predominant direction in ensemble learning continues to center on constructing the committee that achieves the highest accuracy on clean data, which involves the formulation of metrics or training protocols to enhance the diversity within the ensemble models [20], [21], [22]. Nevertheless, recent developments also highlight the efficacy of ensemble learning methods in defending adversarial attacks [13], [14], [15], [16]. Specially, some ensemble defenses aim to boost robustness by enhancing the member diversity from the perspective of optimization, such as differing the output distribution [15] or maximizing gradient divergence [16]. Nonetheless, these defense approaches have not yield satisfactory improvements [10]. To further improve the adversarial robustness, the concept of augmentation data has been introduced in ensemble defenses recently. DVERGE adopts feature distillation [23] to generate a batch of non-robust augmentations for each clean image [17]. TRS uses PGD attacker [11] to produce adversarial augmentations and leverages their gradient information to smooth the ensembles [19]. However, the benefits brought by augmentation techniques do not come for free: they incur much more computational resources to create high-quality data (*i.e.*, $\sim 10\times$ training time). In contrast, generating low-quality data with fewer costs during training significantly reduces robustness, making these methods less practical. Therefore, one might naturally wonder: *is it possible to develop a robust ensemble model with reduced training complexity in practice?*

In this paper, we propose FASTEN, a fast ensemble learning method that achieves higher robustness while implementing

Manuscript received 10 April 2023; revised 1 September 2023; accepted 7 November 2023. Date of publication 28 November 2023; date of current version 16 January 2024. This work was supported in part by the National Key Research and Development Plan in China under Grant 2023YFC3306100, in part by the Major Program of Guangdong Basic and Applied Research under Grant 2019B030302008, in part by the National Natural Science Foundation of China under Grant 62272174, and in part by the Science and Technology Program of Guangzhou under Grant 201902010081. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ali Dehghantanha. *(Corresponding author: Chengying Gao.)*

Lifeng Huang and Qiong Huang are with the College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510000, China.

Peichao Qiu, Shuxin Wei, and Chengying Gao are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: mcsgcy@mail.sysu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2023.3336527, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2023.3336527

TABLE I

THE SUMMARIZED COMPARISONS BETWEEN DIFFERENT ENSEMBLE LEARNING METHODS, INCLUDING TRAINING COST, WHITE-BOX ROBUSTNESS, AND BLACK-BOX ROBUSTNESS. THE SYMBOL↑/↓ INDICATE THE HIGHER/LOWER THE BETTER

Method	Training Cost ↓	White-box ↑	Black-box ↑
Vanilla [18]	★	★	★
ADP [15]	★	★★	★
GAL [16]	★★	★★	★★
DVERGE [17]	★★★	★★★	★★★
TRS [19]	★★★★	★★★	★★
FASTEN (ours)	★	★★★★	★★★★

the method with reduced cost. Particularly, we find that the most costly process in previous works [16], [17], [19] are **(1)** the generation of high-quality augmentation data, which requires multiple forward and backward propagations, and **(2)** the complicated regularizer optimization, which increases the cost for calculating additional information (*e.g.*, gradient similarity, *etc.*). To this end, FASTEN builds robust ensemble models by reconstructing the pipeline as follows:

- **Data Perspective**: While deploying augmentation data is essential to diversify ensemble members [17], [19], we argue that multi-step optimization is not indispensable for generating the high-quality augmentations. Instead, we propose *recurrent augmentation* strategy to accelerate the data generation process in FASTEN. It utilizes a single-step distillation technique to initialize poor augmentation data and then refines its quality by combining the prior knowledge recycled from the training history. This adjustment helps to improve the adversarial effect of augmentations and lower down the training overhead. Although we reduce the number of optimization steps, we empirically find that this strategy sufficiently increase ensemble robustness compared to current methods that use multi-stage optimization in the augmentation process.
- **Optimization Perspective**: Current methods require multiple network propagations to calculate gradient similarity or learn from augmentations, thereby inducing significant computational overhead. To address it, we propose the *contrastive ensemble* regularizer to avoid extreme computational cost. The intuition behind the regularizer is simple: since an ensemble is composed of several members, it is natural to guide the features of the original data and its augmentation to be similar within a member, but also distinguishing them from their counterparts captured by other members. This is comparable to the concept of supervised contrastive learning [2]. Additionally, most the components in the regularizer are extracted passingly from the preceded pipeline, providing a low-cost bonus to encourage ensemble diversity and strengthen the defense against adversarial attacks.

Our empirical results demonstrate that FASTEN exhibits high robustness and scalability, allowing it to be applied in larger ensemble groups, trained on a variety of dataset, and adapted to different network structures. Furthermore, FASTEN speeds up ensemble training by an average of 26×, while achieving higher robustness than state-of-the-art methods under different settings [17], [19], [24]. A summarized comparison of training costs and robustness between current methods and FASTEN is presented in Table I. In summary, the contributions of our work are three-fold:

- We have developed a new technique for generating data more efficiently in the ensemble framework, namely recurrent augmentation. It is able to generate high-quality data at a lower cost, which reduces the training time without sacrificing the benefits of data augmentation.
- We propose a contrastive ensemble regularizer to improve the robustness at nearly free cost. Unlike existing methods that mainly focus on maximizing the inter-model diversity, the proposed regularizer additionally attempts to minimize the intra-model similarity for boosting robustness.
- We conduct extensive experiments to validate the performance of FASTEN under different attacks, datasets, and networks, which demonstrate its superiority compared to baselines. Moreover, FASTEN is compatible with fast adversarial training, enhancing its robustness further.

II. RELATED WORK

A. Adversarial Attacks

The security issue of AI-controlled systems, particularly the vulnerabilities of DNNs, is relating to the researches of adversarial attacks, which have received significant attention recently [25], [26]. Specially, adversary applies an imperceptible perturbation δ to a clean data x to generate an adversarial example $x^{adv} = x+\delta$, which can fool the deep learning model f to flip its groudtruth label y to a wrong prediction, *i.e.*, $f(x^{adv}) \neq y$. Generally, the optimization of searching the adversarial perturbation δ can be formulated as:

$$\underset{\delta}{\text{argmax}} \ \mathcal{L}_\theta(x+\delta, y), \quad \text{s.t. } \|\delta\|_p \leq \epsilon, \tag{1}$$

where $\mathcal{L}$ is the objective function of model f with parameters θ, ϵ is the perturbation magnitude under L_p norm-balls. We follow the popular evaluation procedures to choose cross entropy function as $\mathcal{L}$ and set $p = \infty$ in this paper [15], [16], [17].

In white-box scenarios, adversaries have full knowledge about victim models, and often leverage propagation mechanism to produce adversarial examples, such as Projecting Gradient Descent (PGD) [11], Auto-Attack (AA) [9], *etc.* For black-box settings, adversaries attack substitute models to generate adversarial examples and exploit their transferability to break unseen models, including Momentum Iterative Method (MIM) [27], Diversity Input Method (DIM) [28], Skip Gradient Method (SGM) [29], *etc.* These attacks illustrate excellent fooling rate against DNNs regardless of the architectures or the training dataset, posing a huge threat to DNNs.

B. Adversarial Defenses

The threat of adversarial attacks has motivated the development of various empirical defense mechanisms. Input transformation is a promising technique due to its high efficiency [7], [8]. However, recent studies have raised questions about its robustness, as some attacks have successfully bypassed these defenses [9], [30], [31].

1) Adversarial Training: Utilizing adversarial examples to optimize the network is a reliable way for defending against attacks, and still achieves the highest performance to date [11], [32], [33]. It incorporate adversarial examples as augmentation data and performs min-max optimization to update the model parameters at each step, which is defined as:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}\Big\{\max_{\delta} \mathcal{L}_{\theta}(x+\delta, y)\Big\}, \tag{2}$$

where the inner part is maximized by searching the perturbation within a magnitude ϵ (see Eq. (1)), and the outer part is minimized to depress the adversarial effect for itself.

2) Fast Adversarial Training: The inner maximization process is typically the most computationally expensive component, as it is conducted by white-box attack methods such as PGD. This process generates perturbations through a large number of iterations, resulting in significant computational costs. To reduce the generation overhead, some fast training techniques are introduced in the field. For instance, Shafahi et al. alternatively recycle the gradient information and update model parameters to accelerate the training procedure [34]. Goodfellow et al. combine FGSM [35] with random initialization in adversarial training to avoid the catastrophic overfitting [24]. Although fast training methods indeed reduce the optimization complexity, the poor generalization of different adversarially trained models is also often highlighted in the literature — the accuracy on clean data is significantly degraded to be deployed in the real-world applications [36].

C. Ensemble Learning Defenses

Ensemble trained defenses have demonstrated a better trade-off between accuracy and robustness, which can be roughly divided into two categories: **1)** optimization-based ensembles, which diversify the members from the optimization perspective, and **2)** augmentation-based methods, which generate customized augmentation data for members to boost diversity. For optimization-based ensembles, ADP employs a regularizer to maximizes the divergence of non-maximal outputs between members [15]. GAL enlarges the cosine distance of input gradients to reduce the common adversarial subspace [16]. To construct the optimal ensemble from a network pool, several studies explore disagreement diversity metrics and incorporate data verification techniques, enhancing the robustness to safeguard DNNs [13], [14]. However, they are still vulnerable to advanced attacks in some scenarios [10]. To further boost the adversarial robustness, Strauss et al. [37] demonstrate the efficacy of adopting Gaussian augmentation data in ensemble defenses. DVERGE [17] isolates the non-robust augmentations by using the feature distillation technique [23] and optimizes the ensemble parameters in a cross-model manner. Similarly, TRS [19] leverages PGD method [11] to generate adversarial augmentations and use their gradient information to flat the model boundary. Although they gains obvious improvements over previous works, excessive training overheads for data generation and regularization limit their practicability. To this end, we propose FASTEN, which consists of a fast augmentation technique and a nearly free regularizer, to perform the stronger defense against adversarial examples.

III. Method

A. Background

The general routine of ensemble learning methods is to train multiple well-performed DNNs and then aggregate them together to make the joint predictions. For clarity, given an ensemble model F composed of n sub-models, we denote each member as f_i, where $i \in 1, 2, 3, \ldots, n$. The inference output of the ensemble model F for a data x can be expressed as

$$F(x) = \mathbb{E}\big(f_i(x)\big).\ i \in 1, 2, 3, \ldots, n \tag{3}$$

To establish a robust ensemble model, it is more natural to improve the model diversity than making each individual member robust since the latter choice induces degraded clean accuracy by learning more robust but non-generalizable features [11], [23]. The objective of **optimization-based** algorithms (*i.e.*, GAL [16], ADP [15]) can be formulated as

$$\min_{\theta} \mathcal{L}_{\theta}(x, y) - \beta \cdot \mathcal{R}(F, x), \quad (x, y) \sim D \tag{4}$$

where $\mathcal{L}$ is the cross entropy, θ refers to the weights of F, and the data-label pair (x, y) is sampled from the original dataset D. $\mathcal{R}$ is the regularization term that measures the diversity within the ensemble F on classifying x, and β is a balancing parameter. By minimizing Eq. (4), the clean accuracy and the robustness are improved simultaneously.

Some ensembles additionally leverage the idea of data augmentation to achieve a higher performance, *i.e.*, DVERGE [17] and TRS [19]. Those **augmentation-based** methods generate the augmentation data during the optimization as

$$\min_{\theta} \mathcal{L}_{\theta}(x, y) - \beta \cdot \mathcal{R}(F, x), \quad (x, y) \sim \mathcal{A}(F, D) \tag{5}$$

where $\mathcal{A}$ is an augmentation method that depend on F and D. In particular, DVERGE and TRS chooses the 10-step feature distillation [23] or 6-step PGD [11] to produce the high-quality augmentations, respectively. However, the excessive generation step significantly increases the training cost (Tab. II).

In this paper, we develop a fast and robust ensemble learning method by proposing a novel augmentation method $\boldsymbol{\mathcal{A}}$ with reduced training complexity, as well as a low-cost regularizer $\boldsymbol{\mathcal{R}}$ to enhance the model diversity. $\mathcal{A}$ and $\mathcal{R}$ will be detailed in the Sec. III-C and Sec. III-D, respectively.

B. The Influence of the Augmentation Quality and Quantity

Empirical results show that augmentation-based ensembles perform better than their optimization-based counterparts. We relate their success to the additional augmentation data, and consider which one, quantity or quality, is the more critical factor to affect the robustness. Specially, we consider different setups for DVERGE [17] and TRS [19] as:

- The original DVERGE method, which trains each member f_i over the **n-1** augmentation data from other members $f_j (j \neq i)$ by using **FD-10**, denoted as *DVERGE*;
- The DVERGE method with *weaker* augmentations, which generates the additional data by only **1 Step** FD (*i.e.*, **FD-1**) from the rest of **n-1** members, denoted as *DVERGE-1S*;

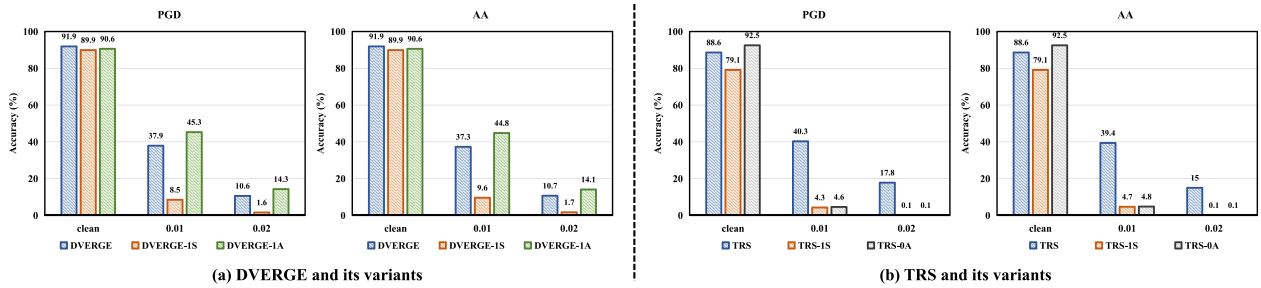


Fig. 1. **The influence of the augmentation quality and quantity**. We separately evaluate the accuracy and robustness of (a) DVERGE [17] and (b) TRS [19], trained with weaker or fewer augmentation data, denoted as *DVERGE-1S*, *DVERGE-1A*, *TRS-1S*, and *TRS-0A*, respectively (detailed in Sec. III-B). The results is tested by using strong attacks PGD-50 [11] (left) and AA [9] (right) under two perturbation magnitudes (*i.e.*, $\epsilon = 0.01/0.02$). Our observations show that the robustness of both DVERGE and TRS severely degraded when weaker augmentation methods are used to generate low-quality training data (see *DVERGE-1S* and *TRS-1S*). By contrast, the quantity of augmentation data for ensemble learning plays a less essential role in boosting adversarial robustness (see *DVERGE-1A* and *TRS-1A*). Furthermore, the effect of weaker or fewer augmentation methods on clean accuracy is minimal.

TABLE II

THE TRAINING TIME (MINUTES) OF THE 3-MEMBERS/5-MEMBERS ENSEMBLE MODELS FOR CIFAR-10 DATASET ON RTX 3080 SINGLE GPU DEVICE. ↓ INDICATES THE LOWER THE BETTER

Method	Time (3 members) ↓	Time (5 members) ↓
Vanilla [18]	77	115
ADP [15]	91	131
GAL [16]	231	370
DVERGE [17]	592	970
TRS [19]	2401	3898
FASTEN (ours)	97	139

- The DVERGE method with *fewer* augmentations, which trains each member by randomly choosing **1 Augmentation** produced by using **FD-10**, denoted as *DVERGE-1A*;
- The original TRS method, which produces the adversarial augmentations by using **PGD-6** attack method to extracts the **n** gradients information, denoted as *TRS*;
- The TRS method with *weaker* augmentation data, which produces the adversarial augmentations by using **1 Step** PGD method (*i.e.*, **PGD-1**), denoted as *TRS-1S*;
- The TRS *without* augmentations, which chooses the original data instead of the augmented data in optimization (*i.e.*, **0 Augmentation**), denoted as *TRS-0A*;

We note that both weaker and fewer setups (*e.g.*, *DVERGE-1S*, *DVERGE-1A*, *etc.*) directly affect the quality and quantity of the training data. We train 3-members DVERGE and TRS with different setups while other configurations are the same in the literature. Two attacks (*i.e.*, PGD-50 [11], and AA [9]) are included to test the robustness. The accuracy of ensemble models for clean data and adversarial examples (*i.e.*, $\epsilon = 0.01/0.02$) is reported in Fig. 1.

The implication is three-fold. Firstly, the robustness of DVERGE and TRS strongly depends on the quality of the augmentation data. Specifically, DVERGE significantly degenerates its performance when we use a weak augmentation technique to generate poor-quality data, as shown in *DVERGE-1S*. TRS exhibits similar trends when trained on poor-quality data, performing even worse than its counterpart with no augmentations at all (see *TRS-1S* and *TRS-0A*). Secondly, the quantity of augmented data has less impact on the defense capability, and fewer augmentations can help maintain high robustness. Interestingly, the performance of DVERGE slightly improves when we take only one augmentation data for training each member (see *DVERGE-1A*). It suggests that introducing too much augmentation data to each member makes it difficult to capture valid features. Thirdly, either weaker or fewer augmentation strategies have little impact on clean accuracy that the ensemble model can still learn enough generalizable features under these harsh conditions.

In summary, we draw two main conclusions from the analysis above: **1)** either *excessive* or *low-quality* augmentation data have a negative effect on training, and **2)** using *fewer* but *high-quality* augmentation data is effective in optimizing a robust ensemble model.

C. Recurrent Augmentation Strategy

The discussion in Sec. III-B inspires us to cut down the training overhead by generating less but high quality augmentation data. Given an image x sampled from the dataset D, current methods usually generate an augmentation set X based on the all members [17], [19], as illustrated in Fig. 2 (a). Since excessive augmentation data is not a necessary condition to boost robustness, we select a small portion of the ensemble to produce the high-quality augmentation data as

$$X^* = \mathcal{A}(\mathcal{M}(F), x) = \{x^A | x^A = \mathcal{A}(f, x)\}, \quad \forall f \in F^* \quad (6)$$

where $\mathcal{M}$ is a mask function to control the propagation flow by picking m members that forms a subset ensemble F^*, *i.e.*, $|F^*| = m, m \ll n$. Similar to the observations in Sec. III-B, we empirically find that producing only *two* augmentation data is enough to achieve a high robustness in the FASTEN framework (*i.e.*, $m = 2$), which is validated in Sec. IV.

Although FASTEN reduces the budgets by $n/2$ fold compared with state-of-the-art ensembles, it still requires multiple network passes to generate high-quality augmentation x^A for each sampled data x. To address this, we design Recurrent Augmentation strategy on accelerating the augmentation process $\mathcal{A}$ through two steps performed in a round-robin manner: **1)** Initialization, and **2)** Refinement.

Initialization, which costs a *single* forward and backward pass to produce a weak augmentation data. Specifically,

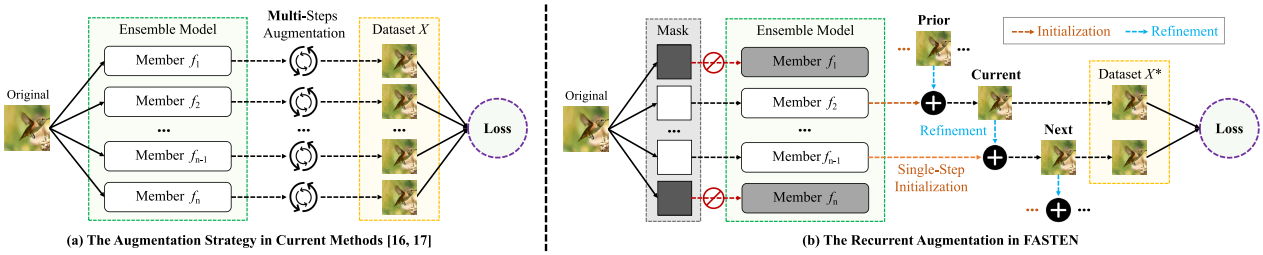


Fig. 2. **The augmentation pipeline of current ensemble methods and FASTEN**. (a) Current methods, such as DVERGE [17] and TRS [19], typically rely on multi-steps augmentation techniques (*e.g.*, FD-10 [23], PGD-6 [11], *etc.*) to generate an augmented dataset X, consisting of n high-quality augmentation data. This paradigm leads to a significant increase in computational costs. (b) The proposed FASTEN uses a single-step optimization process to initialize two augmentation data at first and then polishes their quality by combing the augmentation knowledge recycled from the training trajectory. Thus, the augmentation set X^* in FASTEN merely includes two training data by costing single network pass, thereby significantly reducing computational overhead.

we incorporate Neural Representation Distortion (NRD) [38] to extract feature vulnerabilities as initialized augmentations. The objective is to maximize the perceptual metrics $\mathcal{J}$ on the picked sub-model f from F^*, which can be formulated as

$$\mathcal{J}(f, x) = \|f^l\big(\mathcal{T}(x)\big) - f^l(x)\|_2, \tag{7}$$

where f^l is the extracted intermediate features of the l-th layer, and $\mathcal{T}$ is an input transformation applied to the original data x. We follow [24] to sample a uniform noise to corrupt the inputs, *i.e.*, $\mathcal{T}(x) = x + r$, $r \sim \mathcal{U}(-\epsilon, \epsilon)$. To reduce the propagation cost, we introduce NRD with one-step scheme for producing the initialized augmentation data x^I as

$$x^I = \text{clip}_{x,\epsilon}\big(\mathcal{T}(x) + \epsilon \cdot \text{sign}(\nabla_x \mathcal{J}(f, x))\big), \tag{8}$$

where clip function constrains the visual similarity between the input x and the initialized x^I within L_∞ norm-balls. Intuitively, this process pushes the data towards the closest decision boundaries in the latent space while maintaining it visually similarity to the original input.

Two factors encourage us to choose NRD as augmentation instead of FD [23] or PGD [11]. Firstly, NRD exhibits excellent and generalizable performance in extracting vulnerable feature-level non-robust data that can be utilized to improve the robustness of ensembles. Secondly, the inference process and feature distillation can be merged into the same forward and backward pass (see Eq. (7)). This effectively helps FASTEN to cut down the training budget since the proposed regularizer requires the inference output in the optimization (see Sec. III-D).

Refinement, which enhances the quality of the augmentation data *without* requiring additional propagation budget. However, single-step optimization leads to degraded robustness compared to multi-step augmentations (Sec. III-B). To overcome this limitation, we reinforce the weak initialization by recycling the knowledge from the training trajectory.

We treat the prior augmentations as the teacher and extract their non-robust features that correspond to the original input. These features serve as proper knowledge to guide a new student augmentation data. Thus, the refinement process is defined as the combination of the original data, the initialized augmentation, and its historical augmented data as

$$x_t^A = x + \alpha \cdot g(x_t^I) + (1 - \alpha) \cdot g(x_{t-1}^A) \tag{9}$$

where α is the balancing parameter and we set $\alpha \sim \mathcal{U}(0, 1)$, t refers to the index of the augmentation data for a given input in the training trajectory, and $g(\cdot)$ denotes the the knowledge extraction procedure. Intuitively, $g(x_t^I)$ represents the non-robust feature knowledge derived from the current initialization, and $g(x_{t-1}^A)$ denotes the accumulated augmentation knowledge from the training history, where g is simply designed as

$$g(x_t') = \begin{cases} 0 & t = 1 \\ x_t' - x & t > 1 \end{cases} \tag{10}$$

where $g(\cdot)$ is designed to calculate the residual pixels between the augmentation data x_t' and the original data x, which contain a plentiful of distilled non-robust features. The intuitions behind the refinement are quite simple: **1)** The knowledge $g(x_{t-1}^A)$ probably comes from a different member since the mask $\mathcal{M}$ defined in Eq. (6) is randomly regenerated at each iteration. The stochastic manner facilitates the augmentation data in learning the vulnerabilities across the ensemble members. **2)** This process is performed in a looping manner to progressively improve the quality of the augmentation data as well as to accumulate the non-robust knowledge, thus it can effectively boost the robustness with the training is proceeded. By combining Eq. (9) with Eq. (10), we can polish the poor initialized data by learning from the historical augmentations.

D. Contrastive Ensemble Regularizer

We follow [17] and [39] that only include the augmentation data in the ensemble training. To avoid the overfitting effect, we adopt sequential cross-model training paradigm to optimize each member in turn [17], [18]. The objective in Eq. (4) can be rewritten as the cross-entropy of each member f_i learned on the augmentation data x^A and its label y:

$$\mathcal{L}_{\theta_i}(x, y) \Rightarrow \mathcal{L}_{\theta_i}\big(x^A, y\big) = \mathcal{L}_{\theta_i}\big(\mathcal{A}(f_{j\neq i}, x), y\big), \tag{11}$$

where θ_i denote to the parameters of f_i, and f_j refers to a randomly sampled member in Eq. (6) that differs from the currently optimized one, *i.e.*, $f_j \sim F^*$ and $j \neq i$. By minimizing Eq. (11), the ensemble model increases its capacity to capture the feature from the non-robust augmentation data, which contains the vulnerabilities distilled across the members.

However, relying on Eq. (11) to train ensembles can lead to the dilemma where all members tend to learn similar

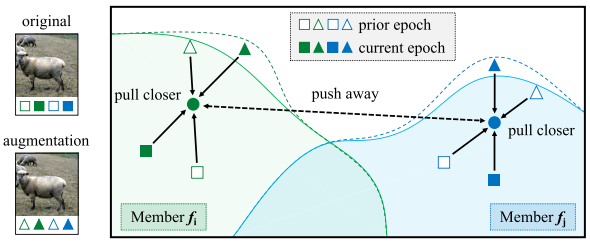


Fig. 3. **The intuition mechanism of the Contrastive Ensemble regularizer**. By learning from the augmentation data (triangles), the member changes its decision boundary since the augmented data contains rich non-robust features and locates far way from the original data (squares). Moreover, the regularizer pulls the original data and its augmentations closer for each ensemble member while it pushes them away from the feature cluster learned by other members simultaneously. The features captured by different members are colored as green and blue, respectively.

non-robust representations from a small augmentation set X^*. It significantly reduces both the diversity and robustness, as highlighted in [23]. To address this issue, we propose a Contrastive Ensemble regularizer, which encourages the learned features to be tightened within each member while also being distinct between members. Two lines of studies provide the inspiration for it. First, recent literature have shown that pushing the original data and its augmentation samples closer together is an effective training strategy [40], [41]. It is beneficial to improve the uncertainty estimation and classification accuracy under data shift scenarios. Second, supervised contrastive learning has led to major advances in various tasks by incorporating label information [2], [42]. Following the similar concept, the fundamental idea behind our approach is to align features belonging to the same class while pushing apart those from different classes, thereby maximizing the distance between feature clusters. It comprises two main components: **1)** Intra-Model Similarity and **2)** Inter-Model Diversity. The intuition behind our regularizer is illustrated in Figure .3.

Intra-Model Similarity, which aligns the input data and its augmentations in the intermediate space. It effectively smooth the model and reduces sensitivity to local noise. Specifically, we follow [40], [43] to employ the Kullback-Leibler Divergence (KL) as a metric to measure the similarity between the feature distributions of a given member f_i as

$$\mathcal{L}_{i,\text{sim}}(f_i, x) = \frac{1}{2}\big(\text{KL}(p_i \| m_i) + \text{KL}(p_i^A \| m_i)\big), \tag{12}$$

where p_i and p_i^A are the output probabilities of the member f_i on classifying the original data x and its augmentation x^A, m_i simulates the center of the latent cluster, *i.e.*, $m_i = \frac{1}{2}(p_i + p_i^A)$. By minimizing Eq. (12), the representations of these augmentation data are collected in tight clusters to be located around the original input even they are produced in different epochs.

Inter-Model Diversity, which pushes the feature cluster of each member away from others. This is similar to the intuition behind supervised contrastive learning, where the representations captured by a member are labeled as positive anchors, while the representations learned by other members are treated as negative samples. Since the original samples and its augmentations would be grouped tightly within each member by optimizing Eq. (12), we choose to maximize the distance between the cluster centers of members as

$$\mathcal{L}_{i,\text{div}}(f_i, x) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \text{KL}(m_i \| m_j). \tag{13}$$

In summary, the regularizer $\mathcal{R}$ defined in Eq. (5) is reformulated by combining the similarity with the diversity objectives (*i.e.*, Eq. (12) and Eq. (13)) for each member f_i as follows

$$\mathcal{R}(f_i, x) = \mathcal{L}_{i,\text{sim}}(f_i, x) - \mathcal{L}_{i,\text{div}}(f_i, x). \tag{14}$$

Unlike previous works [16], [19] that require additional resource to calculate the gradient in the regularization, most of the components in our regularizer have already been obtained at preceded steps, *e.g.*, p_i and p_i^A are derived from Eq. (7) and Eq. (11), *etc.* To qualify the high efficiency of FASTEN, we report the realistic training time (in minutes) for the 3/5-member ensemble models in Tab. II. It shows that FASTEN not only significantly reduces time complexity compared to current methods, but also can be implemented as cheaply as the fastest optimization-based ensemble algorithm ADP, thus enhancing adversarial robustness at a low cost.

E. FASTEN Optimization

FASTEN performs data augmentation and weight optimization in a round-robin fashion. Specifically, FASTEN randomly samples the original data-label pair from the dataset and then generates its corresponded augmentation data depending on the selected members (see Sec. III-C). Subsequently, FASTEN updates the parameters of each member by forcing them to learn from the augmentations and diversify the latent feature clusters within the ensemble model (see Sec. III-D). The overall objective for a single member f_i is

$$\min_{\theta_i} \mathop{\mathbb{E}}_{(x,y)\sim D, l} \Big[\mathcal{L}_{\theta_i}\big(\mathcal{A}(f_j, x), y\big) + \beta \cdot \mathcal{R}(f_i, x)\Big], \tag{15}$$

where $\mathcal{A}$ is the Recurrent Augmentation strategy, f_j is the selected member by masking the ensemble, l is a randomly selected layer, and $\mathcal{R}$ is the Contrastive Ensemble regularizer. More detailed pseudo-code for training an FASTEN ensemble of n members is illustrated in Alg. 1.

Adversarial FASTEN Framework: As FASTEN trains a robust ensemble model with low computational cost, it is natural to explore the potential of using adversarial examples to further enhance robustness. To this end, we propose the Adversarial FASTEN framework (AdvFASTEN), which combines FASTEN and fast adversarial training [24] with only a single propagation overhead. Specifically, AdvFASTEN modifies the original FASTEN framework as follows:

- **Augmentation Procedure**. We treat adversarial examples as another type of augmentation during ensemble training. To save the optimization cost, we generate adversarial examples by using Fast Adversarial Training (FAT) [24] to attack each member with only a single propagation overhead. In particular, we choose to perturb the augmentation data rather than the original data since they are already located near the decision boundaries.

Algorithm 1 FASTEN Framework

Input: Number of members n, Original Dataset D, Batch number b, Maximum training epoch e
Output: The ensemble model F

for $i = 1$ to n **do**
  Initialize the member f_i of the ensemble model F
end for
 # Recurrent Augmentation & Contrastive Ensemble Regularizer
for $k = 1$ to e **do**
  for $j = 1$ to b **do**
    Sample the data-label pair (x, y) from dataset D
    Produce a mask $\mathcal{M}$ to pick a subset ensemble F^* (Eq. (6))
    Randomly choose a layer l for distillation (Eq. (7))
    Produce the initialized data x^I with Eq. (8)
    Generate the refined data x^A from x^I with Eq. (9)
    for $i = 1$ to n **do**
      Calculate cross-model training loss with Eq. (11)
      Calculate Contrastive Ensemble loss with Eq. (15)
      Sum up the loss terms with Eq. (16)
      Update parameters of f_i by using SGD optimizor
    end for
  end for
end for

- **Objective Optimization**. Two modifications are required. First, both the augmentation data and adversarial examples are included to update the ensemble member:

$$\mathcal{L}_{\theta_i}(x, y) \Rightarrow \mathcal{L}_{\theta_i}(x_i^A, y) + \mathcal{L}_{\theta_i}(x_i^{adv}, y). \quad (16)$$

  where x_i^{adv} is the adversarial example which based on the member f_i and data x^A. Second, adversarial examples are also used for regularization, *e.g.*, the center of the latent cluster is updated by $m_i = (p_i + p_i^A + p_i^{adv})/3$, where p_i^{adv} is the output on classifying x_i^{adv}, *etc.* Moreover, its computation can be merged into Eq. (16), which makes the new regularizer cost-effective.

Our empirical results show that AdvFASTEN further boosts the robustness and achieves comparable and even better performance compared to its competitors who uses expensive adversarial training (*i.e.*, PGD-10 optimization) [17], [19].

Why FASTEN Improves the Ensemble Robustness? According to the conclusions in [19], the adversarial transferability among ensemble members is influenced by several factors, including empirical risk, gradient similarity, model smoothness, perturbation size, *etc.* Notably, the proposed FASTEN can effectively reduces gradient similarity among ensemble members and enhances model smoothness for each individual member. Consequently, an adversarial example that can successfully attack a member will be more challenging to transfer against others. This mechanism thereby enhances the overall robustness of the ensemble. For a more comprehensive theoretical analysis, please refer to the supplementary materials.

IV. Experiments

In this section, we present the empirical results to demonstrate the effectiveness of the proposed method. Firstly, we specify the detailed experimental setups in Sec. IV-A. Then, we provide the qualitative analysis of FASTEN on different datasets and network architectures in Sec. IV-B. Afterward, we evaluate the robustness of AdvFASTEN against a wide range of attacks in Sec. IV-C. Lastly, we provide a rich collection of ablation studies for our method in Sec. IV-D and Sec. IV-E.

A. Experimental Settings

1) Network: We follow the setups of previous works [15], [16], [17], [19] in the experiments. Particularly, we mainly consider the ResNet-20 structure to constitute the ensemble model. We also extend the evaluations to deeper and wider architectures (*i.e.*, ResNet-26, ResNet-32, and Vgg-19). To demonstrate the scalability of our method, we establish the ensembles by grouping 3, 5, and 8 members together, respectively.

2) Dataset: We mainly evaluate the ensemble models on CIFAR-10 dataset, which consists of 10 classes with 50000 training data and 10000 test data, respectively. To illustrate the generalizability of the proposed method, we show that FASTEN can also achieve superior performance on other four datasets (*i.e.*, MNIST, CIFAR-100, Tiny ImageNet, and Downsampled ImageNet) compared to state-of-the-art methods.

3) Baselines: We consider several ensemble learning algorithms:

- Vanilla Method [18] trains each member individually by removing all regularization and augmentation data.
- ADP [15] is an optimization-based method. It reduces the adversarial transferability by maximizing the divergence between non-maximal outputs across members.
- GAL [16] designs a novel regularization during the training process, which reduces the cosine similarity between gradient vectors calculated on pairwise members.
- DVERGE [17] is a type of augmentation-based training algorithm. It isolates the adversarial vulnerabilities from each member as the augmentation data and use them to optimize the ensemble model with a cross-model manner.
- TRS [19] combines regularization and augmentation techniques in optimization. It generates the adversarial augmentation data and then utilizes their gradient information to smooth the model boundaries.

Moreover, we include two fast adversarial training methods (*i.e.*, FAT [24] and GradAlign [44]) in our experiments.

4) Attack Models: We test the robustness of different ensemble models under both black-box and white-box attack scenarios.

In general, the black-box adversaries can not access the underlying knowledge of the victim model (*e.g.*, structures, parameters, *etc.*). They generate adversarial examples by attacking the surrogate models and exploit the transferability

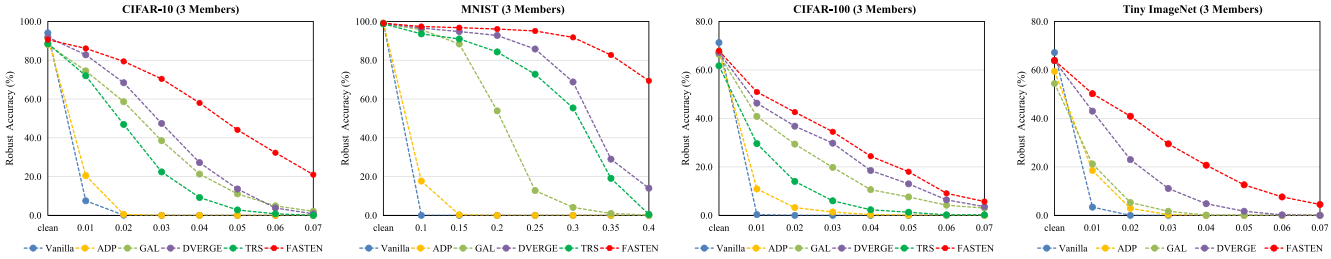


Fig. 4. **The black-box robustness of different methods with 3 members on CIFAR-10, MNIST, CIFAR-100, and Tiny ImageNet datasets**. The results is tested by generating 54 transferable adversarial examples for each clean data. The proposed FASTEN not only achieves the highest robustness compared with current methods, but also illustrates well scalability that the robustness is enhanced with more members included in the ensemble (see supplement).

TABLE III

THE ROBUST ACCURACY (%) OF ENSEMBLE MODELS WITH DIFFERENT GROUP SIZES ON CIFAR-10 DATASET. FOUR WHITE-BOX ATTACKS ARE USED IN THE EVALUATIONS, INCLUDING MIM [27], PGD-10/50 [11], AND AA [9]

Settings	Method	$\epsilon = 0.01$					$\epsilon = 0.02$					$\epsilon = 0.03$				
		MIM	PGD-10	PGD-50	AA	**Avg**	MIM	PGD-10	PGD-50	AA	**Avg**	MIM	PGD-10	PGD-50	AA	**Avg**
3 Members	Vanilla [18]	1.8	8.8	2.6	1.3	3.6	0.1	0.3	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
	ADP [15]	19.0	30.0	9.7	5.7	16.1	1.5	9.6	0.1	0.1	2.8	0.1	2.9	0.0	0.0	0.8
	GAL [16]	10.5	15.3	7.8	2.6	9.1	0.7	1.0	0.5	0.0	0.6	2.1	0.0	0.0	0.0	0.5
	DVERGE [17]	40.5	50.6	37.3	37.3	41.4	14.2	22.4	10.8	10.7	14.5	7.5	7.9	2.4	1.9	4.9
	TRS [19]	41.4	49.4	40.2	39.4	42.6	17.4	28.6	17.8	15.0	19.7	5.9	15.7	6.9	5.5	8.5
	FASTEN (**ours**)	**62.7**	**65.7**	**61.5**	**61.4**	**62.8**	**30.2**	**37.8**	**27.7**	**26.5**	**30.6**	**12.1**	**19.6**	**9.0**	**8.8**	**12.4**
5 Members	Vanilla [18]	3.0	14.6	3.6	1.4	5.7	0.0	0.7	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
	ADP [15]	19.8	28.5	10.0	6.5	16.2	3.4	10.3	0.8	0.2	3.7	0.8	4.3	0.4	0.0	1.4
	GAL [16]	31.1	43.4	29.7	29.8	33.5	6.9	18.8	6.1	5.5	9.3	0.7	7.5	0.5	0.5	2.3
	DVERGE [17]	53.3	59.6	49.3	49.0	52.8	25.2	34.9	21.3	19.3	25.2	9.5	18.5	6.5	5.6	10.0
	TRS [19]	42.4	55.4	42.7	41.9	45.6	20.5	35.3	20.3	20.9	24.3	10.5	23.4	9.3	10.5	13.4
	FASTEN (**ours**)	**66.4**	**69.0**	**66.1**	**65.9**	**66.9**	**36.5**	**47.4**	**33.6**	**35.6**	**38.3**	**17.6**	**27.9**	**14.8**	**14.8**	**18.8**
8 Members	Vanilla [18]	4.6	17.8	4.2	2.4	7.3	0.0	0.9	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
	ADP [15]	18.2	23.3	12.8	9.7	16.0	9.5	11.6	5.8	3.0	7.5	5.9	6.6	2.7	1.7	4.2
	GAL [16]	39.9	59.7	39.6	37.2	44.1	12.4	33.8	10.1	9.0	16.3	2.6	15.9	0.9	0.8	5.1
	DVERGE [17]	59.6	66.0	57.8	57.7	60.3	30.3	42.6	27.0	26.9	31.7	12.5	25.3	10.5	9.2	14.4
	TRS [19]	42.2	49.7	40.8	40.7	43.4	10.5	21.5	9.6	10.2	13.0	1.7	6.2	1.3	1.3	2.6
	FASTEN (**ours**)	**69.5**	**70.6**	**69.1**	**69.0**	**69.6**	**41.9**	**48.3**	**39.5**	**39.4**	**42.3**	**21.0**	**30.6**	**17.7**	**17.2**	**21.6**

to fool the target ensembles. By expanding the configuration in DVERGE [17], five attacks are included: **1)** Fast Gradient Sign Method (FGSM) [35]; **2)** Momentum-based Diversity Input Method (M-DIM) [28]; **3)** Skip Gradient Method (SGM) [29]; and **4)** Momentum-based PGD with 10 and 100 steps for 3 random starts (M-PGD) [11], *i.e.*, M-PGD-10 and the stronger M-PGD-100. We apply cross-entropy loss and C&W loss [45] separately to generate adversarial examples. The step size is set to $\epsilon/5$ and the optimization iteration is set to 100 unless explicitly stated. We use Vanilla ensembles with 3/5/8 members as the surrogate models. A total of **54** transferable adversarial example are created for each clean data, with the attack specifics provided in the supplementary. We follow [17] and [46] to choose the challenging "*all or nothing*" rule in the evaluation. It indicates that an ensemble model gets scores for a clean data if and only if total of its 54 adversarial versions can be correctly classified.

For white-box adversaries, we compare the robustness of different ensembles by using four attacks: **1)** 50-step Momentum method (MIM) [27] with step size $\epsilon/5$; **2)** 10-step and 50-step PGD with 5 random starts and the step size is $\epsilon/5$; and **3)** Auto-Attack (AA) [9], which is an ensemble of parameter-free attacks to fool classifiers. It consists of FAB [47], Square Attack [48] and two variants of PGD.

5) Training Details: We follow the parameter settings in [17] and [19]. All ensemble models are trained for 200 epochs with different group sizes (*e.g.*, 3/5/8 members). For SGD optimizer, the momentum parameter is set to 0.9 and the initial learning rate is 0.1, which is decayed by 10× at the 100-th and 150-th epochs. For the Adam optimizer, its learning rate is 0.1 with a weigh decay of 0.0001. More detailed specific parameters or modifications are reported in supplementary.

B. Experiments of FASTEN

In this section, we compare the performance of various ensemble models (without adversarial training). We start by evaluating the robustness of ResNet-20 ensembles on the CIFAR-10 dataset, and show extension experiments on other datasets (*e.g.*, MNIST, CIFAR-100, *etc.*) and network structures (*i.e.*, ResNet-26, ResNet-32, and Vgg-19). More detailed quantitative results are shown in the supplementary.

1) Performance on CIFAR-10 Dataset: We report the black-box and white-box robustness in Fig. 4 and Tab. III. In particular, we mainly include the results of 3-member ensembles in Fig. 4 (a) and the performance of other sizes in the supplement.

a) Black-box robustness: From Fig. 4 (a), it is unsurprising that the Vanilla ensemble exhibits the best accuracy

TABLE IV

THE ROBUST ACCURACY (%) OF ENSEMBLE MODELS ON MNIST AND CIFAR-100 DATASET. THREE WHITE-BOX ATTACKS ARE USED IN THE EVALUATIONS, INCLUDING MIM [27], PGD-50 [11], AND AA [9]

Settings	Method	MNIST ($\epsilon = 0.1/0.15/0.2$)									CIFAR-100 ($\epsilon = 0.01/0.02$)					
		MIM	PGD-50	AA	MIM	PGD-50	AA	MIM	PGD-50	AA	MIM	PGD-50	AA	MIM	PGD-50	AA
3 Members	Vanilla [18]	3.5	0.5	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.6	0.5	0.4	0.1	0.0	0.0
	ADP [15]	9.6	4.3	2.3	0.4	0.0	0.0	0.0	0.0	0.0	0.6	0.6	0.5	0.0	0.0	0.0
	GAL [16]	0.6	1.1	0.1	3.6	0.0	4.7	6.9	0.0	4.7	7.7	7.0	6.6	0.6	1.2	0.3
	DVERGE [17]	86.4	81.2	81.1	60.1	6.0	6.4	22.5	6.0	6.4	11.6	9.6	9.9	1.7	1.5	1.1
	TRS [19]	87.7	87.4	85.8	73.2	37.4	37.1	49.6	37.4	37.1	4.8	4.3	4.3	1.4	1.3	1.1
	FASTEN (**ours**)	**96.6**	**96.6**	**96.5**	**95.5**	**92.6**	**91.8**	**93.5**	**92.6**	**91.8**	**27.9**	**26.7**	**26.1**	**9.1**	**7.9**	**7.6**
5 Members	Vanilla [18]	3.9	0.6	3.9	0.1	0.0	0.0	0.0	0.0	0.0	1.0	1.1	1.1	0.2	0.1	0.2
	ADP [15]	15.0	9.1	15.0	1.2	0.0	0.0	0.0	0.0	0.0	0.7	0.6	0.3	0.0	0.0	0.0
	GAL [16]	2.1	0.1	2.1	4.8	0.0	5.9	5.9	0.0	5.9	11.7	12.9	10.5	2.3	1.6	1.3
	DVERGE [17]	92.1	91.5	92.1	80.8	27.2	46.4	46.4	27.2	46.4	17.4	16.0	15.9	5.4	4.5	4.4
	TRS [19]	92.5	92.4	92.5	78.8	45.3	54.7	54.7	45.3	54.7	10.7	9.6	9.7	2.9	2.5	2.5
	FASTEN (**ours**)	**97.7**	**97.0**	**97.4**	**96.8**	**95.0**	**94.8**	**95.0**	**95.0**	**94.8**	**31.4**	**30.7**	**30.8**	**12.2**	**10.9**	**10.8**
8 Members	Vanilla [18]	3.1	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0	1.7	1.4	1.3	0.3	0.2	0.2
	ADP [15]	12.8	3.7	1.9	1.5	0.0	0.0	0.1	0.0	0.0	2.0	1.3	0.9	0.2	0.0	0.0
	GAL [16]	6.1	7.2	7.4	6.3	2.2	4.0	7.0	2.1	3.3	18.6	17.4	14.4	4.3	4.1	2.7
	DVERGE [17]	95.4	95.1	95.1	89.8	87.1	85.5	74.8	66.7	62.6	26.3	24.6	23.7	9.5	7.6	7.1
	TRS [19]	92.8	92.7	92.3	79.8	77.6	76.9	57.9	44.3	42.9	12.1	11.0	10.3	2.0	1.7	1.4
	FASTEN (**ours**)	**97.7**	**97.7**	**97.8**	**97.3**	**97.3**	**97.1**	**96.1**	**96.7**	**95.9**	**36.4**	**35.9**	**35.7**	**13.4**	**12.5**	**11.9**

TABLE V

THE ROBUST ACCURACY (%) OF ENSEMBLE MODELS WITH 3 MEMBER3 ON TINY-IMAGENET AND DOWNSAMPLED-IMAGENET DATASET. FOUR WHITE-BOX ATTACKS ARE USED IN THE EVALUATIONS, INCLUDING FGSM [35], MIM [27], PGD-50 [11], AND AA [9]

Method	Tiny-ImageNet ($\epsilon = \frac{1}{255}/\frac{2}{255}/\frac{3}{255}$)												Downsampled-ImageNet ($\epsilon = \frac{1}{255}/\frac{2}{255}$)							
	FGSM	MIM	PGD	AA	FGSM	MIM	PGD	AA	FGSM	MIM	PGD	AA	FGSM	MIM	PGD	AA	FGSM	MIM	PGD	AA
Vanilla [18]	26.1	30.1	29.2	28.9	12.3	8.6	7.7	7.4	8.0	2.5	1.8	1.9	18.2	0.3	0.2	0.1	12.6	0.0	0.0	0.0
ADP [15]	43.6	18.1	18.0	17.9	29.2	3.9	3.2	3.3	19.5	0.3	0.3	0.2	6.2	0.1	0.2	0.1	3.4	0.0	0.0	0.0
GAL [16]	12.5	0.1	0.1	0.0	1.0	0.1	0.0	0.0	0.5	0.1	0.0	0.0	0.4	0.0	0.0	0.0	0.1	0.1	0.0	0.0
DVERGE [17]	45.9	32.6	31.6	31.0	38.0	15.2	13.8	13.8	34.7	7.1	6.6	6.1	26.5	13.1	11.9	11.8	14.7	2.2	1.6	1.6
FASTEN (**ours**)	**51.5**	**48.4**	**48.2**	**48.2**	**43.6**	**36.6**	**36.2**	**36.1**	**39.9**	**26.3**	**24.9**	**25.0**	**32.2**	**29.8**	**29.7**	**29.6**	**21.4**	**12.0**	**11.5**	**10.7**

in classifying clean data. However, the results also indicate its vulnerabilities when black-box adversarial examples are injected, even if the attack strength is weak. Optimization-based methods, including as ADP and GAL, improve their robustness against weak attacks, while their accuracy is still significantly reduced as the perturbation increases (*e.g.*, ADP and GAL are decreased from 91.6% and 88.1% to 0.0% and 38.5% under $\epsilon = 0.03$, *etc.*). Thus, they offer limited defense against strong black-box attacks, even though the additional training overhead is acceptable. DVERGE shows desirable defensive capability against most black-box settings. However, it is important to note that extensive optimization steps are required to produce augmentation data and the member is trained with complex cross-model behavior, *i.e.*, the complexity is $\mathcal{O}(n^2)$. Therefore, extending DVERGE to larger group sizes is not practical due to the exponential growth of training resources (Tab.II).

By comparing with baselines, the proposed FASTEN yields much superior performance in all black-box attack scenarios. For example, 5-member FASTEN can correctly predict 30.9% of transferable adversarial examples even under the most strict black-box setup ($\epsilon = 0.07$), surpassing DVERGE by a large margin of 26.4%. It is primarily due to the fact that FASTEN combines high-quality augmentation data with an effective regularizer to optimize the ensemble members. Additionally, FASTEN exhibits good scalability since its black-box robustness is consistently improving as more members are added into the ensemble model. This is a positive attribute since evidence suggests that as the ensemble group grows larger, FASTEN requires much less additional overhead during the training process than existing methods (Tab. II).

b) White-box robustness: Similar to the trends in Fig. 4 (a), optimization-based methods (*i.e.*, ADP and GAL) often show inferior white-box robustness compared to augmentation-based baselines (*i.e.*, DVERGE and TRS). However, although TRS is the strongest defender for small groups (*i.e.*, 3 members), its robustness is slightly degraded in turn when we expand the ensemble sizes. This interesting phenomenon can also be observed in black-box scenarios (Fig. 4), which may caused by the unbalanced weights between its regularization terms (*i.e.*, model smoothness and gradient similarity). It indicates the poor scalability of TRS for training a larger ensemble model.

In summary, FASTEN substantially outperforms *all* baselines and achieves the best performance regardless of the perturbation magnitudes, attack methods, or ensemble sizes. When defending against strong PGD-50 and AA, FASTEN surpasses the runner-up TRS by 20.4% and 21.1% under small ensembles, while it also exceeds the second-place DVERGE by 11.4% and 11.3% for large groups. Particularly, we notice that the robustness gap continues to grow under both weaker attacks (*i.e.*, PGD-10) and stronger adversaries

(*i.e.*, PGD-50 and AA), demonstrating the stability of our method.

2) Performance on MNIST and CIFAR-100 Datasets: We evaluate the robustness of different ensemble models on MNIST and CIFAR-100 datasets, and record the white-box and black-box robust accuracy in Tab. IV and Fig. 4 (b, c).

Two thought-provoking phenomenon catch our attention. Firstly, it is noticeable that the ensembles trained on the CIFAR-100 dataset are much more vulnerable to adversarial attacks than their counterpart on the CIFAR-10 dataset, *e.g.*, DVERGE ensembles with 3/5/8 members show a large decrease in robustness against small PGD-50 ($\epsilon = 0.01$) by 27.7%, 33.3%, and 33.2%, respectively. It is largely due to the fact that less data and augmentations are collected for each fine-grained class, degenerating the performance for classifying adversarial examples. Secondly, we observed that TRS performed worse than GAL and DVERGE in both black-box and white-box settings, contrary to its excellent robustness on the CIFAR-10 dataset. Similarly, ADP achieved lower resistance than Vanilla ensembles trained normally under a wide range of white-box adversaries. These results partially demonstrate their relatively low generalizability, making them unsuitable for implementing transfer training on other datasets.

Unlike GAL and TRS, FASTEN generalizes well to different data distributions. Specifically, we observe the same trends as in the CIFAR-10 experiments, where FASTEN effectively enhances the robustness and largely outperforms other approaches. For instance, the improvements of FASTEN are 16.3%, 15.5%, 17.1%, and 16.2% under small MIM, PGD-10/50, and AA, respectively. As the group sizes become larger, FASTEN goes a step further to strengthen the adversarial resistance, consistently outperforming all baseline ensembles.

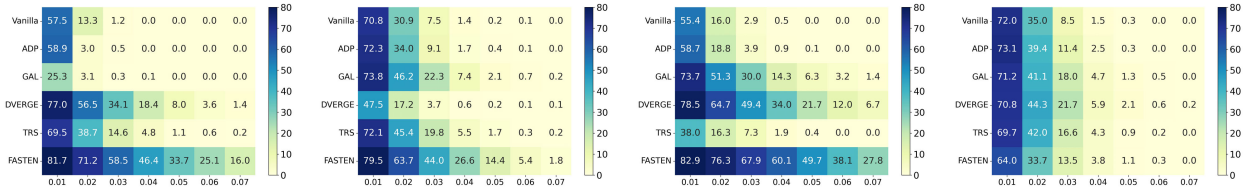
3) Performance on Tiny ImageNet and Downsampled ImageNet: We extend our experiments on two larger and more intricate datasets: Tiny ImageNet and Downsampled ImageNet datasets. Despite the identical resolutions of these two benchmarks (64×64), there exist differences in the data size and the number of classes between them. Tiny ImageNet consists of 100,000 training images categorized into 200 classes, whereas Downsampled ImageNet comprises 130,0000 images distributed across 1000 classes, mirroring the class distribution of the standard ImageNet dataset. We validate the effectiveness of FASTEN using ResNet-18 ensembles with 3 members. The robustness against a variety of white-box and black-box attacks are provided in Tab. V and Fig. 4 (d), respectively. As TRS struggles to effectively train ensembles on these two datasets, we choose not to present its performance in experiments.

From data presented in Tab. V and Fig. 4, it is evident that FASTEN emerges as the front runner in defending against malicious attacks in both white-box and black-box scenarios. Notably, it surpass the best baseline DVERGE by a significant margin of 22.3% and 9.1% under challenging adversary AA. This result stands as compelling evidence of the effectiveness of utilizing FASTEN in the training of large datasets. Another noticeable phenomenon is the consistency between the trends of robustness observed in Tab. V and the conclusions drawn from the CIFAR-100 experiments (Tab. IV). It suggests that training robust ensembles on intricate fine-grained datasets poses a greater challenge due to the perceptual similarity among many classes. Additionally, the results of ADP and GAL show a bit counter-intuitive—their capacities are even worse than Vanilla in certain cases. For instance, ADP exhibits superior resistance against single-step attacks FGSM, while its performance deteriorates relative to Vanilla when confronted with multi-step attacks. While GAL outperforms Vanilla in black-box setups, its robustness is severely degraded to nearly 0% against various white-box attacks. These findings indicate simply employing diversity metrics in a regularizer is ineffective for generalizing on large datasets with expanded classes.

4) Performance on Complicated Structures: Although FASTEN has achieved excellent robust accuracy on various datasets, we aim to investigate its ability to generalize to more complex networks. Specifically, we build ensembles by using the following settings: **1)** ResNet architecture with multi-depths, *i.e.*, ResNet-20, ResNet-26, and ResNet-32; and **2)** other structures different from ResNet, *i.e.*, VggNet-19 networks. The performance on CIFAR-10 dataset is reported in Tab. VI. We attack ResNet-20 Vanilla ensembles to generate transferable adversarial examples, which are used to evaluate the cross-network robustness.

As shown in Tab. VI, almost all baselines show improvements in the black-box robustness metric compared with their ResNet-20 versions. Previous studies confirm that the adversarial effect is significantly impeded when transferring to a different network [27], [49]. We provide evidence that this phenomenon extends to ensemble scenarios, including both network depths and model structures. Particularly, FASTEN not only holds the championship for defending against transferable attacks, but also significantly boosts its performance by training on networks with larger capacity (*i.e.*, VggNet-19), surpassing the second-runner DVERGE by 52.2% under black-box attacks ($\epsilon = 0.07$). Similar trends can also be observed for resisting white-box adversaries. Since deeper networks have a larger capacity to learn the non-robust features from the augmentations, we can see that DVERGE and FASTEN achieve higher robustness than other baselines on ResNet structures. However, DVERGE degenerates the power of VggNet version under large adversarial noises (*i.e.*, $\epsilon = 0.03$) and is overtaken by ADP, which may imply their unbalanced robustness across attack strengths. In contrast, FASTEN retains stability and superiority that enhances ensemble robustness against white-box adversaries. FASTEN exhibits the potential to extend to larger networks that its VggNet version illustrates much better performance compared with the original ResNet ensembles and other methods.

5) Conversion of Defenders and Attackers: Although ensemble models are proposed as robust defense mechanisms, we further explore another dimension of their functionality: whether employing these ensembles (instead of Vanilla ensembles) in attacks can generate more threatening adversarial examples. On the one hand, if the adversarial examples exhibit boosted transferability, it indicates that the attacked ensemble model posses high diversity so that it can be considered as a more potent attacker compared to Vanilla models [10]. On the other hand, leveraging enhanced adversarial examples allows



(a) GAL **(b) DVERGE** **(c) TRS** **(d) FASTEN**

Fig. 5. **The robustness of different methods under more *diverse* surrogate attackers**. The transferable adversarial examples are produced by attacking four ensembles with 5 members (*i.e.*, GAL, DVERGE, TRS, and FASTEN). FASTEN still achieves the highest robustness compared with current methods, but also plays as a stronger attackers to generate adversarial examples with higher transferability. It confirms the high diversity across FASTEN members.

TABLE VI

THE ROBUST ACCURACY (%) OF ENSEMBLE MODELS WITH 3 DIFFERENT MEMBERS ON CIFAR-10 DATASET. TO BUILD THE ENSEMBLE MODEL, WE CONSIDER TWO SETTINGS, INCLUDING: 1) RESNET-20, RESNET-26, AND RESNET-32; OR 2) THREE VGGNET-19 MODELS

Settings	Method	**Black-box** (ϵ)				**White-box** ($\epsilon = 0.01/0.02/0.03$)								
		0.01	0.03	0.05	0.07	MIM	PGD-50	AA	MIM	PGD-50	AA	MIM	PGD-50	AA
ResNet 20+26+32	Vanilla [18]	9.8	0.0	0.0	0.0	2.8	1.8	1.5	0.0	0.0	0.0	0.0	0.0	0.0
	ADP [15]	17.4	0.0	0.0	0.0	21.4	9.2	6.9	2.8	0.5	0.2	0.6	0.0	0.0
	GAL [16]	77.8	47.4	19.6	4.5	4.2	2.4	0.4	2.1	0.1	0.2	3.8	0.0	2.2
	DVERGE [17]	83.9	54.7	20.4	2.2	45.2	41.9	42.6	19.0	14.4	13.9	10.4	3.6	3.0
	TRS [19]	72.3	29.0	5.1	0.5	27.1	24.2	23.7	3.4	3.1	2.5	0.3	0.2	0.0
	FASTEN (**ours**)	**84.6**	**69.2**	**44.6**	**18.8**	**61.3**	**60.0**	**59.8**	**30.8**	**26.6**	**26.5**	**11.2**	**8.7**	**9.0**
VggNet-19 ×3	Vanilla [18]	47.0	0.2	0.0	0.0	12.8	9.9	9.5	0.1	0.0	0.0	0.0	0.0	0.0
	ADP [15]	46.9	0.0	0.0	0.0	42.2	29.6	21.4	19.7	6.1	4.4	10.4	2.3	2.0
	GAL [16]	82.1	38.8	8.7	0.8	39.2	36.2	34.8	16.6	11.6	9.0	6.2	2.8	1.7
	DVERGE [17]	84.3	61.4	28.2	6.4	52.0	49.8	48.8	17.7	16.1	12.7	3.3	2.6	1.2
	TRS [19]	76.1	29.6	5.7	0.6	35.5	31.0	31.5	15.8	11.9	12.0	7.4	5.4	6.3
	FASTEN (**ours**)	**87.4**	**80.9**	**72.0**	**58.6**	**69.9**	**69.4**	**69.3**	**47.4**	**45.7**	**45.2**	**29.5**	**26.5**	**26.0**

us to take a deeper into evaluating the robustness of different ensembles under more challenging scenarios.

Specifically, we treat the ensembles (*i.e.*, GAL, DVERGE, TRS, and FASTEN) as attackers to generate adversarial examples and test the robustness of the remaining methods. The performance is presented in Fig. 5, where the primary axes are the perturbation sizes and the defenders, respectively. Note that the numbers in the cells of the i-th row and j-th column indicate the accuracy of the i-th defender under attacks with j-th perturbation magnitudes. The higher the value, the stronger the defender and the weaker the attacker.

It is natural to observe that the black-box robustness of all methods is degenerated to some extent compared to the results in Fig. 4. It validates that the high diversity of those ensemble surrogates (*e.g.*, GAL, DVERGE, *etc.*) indeed lead to stronger attacks compared to the Vanilla ensembles. Moreover, FASTEN is still the first-place winner in this track, as it performs the best defensive ability to resist more threatening ensemble-model attackers such as GAL and DVERGE (see Fig. 5 (a)∼(c)). It further confirms the efficacy of the proposed method. Additionally, FASTEN can be considered as an outstanding surrogate attacker on the other side, *i.e.*, the adversarial examples generated by attacking FASTEN ensembles demonstrate higher transferability than their counterparts based on other models (see Fig. 5 (d)). This phenomenon also verifies the high diversity of FASTEN members.

6) Key Takeaways: Several important conclusions can be drawn from the above evaluations: **1)** FASTEN is *strong*, exhibiting the highest adversarial robustness and desirable clean accuracy to resist a wide range of attacks (*e.g.*, DIM, PGD, AA, *etc.*); **2)** FASTEN is *generalizable* that it has excellent performance on different datasets (*i.e.*, MNIST, CIFAR-10, and CIFAR-100) and various ensemble structures (*i.e.*, ResNet and VggNet); and **3)** FASTEN is *fast* and *scalable*, with a lower required training overhead than current algorithms, enabling us to efficiently build large ensembles to further improve its robustness in practical scenarios.

C. Experiments of AdvFASTEN

Since FASTEN shows remarkable training efficiency and considerable robustness improvement, we try to develop its potential by integrating FASTEN with adversarial training. Specifically, we pursue the goal of fast training to combine FASTEN with FAT [24], namely AdvFASTEN. Similarly, we apply FAT to baselines and build their adversarially trained ensembles, *e.g.*, FAT-ADP, FAT-GAL, *etc.* Beyond the FAT combined versions, we follow the literature to adopt the expensive PGD-AT for augmentation-based methods (*i.e.*, 10-step PGD-AT for DVERGE and 6-step PGD-AT for TRS) as the reference in the evaluations, denoted as PGD-DVERGE and PGD-TRS. In addition, two advanced fast adversarial training techniques are included in the comparisons (*i.e.*, FAT [24] and GradAlign [44]). The performance of different methods trained on the CIFAR-10 dataset under two strong white-box adversaries (*i.e.*, PGD-50 and AA) is reported in Tab. VII. We separately mark the first rank and the second place in **red** and **blue** on each column, w.r.t. the perturbation magnitude ϵ.

It is evident that GAL and DVERGE cannot be paired with FAT to create their fast adversarially trained versions. Although they enhance white-box robustness slightly in

TABLE VII

THE ROBUST ACCURACY (%) OF ADVERSARIALLY TRAINED ENSEMBLE MODELS ON CIFAR-10 DATASET. TWO WHITE-BOX ATTACKS ARE USED IN THE EVALUATIONS, INCLUDING PGD-50 [11] AND AA [9]. THE FIRST RANK AND THE SECOND PLACE ARE MARKED IN RED AND BLUE, RESPECTIVELY

Settings	Method	PGD-50 (ϵ)							AA (ϵ)						
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.01	0.02	0.03	0.04	0.05	0.06	0.07
3 Members	FAT [24]	**70.6**	55.7	39.0	25.4	15.1	6.8	2.9	**71.0**	56.3	38.6	25.5	13.8	6.3	3.0
	GradAlign [44]	70.6	56.7	38.2	25.8	14.5	6.9	3.7	70.4	56.4	38.0	25.1	13.6	6.7	3.0
	FAT-ADP [15]	69.2	54.8	40.9	27.9	18.0	**10.2**	**4.8**	69.2	54.4	40.6	27.0	17.0	8.7	4.2
	FAT-GAL [16]	46.8	18.1	5.1	0.9	0.0	0.0	0.0	45.1	16.4	4.7	0.5	0.0	0.0	0.0
	FAT-DVERGE [17]	56.8	30.9	13.1	4.9	1.0	0.1	0.1	56.6	30.3	12.8	4.6	1.1	0.4	0.1
	AdvFASTEN (**ours**)	70.5	**57.0**	**42.4**	**29.1**	**18.7**	**10.3**	**4.9**	70.5	**56.9**	**42.3**	**28.8**	**18.5**	**9.7**	**4.8**
	PGD-DVERGE [17]	**72.9**	**59.8**	**44.3**	**29.1**	**18.6**	9.8	4.7	**72.9**	**59.8**	**44.1**	**28.7**	**18.3**	**9.2**	**4.7**
	PGD-TRS [19]	61.5	46.9	31.0	18.3	9.2	4.1	1.6	61.3	45.8	29.9	16.4	7.5	3.1	1.1
5 Members	FAT [24]	71.1	56.6	39.9	26.3	**15.4**	6.7	**3.5**	71.3	56.4	40.5	26.1	15.1	6.6	**3.1**
	GradAlign [44]	**72.1**	**58.5**	40.3	**26.7**	15.2	7.4	3.2	**72.1**	**58.4**	40.0	26.4	14.3	7.0	3.0
	FAT-ADP [15]	69.5	52.0	33.4	20.2	8.8	3.3	1.2	69.4	50.8	32.4	18.6	6.6	2.6	1.1
	FAT-GAL [16]	51.3	23.6	9.3	2.7	0.7	0.3	0.1	49.4	22.9	9.2	2.5	0.6	0.3	0.1
	FAT-DVERGE [17]	63.1	38.2	16.8	5.7	1.6	0.4	0.0	62.6	37.7	16.7	5.8	1.4	0.2	0.1
	AdvFASTEN (**ours**)	70.4	57.7	**44.3**	**30.6**	**20.8**	**11.8**	**6.8**	70.4	57.8	**44.3**	**30.6**	**20.6**	**11.7**	**6.6**
	PGD-DVERGE [17]	**73.4**	**58.8**	**41.4**	26.6	15.3	**7.7**	3.4	**74.8**	**59.6**	**42.0**	**27.0**	**15.4**	**7.0**	2.7
	PGD-TRS [19]	62.3	43.4	28.6	14.3	5.1	1.4	0.6	60.8	41.5	25.0	10.4	4.4	1.3	0.4
8 Members	FAT [24]	71.8	56.8	**40.2**	**27.9**	15.9	7.6	**3.2**	71.9	58.3	**41.9**	**26.2**	**15.1**	**7.4**	**2.9**
	GradAlign [44]	**72.3**	56.6	40.9	27.3	**15.9**	**7.7**	2.9	**72.1**	56.4	40.3	26.8	15.5	7.0	2.6
	FAT-ADP [15]	70.8	52.1	31.4	17.7	8.1	2.6	0.9	70.7	51.7	31.0	17.0	6.9	2.2	0.8
	FAT-GAL [16]	48.8	23.7	11.6	4.5	1.5	0.7	0.1	47.7	22.5	9.4	3.3	0.7	0.2	0.0
	FAT-DVERGE [17]	55.6	25.7	8.9	2.4	0.4	0.2	0.2	55.8	24.4	8.6	1.7	0.3	0.2	0.2
	AdvFASTEN (**ours**)	71.4	**58.9**	**45.0**	**31.4**	**21.0**	**13.1**	**7.0**	71.3	**58.8**	**44.8**	**31.4**	**21.1**	**12.7**	**6.9**
	PGD-DVERGE [17]	**72.5**	**58.3**	39.7	25.9	13.8	5.9	2.7	**72.5**	**58.3**	39.8	25.7	13.6	5.8	2.7
	PGD-TRS [19]	59.3	41.9	27.3	15.2	7.2	2.7	1.2	58.0	40.5	25.1	13.0	5.5	1.3	0.6

TABLE VIII

THE EFFECT OF EACH COMPONENT IN 3 MEMBERS FASTEN, INCLUDING RECURRENT AUGMENTATION STRATEGY (RA) AND CONTRASTIVE ENSEMBLE REGULARIZER (CE). THE SYMBOL $\star$ INDICATES THAT USING A SURROGATE DISTILLATION TECHNIQUE FOR THE INITIALIZATION

Settings	RA		CE		$\epsilon = 0.01$				$\epsilon = 0.02$				$\epsilon = 0.03$			
	init	*refine*	$\mathcal{L}_{\text{sim}}$	$\mathcal{L}_{\text{div}}$	MIM	PGD-10	PGD-50	AA	MIM	PGD-10	PGD-50	AA	MIM	PGD-10	PGD-50	AA
only-init	✓	×	×	×	43.2	53.3	41.2	41.0	15.2	22.8	11.4	11.7	5.3	8.5	2.3	2.1
w/o refine	✓	×	✓	✓	46.3	55.4	43.6	44.7	16.7	26.0	13.4	13.6	6.7	11.1	3.1	3.9
w/o CE	✓	✓	×	×	56.9	61.8	55.9	55.2	22.5	31.9	19.1	19.6	7.3	13.6	4.4	4.2
w/o $\mathcal{L}_{\text{div}}$	✓	✓	✓	×	61.8	64.3	60.8	60.1	29.5	37.0	27.3	25.7	12.0	19.6	8.5	8.4
w/o $\mathcal{L}_{\text{sim}}$	✓	✓	×	✓	56.2	61.7	55.5	55.2	23.3	33.2	20.1	20.5	8.1	14.6	5.5	5.3
FD-init	$\star$	✓	✓	✓	62.5	**66.0**	**61.7**	60.3	29.4	36.5	27.3	26.1	11.6	18.9	8.7	8.5
FASTEN	✓	✓	✓	✓	**62.7**	65.7	61.5	**61.4**	**30.2**	**37.8**	**27.7**	**26.5**	**12.1**	**19.6**	**9.0**	**8.8**

comparison to non-adversarial ensembles, they suffer from the expense of additional training resources. In contrast, the use of multi-step PGD-AT can effectively improve the performance of DVERGE (i.e., PGD-DVERGE). However, it requires dozens of times more computational overhead to create optimal adversarial augmentations relative to the affordable FAT-based ensembles. Additionally, we discovered that even when integrated with PGD-AT, TRS is inferior to DVERGE. This could be due to the conflict between AT and the adversarial gradient in its regularization.

Another notable observation is that the behaviors of these adversarial ensembles display distinct trends for different group sizes. Specifically, by introducing the AT mechanism into a small ensembles (*i.e.*, 3 members), PGD-DVERGE and AdvFASTEN alternately win the championship and runner-up positions under small and large perturbations. However, as the ensemble sizes increased (*i.e.*, 5/8 members), PGD-DVERGE gradually loses its robustness due to the unbalanced weighting effect, resulting in a drop in its superior ranking, as discussed in [17]. Surprisingly, FAT and GradAlign surpass PGD-DVERGE on these tracks, even though they do not benefit much from the larger group setups, *i.e.*, their improvements are minimal despite more members being added. It suggests that simply applying AT techniques to ensemble scenarios is not sufficient. In contrast, we find that AdvFASTEN can take full advantage of ensemble setups. Thanks to the high diversity among members, AdvFASTEN continuously increases the robustness and outperforms both FAT and FASTEN significantly. It shows the superiority of AdvFASTEN in most cases, especially when the perturbation ϵ is large (*i.e.*, $\epsilon \geq 0.02$), demonstrating its better compatibility in comparison with recent elaborate methods.

D. The Effect of Augmentation and Regularizer

In this section, we investigate the impact of the proposed Recurrent Augmentation (RA) strategy and Contrastive Ensemble (CE) regularizer. Concretely, we consider removing or replacing each component as follows:

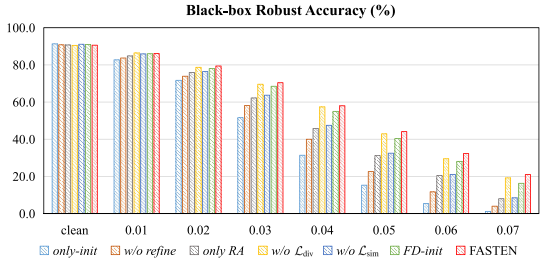


Fig. 6. **The ablation studies of FASTEN against black-box attacks**.

- Using the initialized augmentation in RA to optimize the ensembles, denoted as *only-init*.
- We remove the refinement stage of RA, *i.e.*, *w/o refine*.
- We keep the RA strategy and eliminate the CE regularizer to optimize the ensembles, *i.e.*, *w/o CE*.
- We abandon either Intra-Model Similarity or Inter-Model Diversity objective, denoted as *w/o* $\mathcal{L}_{\text{sim}}$ and *w/o* $\mathcal{L}_{\text{div}}$.
- We choose FD [23] instead of NRD [38] to initialize the non-robust augmentation dataset, denoted as *FD-init*.
- The full version of the proposed method, *i.e.*, *FASTEN*.

Without loss of generality, we trained FASTEN ensembles with different settings above on the CIFAR-10 dataset, and report the white-box and black-box robustness in Tab. VIII and Fig. 6, respectively. It is evident that each component is indispensable. Concretely, we can observe that only using single-step initialized augmentation in the training procedure results in the worst score against a variety of attacks. However, its robustness is noticeably strengthened by introducing the refinement process (see *only-init* and *w/o CE*). It confirms the effectiveness of RA strategy and validates our speculation that high-quality augmentation plays a more crucial role in building a robust ensemble model. This observation is consistent with current augmentation-based techniques DVERGE and TRS, as discussed in Sec. III-B. Furthermore, it is natural that regularizing with either intra-model similarity or inter-model diversity facilitates ensemble diversification. Integrating these two terms together can further mitigate the adversarial effect, showing its efficacy and generality for clustering latent features and benefiting member diversification.

Beyond the NRD initialization [38] described in Sec. III-C, we explore whether using other feature distillation techniques, such as FD [23], is able to improve the performance. By comparing *FD-init* with *FASTEN*, we find that the robustness gap under different white-box attacks is minimal that *FASTEN* is slightly stronger than *FD-init* in most cases. However, when facing black-box adversaries, NRD initialization gives FASTEN a distinct advantage over the FD algorithm (see Fig. 6). Furthermore, NRD can merge the non-robust feature distillation and the input inference into the same forward and backward pass, thereby lowering down the propagation cost for regularization. This property motivates us to choose NRD as a suitable method for the initialization stage.

E. Ablation Studies

In this section, we discuss the number of augmentations generated for each clean data and then search for the optimal layer to distill the non-robust data. Finally, we investigate the balance weight of the proposed regularizer. We trained the ensembles with 5 members on CIFAR-10, and select weak PGD-10 and strong PGD-50 as attackers.

TABLE IX

THE ABLATION STUDY OF AUGMENTATION NUMBER FOR FASTEN ENSEMBLES. WE RECORD THE WHITE-BOX ROBUSTNESS AND TRAINING COSTS (MINUTES) ON RTX 3080 SINGLE GPU DEVICE

Settings	$\epsilon = 0.01$		$\epsilon = 0.02$		Training Time
	PGD-10	PGD-50	PGD-10	PGD-50	
$m = 1$	65.0	60.8	39.9	28.5	120
$m = 2$	69.2	67.2	47.1	36.0	139
$m = 3$	70.5	68.2	47.6	37.4	177
$m = 4$	71.1	69.8	48.5	38.3	215
$m = 5$	69.9	68.9	48.8	38.2	257

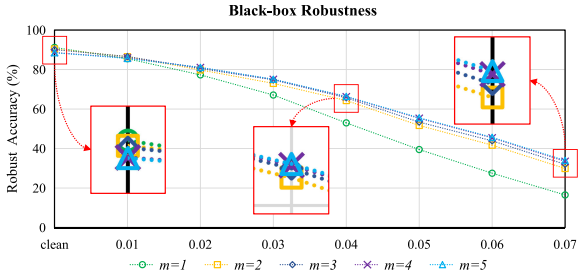


Fig. 7. **The black-box robustness of FASTEN with different augmentations**. We find that more augmentation data indeed brings a tiny improvement against large perturbations while its clean accuracy is continuously dropped.

1) Augmentation Number for Ensembles: RA reduces the training overhead via reducing the augmentation generation by a factor of n/m (see Eq (6)), where n and m are the group sizes of the ensemble model and its randomly masked subset, respectively. We study the effect of augmentation number m ranging from 1 to n, and report the white-box and black-box robust accuracy in Tab. IX and Fig. 7. If we set hyper-parameter $m = n$, RA would utilize all members to produce initialized augmentations like DVERGE and TRS.

We find that selecting a single member in RA significantly degenerates the robustness, particularly when the transferable perturbations are large, *i.e.*, $\epsilon \leq 0.04$. On the other hand, selecting multiple members to generate corresponding augmentation data does indeed improve the performance. However, increasing the number of augmentation data results in a decreasing trend in clean accuracy ($\sim 2\%$), mainly due to the overfull non-robust features distilled during training. It also induces a side effect with increased computational resources (see Tab. IX). Considering the performance/cost trade-off, we set $m = 2$ to implement the fast RA in the FASTEN.

2) Layer Selection for Augmentation: The layers chosen for the feature distillation affects the quality of the augmentation data. Beyond the random scheme, we additionally test 6 fixed layers throughout the network, including the shallow features (*i.e.*, index of 2/6/9) and the deep features (index of 12/15/18). The detailed results are reported in supplementary.

We observe that all schemes can largely improve the performance compared to the baselines, implying the effectiveness of the feature distillation in diversifying the ensemble

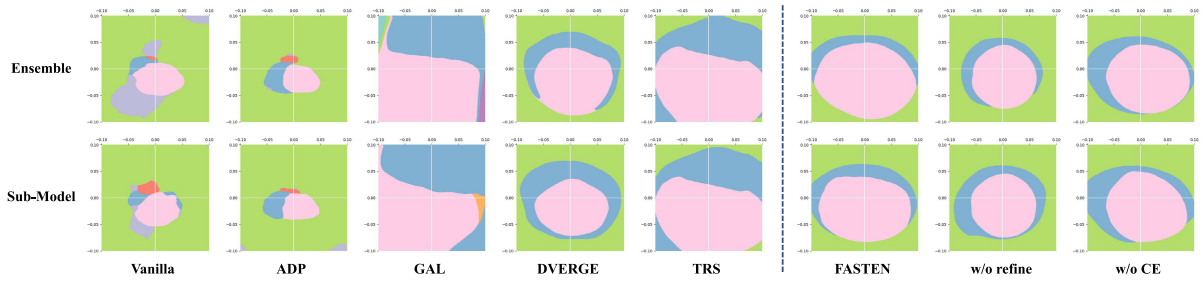


Fig. 8. **The decision regions in L_∞ norm-balls around a randomly sampled data learned by different ensemble methods**. The vertical axis is the gradient direction of the surrogate model and the horizontal axis is a randomly chosen orthogonal vector.

members. Both black-box and white-box experiments exhibit the same trends, indicating that applying the feature distillation process to shallower layers typically results in better clean accuracy, but worse robustness. This finding aligns with the conclusion that low-level features and high-level semantic representations have different impacts on defense capabilities [49]. While generating augmentation data at the deepest layers offers the highest robustness, we have found that it significantly reduces clean accuracy ($\sim 3\%$). In contrast, the random scheme forces the augmentation data to learn benefits across the whole network simultaneously, thus achieving a desirable trade-off in training a robust and well-performing ensemble.

3) Weight of Regularizer: The hyper-parameter β is used to manage the balance between clean accuracy and robustness, and we report its influence in supplementary material. It shows completely different trends against black-box and white-box attacks for small and large parameters. When the weight of the regularizer is increased, we observe that the clean accuracy and robustness against weak black-box attacks gradually deteriorate, while the robustness against strong black-box attacks gradually improves. Regarding white-box settings, larger hyper-parameters are more effective that $\beta = 5.0$ and $\beta = 10.0$ being the optimal choices for weak and strong adversaries, respectively. In this work, we choose the moderate parameter $\beta = 3.0$ to achieve the desired accuracy in classifying the clean data and the adversarial examples.

F. Visualization of Decision Region

To better analyze the intriguing property of different methods, we illustrate the decision regions of all ensemble models in Fig. 8. The visualization results of a data in CIFAR-10 drawn by the ensemble and a member are shown at the top and bottom, respectively. Each color in the plot indicates the prediction of a specific label. The vertical axis is the gradient direction of the surrogate model and the horizontal axis is a random orthogonal vector. We also include two additional setups of FASTEN, *i.e.*, either the refinement stage or the regularizer is removed from FASTEN, as the same definition of "w/o refine" and "w/o CE" mentioned in Sec. IV-D.

Vanilla and ADP exhibit weakness along the gradient due to their decision boundaries being close to the data, allowing for small adversarial perturbations to flip the predictions. The decision region of GAL is relatively large, but it remains fragile along perturbed directions. In contrast, DVERGE and TRS have larger neighborhoods around input data, as they learn vulnerabilities from augmentation data surrounding the original input. Compared to DVERGE and TRS, FASTEN has a longer adversarial distance and larger decision regions, indicating stronger robustness against attacks and random noise. Additionally, the effect of RA and CE can be summarized based on the last two columns. The refinement stage of RA improves the quality of the augmentation data so that FASTEN becomes more robust in various directions. Meanwhile, the CE regularizer smoothens decision boundaries, reducing adversarial transferability efficiently, as demonstrated in [19].

V. Conclusion

In this paper, we focus on adversarial robustness and propose a fast ensemble training method, called FASTEN. It consists of two essential components, *i.e.*, the recurrent augmentation strategy (RA) and the contrastive ensemble regularizer (CE), which reduce the computational overhead from the data and optimization perspectives. Specifically, RA uses single-step feature distillation to initialize the augmentation and refines its quality by recycling knowledge from the training trajectory, thus greatly reducing the resource requirement for generating high-quality data. CE enhances the robustness by simultaneously increasing intra-model similarity and inter-model diversity. Most of its elements are obtained in previous steps, thereby it is cost-saving by avoiding additional network passes. Extensive experiments show that FASTEN not only speeds up the training procedure by $7\times$ and $28\times$ compared to state-of-the-art DVERGE and TRS, but also achieves desirable clean accuracy and much higher robustness for defending against black-box and white-box attacks in different scenarios, including large datasets, complicated structures, and challenging adversarial training setups.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[2] P. Khosla et al., "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.

[3] H. Kwon, Y. Kim, H. Yoon, and D. Choi, "Selective audio adversarial example in evasion attack on speech recognition system," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 526–538, 2020.

[4] J. Wu, B. Chen, W. Luo, and Y. Fang, "Audio steganography based on iterative adversarial attacks against convolutional neural networks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2282–2294, 2020.

[5] L. Huang et al., "Universal physical camouflage attacks on object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 717–726.

[6] B. G. Doan, M. Xue, S. Ma, E. Abbasnejad, and D. C. Ranasinghe, "TnT attacks! Universal naturalistic adversarial patches against deep neural network systems," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 3816–3830, 2022.

[7] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Deflecting adversarial attacks with pixel deflection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8571–8580.

[8] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," 2017, *arXiv:1711.00117*.

[9] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2206–2216.

[10] L. Huang, S. Wei, C. Gao, and N. Liu, "Cyclical adversarial attack pierces black-box deep neural networks," in *Proc. Pattern Recognit.*, vol. 131, Nov. 2022, Art. no. 108831.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[12] J. Liu, C. P. Lau, H. Souri, S. Feizi, and R. Chellappa, "Mutual adversarial training: Learning together is better than going alone," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2364–2377, 2022.

[13] L. Liu et al., "Deep neural network ensembles against deception: Ensemble diversity, accuracy and robustness," in *Proc. IEEE 16th Int. Conf. Mobile Ad Hoc Sensor Syst. (MASS)*, Nov. 2019, pp. 274–282.

[14] W. Wei and L. Liu, "Robust deep learning ensemble against deception," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 4, pp. 1513–1527, Jul. 2021.

[15] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu, "Improving adversarial robustness via promoting ensemble diversity," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4970–4979.

[16] S. Kariyappa and M. K. Qureshi, "Improving adversarial robustness of ensembles with diversity training," 2019, *arXiv:1901.09981*.

[17] H. Yang et al., "DVERGE: Diversifying vulnerabilities for enhanced robust generation of ensembles," 2020, *arXiv:2009.14720*.

[18] M. M. Islam, X. Yao, and K. Murase, "A constructive algorithm for training cooperative neural network ensembles," *IEEE Trans. Neural Netw.*, vol. 14, no. 4, pp. 820–834, Jul. 2003.

[19] Z. Yang et al., "TRS: Transferability reduced ensemble via encouraging gradient diversity and model smoothness," 2021, *arXiv:2104.00671*.

[20] Y. Wu, L. Liu, Z. Xie, K.-H. Chow, and W. Wei, "Boosting ensemble accuracy by revisiting ensemble diversity metrics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16464–16472.

[21] S. Sinha, H. Bharadhwaj, A. Goyal, H. Larochelle, A. Garg, and F. Shkurti, "DIBS: Diversity inducing information bottleneck in model ensembles," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 11, pp. 9666–9674.

[22] Y. Wu et al., "Demystifying learning rate policies for high accuracy training of deep neural networks," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 1971–1980.

[23] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[24] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–10.

[25] Y. Li, L. Li, L. Wang, T. Zhang, and B. Gong, "NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks," 2019, *arXiv:1905.00441*.

[26] C. Chen, X. Zhao, and M. C. Stamm, "Generative adversarial attacks against deep-learning-based camera model identification," *IEEE Trans. Inf. Forensics Security*, early access, Oct. 2, 2019, doi: 10.1109/TIFS.2019.2945198.

[27] Y. Dong et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

[28] C. Xie et al., "Improving transferability of adversarial examples with input diversity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2725–2734.

[29] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, "Skip connections matter: On the transferability of adversarial examples generated with ResNets," 2020, *arXiv:2002.05990*.

[30] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 274–283.

[31] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4307–4316.

[32] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," 2017, *arXiv:1705.07204*.

[33] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7472–7482.

[34] A. Shafahi et al., "Adversarial training for free!" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–13.

[35] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[36] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proc. Int. Conf. Learn. Represent*, 2018, pp. 1–10.

[37] T. Strauss, M. Hanselmann, A. Junginger, and H. Ulmer, "Ensemble methods as a defense to adversarial perturbations against deep neural networks," 2017, *arXiv:1709.03423*.

[38] M. Naseer, S. H. Khan, S. Rahman, and F. Porikli, "Task-generalizable adversarial attack based on perceptual metric," 2018, *arXiv:1811.09020*.

[39] C. Xie and A. Yuille, "Intriguing properties of adversarial training at scale," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–9.

[40] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple data processing method to improve robustness and uncertainty," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–11.

[41] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[42] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–13.

[43] Y. Li, S. Bai, Y. Zhou, C. Xie, Z. Zhang, and A. Yuille, "Learning transferable adversarial examples via ghost networks," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11458–11465.

[44] M. Andriushchenko and N. Flammarion, "Understanding and improving fast adversarial training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16048–16059.

[45] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[46] C. Xie, Y. Wu, L. van der Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 501–509.

[47] F. Croce and M. Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2196–2205.

[48] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 484–501.

[49] L. Huang, C. Gao, and N. Liu, "DEFEAT: Decoupled feature attack across deep neural networks," *Neural Netw.*, vol. 156, pp. 13–28, Dec. 2022.

Lifeng Huang received the Ph.D. degree from Sun Yat-sen University. He is a currently a Lecturer with the College of Mathematics and Informatics, South China Agricultural University. His research interests include adversarial learning and deep learning security and its application.

Qiong Huang is currently a Professor with the College of Mathematics and Informatics, South China Agricultural University. His research interests include information security, cryptography, and cloud computing security and its application.

Peichao Qiu is currently pursuing the master's degree with the School of Computer Science and Engineering, Sun Yat-sen University. His research interests include deep learning security and its application.

Shuxin Wei is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Sun Yat-sen University. Her research interests include adversarial attack and its applications.

Chengying Gao received the Ph.D. degree from Sun Yat-sen University. She is currently an Associate Professor with the School of Computer Science and Engineering, Sun Yat-sen University. Her research interests include computer vision, computer graphics, and deep learning.

Erosion Attack: Harnessing Corruption To Improve Adversarial Examples

Lifeng Huang, Chengying Gao, and Ning Liu

***Abstract*— Although adversarial examples pose a serious threat to deep neural networks, most transferable adversarial attacks are ineffective against black-box defense models. This may lead to the mistaken belief that adversarial examples are not truly threatening. In this paper, we propose a novel transferable attack that can defeat a wide range of black-box defenses and highlight their security limitations. We identify two intrinsic reasons why current attacks may fail, namely data-dependency and network-overfitting. They provide a different perspective on improving the transferability of attacks. To mitigate the data-dependency effect, we propose the Data Erosion method. It involves finding special augmentation data that behave similarly in both vanilla models and defenses, to help attackers fool robustified models with higher chances. In addition, we introduce the Network Erosion method to overcome the network-overfitting dilemma. The idea is conceptually simple: it extends a single surrogate model to an ensemble structure with high diversity, resulting in more transferable adversarial examples. Two proposed methods can be integrated to further enhance the transferability, referred to as Erosion Attack (EA). We evaluate the proposed EA under different defenses that empirical results demonstrate the superiority of EA over existing transferable attacks and reveal the underlying threat to current robust models. The source code is publicly available at https://github.com/mesunhlf/EA.**

***Index Terms*— Adversarial example, adversarial attacks, transferability, adversarial defend, robustness.**

I. Introduction

DESPITE excellent results in a variety of vision tasks, deep neural networks (DNNs) are troubled with adversarial attacks that maliciously crafted adversarial examples can mislead DNNs to predict incorrect outputs [1]. This pose a potential threat to digital applications and may cause real-world systems in uncontrolled and dangerous behaviors [2], [3].

In white-box scenarios, adversaries have full knowledge about victim models so that they can have a high success rate in attacks [4]. Nonetheless, since the internal information of real-world systems is protected and unpublished (*i.e.*, black-box scenarios), those white-box attacks can hardly achieve an effective performance to fool the targeted models [5].

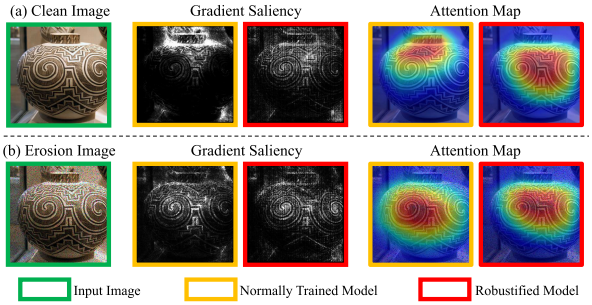


Fig. 1. **The gradient saliency maps [6] and attention regions [7] of the normally trained model compared with robustified model**. We demonstrate the visualization results of a clean & erosion image on normally trained InceptionV3 (colored in yellow) and adversarially trained $\text{InceptionV3}_{\text{ens3}}$ (colored in red), which share the same architecture but include with different weight parameters. For the clean image (top), the robustified model exhibits huge different gradient saliency and attention maps compare with the normal model. Conversely, the artificially crafted erosion images present more structured gradients and similar discriminative regions for both models (bottom).

Exploiting off-the-shelf surrogate models to generate adversarial examples for fooling remote black-box models is a more threatening way, *i.e.*, transferable adversarial attacks. So far several methods have been proposed for boosting the transferability of adversarial examples [5], [8], [9], [10], [11]. Though those transferable attacks can successfully invade most of the normally trained models under black-box settings, recent works show that adversarial examples often perform worse transferability when attacking against defense mechanisms (*e.g.*, adversarial training [12], [13], [14], gradient obfuscation [15], [16], etc.). This might give a false sense of security that adversarial examples are no longer a threat. In this paper, we propose the **Erosion Attack**, a novel and straightforward adversarial attack framework for breaking defenses with higher chances. It shows that a variety of defenses are still fragile to stronger black-box adversarial examples and these defenses are not actually secure enough to be deployed in real-world scenarios.

We have discovered two significant reasons why current transferable attack methods are less effective in defense scenarios. The first one is the data-dependency effect, where the adversarial example is highly correlated with the input data. However, defense models usually behave differently compared with normally trained ones *if the input data is a clean image*, as exemplified in Fig. 1 (a). It may indirectly impact the transferability of adversarial examples since most of methods merely consider clean images and optimize them on normally trained models. This also provides a plausible explanation for

Manuscript received 23 December 2021; revised 13 September 2022 and 20 February 2023; accepted 21 February 2023. Date of publication 14 April 2023; date of current version 29 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62272174 and Grant 61872152, in part by the Major Program of Guangdong Basic and Applied Research under Grant 2019B030302008, and in part by the Science and Technology Program of Guangzhou under Grant 201902010081. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ajmal S. Mian. *(Corresponding author: Chengying Gao.)*

Lifeng Huang is with the College of Mathematics and Informatics, South China Agricultural University, Guangzhou, Guangdong 510275, China.

Chengying Gao and Ning Liu are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong 510000, China (e-mail: mcsgcy@mail.sysu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2023.3251719, provided by the authors.

Digital Object Identifier 10.1109/TIP.2023.3251719

why these adversarial examples can transfer well to other vanilla models, but often fail to fool defenses. The second issue is the network-overfitting effect that current attacks often overfit the surrogate model with a specific decision boundary, resulting in weakened transferability across unseen black-box models. Both factors have varying degrees of impact on existing methods, ultimately leading to decreased adversarial transferability.

The observation of the data-dependency effect motivates us to search for a special augmentation source that shares similar characteristics in both vanilla models and defenses, such as gradient or attention regions. This can provide attackers with more effective guidance against robust models than using only the clean image. By incorporating the augmentation data above in the attack process, adversarial examples become less reliant on the original data and thus yields heightened transferability to fool black-box defenses, referred to as **data erosion** method. Since it differs from the category of loss-preserving augmentation techniques [5], [8], [9], we can easily combine them together to make attacks more robust.

To tackle the problem of network-overfitting, reducing the coupling between adversarial example and the surrogate model is a natural choice. One promising way is ensemble attacks, which adopt multiple networks during optimization to reduce dependence on a single model. However, this method can be computationally expensive or suffer from limited storage capacity [17], [18]. Alternatively, feature-level attacks avoid perturbing the final decision boundary of the given model, but may lose valuable information due to specific network components like pooling [19], [20]. To overcome these challenges, we propose a **network erosion** method that combines the strengths of both approaches while discarding their weaknesses. By leveraging intermediate features, it extends a single surrogate model to a corrupted network and then achieve ensemble attacks at low cost. Consequently, adversarial examples learn from diverse decision boundaries, which enhances its transferability to fool defense models.

Intuitively, the data erosion method can be seen as a *data augmentation* technique that generates good samples. On the other hand, the network erosion method constructs a diverse ensemble to achieve *model augmentation*. By integrating both of these techniques, we can create a unified framework for generating more effective and threatening adversarial examples, which we refer to as **erosion attack** (**EA**).

In summary, the contributions of our work are three-fold:

- We present a data erosion method to reduce the impact of data-dependency. Unlike existing loss-preserving augmentation techniques, our method artificially introduces corruption into clean images by taking into account the differences between normal and robust models, and then serves these erosion images as augmentations.
- We design a network erosion method to mitigate the network-overfitting effect. Instead of using multiple models like traditional methods, it creates a diverse corrupted network relying on a single existing model and achieves the ensemble attacks at little cost and high efficiency, illustrating its practicability in the field.
- Two proposed methods can be combined into a general framework for improving the strength further, namely erosion attack. Experiments demonstrate state-of-the-art results under different black-box scenarios and show the underlying insecurity of the current defenses.

II. Related Work

A. Adversarial Examples

Deep neural networks have been demonstrated to be vulnerable to adversarial attacks [21], [22]. In general, given a model f, clean data x and its groundtruth y, the inference process is defined as $f(x) = y$. The goal of an attack is to find an adversarial example x' under the constraint of p norm that misleads f_t to make the incorrect prediction, *i.e.*, $f_t(x) \neq y$. This optimization problem can be formulated as

$$\underset{x'}{\text{argmax}}\ J(x', y), \quad \text{s.t.}||x - x'||_p < \epsilon, \tag{1}$$

where $J(\cdot)$ is the optimized function (*i.e.*, cross entropy), and ϵ controls the magnitude of perturbations. Following prior attack methods, we mainly focus on fooling a variety of models under L_∞ constraint in the main paper and report performance of L_2 attack in the supplemental material.

B. Black-Box Attack Methods

We assume the attacker has no knowledge of the underlying model. One plausible line is *query-based attacks*, which estimate the gradient direction of the black-box model based on the feedback scores. These algorithms can achieve a high success rate, while they often require thousands of queries to generate a single adversarial example [23], [24], [25].

Another popular line is exploiting the transferability of adversarial examples to fool unknown models, which collectively referred to as *transferable attacks*. However, due to the data-dependency and network-overfitting effects of adversarial examples, the success rate is unsatisfactory. To address this, several methods are proposed for improving transferability.

Diversity Input Method (**DIM**) [5] introduces random transformation operations for the input image x to generate the adversarial example x', which is formalized as:

$$g_{t+1} = \nabla_x J(R(x'_t, p), y), \tag{2}$$

where g_{t+1} is the gradient at iteration $t+1$, $R(\cdot)$ is random transformation with the probability p.

Translation-Invariant Method (**TIM**) [8] calculates the gradients over an ensemble of a set of translated images, which is optimized by convolving a pre-defined kernel as:

$$g_{t+1} = W * \nabla_x J(x'_t, y), \tag{3}$$

where W is a kernel matrix with fixed kernel size k.

Scale-Invariant Method (**SIM**) [9] use the scale copies of input image to compute the gradient for updating adversarial examples, which is formulated as:

$$g_{t+1} = \sum_{i=0}^{m} \nabla_x J(S_i(x'_t), y), \tag{4}$$

where $S_i(\cdot)$ is the scaling operator with a scaled factor $1/2^i$, and m is the number of scaled images.

These methods mainly consider some loss-preserving transformations to save image details [9]. By contrast, data erosion does not belong to them but gains a smaller attention difference between the normally trained models and defenders, *i.e.*, their pixels are drastically corrupted and labels are changed sometimes. They are *different* types of augmentation so that we can compatibly combine them to form stronger methods.

Beyond this, feature-level attacks also improve the strength of adversarial examples. LAF is a strong white-box attack that directly harnesses the weakest layers in defense models [26]. FDA harnesses multiple layers to train several binary models for every attacked class [20]. FIA adopts feature-level attention to boost transferability [27]. Most of them maximize the representation distance between the clean and adversarial data, while our method leverages the most diverse corrupted network for ensemble optimization and perturbs their final outputs simultaneously, more like a kind of gradient-based attack [28]. Moreover, we merely need to train a single corrupt network for disrupting across all classes, which is more efficient and cost-saving compared with [20].

C. Defense Methods

Motivated by the threat of adversarial attacks, many parallel defense mechanisms have been proposed. In general, existing defenses can be divided into the following categories: (1) adversarial training, which treats adversarial examples as augmentation data into training processes to obtain a robustified model [22]. Moreover, ensemble training manner and randomized smoothing can further improve the robustness [12], [29], [30]; (2) gradient obfuscation, which aims to mitigate the adversarial effect by image modifications, including random input transformation [31], data compression [32], and selective feature regeneration [33], neural representation purifier [14], etc. Although these defenses have demonstrated efficacy in resisting some black-box adversarial examples, the experiments show that our proposed attack can break them in most cases.

III. Method

A. Motivation

We aim to design a robust and general transferable attack framework to fool different defenses and reveal the underlying insecurity of current models. Two issues may degrade the power of existing methods. First, the data-dependency effect may weaken transferability in that the adversarial perturbation is highly correlated with the gradient and discriminative region of its input data. However, defense models usually exhibit different attention maps and gradient saliency compared with normally trained ones when classifying a clean image [34], [35]. In general, normal models focuses on local features and presents a noisy saliency map, while the robust models concentrate on other discriminative regions and show a structured gradient aligned with human perception (Fig. 1 (a)). Therefore, current methods that rely solely on using a clean image to attack the normally trained models will significantly degrade their transferability against defenses. To avoid the data-dependency effect, we propose the **data erosion** method.

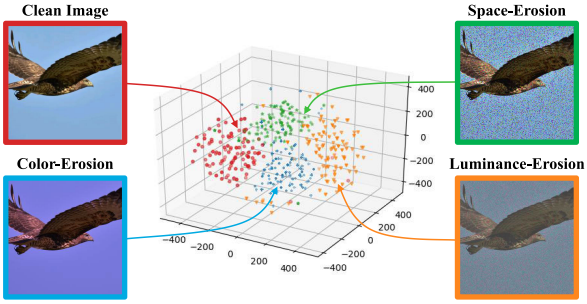


Fig. 2. **t-SNE visualization [37] of the final hidden features extracted using InceptionResNetV2**. For a clean image, we generate 100 samples by using each erosion strategy, and add small magnitude Gaussian noise ($\sigma = 0.005$) to a clean image for better illustration. The plot shows that the artificial corrupted images are drawn from distinguished distributions far away from clean images.

This method is inspired by the observation that artificial corrupted images not only exhibit different gradient saliency and heatmaps from clean images but also show smaller differences between the vanilla models and defenses. Therefore, we serve these erosion images as good augmentation data into attacks to update the perturbation, and the transferability of the generated adversarial example is improved (Sec. III-B).

The Network-overfitting effect of adversarial attacks is another factor in reducing the transferability, *i.e.*, adversarial examples perform highly overfit to the surrogate model with a specific boundary decision, and thus fail to fool other black-box models [5]. Although ensemble attacks can learn from DNNs with different boundaries to mitigate the overfitting effect, it is required to load as many models as possible with limited storage, presenting its scalability challenges [17]. Moreover, most gradient-based attacks merely consider the final prediction distribution to calculate the gradients but ignore some lost valuable features caused by the forward propagation operations [36]. To learn a variety of classified boundaries as well as gather rich intermediate representation information at little cost, we simply modify the structure of an existing model to create a corrupted network and apply it to achieve an ensemble optimization to avoid network-overfitting, namely **network erosion** method (Sec. III-C).

B. Data Erosion Method

If augmentation data demonstrates diverse gradients and attention maps, the adversarial example generated for an ensemble composed of a clean image and its augmentations will become less correlated with the original input, thus boosting its transferability to fool black-box models. Concretely, we corrupt the image to produce erosion augmentation data in the optimization as a novel attack object function for Eq. (1):

$$\underset{x'}{\operatorname{argmax}} \left\{ J(x', y) + \sum_{i=1}^{n} J(x'_{e_i}, y) \right\} \tag{5}$$

where x'_{e_i} is the erosion image, and it is produced by corrupting the current x' at every step, n is the sampling number, we follow prior works to control the perturbation under L_∞ norm constraint [5], [25], *i.e.*, $||x - x'||_\infty < \epsilon$.

In this work, we aim to study the property of different types of erosion data and exploit a bunch of augmentation schemes for transferable attacks (detailed in Secc. IV-B and Sec. IV-E). We empirically find that the best-performed augmentation data is generated by the *space-erosion* strategy that randomly damages the pixel continuity of the input image. Apart from the space-erosion strategy, we also explore other types of augmentations, such as color-erosion and luminance-erosion. Though the performance of these two strategies is slightly lower than space-erosion, they may become promising directions for future research in the field. The generated augmentation examples are shown in Fig. 2.

1) Space-Erosion: We empirically observe an interesting phenomenon that if the spatial continuity of the neighboring pixels emerged in an image has been destroyed, the normally trained models may change their discriminative region, and pay more attention to structure and general features than local details, which is analogous to the behaviors of robustified models (Fig. 1 (b)). It inspires us to damage the space continuity of the current image x' to sample the erosion data x'_e as:

$$x'_e = M \cdot x', \quad M \sim \text{Bernoulli}(1-\xi) \tag{6}$$

where M is a binary mask randomly generated in each iteration, and its elements are sampled from Bernoulli distribution, ξ is a pre-defined probability to control the erosion level. If $\xi = 0$, there is no pixel to be corrupted, and if $\xi = 1$, all element values are damaged to 0.

Note that the proposed space-erosion is similar to applying dropout operation [38] for input data, while the remained pixels are not scaled up by $1/(1-\xi)$ in our strategy. We recall that the intuitions behind dropout operation [38] and space-erosion are quite similar: dropout provides an efficient way to sample different network structures to prevent overfitting, while space-erosion distorts image details and samples good augmentations for mitigating the data-dependency effect.

2) Color-Erosion: We find that the color corrupted images have a similar property to space-erosion data. This discovery motivates our color-erosion strategy, which produces the erosion images by randomly linear combining the basic color images:

$$x'_e = \alpha_1 \cdot x'_{gb} + \alpha_2 \cdot x'_{rg} + \alpha_3 \cdot x'_{rb}, \tag{7}$$

where x'_{gb}, x'_{rg}, x'_{rb} are denoted as the basic color images for the adversarial example x' (*e.g.,* x'_{gb} is kept the elements of G and B channels, and corrupt the values of R channel, etc), $\widehat{\alpha}_i$ is a positive weight drawn from the uniform distribution U(0, 1), and the weight $\alpha_i = \widehat{\alpha}_i / \sum_j \widehat{\alpha}_j$.

3) Luminance-Erosion: The brightness perturbed images have the adversarial effect of fooling different models [39], which suggests us to explore luminance-erosion augmentation. In practice, we first transform the input image from RGB into YUV space, and then perturb the components in Y (luminance) and V (saturation) channels randomly as:

$$x'_e = \gamma_1 \cdot x'_y + x'_u + \gamma_2 \cdot x'_v, \quad \gamma_i \sim \mathcal{N}(I, \sigma_e^2) \tag{8}$$

where x'_y, x'_u, x'_v are denoted as basic luminance images (*e.g.,* x'_y preserves the components of Y channel from x', etc). γ_i is a scaling vector sampled from a normal distributions $\mathcal{N}(I, \sigma_e^2)$.

4) Difference Between Prior Augmentations and Ours: Prior methods [5], [8], [9] also use augmentations in attacks. The main differences are two-fold: (1) they adopt augmentations to avoid the local maxima, while our augmentation data is used to narrow the gaps between the normal models and defenders; (2) they select loss-preserving transformations to maintain the image details and achieve model augmentation [9], while data erosion does not belong to this type of operation since it forces models to focus on general features by corrupting the pixels, as discussed in Sec. IV-B. Therefore, it is a special augmentation designed for breaking defenses, and thus we can naturally combine them together to improve the efficiency of attacks.

C. Network Erosion Method

Although standard ensemble attacks are available to mitigate the network-overfitting effect, they require expensive computational and storage costs. To learn diverse decision boundaries by leveraging intermediate features, we attack the ensemble of an original network and its corrupted version to generate more threatening adversarial examples at little cost (Fig. 3 (a)). The objective function in Eq. (1) is reformed as:

$$\underset{x'}{\text{argmax}} \left\{ J(x', y) + J_e(x', y) \right\} \tag{9}$$

where J_e is the objective function of built corrupted network.

As illustrated in Fig. 3 (b), the main differences between the original network and its corrupted one are: (1) we corrupt the structure of the original model to build an *intermediate feature classifier* with diverse boundaries by a few fine-tuned overhead; (2) we add *stochastic layers* before convolutional blocks to further improve the model diversity and robustness.

1) Intermediate Feature Classifier: It is reasonable to assume that adversaries have a small set of images that can be correctly labeled by the surrogate model, even if those images are not belonging to both of the training sets of surrogate and black-box networks. To this end, we aim to extract the representations of these images for fine-tuning an intermediate feature classifier (IFC) at little cost. Particularly, for a given image x, we can get its predicted label y and intermediate feature z_l at l^{th} layer from a surrogate model f, *i.e.*, $y = f(x), z_l = f_l(x)$. Therefore, utilizing the standard cross entropy loss (CE) to train the feature classification head f'_c on the small substitute dataset $\mathcal{D}'$ can be expressed as:

$$\begin{aligned} J_{\text{CE}}(x, \theta') &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}'} \left[-y^{\text{T}} \log f'_c(z_l) \right] \\ &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}'} \left[-f(x)^{\text{T}} \log f'_c\big(f_l(x)\big) \right] \end{aligned} \tag{10}$$

where f'_c is the fine-tuned block of IFC, which shares the same architecture with the classification head f_c of the white-box model f (see Fig. 3 (b)). Notice, we only need to spend a few computational resources and training epochs on optimizing the weights θ' of the designed block f'_c (discussed in Sec. IV-F).

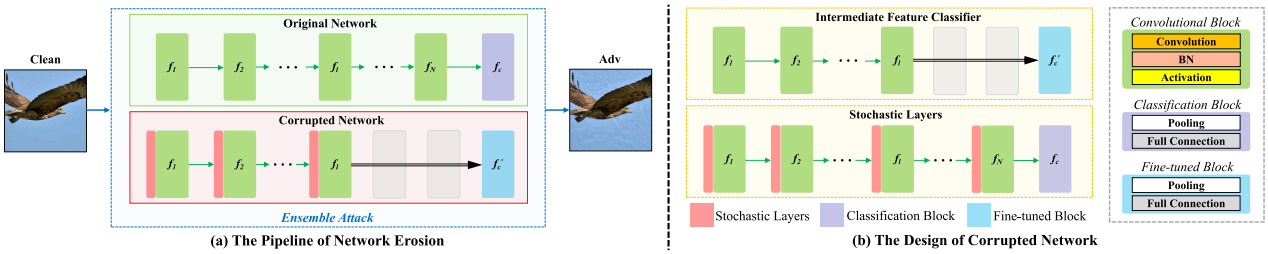


Fig. 3. **The mechanism of network erosion method**. (a) The adversarial examples are generated by attacking the ensemble of an original network (top) and a corrupted network (bottom); (b) Two essential design for the corrupted network: intermediate feature classifier, which is designed as directly connecting the feature maps f_l with a fine-tuned block f_c' (top); and stochastic layers, which are densely embedded before the convolutional blocks (bottom).

Algorithm 1 Erosion Attack Framework

Input: A clean images x; true label y; deep learning model $f(\cdot)$; loss function (*i.e.*, cross entropy) $J(\cdot)$; maximum magnitude of perturbation ϵ and iteration number T; decay factor μ;
Input: erosion transformation $E(\cdot)$ and sampling number n (for Data Erosion method); the loss function of the corrupted network $J_e(\cdot)$ (for Network Erosion method).
Output: Adversarial Example x'_T

$\alpha \leftarrow \epsilon/T$, $g_0 \leftarrow 0$, $x'_0 \leftarrow x$, $t \leftarrow 0$
while $t < T$ **do**
  $g \leftarrow 0$, $X_t \leftarrow \varnothing$
  $X_t \leftarrow X_t \cup x'_t$
  for $i = 0$ to n **do**
    $x_e \leftarrow E(x'_t)$ // proposed **Data Erosion**
    $X_t \leftarrow X_t \cup x_e$
  end for
  $\widetilde{J} \leftarrow J(X_t, y) + J_e(X_t, y)$ // proposed **Network Erosion**
  $g \leftarrow \nabla_x \widetilde{J}$
  $g_{t+1} \leftarrow \mu \cdot g_t + g/(n \cdot \|g\|_1)$
  $x'_{t+1} \leftarrow x'_t + \text{sign}(g)$
  $t \leftarrow t + 1$
end while

Intuitively, the structure of an IFC can be viewed as building a skip connection between a specific intermediate layer and the fine-tuned classification block, while the skipped layers have been "corrupted" to be out of work.

2) Stochastic Layers: We aim make the corrupted network become more diverse and robust *without* any additional training costs. Concretely, we embed the stochastic layers before each convolutional block throughout the network during the *inference phase*, which makes the features diverse when an image passing through the layers (Fig. 3 (b)). It makes the diversity of output probabilities becomes larger while keeping the classification accuracies remains high, which is validated in Sec. IV-B and supplementary. We design three different stochastic layers:

- *Random Scaling Layer*. During the training stage, directly adding Gaussian noise to the layers can improve the robustness of the trained model [40]. However, it is difficult to quantitatively measure the magnitude of features across the layers to control the noise level. Instead, we propose to scale the features within a small range as:

$$z_{l+1} = g_l^{\mathfrak{r}}(z_l) = s \cdot z_l, \quad s \sim \mathcal{N}(I, \sigma^2) \tag{11}$$

where $g_l^{\mathfrak{r}}(\cdot)$ is random scaling operation at layer l, z_{l+1} is the feature map for next layer, s is a vector sampled from the same dimensional space of x_l, and we set $\sigma = 0.025$.

- *Feature Smoothing Layer*. Generally, perturbations are small at the pixel level and gradually increase at feature space, *i.e.*, the noises are activated as the image propagates through the network [41]. Inspired by [41], we introduce a Gaussian kernel to smooth the features and improve the robustness of the model to be attacked:

$$z_{l+1} = g_l^{\mathfrak{f}}(z_l) = W_d * z_l \tag{12}$$

where $g_l^{\mathfrak{f}}(\cdot)$ is denoted as feature smoothing operation, W_d is a normalized Gaussian filter with 3×3 kernel size to implement convolution operation with feature map z_l.

- *Cascade Group Layer*. For further improving the model diversity at inference stage, we consider to integrate two proposed layers into a single module as:

$$z_{l+1} = g_l^{\mathfrak{s}}(z_l) = g_l^{\mathfrak{r}} \circ g_l^{\mathfrak{f}}(z_l) \tag{13}$$

where $g^{\mathfrak{s}}$ denoted as cascade group, $g_l^{\mathfrak{r}}$ and $g_l^{\mathfrak{f}}$ are random scaling (Eq. (11)) and feature smoothing (Eq. (12)), respectively, symbol $\circ$ is defined as composite function, *i.e.*, $g_l^{\mathfrak{r}} \circ g_l^{\mathfrak{f}}(z_l) = g_l^{\mathfrak{r}}\big(g_l^{\mathfrak{f}}(z_l)\big)$.

Note that the idea of random scaling and feature smoothing are proposed for resisting adversarial attacks, while we first explore their potential capability for enhancing transferability. Moreover, their implemental details are different from existing methods [40], [41] to avoid tremendous training costs (*e.g.*, removing 1×1 convolution embeddings, only adopting at the inference stage, etc.). We experiment with these layers in Sec. III-C and supplemental material, and find that the best one is the *cascade group* layer.

D. Erosion Attack Framework

In summary, data erosion introduces good augmentations with different attention regions for mitigating the data-dependency effect. Network erosion creates a corrupted network with diverse decision boundaries to prevent the network-overfitting effect. We combine them into a general framework to boost the adversarial transferability of attacks further, namely **erosion attack** (EA). The algorithm is described in Alg. 1.

It is noteworthy that current transferable attacks can be naturally integrated into the proposed erosion attack framework

to generate stronger adversarial examples. For example, the integration version of TIM [8] and EA (denoted as EA-TIM) only need to update line 10 in Alg. 1 as:

$$g = W * \nabla_x \widetilde{J} \tag{14}$$

Similarly, we can establish the combination versions for DIM [5] and SIM [9], denoted as EA-DIM and EA-SIM, respectively. More details are provided in supplementary.

IV. Experiments

In this section, we start with the experimental setup in Sec. IV-A, and verify the intriguing properties of our method in Sec. IV-B. Then we report the results of the proposed erosion attack method under a single model or ensemble setting in Sec. IV-C and Sec. IV-D. Finally, we provide a rich collection of ablation studies for our proposed data & network erosion methods in Sec. IV-E and Sec. IV-F.

A. Experimental Setting

1) Models and Defenses: Following [11] and [27], we attack white-box normally trained models to generate adversarial examples, and evaluate the success rates of fooling the black-box models. The normally trained models include InceptionV3 (I3), InceptionV4 (I4) [42], Inception-ResNetV2 (IR2) [43], ResNet50 (R50) [44]. We also include DenseNet161 (D161) [45] as a black-box model. We report the results of I3 and IR2 in the main paper and others in supplemental material.

We consider thirteen robust models:

- Robustified InceptionV3$_{ens3}$ (I3$_{ens3}$), InceptionV3$_{ens4}$ (I3$_{ens4}$), and InceptionResNetV2$_{ens}$ (IR2$_{ens}$) [12], [22].
- High-level Representation Guided Denoiser (HGD) [15].
- Input transformation with bit depth reduction (BDR) [46].
- The combinations of pixel deflection and wavelet denoising or total variance minimization (PDW and PDT) [16].
- Randomly resizing and padding (R&P) [31].
- Compression and reconstruction models (COM) [32].
- Randomized smoothing (RS) [29].
- Selective feature regeneration for defense (SFR) [33].
- Neural representation purifier (NRP) [14].
- NAS searched Robust Architectures (ROB) [47].

For those of test-time "plug-in" defenses, we choose adversarial I3$_{ens3}$ as underlying model to form a more robust defense.

2) Dataset: Following [9], we randomly selected 1000 images from ImageNet `validation`. We note that these images also be collected in the 2nd SACP2019 dataset (Tianchi Security AI Challenger Program Competition). Almost all images can be correctly classified by normally trained models.

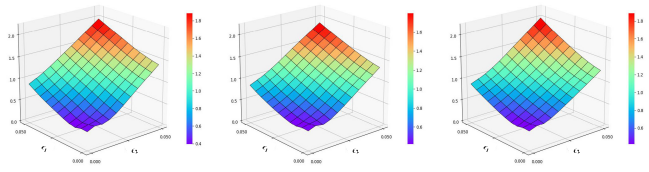
3) Baselines: We include three art transferable attacks DIM, TIM, and SIM [5], [8], [9]. In our experiments, we select the *strongest combination* versions reported in literature [5], [8], [9] as default methods unless explicitly stated, denoted as DIM*, TIM*, SIM*, respectively (*e.g.*, TI-DIM for TIM*, SI-NI-TI-DIM for SIM*, etc). To demonstrate the efficacy of our method, we can integrate the proposed erosion attack (EA) with above methods as EA-DIM*, EA-TIM*, and EA-SIM*, respectively. In addition, ResNet specific SGM [10], the current best combination attack VT [11], and two state-of-the-art feature-level attacks LAF [26] and FIA [27] are considered in the experiments. Since SGM only serves ResNets structures, we compare with it on R50 and report the results in the supplemental material.

4) Parameter Setup: We set the maximum perturbation $\epsilon = 0.05$ in normalized pixel [0, 1], total iterations $T = 10$ with the step size $\alpha = 0.005$. For baselines, we follow [8] and [27] to set momentum decay $\mu = 1.0$, transformation probability $p = 0.7$ for DIM, Gaussian kernel size 11×11 for TIM, the number of scaled images $m = 4$ for SIM, skip decay $\gamma = 0.2$ for SGM, variance tuned $\beta = 1.5$ and $N = 20$ for VT. For LAFEAT, we use its proposed DLR loss combined with penultimate trained logits layers. For FIA, we choose the optimal layers, *i.e.*, $Mixed\_5b$, $Mixed\_6f$, $Conv\_4a$, and $block2\_unit6$ for I3, I4, IR2, and R50, respectively. The parameters of our method are discussed in Sec. IV-E and Sec. IV-F.

B. Analysis of Erosion Attack

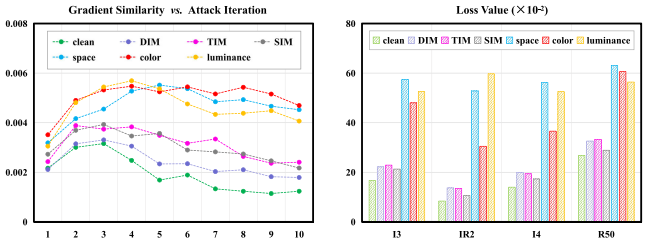
1) Data Erosion Method: To verify the qualification of erosion images as good samples for transferable attacks, we use IR2 as a surrogate model to visualize the loss surface of the black-box defense I3$_{ens3}$ in Fig. 4. The results are computed over 1000 images selected from ImageNet `validation`. For the proposed data erosion method, we observe that the incremental losses along with the gradient calculated by all of the data erosion strategies (w/ erosion augmentations) greatly exceed those along with the clean image (w/o erosion augmentations), which validates our generated erosion data can be used as good samples to facilitate transferable attacks.

We further plot the average cosine similarity of gradient between normal and robust models to discuss the different behaviors of several types of augmentation techniques, where the images generated by using our method are denoted as space, color, and luminance, respectively (see Fig. 5 (a)). Compared with clean images and existing techniques (*i.e.*, DIM, TIM, and SIM), the gradients of erosion images computed at the normally trained model are much more similar to their gradients at the robust model (*i.e.*, they have higher cosine values). It is largely because prior methods adopt conventional transformations like resizing or scaling to maintain the image details, but data erosion strategies severely distort the pixels, guiding normally trained models to concentrate on the general structure information and object-related features. Moreover, the gradient similarities decrease with the attack iteration increased for clean images and prior methods, while they keep relatively stable across optimization iterations for erosion images. It confirms that artificial corrupted data can guide the vanilla model in calculating more transferable perturbations against robust defenses. We also produce two sets of adversarial examples, *i.e.*, one is augmented with erosion images, and another one is merely relying on the clean images. We report the cosine similarity between these two sets in the supplemental material. It illustrates that the correlation of gradients



(a) Space-Erosion (b) Color-Erosion (c) Luminance-Erosion

Fig. 4. **The loss surface of robustified $I3_{ens3}$ over 1000 images**. We plot the loss surface on points $x' = x + \epsilon_1 \cdot g_1 + \epsilon_2 \cdot g_2$, where g_1 is the signed gradient of IR2 on clean image x (w/o erosion augmentations) calculated by Eq. (1); and g_2 is the signed gradient computed by erosion strategies (w/ erosion images as augmentations) in Eq. (5).



(a) Similarity Between Normal & Robust Models (b) Loss Value of Different Augmentations

Fig. 5. **Analysis of data erosion method.** We select 1000 images to measure the metrics, including: (a) the gradient similarity of different types of augmentation images between the robustified $I3_{ens3}$ and the normally trained IR2 across attack iterations; (b) the average cross entropy values of different types of augmentation images on four normally trained models.

	IR2	I3	R50	IFC-A	IFC-B	IR2-RS	IR2-FS	IR2-SC
IR2	0.0	7.9	10.1	10.6	7.4	2.9	2.2	4.3
I3	7.9	0.0	9.7	13.9	11.5	8.5	8.0	8.8
R50	10.1	9.7	0.0	10.1	11.7	10.5	9.9	10.4
IFC-A	10.6	13.9	10.1	0.0	6.7	11.3	10.7	11.5
IFC-B	7.4	11.5	11.7	6.7	0.0	8.7	8.0	9.5
IR2-RS	2.9	8.5	10.5	11.3	8.7	0.0	4.1	3.6
IR2-FS	2.2	8.0	9.9	10.7	8.0	4.1	0.0	2.9
IR2-SC	4.3	8.8	10.4	11.5	9.5	3.6	2.9	0.0

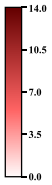


Fig. 6. **Average pairwise JSD ($\times 10^{-2}$) values**. The JSD are calculated between normally trained models and corrupted models over whole ImageNet `validation` dataset. The higher value, the larger diversity.

between two sets is decreasing dramatically, which provides evidence that introducing erosion data as augmentations can effectively alleviate the data-dependency effect.

In addition, we record the average losses and the prediction accuracy between erosion images and prior augmentations in Fig. 5 (b) and supplemental material. The losses of erosion images are largely increased across all models, which sometimes flips the label and decreases the prediction accuracy. By contrast, those existing methods almost keep the same accuracy and losses as clean images. This observation is consistent with the tendency in Fig. 5 (a) that drastic corruptions of the input image lead to higher gradient similarities and loss values. It provides evidence that our method does not belong to those loss (label)-preserving transformations, and thereby we can combine it with existing augmentation naturally to enhance the transferability further, as validated in Sec. IV-E.

2) Network Erosion Method: We aim to build a diverse corrupted network that has different decision boundaries compared with the original model, *i.e.*, both intermediate feature classifier and stochastic layers can improve the model diversity. Following [18], we adopt the Jensen-Shannon Divergence

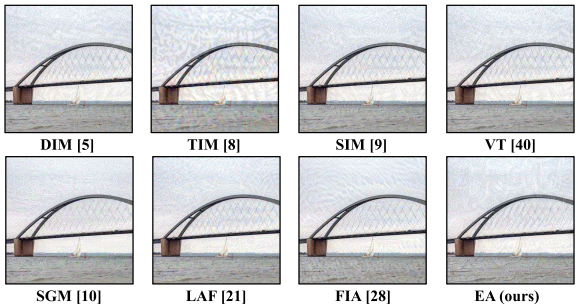


Fig. 7. **Visualized Adversarial examples**. We attack R50 by using baselines and the proposed EA. Though their visual imperceptibility are similar, EA can lead to much higher transferable success rate than other attacks.

(JSD) to evaluate the diversity, and report the pairwise JSD of the output probability distribution of networks in Fig. 6.

Without loss of generality, we consider eight networks, including: (1) three normally trained models IR2, I3, R50; (2) two intermediate feature classifiers fine-tuned by *Block17* and *Mixed_7a* layers of IR2, namely IFC-A, IFC-B, respectively; (3) separately embedding stochastic layers (*i.e.*, random scaling, feature smoothing, and cascade group) into IR2 structure, denoted as IR2-RS, IR2-FS, IR2-SC, respectively.

The implications of Fig. 6 are three-fold: (1) the diversity between the original model IR2 and its corrupted versions is large and even comparable with other vanilla models (*i.e.*, I3, R50); (2) intermediate feature classifiers and the network embedded with stochastic layers also exhibit significant diversity among each other; (3) cascade group layer can bring larger diversity to models than random scaling and feature smoothing in most cases, which may affect the performance of transferability (see Sec. IV-F). These results suggest corrupted network is qualified to enhance the adversarial examples.

C. Single-Model Attack Experiment

We evaluate the performance of the proposed erosion attack (EA). Concretely, we adopt space-erosion to generate erosion data and choose the most diverse feature classifier embedded with cascade group layers to build a corrupted network. The EA-based versions for DIM*, TIM*, and SIM* are denoted as EA-DIM*, EA-TIM*, and EA-SIM*, respectively (Sec. III-D). We also include the FIA-based attacks [27] and the current strongest combination VT* [11] in the comparisons.

1) Standalone Experiment: To demonstrate the efficacy of EA, we show the standalone performance (*not* the combination versions) of baselines and EA under following setups (see Tab. I): (1) three black-box normally trained models (3 $\sim$ 5 columns); and (2) five robustified defenses (6 $\sim$ 10 columns).

Among standalone baselines, FIA and DIM are the best attackers to mislead black-box normally trained models, while TIM and SIM perform better than the others under most defense scenarios. It is reasonable that LAF gains low transferability due to it is designed for white-box scenarios. Notice, the proposed EA outperforms all baselines by a large margin against both normally trained models and robust defenses (*e.g.*, adversarial examples crafted on IR2: average improvements

TABLE I
STANDALONE PERFORMANCE EXPERIMENTS

Model	Method	I4	R50	D161	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Avg**
I3	DIM [5]	56.9	45.3	46.3	13.5	4.1	14.4	14.1	15.8	26.3
	TIM [8]	28.7	21.5	30.7	16.6	8.7	16.9	15.4	16.6	19.4
	SIM [9]	42.7	36.4	37.2	16.0	4.9	16.4	14.9	15.5	23.0
	VT [11]	51.7	38.5	41.3	15.7	5.5	16.2	14.2	15.9	24.9
	LAF [26]	43.3	34.0	32.8	9.3	3.6	9.2	11.8	11.9	19.5
	FIA [27]	61.9	53.9	49.1	17.1	6.9	16.5	**22.4**	19.4	31.0
	EA (ours)	**70.7**	**57.8**	**59.1**	**21.4**	**9.3**	**22.5**	21.5	**22.5**	**35.6**
IR2	DIM [5]	64.6	56.8	51.4	27.2	15.2	28.0	17.3	29.2	36.2
	TIM [8]	45.0	35.8	47.5	29.9	25.2	30.0	21.5	28.8	33.0
	SIM [9]	56.9	50.0	52.1	27.8	17.2	28.7	21.1	26.5	35.0
	VT [11]	53.3	45.9	44.2	26.0	14.6	25.7	21.5	26.4	32.2
	LAF [26]	40.7	32.5	31.7	13.5	6.0	13.2	13.0	12.0	20.3
	FIA [27]	52.5	50.6	45.9	27.0	17.1	26.5	26.0	26.8	34.2
	EA (ours)	**82.6**	**71.7**	**73.9**	**45.5**	**28.7**	**43.8**	**32.2**	**42.8**	**52.7**
Model	**Method**	**I3**	**IR2**	**D161**	$I3_{ens3}$	$IR2_{ens}$	**HGD**	**BDR**	**R&P**	**Avg**
I4	DIM [5]	68.4	56.0	51.1	15.1	6.6	15.2	15.7	17.6	26.3
	TIM [8]	38.9	26.0	33.2	18.0	12.1	18.3	16.9	18.2	22.7
	SIM [9]	56.6	48.9	45.4	20.9	9.6	21.8	17.2	20.3	30.1
	VT [11]	61.0	50.6	42.6	16.1	7.5	16.0	13.1	16.7	28.0
	LAF [26]	49.4	39.0	31.1	9.7	4.2	9.8	11.5	11.5	20.8
	FIA [27]	57.3	46.5	40.6	18.6	13.3	17.8	19.9	18.4	29.1
	EA (ours)	**80.1**	**69.6**	**66.1**	**31.1**	**13.4**	**31.6**	**26.2**	**31.6**	**43.7**
R50	DIM [5]	56.7	49.2	57.2	19.3	8.3	16.9	17.1	20.9	30.7
	TIM [8]	29.5	20.1	38.0	17.6	11.7	17.8	17.6	18.2	21.3
	SIM [9]	49.5	42.0	52.3	19.2	8.2	20.0	18.3	20.5	28.8
	VT [11]	53.1	47.2	52.2	21.2	9.9	20.7	17.3	20.4	30.3
	SGM [10]	59.4	50.3	62.2	19.3	8.6	16.8	20.0	22.8	32.4
	LAF [26]	46.4	34.6	43.9	13.4	5.6	12.3	13.4	14.6	23.0
	FIA [27]	62.7	54.1	59.5	22.9	10.2	21.4	21.9	23.1	34.5
	EA (ours)	**67.4**	**56.6**	**64.4**	**24.2**	**12.6**	**27.0**	**23.9**	**27.0**	**37.9**

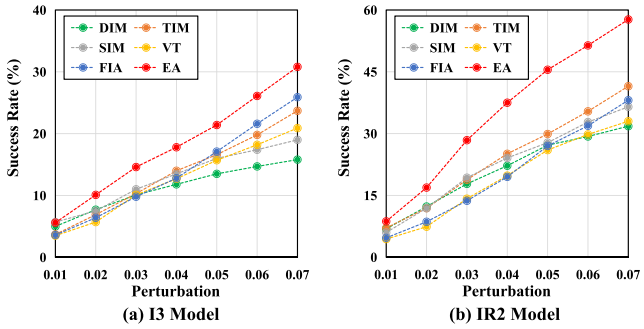


Fig. 8. **Transferability of attacks on different budgets**. We attack I3 and IR2 under various perturbation magnitudes to generate adversarial examples and report their transferability against the $I3_{ens3}$ defense.

are 16.5% and 19.7% compared with DIM and TIM, etc.), thereby establish state-of-the-art for transferable adversarial attacks. Some generated results are shown in Fig. (7) and supplemental material.

In addition, we generate adversarial examples under various perturbation budgets (*i.e.*, $0.01 \sim 0.07$ with step size 0.01) and report the transferable success rates against $I3_{ens3}$ defense in Fig. 8. An interesting phenomenon is that attacks may exhibit different trends, *e.g.*, DIM is strong when the budget is small, but it becomes weaker than other baselines with the adversarial magnitude increased. Among them, the proposed EA consistently achieves the best performance, further suggesting the generalizability of the proposed method.

2) Cross-Network Experiment: We report the combination versions of attacks against six undefended models

TABLE II
COMBINATION PERFORMANCE OF CROSS-NETWORK EXPERIMENTS

Model	Method	I3	IR2	I4	R50	R152	D161	**Avg**
I3	DIM* [5]	98.0*	50.4	56.9	45.3	39.2	46.3	56.0
	FIA-DIM* [27]	96.7*	58.7	63.0	55.9	52.1	53.3	63.3
	EA-DIM* (ours)	**98.8***	**77.4**	**80.8**	**72.6**	**67.5**	**75.6**	**78.8**
	TIM* [8]	96.8*	39.3	49.3	36.4	32.3	49.4	50.6
	FIA-TIM* [27]	96.6*	44.2	52.3	44.4	39.4	54.5	55.2
	EA-TIM* (ours)	**98.9***	**63.0**	**72.3**	**60.0**	**55.4**	**76.7**	**71.1**
	SIM* [9]	97.6*	50.4	59.0	46.2	44.1	62.8	60.0
	VT* [11]	97.7*	60.1	61.5	49.7	49.1	64.9	63.8
	FIA-SIM* [27]	97.0*	51.3	60.4	55.6	48.8	70.1	63.9
	EA-SIM* (ours)	**98.3***	**70.5**	**76.9**	**67.2**	**62.3**	**85.7**	**76.8**
IR2	DIM* [5]	69.9	97.1*	64.6	56.8	51.4	57.8	66.3
	FIA-DIM* [27]	65.8	93.0*	58.2	58.4	54.1	55.4	64.2
	EA-DIM* (ours)	**87.5**	**98.6***	**85.0**	**80.6**	**77.9**	**83.7**	**85.6**
	TIM* [8]	63.2	96.1*	61.8	50.9	48.0	61.8	63.6
	FIA-TIM* [27]	51.0	92.7*	44.8	44.8	40.1	52.6	54.3
	EA-TIM* (ours)	**79.4**	**97.6***	**76.5**	**70.2**	**67.8**	**82.9**	**79.1**
	SIM* [9]	63.7	92.1*	61.4	55.8	51.9	68.6	65.6
	VT* [11]	66.4	93.7*	64.2	57.9	56.4	70.9	68.3
	FIA-SIM* [27]	64.6	93.2*	58.8	60.8	56.8	63.9	66.3
	EA-SIM* (ours)	**81.7**	**97.7***	**77.2**	**73.8**	**69.3**	**89.8**	**81.6**

(*e.g.*, I3, IR2, I4, R50, R152, and D161) in Tab. II, qualifying the efficiency of EA framework. We compare EA with SGM on R50 in supplementary. The results of LAF are not listed because it mainly focuses on breaking white-box defenses, and shows degraded transferability against black-box models.

We observe that our erosion attack improves the success rates under the white-box scenarios (labeled as "*"). For more concerning black-box settings, we notice erosion attack based methods consistently outperform their baseline versions by a large margin (*e.g.*, the average increments are 22.8%/20.5%/16.8% on I3 by using EA-based methods, etc.). It is noteworthy to mention that EA-DIM* is the *strongest* transferable attack against the normally trained models, and it may be caused by the dual data augmentation effect from both DIM and EA. The quantitative results suggest the superiority of our proposed EA for fooling different undefended networks under both white-box and black-box scenarios.

3) Defenses Experiment: The success rates of combination attacks against eleven defenses are recorded in Tab. III. We can see the performances of proposed methods largely outperform the baselines across all the defenses regardless of the attacked surrogate models or defense mechanisms. For instance, the average success rates of TIM* crafted on I3/IR2 are improved by 28.1%/25.9% when integrating it into the proposed EA framework. In particular, we find that EA-SIM* achieves the best performance against defenses by attacking the IR2 model (*i.e.*, average success rate of 74.5%). The results under various defenses and undefended models (see supplementary) qualify the efficacy of EA to trick current black-box models.

D. Ensemble-Model Attack Experiment

Following [8] and [27], we further provide the results of baselines and proposed EA under ensemble-model settings, *i.e.*, simultaneously attacking multiple vanilla networks. Specifically, four normally trained models are considered in ensemble paradigms, including I3, IR2, I4, and R50. The comparisons are conducted under the following settings:

TABLE III

THE SUCCESS RATES (%) OF COMBINATION ATTACKS USING BASELINES AND PROPOSED EROSION ATTACK UNDER $\epsilon = 0.05$

Model	Method	$I3_{ens3}$	$I3_{ens4}$	$IR2_{ens}$	HGD	BDR	PDW	PDT	R&P	COM	RS	SFR	NRP	**Average**
I3	DIM* [5]	13.2	14.3	4.3	15.8	14.5	31.2	23.7	16.3	21.3	5.1	23.1	8.6	16.0
	FIA-DIM* [27]	20.9	24.1	7.8	20.9	24.2	30.5	28.4	24.5	35.7	9.7	29.2	9.9	22.2
	EA-DIM* (ours)	**30.9**	**32.8**	**14.5**	**30.6**	**28.3**	**50.6**	**44.4**	**35.1**	**51.0**	**17.7**	**54.4**	**14.2**	**33.7**
	TIM* [8]	29.1	32.3	17.7	29.0	20.2	34.3	34.2	32.2	31.5	16.9	29.7	11.7	26.6
	FIA-TIM* [27]	31.6	34.2	22.6	31.2	32.8	36.3	36.0	33.3	42.4	21.6	33.3	18.6	31.2
	EA-TIM* (ours)	**60.4**	**59.5**	**43.1**	**60.3**	**45.8**	**65.4**	**62.2**	**58.6**	**62.6**	**40.0**	**64.2**	**34.0**	**54.7**
	SIM* [9]	44.3	43.9	30.9	44.9	31.0	49.8	44.6	44.4	44.3	21.8	41.4	18.2	38.3
	VT* [11]	52.6	52.2	38.0	53.2	42.3	59.7	60.7	53.9	53.8	30.7	55.1	29.3	48.5
	FIA-SIM* [27]	42.8	45.0	30.3	43.4	44.1	48.0	46.8	43.2	54.2	29.3	46.6	23.8	41.5
	EA-SIM* (ours)	**70.4**	**70.2**	**54.3**	**70.5**	**60.7**	**76.1**	**75.4**	**68.7**	**75.7**	**51.5**	**71.0**	**44.2**	**65.7**
IR2	DIM* [5]	25.3	24.3	13.3	28.7	17.1	41.5	30.1	28.8	31.4	8.2	38.8	9.6	24.8
	FIA-DIM* [27]	26.7	26.1	15.6	25.7	26.0	31.1	32.6	24.9	36.1	14.4	34.5	13.0	25.6
	EA-DIM* (ours)	**54.3**	**49.8**	**36.0**	**55.6**	**37.7**	**61.4**	**53.8**	**54.8**	**64.3**	**23.5**	**62.8**	**20.4**	**47.9**
	TIM* [8]	44.6	43.1	42.3	45.4	29.9	50.3	43.6	45.5	45.6	21.6	47.0	17.0	39.7
	FIA-TIM* [27]	35.5	37.3	30.4	35.6	35.2	39.1	38.0	39.4	44.2	28.1	40.4	24.7	35.7
	EA-TIM* (ours)	**70.3**	**68.6**	**67.5**	**69.9**	**58.5**	**73.7**	**67.7**	**68.6**	**73.9**	**49.2**	**73.6**	**42.3**	**65.6**
	SIM* [9]	54.1	53.1	53.6	54.1	43.1	58.1	54.4	53.3	55.8	30.1	52.9	29.5	49.3
	VT* [11]	59.5	58.5	60.5	61.6	53.6	67.8	70.9	64.6	67.7	41.4	65.3	37.6	59.1
	FIA-SIM* [27]	48.4	51.4	42.8	48.1	52.0	54.2	53.2	59.9	57.6	43.3	51.7	29.3	49.3
	EA-SIM* (ours)	**78.9**	**76.9**	**74.2**	**78.8**	**72.1**	**82.9**	**81.1**	**76.4**	**82.8**	**59.1**	**78.5**	**52.7**	**74.5**

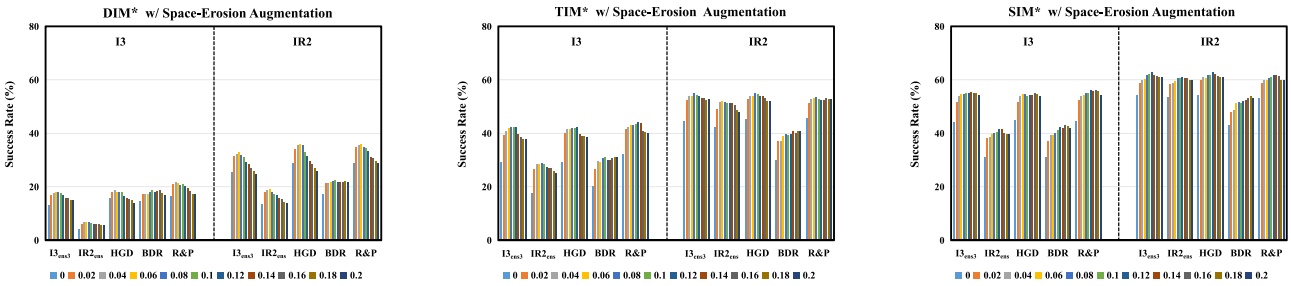


Fig. 9. **Ablation studies of space-erosion on I3 and IR2**. The adversarial examples are crafted by DIM*, TIM* and SIM* augmented with a space-erosion image on I3 and IR2. We can observe that the transferability is obviously boosted even a small portion of pixels to be damaged. Better viewed with zoom-in.

TABLE IV

THE SUCCESS RATES (%) OF ENSEMBLE-MODEL ATTACKS USING BASELINES AND PROPOSED EROSION ATTACK UNDER $\epsilon = 0.05$

Setup	Method	$I3_{ens3}$	$I3_{ens4}$	$IR2_{ens}$	HGD	BDR	PDW	PDT	R&P	COM	RS	SFR	NRP	ROB	**Average**
4-models	DIM* [5]	48.1	45.8	26.9	52.7	26.8	59.8	43.5	54.9	54.7	13.9	58.3	14.3	10.8	39.3
4-models	FIA-DIM* [27]	48.7	46.3	21.3	47.9	39.1	57.3	47.1	56.5	69.9	22.4	64.1	15.7	12.9	42.2
2-models	EA-DIM* (ours)	62.2	60.0	39.8	63.1	45.3	74.5	63.3	64.6	72.6	28.3	72.4	24.0	14.4	52.7
4-models	EA-DIM* (ours)	**75.1**	**72.1**	**55.7**	**75.4**	**57.4**	**79.1**	**71.3**	**77.1**	**79.7**	**38.3**	**78.9**	**30.1**	**16.0**	**62.0**
4-models	TIM* [8]	68.8	68.4	60.6	69.0	43.3	68.4	59.9	69.3	66.1	34.3	65.8	28.7	14.8	55.2
4-models	FIA-TIM* [27]	71.0	68.0	53.8	70.8	64.1	74.5	69.2	70.7	82.4	51.3	68.4	35.9	16.2	61.3
2-models	EA-TIM* (ours)	79.1	78.0	69.7	78.9	63.4	79.6	76.6	77.1	78.3	56.0	80.9	51.0	19.0	68.3
4-models	EA-TIM* (ours)	**83.0**	**82.1**	**75.4**	**83.1**	**70.9**	**84.8**	**81.1**	**82.3**	**84.5**	**63.2**	**84.1**	**54.7**	**21.5**	**73.1**
4-models	SIM* [9]	76.7	75.9	70.7	76.9	60.5	78.0	71.1	76.7	75.3	46.1	78.8	43.7	16.6	65.2
4-models	VT* [11]	83.0	81.1	75.2	84.4	71.9	82.9	78.3	82.5	84.6	55.4	83.3	55.9	20.3	72.2
4-models	FIA-SIM* [27]	70.4	68.2	54.1	70.3	64.5	71.8	68.4	70.7	81.7	50.1	67.6	39.7	16.9	61.1
2-models	EA-SIM* (ours)	85.0	85.6	77.7	84.4	78.1	86.0	85.1	84.5	87.3	67.4	84.1	65.1	25.3	76.6
4-models	EA-SIM* (ours)	**91.0**	**88.1**	**81.1**	**90.6**	**80.7**	**90.9**	**88.1**	**90.2**	**91.3**	**69.1**	**89.9**	**71.4**	**28.7**	**80.9**

(1) the adversarial images are crafted on the ensemble of all of the four surrogate models, denoted as *4-models*; and (2) only the ensemble of I3 and IR2 to be attacked, denoted as *2-models*. We choose the same parameters as Sec. IV-C. Note that ROB is the current strongest defense, therefore we trick it in the ensemble setup.

From Tab. IV, we observe that the proposed EA significantly improves the transferability of adversarial examples under ensemble settings. Our strongest attack EA-SIM* achieves a high average performance of 80.9% under the *4-models* setup, which surpasses the most challenging baseline VT* by a large margin of 8.7%. Moreover, we notice that EA-based methods are better than *4-models* strongest baselines even though they only use two surrogate models in the ensemble optimization (*i.e.*, *2-models* setting), which demonstrates the superiority of the proposed EA. Another non-trivial phenomenon is that the improvements of EA-based methods between *2-models* and *4-models* setups are relatively small (*i.e.*, 4.8% and 4.3% for EA-TIM* and EA-SIM*). This confirms that our method can use a few existing networks with low fine-tuning costs to improve the model diversity, rather than simply training the new models with different architectures and applying them into standard ensemble attacks to reach this goal.

E. Ablation Studies for Data Erosion Method

In this section, we provide ablation studies for data erosion strategies. Particularly, we first introduce *only one* corrupted image as an augmentation (*i.e.*, $n = 1$ in Eq. (5)) to find

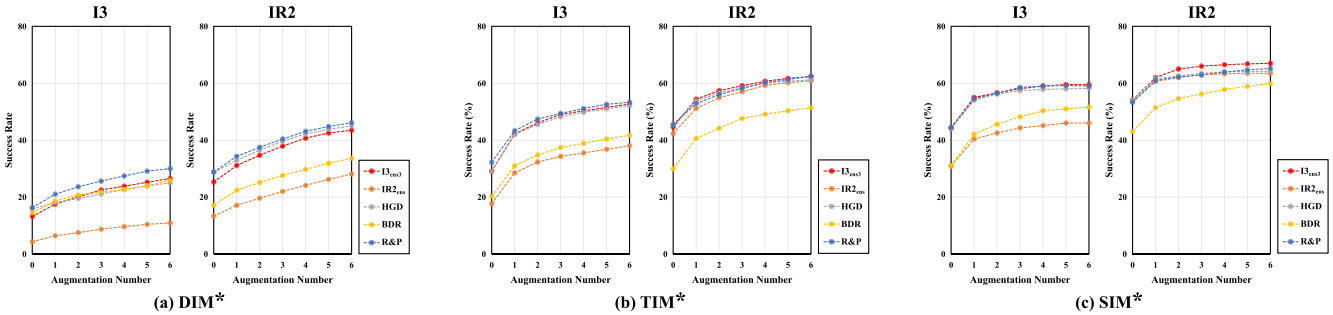


Fig. 10. **Ablation studies of augmentation number against five defenses**. The attacks are crafted on I3 and IR2 model by (a) DIM*, (b) TIM* and (c) SIM*. More augmentation images lead to better performance. Better viewed with zoom-in.

TABLE V

THE SUCCESS RATES (%) OF DIM*, TIM* AND SIM* ATTACKS WITH *Only One* AUGMENTATION AGAINST FIVE DEFENSE MODELS

(a) Comparisons of different augmentation schemes for DIM* by attacking white-box I3/IR2 model

Model	Augmentation	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**	Model	Augmentation	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**
I3	DIM* [5]	13.2	4.3	15.8	14.5	16.3	12.8	IR2	DIM* [5]	25.3	13.3	28.7	17.1	28.8	22.6
	+ *space* (**ours**)	17.5	6.4	18.0	**18.6**	21.0	**16.3 (+3.5)**		+ *space* (**ours**)	31.1	17.1	32.9	**22.4**	34.3	27.6 (+5.0)
	+ *color* (**ours**)	**17.7**	**6.5**	**19.1**	16.2	**21.1**	16.1 (+3.3)		+ *color* (**ours**)	**33.0**	**18.5**	**35.5**	18.9	**37.1**	**28.6 (+6.0)**
	+ *luminance* (**ours**)	16.5	5.8	16.9	17.4	19.4	15.2 (+2.4)		+ *luminance* (**ours**)	28.0	15.7	32.7	21.0	32.3	25.9 (+3.3)
	+ *random* (**ours**)	16.7	6.1	16.7	18.2	18.8	15.3 (+2.5)		+ *random* (**ours**)	30.5	16.3	30.5	23.0	30.3	26.1 (+3.5)
	+ *mixup* (**ours**)	17.1	6.6	17.1	17.8	20.1	15.7 (+2.9)		+ *mixup* (**ours**)	31.5	18.5	31.5	21.1	34.7	27.5 (+4.9)

(b) Comparisons of different augmentation schemes for TIM* [8] by attacking white-box I3/IR2 model

Model	Augmentation	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**	Model	Augmentation	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**
I3	TIM* [8]	29.1	17.7	29.0	20.2	32.2	25.6	IR2	TIM* [8]	44.6	42.3	45.4	29.9	45.5	41.5
	+ *space* (**ours**)	**42.2**	**28.5**	**42.0**	**31.0**	**43.3**	**37.4 (+11.8)**		+ *space* (**ours**)	54.4	**51.2**	54.6	**39.4**	52.8	**50.5 (+9.0)**
	+ *color* (**ours**)	40.7	26.2	40.6	28.0	42.8	35.7 (+10.1)		+ *color* (**ours**)	**54.7**	51.1	**55.0**	36.1	**53.9**	50.2 (+8.7)
	+ *luminance* (**ours**)	38.8	26.0	38.2	28.0	40.7	34.3 (+8.7)		+ *luminance* (**ours**)	51.6	48.5	50.8	37.8	49.8	47.7 (+6.2)
	+ *random* (**ours**)	39.1	26.0	39.1	28.1	41.5	34.8 (+9.2)		+ *random* (**ours**)	53.3	50.1	53.3	37.9	53.2	49.6 (+8.1)
	+ *mixup* (**ours**)	41.3	27.3	40.7	28.3	43.0	36.1 (+10.5)		+ *mixup* (**ours**)	53.6	50.4	51.2	38.0	53.1	49.7 (+8.2)

(c) Comparisons of different augmentation schemes for SIM* [9] by attacking white-box I3/IR2 model

Model	Augmentation	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**	Model	Augmentation	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**
I3	SIM* [9]	44.3	30.9	44.9	31.0	44.4	39.1	IR2	SIM* [9]	54.1	53.6	54.1	43.1	53.3	51.6
	+ *space* (**ours**)	**55.0**	**40.4**	54.0	**42.1**	**55.0**	**49.3 (+10.2)**		+ *space* (**ours**)	**62.0**	**60.6**	**61.6**	**51.4**	**61.0**	**59.3 (+7.7)**
	+ *color* (**ours**)	52.3	37.5	52.9	37.8	52.5	46.6 (+7.5)		+ *color* (**ours**)	60.5	58.8	60.7	51.0	60.4	58.3 (+6.7)
	+ *luminance* (**ours**)	52.6	37.3	53.0	38.1	52.4	46.7 (+7.6)		+ *luminance* (**ours**)	60.0	59.0	59.9	50.0	58.8	57.5 (+5.9)
	+ *random* (**ours**)	54.1	37.9	51.9	41.0	53.7	47.7 (+8.6)		+ *random* (**ours**)	60.6	59.1	60.5	50.2	59.2	57.9 (+6.3)
	+ *mixup* (**ours**)	52.2	37.7	**54.7**	37.1	53.8	47.1 (+8.0)		+ *mixup* (**ours**)	60.4	59.6	60.4	49.8	59.2	57.9 (+6.3)

the optimal hyper-parameter, and then explore the best augmentation scheme for attacks. Lastly, we discuss the effect of sampling numbers. The adversarial examples are crafted with erosion image on I3 and IR2 by DIM*, TIM* and SIM*, and test the transfer success rates against five defenses, including $I3_{ens3}$, $IR2_{ens}$, HGD, TVM, and R&P. More experiments are recorded in supplementary.

1) Space-Erosion: The parameter ξ defined in Eq. (6) controls the erosion level of space continuity for a given image. If we set $\xi = 0$, there are not any pixels to be damaged and the erosion image restores to its clean version. We evaluate the erosion parameter ξ ranged from 0.0 to 0.2 with step 0.02, and report the performance in Fig. 9 (a). We can see that introducing space-erosion images as augmentations is effective for all three baseline methods. Even with a small proportion of corrupted pixels, it is surprising that the augmentation data can facilitate transferable attacks (*i.e.*, the improvements of DIM*, TIM* and SIM* across all defenses are nearly $3\% \sim 10\%$ when we set $\xi = 0.02$). With the erosion probability increased, the success rates continued increasing until $\xi = 0.1$. Therefore, we adopt the probability $\xi = 0.1$ in the following experiments.

2) Luminance-Erosion: The standard deviation σ_e influences the brightness of an image (Eq. (8)). We test σ_e in the range of $[0.0, 0.2]$ with a granularity 0.02 to study its impact, and report the results in supplemental material. The standard deviation exhibits similar trends under all baseline methods that the transfer success rates increase at first and then drop rapidly, which indicates that dramatically adjusting the luminance of images may degrade the transfer performance. To this end, we set $\sigma_e = 0.04/0.1/0.1$ for DIM*/TIM*/SIM*, respectively.

3) Augmentation Schemes: In this section, we aim to find the best augmentation scheme for transferable attacks. We test the performance under the following five setups:

(1) holding a single strategy to produce the corrupted data during the whole attacking process, denoted as *space*, *color*, and *luminance*, respectively ($1^{st} \sim 3^{rd}$ rows);

(2) randomly rechoosing a strategy at each iteration to generate augmentations, denoted as *random* (4^{th} row);

(3) following the mixup operation [48], we utilize the linear mechanism to mix three types of generated erosion images with equal weights, denoted as *mixup* (5^{th} row).

We introduce an erosion image in DIM*, TIM*, and SIM* methods to attack I3 and IR2, and show the results in Tab. V. By applying the augmentation data, we can observe that all the schemes boost the transferability of generated adversarial examples. Most surprisingly, *space* scheme (*i.e.*, only adopting space-erosion strategy) significantly improves the success rates against all the defenses in most cases (*i.e.*, I3 model: the average improvements are 3.3%/11.8%/10.2% for baseline DIM*, TIM*, and SIM*). Besides, the average performance of *luminance* is the worst one among all the schemes. Thanks to considering different types of augmentations simultaneously, *random* and *mixup* also show their efficacy to facilitate transferable attacks. However, they do not exhibit better performance over *space* in most cases. Therefore, we simply select *space* scheme (*i.e.*, space-erosion) in our method.

4) Augmentation Number: The above discussions are based on a single augmentation image for baseline methods (*i.e.*, $n = 1$). Here, we study the influence of augmentation number n of the space-erosion strategy ranged from 0 to 6 (Eq. (5)). Notice, the attack degrades to its baseline version when we abandon any augmentations *i.e.*, $n = 0$. From Fig. 10, it is natural to observe that drawing more corrupted data leads to better results (*e.g.*, the improvements of TIM crafted on I3 are over 15% across defenses when setting $n = 6$). Nonetheless, the computational costs are growing linearly with the augmentation number increased. To better balance the performance/cost trade-off, we employ $n = 3$ in the experiments.[1]

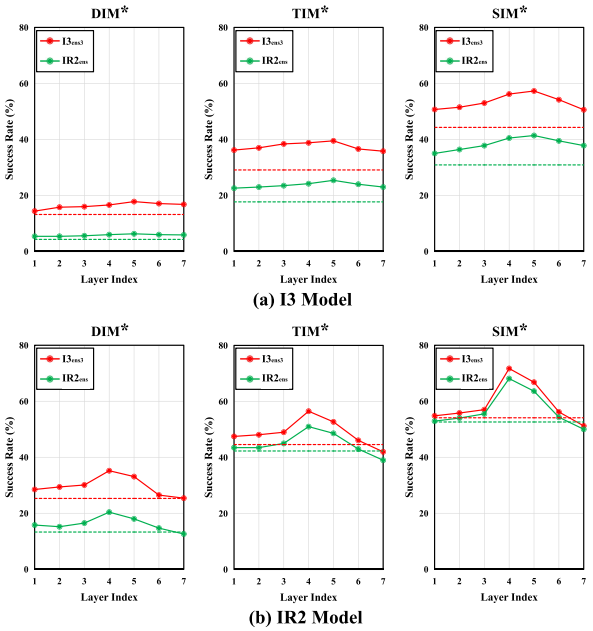


Fig. 11. **Ablation study of the optimal layer for intermediate feature classifiers (IFC)**. The adversarial examples are crafted on I3 and IR2 against two adversarial models (*i.e.*, $I3_{ens3}$ and $IR2_{ens}$). We plot the success rates of baselines DIM*, TIM* and SIM* as dashed lines, and plot the performance of those optimized over an intermediate feature classifier as solid lines. We find that the transferability is improved significantly by applying an IFC into ensemble optimization (*i.e.*, $I3_{L5}$ and $IR2_{L4}$). Moreover, the transferability trends are similar for a given surrogate model under different attack methods.

F. Ablation Studies for Network Erosion Method

In this section, we separately report the ablation experiments of intermediate feature classifier (IFC) and stochastic layers. First, we study the optimal layer scheme and the training epoch setups for IFC. Second, we learn the appropriate proportion and dependent performance for stochastic layers.

In practice, we reuse the feature extractors from the original model and fine-tune IFCs on the ImageNet `test` dataset, *i.e.*, a small dataset that contains 10^5 images. All the images are *excluded* from the training set of normally trained models and defenses, which meet our assumption mentioned in Sec. III-C.

1) Optimal Layer for Intermediate Feature Classifier: Since the proposed Intermediate feature classifier (IFC) is designed by connecting a specific layer with a fine-tuned classification head (see Fig. 3), we aim to search the optimal layers in this section. Without loss of generality, we choose 7 layers with the same interval step across the deeper half architecture for a given surrogate network. The notations of each IFC are relative to the index of the layer. For instance, $I3_{L1}$ is fine-tuned by the features at the first selected layer (*i.e.*, *Mixed_6a*). The notations are detailed in the supplemental material.

Since we obtain the IFCs fine-tuned by different layers, we can craft transferable adversarial examples on the ensemble of a feature classifier and the original model. The transfer success rates of each fine-tuned IFC are demonstrated in Fig. 11. The implications are three-fold. First, the transferability is improved significantly by applying a specific feature classifier into ensemble optimization, *i.e.*, $I3_{L5}$ and $IR2_{L4}$ achieve the best performance (*e.g.*, for attacks crafted on I3 by SIM*, the improvements are 13.0%/10.5% against $I3_{ens3}$/$IR2_{ens}$ defenses when applying IFC into optimization, etc). The results validate the effectiveness of the proposed IFC for adopting ensemble attacks rather than merely based on the original model. Second, the transferability curves from a given model exhibit highly similar trends across defenses regardless of the methods. This observation is similar to the [19], which inspires us to find the optimal layer to fine-tune a classification block offline and facilitate strong attacks for any black-box defenses. Third, the transferability increases continually at first and then gradually drops at the last layers. This tendency is partially in coincidence with the results in [26] and [49]. It may indicate that the features in these deeper layers are richer than in other layers, which is easy to be disrupted or used to train the IFC with high diversity. And some valuable features will be lost after following pooling operations, as we speculated in Sec. III. To this end, we adopt the best intermediate feature classifier, *i.e.*, $I3_{L5}$, $IR2_{L4}$, $I4_{L4}$ and $R50_{L3}$ (I4 and R50 are reported in supplemental material).

2) Training Epoch: We have studied the influence of fine-tuned epochs on transferability. We train two IFCs (*i.e.*, $I3_{L5}$ and $IR2_{L4}$) on different epochs, and report the performance under defenses $I3_{ens3}$ and $IR2_{ens}$ in supplementary. We find that at most ten epochs are enough for success rates to converge. Therefore, we set the maximum number of epochs

[1]We observe that augmentation number $n = 3$ is enough for the strongest combination SIM* to converge.

TABLE VI

THE SUCCESS RATES (%) OF PROPOSED NETWORK EROSION. WE REPORT THE SUCCESS RATES OF THE ENSEMBLE WITH INTERMEDIATE FEATURE CLASSIFIER (IFC) OR STOCHASTIC LAYERS, INCLUDING RANDOM SCALING (RS), FEATURE SMOOTHING (FS), AND CASCADE GROUP (CG)

(a) The ensemble of Intermediate Feature Classifier or Stochastic Layers by DIM* [5]

Model	IFC	RS	FS	CG	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**	Model	IFC	RS	FS	CG	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**
I3	-	-	-	-	13.2	4.3	15.8	14.5	16.3	12.8	IR2	-	-	-	-	25.3	13.3	28.7	17.1	28.8	22.6
	✓	-	-	-	16.6	6.5	19.0	16.5	21.3	16.0 (+3.2)		✓	-	-	-	35.9	19.2	38.7	21.5	40.1	31.1 (+8.5)
	-	✓	-	-	18.0	5.7	19.2	17.3	21.8	16.4 (+3.6)		-	✓	-	-	33.0	19.3	34.8	21.1	36.9	29.0 (+6.4)
	-	-	✓	-	20.4	7.3	22.3	17.4	24.4	18.4 (+5.6)		-	-	✓	-	37.8	21.8	40.1	21.4	40.0	32.2 (+9.6)
	-	-	-	✓	19.2	7.3	21.1	17.9	23.9	17.9 (+5.1)		-	-	-	✓	37.1	22.7	39.2	21.8	41.1	32.4 (+9.8)
	✓	✓	-	-	17.8	6.4	19.8	16.6	21.7	16.5 (+3.7)		✓	✓	-	-	33.1	19.4	35.6	21.4	36.8	29.3 (+6.7)
	✓	-	✓	-	**20.4**	7.3	22.3	**17.4**	24.4	18.4 (+5.6)		✓	-	✓	-	41.7	21.4	43.9	23.7	43.2	34.8 (+12.2)
	✓	-	-	✓	19.5	**8.9**	**22.4**	17.1	**24.5**	18.5 (**+5.7**)		✓	-	-	✓	40.7	23.4	43.4	24.3	44.1	35.2 (**+12.6**)

(b) The ensemble of Intermediate Feature Classifier or Stochastic Layers by TIM* [8]

Model	IFC	RS	FS	CG	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**	Model	IFC	RS	FS	CG	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**
I3	-	-	-	-	29.1	17.7	29.0	20.2	32.2	25.6	IR2	-	-	-	-	44.6	42.3	45.4	29.9	45.5	41.5
	✓	-	-	-	39.5	25.4	39.3	24.8	42.3	34.3 (+8.7)		✓	-	-	-	56.6	51.8	56.5	36.1	56.8	51.6 (+10.1)
	-	✓	-	-	37.1	24.3	37.2	23.5	38.7	32.2 (+6.6)		-	✓	-	-	52.3	48.4	52.4	32.6	51.8	47.5 (+6.0)
	-	-	✓	-	39.4	26.1	38.8	24.8	41.4	34.1 (+8.5)		-	-	✓	-	55.1	51.4	55.6	36.3	53.5	50.4 (+8.9)
	-	-	-	✓	40.8	27.2	40.5	26.3	42.5	35.5 (+9.9)		-	-	-	✓	55.4	53.2	54.8	36.7	54.9	51.0 (+9.5)
	✓	✓	-	-	40.3	26.6	40.5	25.3	42.5	35.0 (+9.4)		✓	✓	-	-	57.2	52.3	56.9	37.5	57.0	52.2 (+10.7)
	✓	-	✓	-	41.8	27.1	42.3	**27.4**	43.3	36.4 (+10.8)		✓	-	✓	-	59.1	53.7	59.6	37.8	58.1	53.7 (+12.2)
	✓	-	-	✓	**43.1**	**28.0**	**42.8**	26.9	**44.7**	**37.1** (**+11.5**)		✓	-	-	✓	**60.1**	**54.5**	**60.2**	**38.3**	**59.0**	**54.4** (**+12.9**)

(c) The ensemble of Intermediate Feature Classifier or Stochastic Layers by SIM* [9]

Model	IFC	RS	FS	CG	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**	Model	IFC	RS	FS	CG	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**
I3	-	-	-	-	44.3	30.9	44.9	31.0	44.4	39.1	IR2	-	-	-	-	54.1	53.6	54.1	43.1	53.3	51.6
	✓	-	-	-	57.3	41.4	57.7	39.3	56.7	50.5 (+11.4)		✓	-	-	-	71.7	68.1	71.3	57.4	70.9	67.9 (+16.3)
	-	✓	-	-	49.8	35.5	49.8	36.2	51.0	44.5 (+5.4)		-	✓	-	-	60.3	57.8	60.2	48.3	59.0	57.1 (+5.5)
	-	-	✓	-	51.7	37.4	51.3	38.1	52.1	46.1 (+7.0)		-	-	✓	-	61.7	60.0	61.0	49.6	59.4	58.3 (+6.7)
	-	-	-	✓	51.9	37.0	52.5	39.2	52.9	46.7 (+7.6)		-	-	-	✓	63.3	61.8	63.1	50.8	61.9	60.2 (+8.6)
	✓	✓	-	-	58.8	42.4	58.7	42.5	58.6	52.2 (+13.1)		✓	✓	-	-	72.7	68.3	72.9	59.9	71.6	69.1 (+17.5)
	✓	-	✓	-	59.5	41.9	**60.7**	44.1	**59.2**	53.1 (+14.0)		✓	-	✓	-	74.5	67.9	**74.7**	61.6	**72.2**	70.2 (+18.6)
	✓	-	-	✓	**60.6**	**43.3**	60.3	**45.1**	58.7	**53.6** (**+14.5**)		✓	-	-	✓	**74.5**	**68.5**	74.5	**62.4**	72.1	70.4 (**+18.8**)

$N = 10$. Moreover, our method has surpassed the baseline, even trained with a few epochs ($N = 2$). This satisfies our requirement that suffering much fewer fine-tuning resources on a small dataset can boost the transferability significantly (see Sec. III-C).

3) The Proportion of Stochastic Layers: We discuss the effect of the embedding ratio of stochastic layers in the network structure, *i.e.*, whether adding more stochastic layers leads to better results. We embed the proposed cascade group (CG) into the original networks for adopting ensemble attacks. The embedded layers are abandoned with probability Λ as:

$$z_{l*1} = p \cdot g_l^5(x_l), \quad p \sim \text{Bernoulli}(\Lambda) \tag{15}$$

where p is sampled from Bernoulli distribution with parameter Λ. If $\Lambda = 1$, all the cascade group layers are preserved. Conversely, all the stochastic layers are removed when we set $\Lambda = 0$, *i.e.*, merely attacking the original model.

We test the parameter Λ ranged in $[0.0, 1.0]$ with step size 0.1, and report the success rates under five defenses in Fig. 12. It is noteworthy that the performance of all attacks is improved significantly even by adding a small proportion of cascade group layers (*i.e.*, $\Lambda = 0.1$). For the DIM*, the performance continues to improve at first, and then keeps stable after the proportion surpasses 0.6. On the other hand, we observe that adding more stochastic layers leads to stronger transferability for TIM* and SIM*, and it achieves the best performance when the stochastic layers are embedded throughout the network (*i.e.*, $\Lambda = 1.0$). This phenomenon is caused by the ensemble optimization over a set of translated images proposed in TIM [8], which brings more randomness capacity to corrupted networks. Therefore, we simply adopt adding all stochastic layers, *i.e.*, $\Lambda = 1.0$.

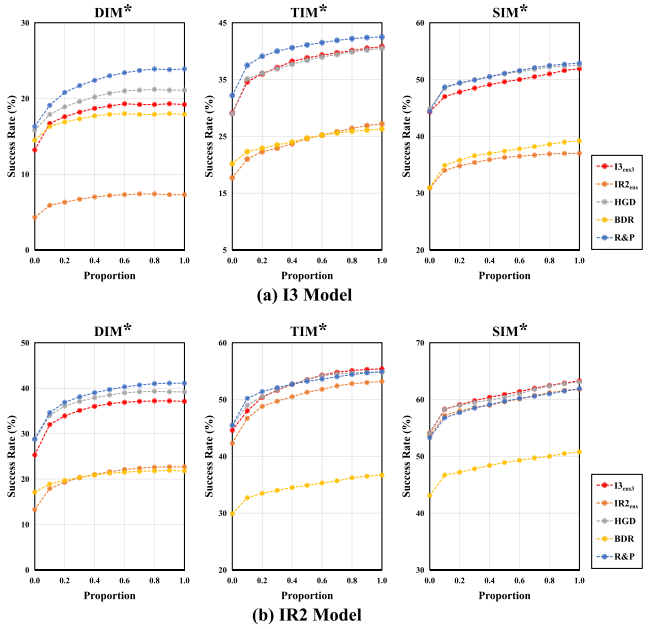


Fig. 12. **The success rates (%) of the different proportion of cascade group layers**. We use DIM*, TIM* and SIM* to attack the ensemble of a normally trained models with its embedded version, including: (a) I3, and (b) IR2. In general, more stochastic layers result in better performance regardless of attack methods or attacked surrogate models.

TABLE VII

THE COMPARISON BETWEEN GHOSTNET [18] AND PROPOSED CASCADE GROUP (CG) IN ENSEMBLE OPTIMIZATION

(a) Comparisons between GhostNet and cascade group (CG) on I3

Attack	Ensemble Setup	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**
DIM*	w/ GhostNet [18]	17.2	6.2	18.1	15.9	19.9	15.5
	w/ CG (**ours**)	**19.2**	**7.3**	**21.1**	**17.9**	**23.9**	**17.9** (+2.4)
TIM*	w/ GhostNet [18]	34.2	20.1	34.9	24.6	37.6	30.3
	w/ CG (**ours**)	**40.8**	**27.2**	**40.5**	**26.3**	**42.5**	**35.5** (+5.2)
SIM*	w/ GhostNet [18]	44.7	30.1	44.5	30.7	45.9	39.2
	w/ CG (**ours**)	**51.9**	**37.0**	**52.5**	**39.2**	**52.9**	**46.7** (+7.5)

(b) Comparisons between GhostNet and cascade group (CG) on IR2

Attack	Ensemble Setup	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	**Average**
DIM*	w/ GhostNet [18]	30.2	18.8	33.9	19.8	35.5	27.6
	w/ CG (**ours**)	**37.1**	**22.7**	**39.2**	**21.8**	**41.1**	**32.4** (+4.8)
TIM*	w/ GhostNet [18]	51.7	47.1	51.6	31.6	51.1	46.6
	w/ CG (**ours**)	**55.4**	**53.2**	**54.8**	**36.7**	**54.9**	**51.0** (+4.4)
SIM*	w/ GhostNet [18]	62.8	60.6	**63.2**	50.2	**62.7**	59.9
	w/ CG (**ours**)	**63.3**	**61.8**	63.1	**50.8**	61.9	**60.2** (+0.3)

4) The Effect of Stochastic Layers: We study the impact of stochastic layers under the following settings (see Tab. VI):

(1) baselines without ensemble optimization (1^{st} row);

(2) applying the IFC into ensemble optimization (2^{nd} row);

(3) embedding stochastic layers (*i.e.*, random scaling, feature smoothing and cascade group) into the original model to construct a corrupted network for ensemble paradigms, denoted as RS, FS and CG, respectively ($3^{rd} \sim 5^{th}$ rows);

(4) using both of the original model and the corrupted network (*i.e.*, the IFC embedded with stochastic layers) for ensemble optimization ($6^{th} \sim 8^{th}$ rows).

The performances are reported in Tab. VI. We observe that merely adding stochastic layers into an existing network for ensemble optimization also improves the success rates across the defenses (3∼5 rows). Among them, the cascade group (CG) often gets higher success rates than random scaling (RS) and feature smoothing (FS), which might be caused by its larger model diversity than the other two layers (Fig. 6). It verifies our hypothesis that improving the robustness and diversity of the surrogate model may enhance transferability, as discussed in Sec. IV-B. Moreover, using a corrupted network (*i.e.*, densely embedding stochastic layers into the IFC) for the ensemble paradigm largely improves the strength further instead of only utilizing IFC or stochastic layers (6∼8 rows). These results validate the high efficiency of the proposed network erosion method. Empirically, we find that IFC embedded with cascade group achieves the best results.

5) Compare With Other Model Augmentations: Based on different network structures, GhostNet [18] applies feature dropout or skip connection perturbation to achieve the augmentation effect. We separately adopt the feature dropout and skip connection perturbation for I3 and IR2 to construct the GhostNet, and attack an ensemble of the original model and GhostNet to generate adversarial examples. We choose the same parameters reported in [18] in the comparisons.

From tab. VII, we observe that the cascade group (CG) is superior to GhostNet in most cases for ensemble optimization. Especially, our method surpasses GhostNet largely on attacking I3 models (*e.g.*, for I3, the improvements are 2.4%, 5.2%, 7.5% for DIM*, TIM*, and SIM*, respectively, etc.). The results verify the qualification of the cascade group layer for transferable attacks. What's more, we emphasize that the proposed stochastic layers are much more general than [10], [18] that they can apply to *arbitrary* network architectures.

V. CONCLUSION

This paper focuses on studying transferable attacks against unseen defenses under black-box scenarios. To mitigate the data-dependency effect, we propose a data erosion method that introduces artificial erosion images as augmentations during the attacks. Different from prior works, the proposed data erosion method is a specific augmentation designed to narrow the gap between the normally trained models and defenders. We experiment with different types of erosion images and find that the space-erosion data achieves the best performance to boost transferability. To avoid the network-dependency effect, the proposed network erosion method leverages an original model to create a corrupted network with diverse decision boundaries, which is applied to the ensemble optimization with litter cost, exhibiting better generalizability and higher efficiency compared with current attacks. The proposed data & network erosion can be naturally integrated into a general framework to further promote the threatening of adversarial examples, referred to as erosion attack (EA). Extensive experiments not only demonstrate the superiority of our proposed method compared with existing transferable attacks but also remind the security issues of the current defenses.

REFERENCES

[1] C. Szegedy et al., "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[2] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155161–155196, 2021.

[3] L. Huang et al., "Universal physical camouflage attacks on object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 720–729.

[4] Y. Zhang et al., "Principal component adversarial example," *IEEE Trans. Image Process.*, vol. 29, pp. 4804–4815, 2020.

[5] C. Xie et al., "Improving transferability of adversarial examples with input diversity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2730–2739.

[6] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[7] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[8] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4312–4321.

[9] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–12.

[10] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, "Skip connections matter: On the transferability of adversarial examples generated with ResNets," 2020, *arXiv:2002.05990*.

[11] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1924–1933.

[12] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," 2017, *arXiv:1705.07204*.

[13] N. Akhtar, J. Liu, and A. Mian, "Defense against universal adversarial perturbations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3389–3398.

[14] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "A self-supervised approach for adversarial robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 262–271.

[15] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1778–1787.

[16] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Deflecting adversarial attacks with pixel deflection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8571–8580.

[17] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," 2016, *arXiv:1611.02770*.

[18] Y. Li, S. Bai, Y. Zhou, C. Xie, Z. Zhang, and A. Yuille, "Learning transferable adversarial examples via ghost networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11458–11465.

[19] N. Inkawhich, W. Wen, H. H. Li, and Y. Chen, "Feature space perturbations yield more transferable adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7066–7074.

[20] N. Inkawhich, K. J. Liang, B. Wang, M. Inkawhich, L. Carin, and Y. Chen, "Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability," 2020, *arXiv:2004.14861*.

[21] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[23] A. N. Bhagoji, W. He, B. Li, and D. Song, "Practical black-box attacks on deep neural networks using efficient query mechanisms," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 154–169.

[24] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2137–2146.

[25] Y. Dong et al., "Efficient decision-based black-box adversarial attacks on face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7714–7722.

[26] Y. Yu, X. Gao, and C.-Z. Xu, "LAFEAT: Piercing through adversarial defenses with latent features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5735–5745.

[27] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature importance-aware transferable adversarial attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7639–7648.

[28] Y. Dong et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

[29] J. M Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," 2019, *arXiv:1902.02918*.

[30] H. Yang et al., "DVERGE: Diversifying vulnerabilities for enhanced robust generation of ensembles," 2020, *arXiv:2009.14720*.

[31] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," 2017, *arXiv:1711.01991*.

[32] X. Jia, X. Wei, X. Cao, and H. Foroosh, "ComDefend: An efficient image compression model to defend adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6084–6092.

[33] T. Borkar, F. Heide, and L. Karam, "Defending against universal attacks through selective feature regeneration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 709–719.

[34] C. Etmann, S. Lunz, P. Maass, and C.-B. Schönlieb, "On the connection between adversarial robustness and saliency map interpretability," 2019, *arXiv:1905.04172*.

[35] A. Noack, I. Ahern, D. Dou, and B. Li, "An empirical study on the relation between network interpretability and adversarial robustness," *Social Netw. Comput. Sci.*, vol. 2, no. 1, pp. 1–13, Feb. 2021.

[36] A. Ruderman, N. C. Rabinowitz, A. S. Morcos, and D. Zoran, "Pooling is neither necessary nor sufficient for appropriate deformation stability in CNNs," 2018, *arXiv:1804.04438*.

[37] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[39] C. Pestana, N. Akhtar, W. Liu, D. Glance, and A. Mian, "Adversarial perturbations prevail in the Y-channel of the YCbCr color space," 2020, *arXiv:2003.00883*.

[40] X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh, "Towards robust neural networks via random self-ensemble," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 369–385.

[41] C. Xie, Y. Wu, L. van der Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 501–509.

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[43] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-V4, Inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[45] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[46] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," 2017, *arXiv:1711.00117*.

[47] M. Guo, Y. Yang, R. Xu, Z. Liu, and D. Lin, "When NAS meets robustness: In search of robust architectures against adversarial attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 631–640.

[48] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[49] N. Inkawhich, K. Liang, L. Carin, and Y. Chen, "Transferable perturbations of deep feature distributions," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–14.

Lifeng Huang is currently a Lecturer with the College of Mathematics and Informatics, South China Agricultural University. He works in areas of adversarial learning, deep learning security, and its application.

Chengying Gao received the Ph.D. degree from Sun Yat-sen University.

She is currently an Associate Professor with the School of Computer Science and Engineering, Sun Yat-sen University. She works in areas of computer vision, computer graphics, and deep learning.

Ning Liu received the Ph.D. degree from Sun Yat-sen University. He is currently a Professor with the School of Computer Science and Engineering, Sun Yat-sen University. His research interests include computer vision, cyberspace security, and deep learning.

Pattern Recognition 150 (2024) 110225

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

LAFED: Towards robust ensemble models via Latent Feature Diversification

Wenzi Zhuang [a,1], Lifeng Huang [b,1], Chengying Gao [a], Ning Liu [a,c,*]

[a] School of Computer Science and Engineering, Sun Yet-Sen University, Guangzhou, Guangdong, China
[b] College of Mathematics and Informatics, South China Agricultural University, Guangzhou, Guangdong, China
[c] Guangdong Key Laboratory of Information Security Technology, China

ARTICLE INFO

Keywords:
Adversarial example
Adversarial defense
Ensemble model
Robustness

ABSTRACT

Adversarial examples pose a significant challenge to the security of deep neural networks (DNNs). In order to defend against malicious attacks, adversarial training forces DNNs to learn more robust features by suppressing generalizable but non-robust features, which boosts the robustness while suffering from significant accuracy drops on clean images. Ensemble training, on the other hand, trains multiple sub-models to predict data for improved robustness and still achieves desirable accuracy on clean data. Despite these efforts, previous ensemble methods are still susceptible to attacks and fail to increase model diversity as the size of the ensemble group increases. In this work, we revisit the model diversity from the perspective of data and discover that high similarity between training batches decreases feature diversity and weakens ensemble robustness. To this end, we propose **Latent Feature Diversification (LAFED)**, which reconstructs training sets with diverse features during the optimization, enhancing the overall robustness of an ensemble. For each sub-model, LAFED treats the vulnerability extracted from other sub-models as raw data, which is then combined with round-changed weights with a stochastic manner in the latent space. This results in the formation of new features, remarkably reducing the similarity of learned representations between the sub-models. Furthermore, LAFED enhances feature diversity within the ensemble model by utilizing hierarchical smoothed labels. Extensive experiments illustrate that LAFED significantly improves diversity among sub-models and enhances robustness against adversarial attacks compared to current methods. The code is publicly available at https://github.com/zhuangwz/LAFED.

1. Introduction

The discovery of adversarial examples poses a challenge to the security of DNNs. By applying imperceptible perturbation to clean images, adversarial attacks have caused great trouble to DNN-based applications in both the digital space [1] and the real world [2]. More significantly, attacks often utilize white-box [3] models to generate adversarial examples, which usually exhibit high transferability across black-box models with different architectures [4–6]. One plausible explanation for this phenomenon is that humans rely on abstract and robust features to recognize objects in images, while DNNs attempt to learn more generalizable but non-robust features [7]. As a result, DNNs have high accuracy on unseen images while also being vulnerable to adversarial noise manipulated by attackers. Consequently, attack methods can leverage the vulnerability of non-robust features to craft malicious input data and transfer its adversarial effect to fool remote black-box models. This leads one to wonder: how can DNNs enhance robustness while balancing the trade-off between robust and non-robust features?

To answer the question, several defense methods have been proposed. One line is adversarial training [8], which involves minimizing the optimization loss on adversarial examples during each training step. This process compels DNNs to learn robust features from adversarial examples instead of non-robust features from clean data. As a result, adversarially trained models become more robust but significantly decrease classification accuracy due to a lack of generalizability on unseen clean data. [9]. Although some subsequent works have demonstrated a better balance between accuracy and robustness, their performance on clean images is still much lower than normally trained models [10,11]. Another popular way for boosting adversarial resistance is ensemble training [12–15]. Intuitively, they adopt several sub-models for ensemble predictions. Each sub-model in the ensemble captures sufficient generalizable and non-robust features during optimization, thereby exhibiting high accuracy on clean data. Moreover, sub-models are deliberately trained with different latent representations or diverse decision boundaries compared to each other. Therefore, ensemble models

* Corresponding author at: School of Computer Science and Engineering, Sun Yet-Sen University, Guangzhou, Guangdong, China.
E-mail address: liuning2@mail.sysu.edu.cn (N. Liu).
[1] First author and second author contribute equally to this work.

https://doi.org/10.1016/j.patcog.2023.110225
Received 17 June 2022; Received in revised form 28 September 2023; Accepted 26 December 2023
Available online 5 January 2024

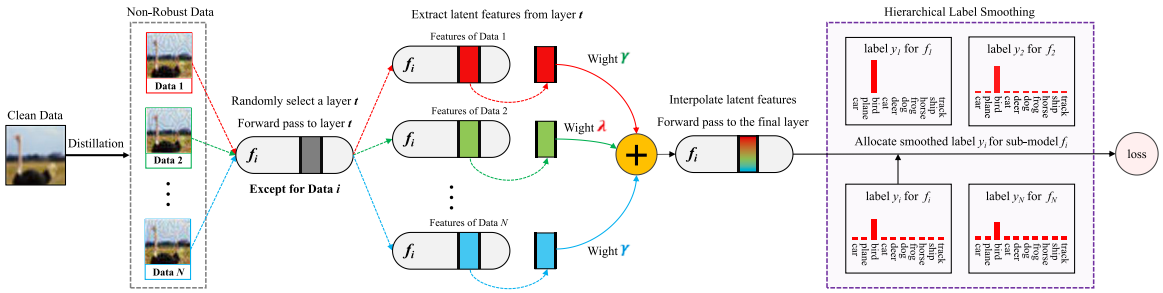


Fig. 1. The intuitive pipeline of LAFED for training an ensemble member f_i. Concretely, for each clean image, LAFED generates a batch of non-robust data as initialized training data. To further diversify the sub-models, LAFED combines non-robust images with unbalanced weights (e.g., majority λ and minority γ) in a stochastic manner to reconstruct training data in the latent feature space, forcing each sub-model to learn different representations. Moreover, LAFED adopts customized labels for sub-models and guides them to form diverse latent spaces with different feature clusters.

are much more robust as attackers cannot successfully fool the model by tricking a majority of the ensemble members to predict incorrect results.

Generally, ensemble training methods are designed to boost diversity among sub-models, including varying the output probabilities [12], maximizing the gradient divergence [13], isolating the non-robust features [14], and encouraging smooth model and diverse gradient [15], etc. Although they have desirable resistance against adversarial attacks, they cannot effectively improve robustness against both black-box and white-box attacks simultaneously or enhance robustness as the number of sub-models increases. Inspired by the hypothesis that deep learning models tend to capture similar non-robust features when training on the same dataset [7], we revisit the model diversity from the perspective of data. We surprisingly find that a low degree of similarity between training sets of sub-models usually results in poor adversarial transferability within the ensemble models. This means the adversarial examples generated from one sub-model are difficult to fool other sub-models, indicating that the ensemble model has high diversity and robustness against attacks. Following [16], we further utilize Pearson Correlation Coefficient (PCC) values as an indicator to measure the degree of data similarity. Specifically, we treat the network outputs as the response to features in inputs and measure the correlation between the features of data using PCC values. Prior methods have shown high similarity between training images among different sub-models, and PCC values continuously increase as the number of sub-models becomes large, accompanied by a rise in transferability among sub-models [14]. It may limit the robustness of ensemble models.

This intriguing property motivates us to decrease the similarity of latent features in training data among sub-models to boost ensemble robustness. To achieve this goal, two lines of work catch our attention. The first is data augmentation, which is achieved by interpolation at the pixel or intermediate layer space [17,18], which facilitates the network in learning more generalizable representations. The second one is the label smoothing technique, which is widely used to prevent the model from becoming over-confident, and to ensure the representations congregate in tight clusters [19]. Both of them are proposed to improve the generalizability of the model on classifying unseen clean images. Our investigation has inspired the development of **Latent Feature Diversification (LAFED)**. It reconstructs customized training data for sub-models and boosts the diversity of features learned by ensemble members, ultimately improving the adversarial robustness. The pipeline of LAFED for training a single sub-model is illustrated in Fig. 1. Especially, LAFED follows [14] to treat non-robust features captured by DNNs as a vulnerability of their corresponding data, and extracts these adversarial vulnerabilities from each sub-model to generate a batch of non-robust data, viewing them as desirable initialized training datasets after adding special noises. To further promote the diversity of learned features among ensemble members, LAFED then stochastically combines these non-robust images with unbalanced weights to reconstruct a new batch of training data. This operation is extended throughout all eligible spaces to substantially intensify the diversity of learned representations. Furthermore, LAFED adopts hierarchical smoothed labels instead of original one-hot vectors during the optimization. This promotes the divergence of representation clusters with varying degrees for different sub-models, further diversifying the latent space of ensemble members. LAFED performs these optimizations to all the sub-models in a round-robin manner, significantly enhancing the overall robustness of the ensemble against adversarial attacks.

In summary, the contributions of our work are three-fold:

- We revisit the model diversity from the perspective of data and identify that the high correlation of features in training data is an essential factor to affect the diversity and robustness within the ensemble models.
- We propose a novel ensemble training method, referred to as **Latent Feature Diversification (LAFED)**, which decreases the similarity of training batches and enhances the diversity of learned features among ensemble members.
- The empirical results not only demonstrate the superiority of the proposed method over state-of-the-art ensemble models but also suggest its generalizability that LAFED can effectively enhance the robustness of ensembles by simply including more sub-models.

2. Related work

2.1. Adversarial attacks and adversarial training

DNNs have been widely used in many fields. Nonetheless, their vulnerabilities have also been discovered, which brings security issues to real-world systems [20–23]. By applying imperceptible noises to clean images, these maliciously crafted data can easily mislead models to output wrong predictions [24–26]. The model used to generate adversarial examples is usually a common pre-trained model, and the attacker uses the gradient information of the model to generate the adversarial examples. Specifically, the attacker aims to find a small perturbation δ for the specific clean image x to produce an adversarial example by $x_{\text{adv}} = x + \delta$ that fools DNNs under the norm constraint $\| \delta \|_p \leq \epsilon$. The formulation can be defined as optimizing over δ to maximize the objective loss $\mathcal{L}_\theta(\cdot)$ for the classifier with its parameters θ on the input-label pair (x, y), i.e., maximizing $\mathcal{L}_\theta(x + \delta, y)$.

In general, most works select l_∞ norm constraint to control the adversarial perturbation magnitude and measure the robustness of trained models [8,12,15]. Various adversarial attack methods have been investigated to improve the fooling rates, such as projecting gradient descent (PGD) [8], diversity input transformation (DIM) [27], skip connections corruption (SGM) [28], or using an ensemble of multi-attacks and multi-models [29–31]. They often achieve high performance under both white-box and black-box scenarios, which exhibit huge security threats to current DNNs.

To resist adversarial attacks, many empirical defenses have been established. A popular way is adversarial training (AT), which gains remarkable efficacy in boosting the robustness of adversarially trained networks [8]. Generally, it performs a min–max optimization at each step, where the inner maximization is implemented by strong attacks and the outer minimization is used to depress the adversarial effect against itself, which is formulated as follows:

$$\min_{\theta}\mathbb{E}_{(x,y)\sim D}[max_{\delta\in S}L_{\theta}(x+\delta,y)]], \tag{1}$$

where the input-label pair (x, y) in training data is sampled from distribution D and S is the set of allowed perturbations under constraint. Based on the basic AT, Zhang et al. propose TRADES to better trade-off between robustness and clean accuracy [10], which can be bounded using natural error and boundary error. Such decomposition is proved to be effective in balancing accuracy and robustness as well as providing theoretical guarantees. Atzmon et al. adds a fixed linear layer to the original network to build a sample network at each training epoch and incorporates it in a loss function as a proxy, significantly enhancing the adversarial robustness against attacks [32]. However, this type of training paradigm forces the network to learn robust features rather than capturing non-robust but generalizable representations. Thus, it usually leads to a drop in accuracy for unseen clean data [9].

2.2. Ensemble training for adversarial defense

Most traditional ensemble learning methods are proposed to enhance the generalizability of DNNs to the unknown testing data, such as bagging [33] and boosting [34]. Moreover, Kuncheva et al. have shown that improving the diversity within bayesian neural networks leads to better ensemble performance [35]. Following this line, the concept of ensemble learning is introduced to improve the adversarial robustness without dropping the clean accuracy.

Several ensemble training methods have been proposed to train diverse sub-models, which aim to decrease the internal transferability within ensemble models and improve the overall robustness. Pang et al. regard the divergence among the non-maximal predictions of individual sub-models as a part of the ensemble diversity in adversarial defense, thus proposing adaptive diversity promoting (ADP) regularizer during optimization, which diversifies the classification predictions of models [12]. Kariyappa et al. believe that the common adversarial subspace affects the model diversity and thereby propose gradient alignment loss (GAL) maximizes the dissimilarity within the ensembles and enhances robustness [13]. TRS finds that the gradient orthogonality among sub-models and classification boundary smoothness facilitate the reduction of transferability against adversarial examples [15]. DVERGE successfully isolates the vulnerabilities from each sub-model by distilling non-robust features. It then adopts them to optimize each ensemble member separately, achieving high robustness against a variety of black-box attacks [14]. Although these methods substantially improve the model diversity, unfortunately, they show poor generalizability in that the ensemble robustness is not enhanced reasonably as the group size becomes large. Moreover, they cannot effectively defend against white-box and black-box attacks simultaneously. This phenomenon is primarily attributed to the high correlation of training data provided for each sub-model. As a result, they tend to capture similar features in the latent space. To address this issue, we propose the LAFED method, which diversifies both training data and learned representation to optimize ensemble models.

2.3. Mixup training and label smoothing

Mixup technique is utilized to create synthesized data by interpolating two clean images sampled from a training batch in the input space [17]. Optimizing over those synthesized data is a plausible way to improve the generalizability of trained networks. Verma et al. have demonstrated that using manifold interpolation at the feature space is more effective than standard Mixup which achieves better performance while consuming less computational resources [18]. Additionally, the mixup technique is also applied in adversarial defenses. For instance, Lee et al. propose Adversarial Vertex Mixup (Vertex) to mitigate the adversarial feature overfitting effect, which mixes the vertex sample with the original data during the adversarial training [36]. Likewise, Pang et al. develop mixup inference (MI), which shrinks and transfers adversarial perturbations to resist attacks [37]. Experimental results illustrate that both Vertex and MI improve the adversarially robust generalization.

Label smoothing (LS) is an effective method to reduce overfitting and improve the generalizability of trained models. It generates a soft label y^s by applying a uniform vector to the original hard label y, i.e., $y^s = y(1-\eta)+\eta/K$, where K is the number of classes. As a result, implementing LS can noticeably improve the accuracy of image classifiers and language translators [38–40]. An explanation is that it shrinks the feature norms and tightens the clusters of each class data, thereby boosting the performance of models across different tasks [19]. However, in practice, the model trained with LS is found to be more vulnerable to black-box attacks [41]. To address this issue, instead of applying the LS technique directly, we propose a novel hierarchical label smoothing strategy to diversify the feature distribution of sub-models. It has been shown to significantly enhance the robustness against white-box and black-box attacks.

3. Method

3.1. Data similarity and feature diversity

Most prior works mainly treat ensemble training as an optimization problem. Therefore, they propose different optimization objectives to improve the robustness, including ADP [12], GAL [13], and TRS [15]. In this work, we revisit the ensemble method from the perspective of data at first. Motivated by the assumption that identical training data may force different networks to learn similar features [14], we begin by studying the relationship between the data similarity and feature diversity through a toy example. Concretely, we train two models individually using half of the CIFAR-10 dataset, but with overlapping training data by 0%, 20%, 40%, 60%, 80%, and 100%. Then we follow the setups in [14] to craft adversarial examples by using 50-step PGD to attack each model, and then record their pairwise transferability as a proper indicator to measure the feature diversity between models. The results are presented in Fig. 2.

It is evident that the pairwise transferability exhibits similar trends across different attack strength (e.g., $\epsilon = 0.01/0.02/0.03$). As the overlap ratio increases, the transferability of adversarial examples between two models noticeably improves. The highest transferability can be reached by completely using the same dataset to train two classifiers, indicating that adversarial examples generated from one model can easily fool another one. This phenomenon implies higher similarity of training data can lower the diversity of learned features for networks, resulting in worse robustness of the ensemble of these two classifiers against transferable attacks.

3.2. Latent Feature Diversification method

Inspired by the analysis above, we propose **Latent Feature Diversification (LAFED)** method, which is designed to reconstruct new training batches with diverse features for each sub-model. By doing so, it enables every sub-model to learn diversified representations. Additionally, we recognize the important role that ground-truth labels play in guiding networks during optimization. Intuitively, LAFED simultaneously enhances the diversity of representations and labels, thereby achieving greater robustness for ensemble models.

As shown in Fig. 1, LAFED consists of 4 steps for training each sub-model within an ensemble model:

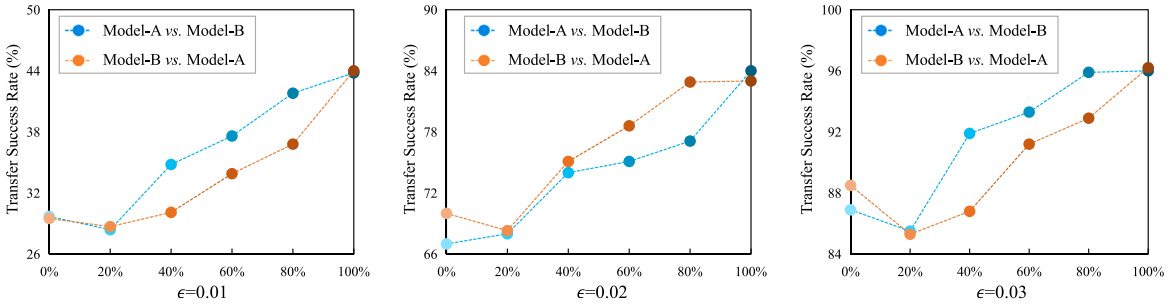


Fig. 2. The adversarial transferability between two models (Model-A and Model-B). We note that these two models share the same structure while they are trained with different ratios of overlapped training data. The larger overlap ratio results in the higher transfer rate of adversarial examples generated from one model to another one, indicating a lower diversity between these two models.

- **Step 1**. Initializing raw training data by distilling the vulnerability from other sub-models and treating them as non-robust data.
- **Step 2**. Generating diversified representations by interpolating raw data with round-changed weights at the latent feature space using an unbalanced feature combination strategy.
- **Step 3**. Allocating customized labels for the generated representations based on the hierarchical label smoothing strategy.
- **Step 4**. Forward passing the mixed representations through the network to calculate objective loss, and updating the parameters of the sub-model based on back-propagation.

These four steps are implemented in a round-robin manner until the training epoch reaches the maximum.

Raw data initialization. Since most methods merely depend on the identical clean dataset to train networks, we treat the vulnerable data with diverse features distilled from other sub-models as a proper initialization of the raw data. For the ensemble model F composed of N sub-models, we denote each sub-model as f_i, where $i \in 1, 2, \dots, N$. When an image x is propagating through the network f_i, the output before the activation module at the lth layer is formulated as $f_i^l(x)$.

To initialize the raw dataset for each sub-model f_i, we follow [7] to sample a target input-label pair (x, y) and another randomly-chosen source pair (x_s, y_s) from the clean data distribution D. The distillation of the vulnerability at lth layer for the target image x corresponding to the source image x_s can be formulated as:

$$\begin{aligned} z_{i,l} = \arg\min_{z} \left\| f_i^l(z) - f_i^l(x) \right\|_2^2, \quad (x, y), (x_s, y_s) \sim D, \\ \text{s.t. } \left\| z - x_s \right\|_\infty \le \epsilon, \end{aligned} \tag{2}$$

where $z_{i,l}$ denotes the generated non-robust data. Intuitively, the distillation process aims to find a sample $z_{i,l}$, which is visually similar to the source data x_s at the pixel space, but its latent features are close to the target data x at the intermediate space, i.e., $z_{i,l} \approx x_s$, $f_i^l(z_{i,l}) \approx f_i^l(x)$ (see Fig. 3(a)). In other words, the non-robust data $z_{i,l}$ represents the vulnerability of the sub-model f_i on classifying a clean image x_s. During the optimization, LAFED performs the distillation procedure iteratively to initialize the raw dataset for every sub-model.

Unbalanced feature combination. Although DVERGE attempt to diversify the training data by generating the vulnerable data, the provided dataset for each sub-model consists of non-robust data extracted from other sub-models [14]. It leads to a high similarity among captured features and thus degrades the diversity within the ensemble, as we discussed above. To this end, we propose to train each sub-model based on the composition of those raw data by different proportions, which effectively decreases the similarity of training data among sub-models. Concretely, we use interpolation to achieve augmentation for their respective dataset. Instead of merely blending the raw data at the input space, we extend this operation throughout the network to encourage the ensemble members to learn distinctive features from one another, which is formulated as:

$$\mathcal{F}_i^t = \lambda \cdot f_i^t(z_{j,l}) + \sum_{k \neq i, k \neq j} \gamma \cdot f_i^t(z_{k,l}), \tag{3}$$

where $\mathcal{F}_i^t$ denotes the interpolated latent feature at the tth layer of the sub-model f_i, and t is a randomly-chosen intermediate layer. $z_{j,l}$ denotes the raw data distilled at the lth layer of another sub-model f_j, where the j is a randomly selected sub-model index except for the currently trained model f_i, i.e., $j \in 1, 2, 3, \dots, N, j \neq i$. We note that both t and j are rechoosed layer indexes at each optimization step. Following [17], we sample the major interpolation coefficient λ from the Beta distribution with hyper-parameter α, and the secondary weight γ is assigned for other raw data $z_{k,l}$ as:

$$\gamma = (1 - \lambda) \big/ (N - 2). \quad \lambda \sim \text{Beta}(\alpha, \alpha) \tag{4}$$

Intuitively, LAFED creates a reconstructed feature ($\mathcal{F}_i^t$) based on vulnerability learned from multiple raw data. In particular, LAFED adopts the feature of $z_{j,l}$ as the majority component, and other features as the minority part to construct $\mathcal{F}_i^t$. Due to the unbalanced manner of the sampling weights process, LAFED combines the vulnerable data to generate diverse intermediate features during every optimization, which considerably reduces the similarity among sub-models, enhancing the overall robustness.

Hierarchical label smoothing. Label Smoothing can improve the performance of DNNs across a variety of tasks by grouping learned features in tight clusters for a total of K classes [19]. However, prior works show that the network optimized with smoothed labels may degrade its resistance against black-box adversarial examples [41]. To address this problem, we do not directly utilize label smoothing to train the ensemble, as it can result in relatively weak robustness. Instead, we use it to increase the diversity of ensemble members by designing a Hierarchical Label Smoothing strategy, which encourages the feature clusters scattered in varying degrees for different sub-models. Rather than adopting standard one-hot labels or default uniform smoothing policy, LAFED allocates customized smoothed labels for different ensemble members. The data-label pair $(x_s, y_{s,i}^{ls})$ applied for training sub-model f_i is defined as:

$$y_{s,i}^{ls} = y_s(1 - \eta_i) + \frac{\eta_i}{K} \tag{5}$$

where y_s is the original one-hot label vector of data x_s, η_i is varied for each sub-model, i.e., $\eta_i = p \cdot (i - 1)$, and p is a pre-defined parameter. Note that $y_{s,i}^{ls}$ degenerates to the original one-hot label if the model is the first sub-model, i.e., $i = 1$ and $\eta_i = 0$. Moreover, hierarchical label smoothing degrades to the standard label smoothing technique if we select the unchanged weight η for all sub-models, i.e., $\eta_i = \eta$, where $i = 1, 2, 3, \dots, N$. Intuitively, by using hierarchical strategy during the optimization, the clusters of latent representations become tighter with the index of sub-model increased, as discussed in Section 4.2.

Algorithm 1 LAFED Training Algorithm.

Input: the number of sub-models N, the number of training batch B

for $i = 1$ to N **do**
  Initialize and pretrain sub-model f_i
end for
/* *Diversifying learned features at the latent space* */
for $i = 1$ to N **do**
  Randomly choose layer l for distillation
  Randomly choose layer t for feature mixup
  for $b = 1$ to B **do**
    Sample the source and target data pairs
    Initializing the raw training batch with Eq. (2)
    Uniformly randomly choose j where $j \neq i$
    Diversifying latent features with Eq. (3)
    Multi-level label smoothing with Eq. (5)
    Calculate loss with Eq. (8)
    Update sub-model f_i by using SGD optimizer
  end for
end for

Ensemble optimization. The objective of each sub-model is to capture diverse features by leveraging the unbalanced combination of non-robust data extracted from other sub-models. Therefore, LAFED randomly samples clean data-label pairs (x, y) and (x_s, y_s) to construct the combined intermediate input $\mathcal{F}_i^t$ (see Eq. (3)), and then calculates the objective function for updating the parameters θ_i of sub-model f_i in a single step as:

$$\min_{\theta_i}\left[\mathcal{L}\left(f_i(\mathcal{F}_i^t), y_{s,i}^{ls}\right)\right], \quad (x, y), (x_s, y_s) \sim D, \tag{6}$$

where $\mathcal{L}(\cdot)$ denotes the cross-entropy loss, and $y_{s,i}^{ls}$ denotes the customized label of source image x_s for f_i (see Eq. (5)). By minimizing the loss $\mathcal{L}_i^t$, the current trained f_i has learned the diverse adversarial features from others on predicting x_s. In summary, the overall objective of LAFED for sub-model f_i can be formulated as:

$$\min_{\theta_i}\mathbb{E}_{(x,y),(x_s,y_s),t}\left[\mathcal{L}\left(f_i(\mathcal{F}_i^t), y_{s,i}^{ls}\right)\right]. \tag{7}$$

Algori 1 demonstrates the pseudo-code for training an ensemble with N members. Note that the proposed LAFED is similar to manifold mixup [18], which combines the data at intermediate layers. However, there are three key differences. Firstly, manifold mixup combines two different clean data at intermediate feature layers, while our method blends multiple vulnerabilities distilled from other sub-models with round-changed weights. Secondly, manifold mixup interpolates the one-hot labels of the sampled minibatch, while our method creates customized smoothed labels based on the source image and sub-model indexes. Thirdly, manifold mixup is used to improve the generalizability of the network, while our method aims to lower the feature similarity as well as increase the diversity within ensembles.

4. Experiment

In this section, we present the empirical results to illustrate the efficacy of the proposed method in defending against adversarial attacks. We first specify the detailed experimental setups in Section 4.1. Then we provide the qualitative analysis of the proposed LAFED in Section 4.2. Afterward, we evaluate the robustness of our methods against various attacks on different datasets and settings in Sections 4.3, 4.4, and 4.5. Lastly, we provide a rich collection of ablation studies for the proposed LAFED in Section 4.6.

4.1. Settings

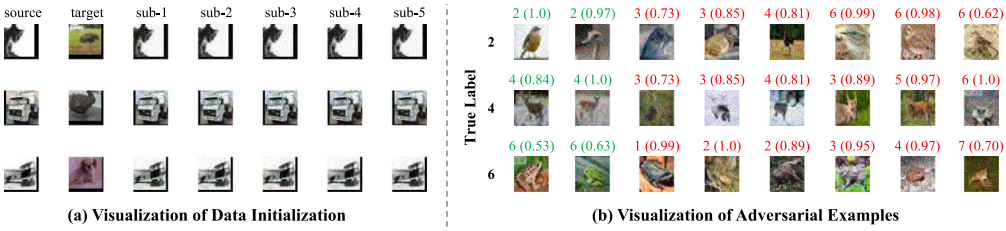
Benchmark. We follow [14] to use ResNet-20 [42] as sub-models and evaluate experiments on the CIFAR-10 dataset. Moreover, we further extend experiments on the CIFAR-100 dataset and VggNet-19 (in the supplemental material) to verify the generality of our method. We compare LAFED with six counterparts, including: (1) baseline, which trains an ensemble in a standard way; (2) four prior robust ensemble training methods ADP [12], GAL [13], DVERGE [14], and TRS [15], and (3) Art adversarially trained TRADES [10]. All ensemble methods are trained by Pytorch framework on a single GeForce RTX 2080 Ti GPU.

Training details. For a fair comparison, all the settings are the same as the work [14]. For ensemble methods, the pre-trained models of baseline, ADP, GAL, and DVERGE are downloaded from the released public repository in [14]. Since TRS does not release the public pre-trained models, we re-train TRS ensemble models by implementing the code with the recommended hyper-parameters reported in the paper. Concretely, the baseline and DVERGE ensembles are trained 200 epochs by using SGD optimizer with momentum 0.9 and weight decay 0.0001. Besides, the ensemble models of ADP, GAL and TRS are trained by using Adam optimizer with an initial learning rate of 0.001.

We following [14] to choose the same settings for the proposed LAFED. Concretely, we train the LAFED ensemble for 200 epochs by using SGD with momentum 0.9 and weight decay 0.0001. The initial learning rate is 0.1 and decayed by $10\times$ at the 100-th and 150-th epochs for baseline, DVERGE, and the proposed LAFED. For DVERGE and LAFED, we distill non-robust dataset by momentum PGD [8] with 10 steps and set the step size equal to $\epsilon/10$. The choice of ϵ is 0.07, 0.05, 0.05, 0.05 and 0.05 for 3, 5, 8, 10, and 15 sub-models to train DVERGE ensemble. As for the proposed LAFED, we select $\epsilon = 0.07$ for ensembles with different number of sub-models, which is discussed in Section 4.6. The eligible layers of unbalanced mixing operation for LAFED include: the input layer, the convolution before the first block, and the final output of all blocks. Moreover, DVERGE, TRS and LAFED use the pre-trained baseline models to retrain the ensemble. GAL replaces the ReLU function with leaky ReLU function to prevent gradient vanishing. All other settings are the same for different methods. For fair comparisons, we select the same settings are the same for both CIFAR-10 and CIFAR-100 datasets to reflect the generality of various methods.

Attack models. We use both clean images and adversarial examples as input data to test the clean accuracy and adversarial robustness of the ensemble models. Specially, we consider both black-box transferable attacks and white-box attacks in the experiments, and the perturbation strength is ranged from 0.01 to 0.07, covering the effective perturbation of most current attack methods.

We utilize hold-out ResNet-20 ensembles trained with 3/5/8 sub-models as the surrogate models to generate transferable adversarial examples. The **black-box attacks** include: (1) momentum PGD with three random starts [8]; (2) M-DI^2-FGSM with transformation probability 0.5 [27]; and (3) skip gradient method (SGM) [28]. The iteration number is set to 100 and step size defaults as $\epsilon/5$ for all attacks. The input transformation probability for M-DI^2-FGSM is set to 0.5, and the γ for SGM is selected as 0.2. To generate strong and diverse adversarial examples, we apply both cross-entropy loss and C&W loss [1] to incorporate with the attacks. In summary, a total of 30 transferable adversarial examples are generated for each clean image to test black-box robustness. For **white-box attacks**, we mainly consider two art methods: (1) 50-steps PGD with the step size of $\epsilon/5$ with five random starts; and (2) Auto-Attack (AA) [30], which is one of the most effectively white-box attacks.



(a) Visualization of Data Initialization (b) Visualization of Adversarial Examples

Fig. 3. Visualization results of data initialization and adversarial examples generated by using LAFED ensemble with 5 members (CIFAR-10). (a) We demonstrate source images, target images, and the distilled data from each sub-model, with randomly applied resizing and padding operations. (b) We illustrate the predictions of transferable adversarial examples ($\epsilon = 0.07$), where the true labels are displayed on the left, while the predicted labels and confidence scores are displayed on top. Notably, the correct and misclassified results are differentiated by green and red colors, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

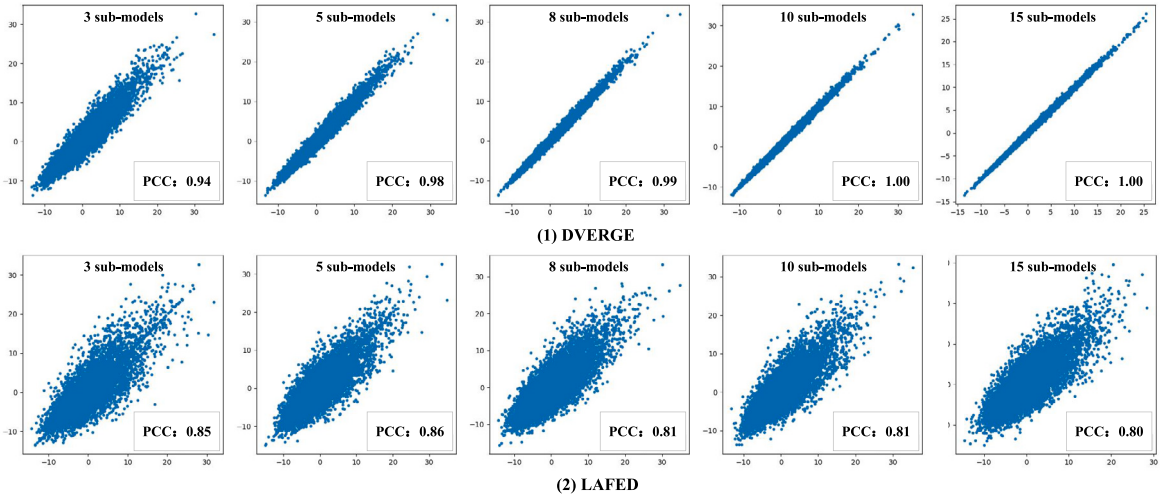


Fig. 4. PCC values of optimized data between 3/5/8/10/15 sub-models.

4.2. Data similarity and model diversity

Data similarity. Prior method DVERGE [14] also performs the optimization based on the vulnerability distilled from other members, while the features in these data provided for different sub-models are quite similar. To verify our speculation, we follow [16] to utilize Person Correlation Coefficient (PCC) value as an indicator to measure the similarity of features in different data batches. Concretely, we randomly select two sub-models to reconstruct the training batch by using DVERGE and our method, and observe their pairwise PCC values of outputs on an individually trained model. We sample 1000 images from CIFAR-10 dataset, and report the PCC values of DVERGE (top) and the proposed LAFED (bottom) in Fig. 4.

We find that the average PCC values of DVERGE are relatively high and consistently increase as the group size becomes large, i.e., the PCC is 0.94 for 3 sub-models and 1.0 for 10 sub-models. It implies the high correlation between their distilled training data as well as limits the improvement of robustness when adding new ensemble members. By contrast, LAFED noticeably reduces the correlation between reconstructed training batches, and it can continuously decrease PCC values by appending new members to the ensemble, which provides evidence that LAFED is qualified for ensemble training to effectively diversify the training batches.

Representation diversity. Another perspective to evaluate the model diversity is the statistics of learned representations. Specially, we select seven intermediate layers with the same interval to summarize the variance of representations. As illustrated in Fig. 5, we use different colors to present the statistic of each sub-model.

For Baseline, ADP, and DVERGE, it is clear that the feature variance of each ensemble member is quite close across the intermediate layers, which implies the data features are assembled in similar clusters. We also observe that the ensembles of GAL and TRS gradually diversified their feature distribution from shallow to deep blocks. However, the divergences between sub-models are not uniform and the variance of learned representations exhibits a polarizing phenomenon, i.e., the variances of some members are relative high (or low) but get close to each other at the last several layers, as demonstrated in Fig. 5(e). It indicates these models may form similar feature distribution at the intermediate space, thus weakening the ensemble diversity.

As for the proposed LAFED, we find that the variance shows progressive varied trends across sub-models by using hierarchical label smoothing strategy to allocate data labels. With the index of sub-model increased, the variance of representations is continuously reduced, which implies the feature clusters become tighter. It verifies that LAFED enforces the learned features to group with different divergence, i.e., smoother labels lead to tighter clusters and smaller variance (see Eq. (5)), thus the diversity of learned representations becomes large within the ensembles.

Transferability within ensemble. As discussed in [14], the transferability within sub-models is a proper measurement to evaluate the diversity and robustness of the ensemble. We report the pairwise transferability of an ensemble with 5 sub-models tested under the attacking strength $\epsilon = 0.05$ in Fig. 6. The number located in the ith row and jth column represents the transfer success rate of the adversarial examples generated from the ith member to fool the jth sub-model. The white-box results are recorded in the diagonal locations.

As shown in Fig. 6, the transferability within the ensemble of ADP and GAL is relatively high, which indicates their inefficacy to improve the model diversity. We find that TRS introduces bias within the ensemble that the transferability in some cells is significantly

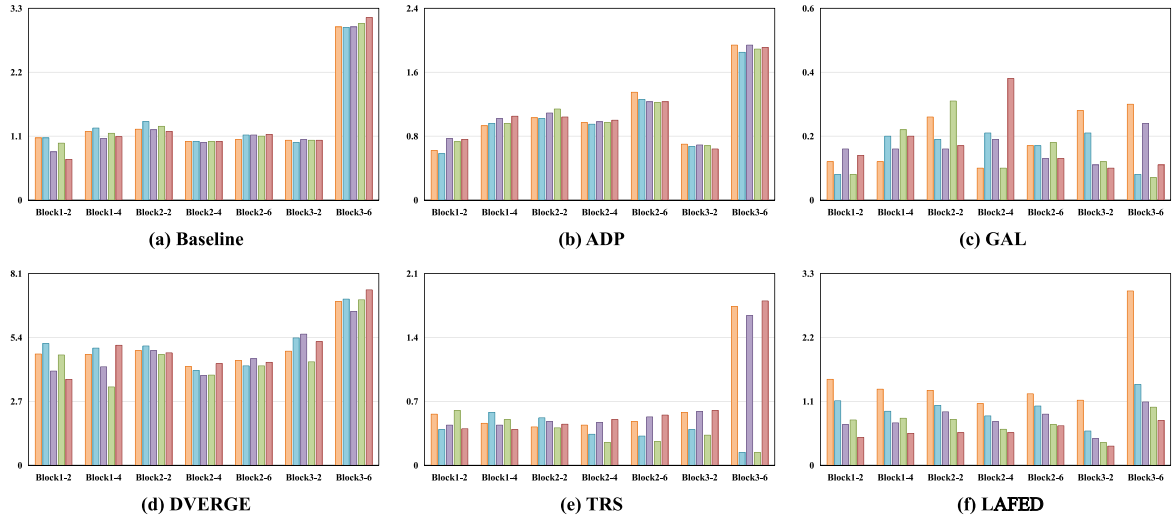


Fig. 5. Statistics of feature variance of 5 ensemble members.

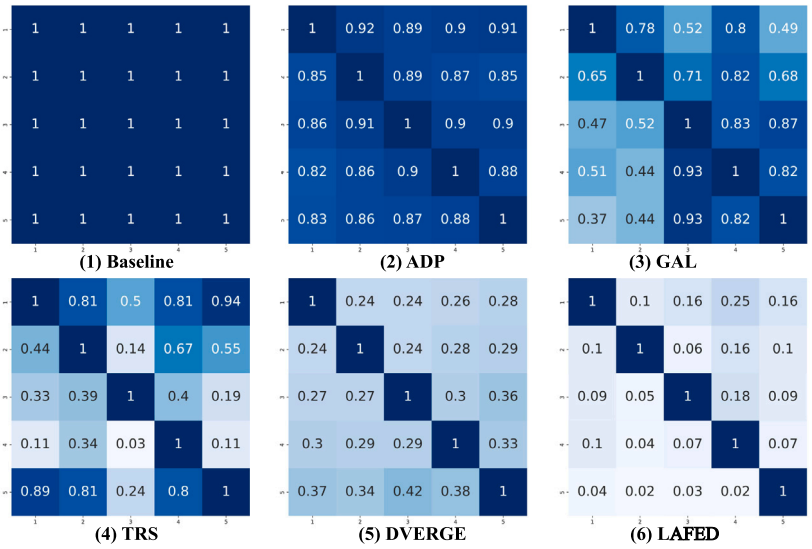


Fig. 6. Transferability among 5 sub-models for different ensemble methods under perturbation $\epsilon = 0.05$.

higher than in others, which reflects its unbalanced diversification property. Though DVERGE has lower transferability compared with other baselines by isolating vulnerability from sub-models, its average transfer rate is still close to 0.3. By applying unbalanced feature combination and hierarchical label smoothing, the proposed LAFED further diversifies the training batch and tightens the feature clusters. Therefore, it achieves the lowest average adversarial transferability and potentially enhances the adversarial resistance. The evaluation results demonstrate that LAFED can be used as an available method to boost model diversification and ensemble robustness.

Decision region. We illustrate the decision region of all ensemble methods with 3 sub-models in Fig. 7. The first row shows the drawing result of the ensemble prediction, and the last three rows illustrate the classifying region of each member, where the same color represents the same model prediction. The vertical axis indicates the gradient direction and the horizontal axis is randomly chosen direction.

As shown in Fig. 7, the sub-models of Baseline, ADP, and GAL are easily fooled by the adversarial examples generated along gradient direction, which leads to weak white-box robustness. TRS exhibits an unbalance property among ensembles that some sub-models are easily tricked by gradient-based attacks while others are not. It is consistent with the observation we discussed in Fig. 6. As for DVERGE and the proposed LAFED, we find ensemble models are more difficult to be flipped the label along the gradient direction as well as other random directions, which indicates their superior robustness compared with other methods.

4.3. Evaluation on CIFAR-10 dataset

We evaluate the robustness under three black-box transfer attacks (i.e., momentum PGD, M-DI^2-FGSM, and SGM) and two white-box attacks (i.e., PGD-50 and AA). We report the averaged classification success rate of ensembles with different group sizes in Fig. 8. Since the results of AA show similar trends with PGD-50, we report AA in supplementary. We note that the black-box testing is challenged because of the "all-or-nothing" rule: the result of one data is correctly predicted only if all of 30 transferable adversarial examples fail to fool the ensemble model. The detailed attack settings and more numerical results are shown in the supplemental material.

In this section, we train the ensemble models on ResNet-20 network and record the results under 3/5/8/10/15 members in Fig. 8. We observe that TRS and DVERGE are the best white-box and black-box defenders among the baselines when the ensemble is the group with 3

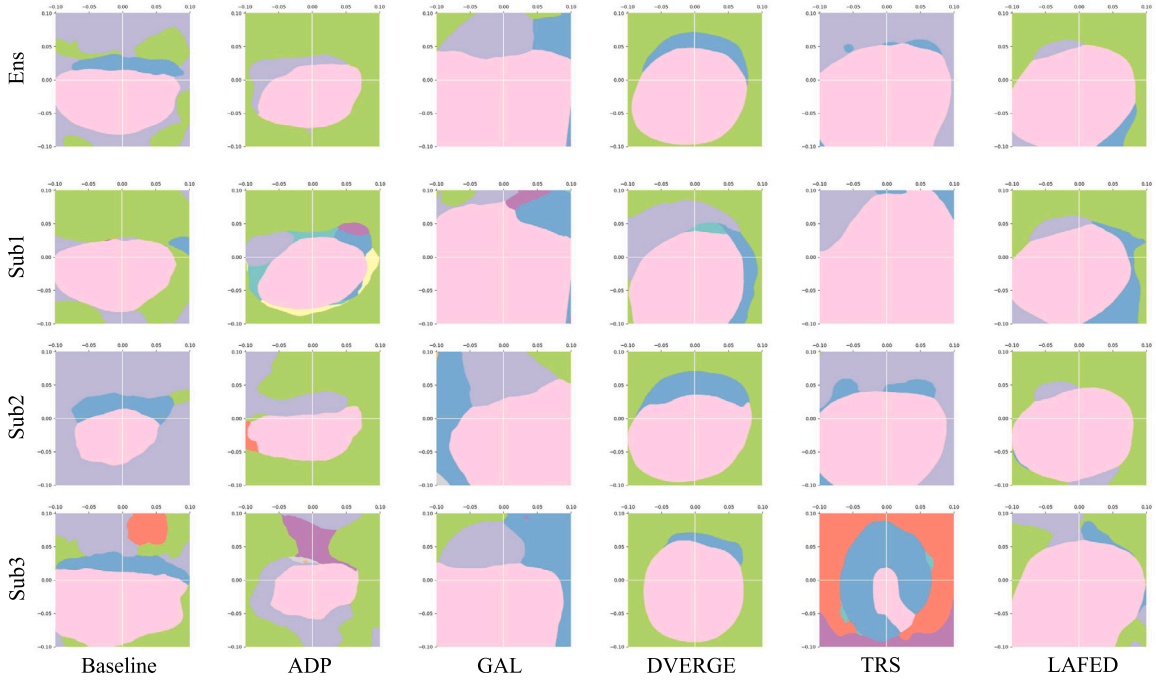


Fig. 7. The decision boundary of different methods with 3 sub-models on CIFAR-10 dataset.

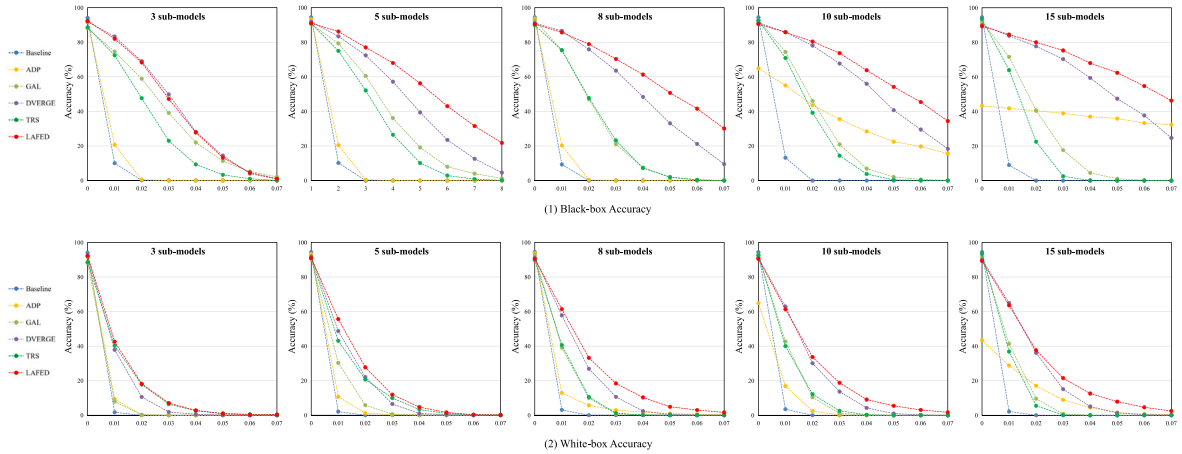


Fig. 8. Robustness of different methods with 3/5/8/10/15 sub-models (ResNet-20) on Cifar-10. (1) Accuracy against black-box transferable attacks; (2) Accuracy against white-box PGD-50 attacks.

sub-models. However, we notice an interesting phenomenon that the robustness of TRS is degraded if more members are included in the ensemble, which may imply its poor generalizability for optimizing a large ensemble models (see Fig. 8 (2)). Moreover, we observe abnormal behaviors in ADP when the ensemble includes 10 or 15 sub-models. Specifically, its clean accuracy experiences a significant drop to 64.9% and 43.3%, respectively. This is primarily because its regularization is dependent on the number of classes and the group size of the ensembles. When the group size equals or exceeds the number of classes, we can see the training procedure may collapse.

By comparing with baselines, the proposed LAFED (red line) separately achieves comparable robust accuracy compared with the best white-box defender TRS and the strongest black-box defender DVERGE when group size is small (i.e., 3 sub-models), and LAFED substantially outperforms *all* baselines when adding more members in the ensemble regardless of the perturbation magnitude (i.e., 5/8/10/15 sub-models). Moreover, current ensemble methods demonstrate weak resistance under big adversarial perturbations. For instance, the accuracy of 8 members DVERGE is merely 9.5% under black-box attacks with perturbations $\epsilon = 0.07$. By contrast, LAFED successfully improves the robustness of 30.1%, which surpasses current state-of-the-art DVERGE by a large margin of 20.6% without dropping the clean accuracy, as shown in Fig. 8 (1). The findings indicate that FASTEN is highly generalizable, and it has the ability to boost the resilience of ensembles by incorporating more sub-models, e.g., the improvements are 20.8%/28.4%/29.8% between 3-sub and 5-sub models under black-box attacks ($\epsilon = 0.03/0.04/0.05$) and 13.0%/9.5%/4.8%/ against white-box attacks ($\epsilon = 0.02/0.03/0.04$). We also report the experiment results on VggNet-19 in the supplemental material, and LAFED still achieves the highest robustness compared with baselines. These quantitative results demonstrate the high efficacy of the proposed LAFED against both white-box and black-box attacks.

4.4. Evaluation on CIFAR-100 dataset

We follow the settings in Section 4.3 to test the robustness of different methods on CIFAR-100 dataset, and report the comparison results

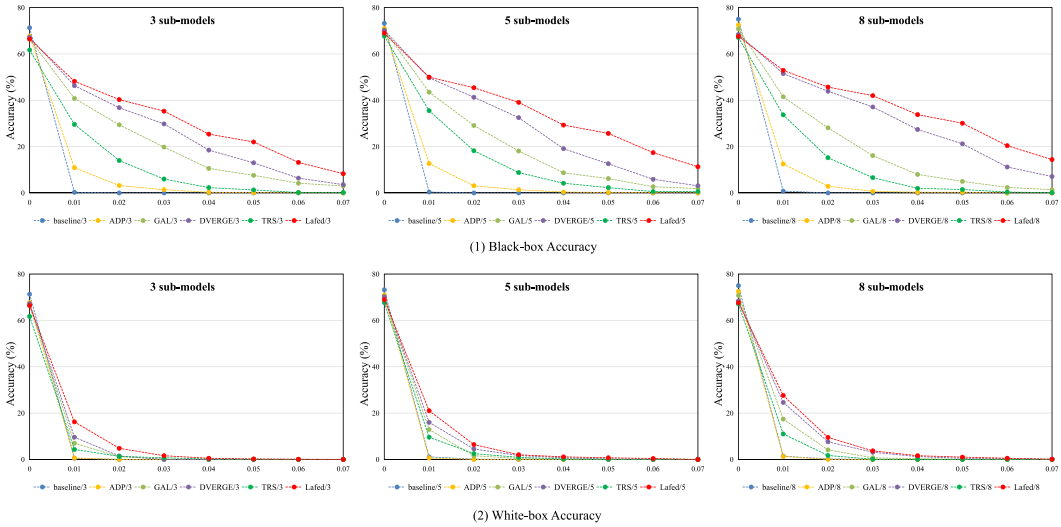


Fig. 9. Robustness of different methods with 3/5/8 sub-models on Cifar-100. (1) Accuracy against black-box transferable attacks; (2) Accuracy against white-box PGD-50 attacks.

in Fig. 9. It is no surprise to observe the same trends as the results recorded in CIFAR-10 experiments. More surprisingly, the proposed LAFED not only performs better than other models in large ensembles (i.e., 5/8 members), but also has the strongest resistance under small group (i.e., 3 sub-models), which further verifies its effectiveness to diversify latent features, which noticeable boost its defensive ability against both black-box and white-box attacks.

In summary, we emphasize several crucial conclusions:

- LAFED is *strong* that it has the best robustness against a variety of attacks without sacrificing clean accuracy compared with the current ensemble methods.
- LAFED is *generalizable* that it performs well on both small datasets (i.e., CIFAR-10) and more complex datasets (i.e., CIFAR-100), but also can be easily extended to a variety of network architectures (i.e., ResNet-20 and Vgg-19).
- LAFED is *universal* for ensemble training optimization so that it can effectively take the advantage of ensemble settings to achieve a great promotion if we include more members for classification inferences.

4.5. LAFED with adversarial training

Since LAFED has achieved the highest accuracy against various attacks by diversifying latent features among sub-models, it still cannot achieve desirable accuracy against strong white-box attacks under $\epsilon = 0.06/0.07$. To this end, we explore whether integrating adversarial training into the proposed LAFED framework can further improve the robustness against white-box attacks with big perturbations. Concretely, we apply adversarial training (AT) loss with LAFED training to optimize the stronger ensembles (LAFED+AT), and compare both its white-box and black-box robustness with state-of-the-art adversarial training TRADES ensembles. In short, the objective function of LAFED+AT is formulated as:

$$\min_{f_i} \mathbb{E}_{(x,y),(x_s,y_s),t}\left[\underbrace{\mu \cdot \mathcal{L}\left(f_i(\mathcal{F}_i^t), y_{s,i}^{ls}\right)}_{Lafedloss} + \underbrace{\mathcal{L}\left(x_s', y_s\right)}_{ATloss}\right], \tag{8}$$

where μ is a hyper-parameter for balancing two losses, i.e., the ensemble training loss (LAFED loss) and the adversarial training loss (AT loss), and x_s' is the adversarial example generated by using PGD to attack the source image x_s under the white-box scenario.

We first explore the performance of LAFED ensembles with adversarial training under different coefficients μ. Three types of data are considered to evaluate the accuracy and robustness for the ensemble model with adversarial training, including: **(1)** clean images without modifying pixels; **(2)** black-box transfer examples with perturbation strength $\epsilon = 0.07$; and **(3)** white-box attack examples with perturbation strength $\epsilon = 0.05$.

As shown in Table 1, a smaller μ degenerates the impact of LAFED objective, leading to higher white-box robustness but lower the clean accuracy. In addition, we can obtain that the best black-box robustness appears when $\mu = 1.0$. Meanwhile, training with $\mu = 1.0$ can balance the effect from both LAFED loss and AT loss, achieving both robustness improvement and considerable clean accuracy. Due to the effect of LAFED loss keeping the same when the number of sub-models changes, we do not need to re-select the μ under 5 and 8 sub-models.

To make a fair comparison, we adopt adversarial training framework to optimize different ensemble models with the recommended hyper-parameters reported in the paper, and report the robustness under white-box attacks in Table 2. Moreover, the performance of various methods under black-box scenarios are recorded in the supplemental material. Since TRADES is treated as one of the most effective adversarial training methods to balance the clean accuracy and robustness, it has slightly higher accuracy in small ensembles for clean data (i.e., 3 members). However, its adversarial robustness is weaker than LAFED+AT under most of the attacking scenarios (i.e., $\epsilon \geq 0.02$). More significantly, LAFED+AT ensemble consistently outperforms TRADES ensemble when we add extra members regardless of the strength of attacks (i.e., $0.01 \leq \epsilon \leq 0.07$). As for DVERGE and TRS, we observe that their white-box robustness decrease as the number of sub-models increases, which is caused by the imbalance between adversarial loss and ensemble loss as the ensemble size changes. On the contrary, LAFED+AT has well generalizability by combining with adversarial training that successfully promotes the white-box robustness as the number of sub-models increases. It is largely attributed to the proposed unbalanced feature combination strategy that it does not change the magnitude of LAFED loss even if we consider more members included in the ensembles. In summary, compared with TRADES and other ensemble methods with adversarial training, LAFED+AT can simultaneously promote the overall robustness and keeps comparable clean accuracy, which validates the efficacy of the proposed LAFED.

4.6. Ablation experiment

Analysis of layer selection strategy. The training phase (Algorithm 1) includes two random steps, namely non-robust features distillation layer l and feature combination layer t. The sensitivity of layer index for distillation has been discussed in [14], and it shows that random

Table 1
Robustness of LAFED ensembles trained with different μ. Each number in the table block represents (clean accuracy)/(black-box transfer accuracy under $\epsilon = 0.07$)/(white-box accuracy under $\epsilon = 0.05$).

Settings	$\mu = 0.5$	$\mu = 1.0$	$\mu = 1.5$
3 sub-models	78.4%/62.3%/24.5%	79.4%/62.8%/23.6%	81.8%/60.5%/20.5%

Table 2
White-box robustness of different ensemble methods combined with adversarial training under 3/5/8 sub-models (%).

ϵ	clean	0.01	0.02	0.03	0.04	0.05	0.06	0.07
TRADES/3	80.4	70.9	57.7	45.6	31.1	19.9	11.2	6.6
ADP+AT/3	81.2	70.5	57.2	42.0	27.7	18.0	8.9	5.1
GAL+AT/3	78.8	68.0	54.2	38.9	23.0	12.3	5.6	2.2
DVERGE+AT/3	83.0	**72.9**	59.8	44.4	29.0	19.0	10.0	4.8
TRS+AT/3	**86.2**	63.7	36.8	17.0	5.6	1.6	0.5	0.3
LAFED+AT/3	79.4	69.6	**61.3**	**47.4**	**34.7**	**23.6**	**14.2**	**7.4**
TRADES/5	80.4	70.9	59.0	46.0	33.5	23.0	13.4	7.0
ADP+AT/5	82.6	71.1	56.8	42.3	27.8	16.9	8.9	4.3
GAL+AT/5	80.9	70.5	56.6	42.9	29.6	16.8	8.9	4.2
DVERGE+AT/5	84.2	**73.4**	58.8	41.4	26.6	15.3	7.7	3.4
TRS+AT/5	**86.6**	69.3	45.6	23.2	9.6	2.3	1.3	0.5
LAFED+AT/5	80.3	71.7	**61.6**	**49.7**	**36.1**	**25.5**	**15.8**	**8.4**
TRADES/8	79.9	71.1	60.0	46.8	35.3	23.8	14.2	7.5
ADP+AT/8	84.5	72.5	58.6	43.8	29.5	17.8	8.9	4.1
GAL+AT/8	82.6	72.3	59.2	44.9	30.6	20.3	10.5	5.2
DVERGE+AT/8	85.4	**73.3**	58.3	40.6	25.0	14.1	5.7	2.3
TRS+AT/8	**89.9**	63.7	32.9	14.9	5.1	1.2	0.4	0.2
LAFED+AT/8	80.7	71.8	**61.1**	**48.7**	**35.9**	**24.7**	**15.5**	**9.3**

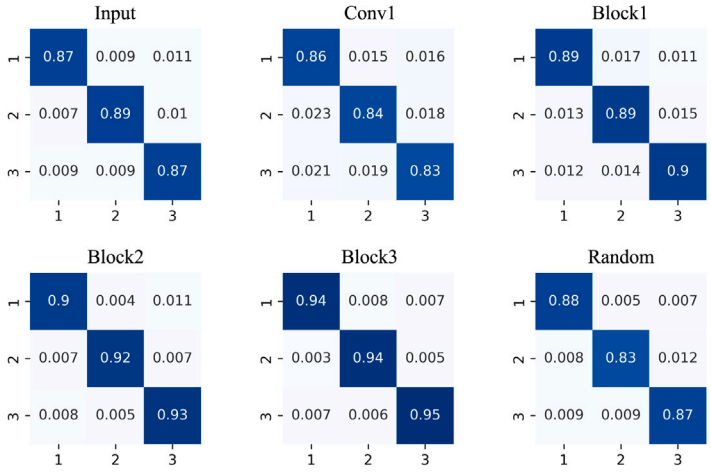


Fig. 10. Transferability results for different layer selection under the attacking perturbation $\epsilon = 0.01$.

Table 3
Robustness of different layer selection under white-box attacks ($\epsilon = 0.01$) and black-box attacks ($\epsilon = 0.03$).

Attacker	Input	Conv1	Block1	Block2	Block3	Random
White-box	36.1	37.9	35.4	32.0	31.0	**40.2**
Black-box	37.9	**49.7**	39.2	35.6	32.3	48.3
Average	37.0	43.8	37.3	33.8	31.2	**44.3**

selection achieves higher performance. In this section, we further study the effect to feature mixup operation with different layer selections. Concretely, we shrink its range to fix cases, including: **(1)** the input layer, denoted as "Input"; **(2)** the first convolution layer, denoted as "Conv1"; **(3)** the outputs of Block 1/2/3 (i.e., the 7-th, 13-th and 19-th layer of the ResNet20); and **(4)** randomly samples a layer from candidate layers above, denoted as "Random". Fig. 10 reports the within-ensemble transferability under 3 sub-models case, and Table 3 shows the results of white-box and black-box robustness.

Several findings catch our attention. First, all selections have a low pairwise transferability (Fig. 10), which verifies the efficacy of LAFED for diversifying sub-models. Second, mixing features at shallow layers (i.e., the first convolution) improves the robustness of each member against white-box attacks (e.g., the diagonal numbers in the "Input" and "Conv1" is lower than others in Fig. 10), while combining features at deep layers (i.e., the outputs of Block 2/3) for mixup can lower the transferability among sub-models (see off-diagonal numbers in Fig. 10). The main reason is that the robustness of each sub-model itself plays a more important role in the small size ensemble, thus shallow layer selection shows better performance than mixing deep features. As the number of sub-models increasing, the diversity among sub-models dominates the importance of overall robustness [14]. Third, the randomly sampling strategy learn the advantages of various layers synchronously, which improves the overall robustness and reduces the transferability among sub-models as well. As a result, it achieves the best white-box robustness than other choices, and has slightly worse black-box accuracy than "Block3" (see Fig. 10). This observation confirms that applying mixup at the feature space exhibit more effectiveness in further diversifying sub-models of the ensemble.

Analysis of distillation magnitude. The ϵ used for generating non-robust data is important to LAFED. We evaluate the impact of different ϵ on

Table 4
Robustness of LAFED ensembles trained with different ϵ. Each number in the table block represents (clean accuracy)/(black-box transfer accuracy under perturbation strength 0.03)/(white-box accuracy under perturbation strength 0.01).

Settings	$\epsilon = 0.03$	$\epsilon = 0.05$	$\epsilon = 0.07$
3 sub-models	**93.4%**/4.2%/22.0%	93.1%/23.2%/30.1%	91.3%/**48.3%** /**40.2%**
5 sub-models	**93.9%**/25.8%/41.0%	92.2%/55.1%/50.7%	91.4%/**67.1%**/**51.3%**
8 sub-models	**92.7%**/41.3%/51.7%	91.9%/64.3%/**58.3%**	91.0%/**71.4%**/56.5%

Table 5
Robustness of LAFED ensembles trained with different p. Each number in the table block represents (clean accuracy)/(black-box transfer accuracy under $\epsilon = 0.04$)/(white-box accuracy under $\epsilon = 0.02$).

Settings	$p = 0.025$	$p = 0.05$	$p = 0.075$
3 sub-models	91.5%/21.2%/13.0%	**92.4%**/**27.8%**/18.2%	91.1%/27.5%/**20.2%**

Table 6
Robustness of each strategy with 5 sub-models under PGD-50 white-box attacks.

	UFC	HLS	clean	0.01	0.02	0.03	0.04	0.05	0.06	0.07
LAFED	✓	×	91.4%	51.3%	23.4%	9.1%	2.7%	0.6%	0.2%	0%
	×	✓	90.6%	46.1%	18.5%	5.8%	1.1%	0.3%	0%	0%
	✓	⋆	91.4%	43.2%	15.6%	5.6%	1.5%	0.4%	0.1%	0%
	⋆	✓	**91.9%**	47.3%	20.4%	7.3%	3.0%	0.9%	0.3%	0.1%
	✓	✓	91.0%	**55.6%**	**27.7%**	**11.9%**	**4.7%**	**1.6%**	**0.3%**	**0.1%**

adversarial robustness. We test the performance of LAFED ensembles trained with $\epsilon = 0.03/0.05/0.07$ under 3, 5, and 8 sub-models separately. Table 4 reports the accuracy of LAFED ensembles trained with different ϵ and evaluated on (1) clean data; (2) black-box transfer examples with perturbation strength 0.03; and (3) white-box attack examples with perturbation strength 0.01. Moreover, we test the transferability among sub-models of LAFED ensembles trained with different ϵ, which is another essential metric to measure the effect of the perturbation magnitude (see supplemental material).

As shown in Table 4, we find that a larger ϵ encourages LAFED ensembles to improve both white-box and black-box robustness. On the contrary, a smaller ϵ leads to higher clean accuracy. However, the improvement of robustness is much higher than the decrease of clean accuracy under $\epsilon = 0.07$ compared with the small perturbation $\epsilon = 0.03$. Meanwhile, the LAFED ensembles trained with $\epsilon = 0.07$ is also achieves the lowest transferability under 3,5, and 8 sub-models, which implies the strongest resistance against black-box adversarial examples (see supplementary). It is consistent with the quantitative results in Table 4. To this end, we choose $\epsilon = 0.07$ as the optimal parameter for the proposed LAFED method.

The interval of hierarchical label smoothing. Following [19], we select uniform distribution to allocate the weight of smoothed label (see Eq. (5)). The smoothed degree of label vectors is controlled by $\eta_i = p \cdot (i - 1)$, where i is the index of sub-models and p implies the interval of hierarchical label smoothing. We set up 3 groups of ablation experiments to study the effect of p, corresponding to 0.025, 0.05 and 0.075 under the ensemble of 3 members case. Table 5 represents the performance of LAFED ensembles under different settings, including (1) clean data; (2) black-box transferable adversarial examples with perturbation magnitude $\epsilon = 0.04$; and (3) white-box methods with attacking strength $\epsilon = 0.02$. As shown in Table 5, LAFED achieves the highest clean accuracy and black-box robustness under $p = 0.05$, but the highest white-box accuracy appears under $p = 0.075$. To better balance the clean accuracy and overall robustness of LAFED, we choose $p = 0.05$ to implement Hierarchical Label Smoothing.

Effectiveness of unbalanced feature combination and hierarchical label smoothing. Unbalanced Feature Combination (UFC) and Hierarchical Label Smoothing (HLS) are the essential strategies for diversifying latent features. To highlight the efficacy of each strategy in the proposed LAFED method, we conduct the ablation study to compare the robustness if we **removing** (denoted as "×") or **replacing** one of them by standard forms (denoted as "⋆").

Concretely, we consider two schemes for replacement scenarios:

- Instead of sampling unbalanced weights in the proposed UFC, we use standard uniform weights for feature combination.
- Rather than adopting the proposed HLS, we select standard label smoothing strategy with the same weight for training all the ensemble members.

The comparisons under different settings are shown in Table 6. On the one hand, we can see both UFC and HLS are important to promote the overall robustness of the proposed LAFED, and each of them boosts the performance for classifying clean data and adversarial examples. On the other hand, by replacing the proposed strategies with the standard module, LAFED can be easily generalized to those scenarios that require higher accuracy with relatively lower robustness. These results demonstrate that UFC and HLS are elaborately designed for **ensemble scenarios** thereby applying them simultaneously can substantially improve robustness against adversarial attacks.

Table 7
Training time (hours) comparison between methods..

Methods	Baseline	ADP	GAL	DVERGE	TRS	TRADES	LAFED (ours)
Train time	1.3	2.0	6.5	12.5	28.0	14.5	11.5

4.7. Training time

We also report the comparison of training time in Table 7. As Table 7 shows, training the LAFED ensemble is faster than DVERGE and TRS, which are state-of-the-art ensemble methods. Although both LAFED and DVERGE distill the non-robust data as training dataset, DVERGE has to back propagate several non-robust data independently, while LAFED merely optimizes a mixup feature based on back-propagation that costs less training resources. In addition, training LAFED ensembles is also faster than TRADES. Though ADP needs the least time compared with LAFED, it cannot effectively promote robustness as shown in Fig. 6 of the main paper. Due to the significant robustness improvements of LAFED, we believe that it is worth to spent extra training time compared with GAL. In summary, LAFED not only improves the robustness but also shortens the training time compared with the state-of-the-art method.

5. Conclusion

In this work, we focus on the problem of ensemble training methods. We start by identifying the fact that a high correlation of training

data can affect the diversity and transferability between individually trained models. It motivates us to propose Latent Feature Diversification (LAFED), aiming to diversify the latent feature among sub-models and boost the robustness of the overall model. Concretely, LAFED introduces unbalanced feature combination and hierarchical label smoothing to decrease the correlation of training batch as well as increase the diversity of learned representations among sub-models. Experiments demonstrate that training with LAFED objective can effectively decrease the transferability among sub-models and achieve the best performance under both black-box and white-box attacks compared to existing ensemble methods without sacrificing clean accuracy. Moreover, LAFED consistently improves robustness as the number of ensemble members increases.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work was supported by National Key Research and Development Plan in China (2023YFC3306100).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patcog.2023.110225.

References

[1] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 Ieee Symposium on Security and Privacy (Sp), IEEE, 2017, pp. 39–57.
[2] L. Huang, C. Gao, Y. Zhou, C. Xie, A.L. Yuille, C. Zou, N. Liu, Universal physical camouflage attacks on object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 720–729.
[3] Y. Shi, Y. Han, Q. Zhang, X. Kuang, Adaptive iterative attack towards explainable adversarial robustness, Pattern Recognit. 105 (2020) 107309.
[4] A. Ghosh, S.S. Mullick, S. Datta, S. Das, A.K. Das, R. Mallipeddi, A black-box adversarial attack strategy with adjustable sparsity and generalizability for deep image classifiers, Pattern Recognit. 122 (2022) 108279.
[5] A.E. Cinà, A. Torcinovich, M. Pelillo, A black-box adversarial attack for poisoning clustering, Pattern Recognit. 122 (2022) 108306.
[6] L. Huang, S. Wei, C. Gao, N. Liu, Cyclical adversarial attack pierces black-box deep neural networks, Pattern Recognit. (2022) 108831.
[7] A. Ilyas, S. Santurkar, L. Engstrom, B. Tran, A. Madry, Adversarial examples are not bugs, they are features, Adv. Neural Inf. Process. Syst. 32 (2019).
[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, 2018.
[9] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, Robustness may be at odds with accuracy, in: International Conference on Learning Representations, (2019) 2019.
[10] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, M. Jordan, Theoretically principled trade-off between robustness and accuracy, in: International Conference on Machine Learning, PMLR, 2019, pp. 7472–7482.
[11] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, Q. Gu, Improving adversarial robustness requires revisiting misclassified examples, in: International Conference on Learning Representations, 2019.
[12] T. Pang, K. Xu, C. Du, N. Chen, J. Zhu, Improving adversarial robustness via promoting ensemble diversity, in: International Conference on Machine Learning, PMLR, 2019, pp. 4970–4979.
[13] S. Kariyappa, M.K. Qureshi, Improving adversarial robustness of ensembles with diversity training, 2019, arXiv preprint arXiv:1901.09981.
[14] H. Yang, J. Zhang, H. Dong, N. Inkawhich, A. Gardner, A. Touchet, W. Wilkes, H. Berry, H. Li, DVERGE: Diversifying vulnerabilities for enhanced robust generation of ensembles, Adv. Neural Inf. Process. Syst. 33 (2020).
[15] Z. Yang, L. Li, X. Xu, S. Zuo, Q. Chen, P. Zhou, B.I. Rubinstein, C. Zhang, B. Li, TRS: Transferability reduced ensemble via promoting gradient diversity and model smoothness, in: Thirty-Fifth Conference on Neural Information Processing Systems, 2021.
[16] C. Zhang, P. Benz, T. Imtiaz, I.S. Kweon, Understanding adversarial examples from the mutual influence of images and perturbations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14521–14530.
[17] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations, 2018.
[18] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, Y. Bengio, Manifold mixup: Better representations by interpolating hidden states, in: International Conference on Machine Learning, PMLR, 2019, pp. 6438–6447.
[19] R. Müller, S. Kornblith, G.E. Hinton, When does label smoothing help? Adv. Neural Inf. Process. Syst. 32 (2019) 4694–4703.
[20] Y. Xiao, C.-M. Pun, B. Liu, Fooling deep neural detection networks with adaptive object-oriented adversarial perturbation, Pattern Recognit. 115 (2021) 107903.
[21] X. Sun, G. Cheng, L. Pei, J. Han, Query-efficient decision-based attack via sampling distribution reshaping, Pattern Recognit. 129 (2022) 108728.
[22] R. Duan, X. Ma, Y. Wang, J. Bailey, A.K. Qin, Y. Yang, Adversarial camouflage: Hiding physical-world attacks with natural styles, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1000–1008.
[23] R. Duan, X. Mao, A.K. Qin, Y. Chen, S. Ye, Y. He, Y. Yang, Adversarial laser beam: Effective physical-world attack to DNNs in a blink, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16062–16071.
[24] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, 2013, arXiv preprint arXiv:1312.6199.
[25] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, 2014, arXiv preprint arXiv:1412.6572.
[26] N. Papernot, P. McDaniel, I. Goodfellow, Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, 2016, arXiv preprint arXiv:1605.07277.
[27] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, A.L. Yuille, Improving transferability of adversarial examples with input diversity, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2730–2739.
[28] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, X. Ma, Skip connections matter: On the transferability of adversarial examples generated with ResNets, in: International Conference on Learning Representations, 2019.
[29] J. Hang, K. Han, H. Chen, Y. Li, Ensemble adversarial black-box attacks against deep learning systems, Pattern Recognit. 101 (2020) 107184.
[30] F. Croce, M. Hein, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, in: International Conference on Machine Learning, PMLR, 2020, pp. 2206–2216.
[31] Z. Hu, H. Li, L. Yuan, Z. Cheng, W. Yuan, M. Zhu, Model scheduling and sample selection for ensemble adversarial example attacks, Pattern Recognit. (2022) 108824.
[32] M. Atzmon, N. Haim, L. Yariv, O. Israelov, H. Maron, Y. Lipman, Controlling neural level sets, Adv. Neural Inf. Process. Syst. 32 (2019).
[33] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[34] T.G. Dietterich, Ensemble methods in machine learning, in: International Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15.
[35] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learn. 51 (2) (2003) 181–207.
[36] S. Lee, H. Lee, S. Yoon, Adversarial vertex mixup: Toward better adversarially robust generalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 272–281.
[37] T. Pang, K. Xu, J. Zhu, Mixup inference: Better exploiting mixup to defend adversarial attacks, in: International Conference on Learning Representations, 2019.
[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[40] C.-B. Zhang, P.-T. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, M.-M. Cheng, Delving deep into label smoothing, IEEE Trans. Image Process. 30 (2021) 5984–5996.
[41] C. Fu, H. Chen, N. Ruan, W. Jia, Label smoothing and adversarial robustness, 2020, arXiv preprint arXiv:2009.08233.
[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

Wenzi Zhuang is currently working at the School of Computer Science and Engineering, Sun Yat-sen University. His research interests are adversarial learning, computer graphics and its application.

Lifeng Huang is currently working at the College of Mathematics and Informatics, South China Agricultural University. He works in areas of adversarial learning, deep learning and its application.

Chengying Gao is an Associate Professor at the School of Computer Science and Engineering, Sun Yat-sen University. She got her Ph.D. degree from Sun Yat-sen University. She works in areas of computer vision, computer graphics, and deep learning.

Ning Liu received the Ph.D. degree from Sun Yat-sen University. He is currently a Professor with the School of Computer Science and Engineering, Sun Yat-sen University. His research interests include computer vision, cyberspace security, and deep learning.

Neural Networks 156 (2022) 13–28

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

DEFEAT: Decoupled feature attack across deep neural networks

Lifeng Huang [a], Chengying Gao [b], Ning Liu [b,c,*]

[a] College of Mathematics and Informatics, South China Agricultural University, China
[b] School of Computer Science and Engineering, Sun Yet-Sen University, China
[c] Guangdong Key Laboratory of Information Security Technology, China

ARTICLE INFO

Article history:
Received 22 June 2022
Received in revised form 31 August 2022
Accepted 8 September 2022
Available online 20 September 2022

Keywords:
Adversarial example
Transferability
Black-box
Feature-level attack
Defenses

ABSTRACT

Adversarial attacks pose a security challenge for deep neural networks, motivating researchers to build various defense methods. Consequently, the performance of black-box attacks turns down under defense scenarios. A significant observation is that some feature-level attacks achieve an excellent success rate to fool undefended models, while their transferability is severely degraded when encountering defenses, which give a false sense of security. In this paper, we explain one possible reason caused this phenomenon is the domain-overfitting effect, which degrades the capabilities of feature perturbed images and makes them hardly fool adversarially trained defenses. To this end, we study a novel feature-level method, referred to as **De**coupled **Fe**ature **At**tack (DEFEAT). Unlike the current attacks that use a round-robin procedure to estimate gradient estimation and update perturbation, DEFEAT decouples adversarial example generation from the optimization process. In the first stage, DEFEAT learns an distribution full of perturbations with high adversarial effects. And it then iteratively samples the noises from learned distribution to assemble adversarial examples. On top of that, we can apply transformations of existing methods into the DEFEAT framework to produce more robust perturbations. We also provide insights into the relationship between transferability and latent features that helps the community to understand the intrinsic mechanism of adversarial attacks. Extensive experiments evaluated on a variety of black-box models suggest the superiority of DEFEAT, i.e., our method fools defenses at an average success rate of 88.4%, remarkably outperforming state-of-the-art transferable attacks by a large margin of 11.5%. The code is publicly available at https://github.com/mesunhlf/DEFEAT.

1. Introduction

Deep neural networks (DNNs) have provided state-of-the-art results for various computer vision tasks (Ito, Nakae, Hata, Okano, & Ishii, 2019; Li, Li, Liu, & Hong, 2021; Lin, Jia, Huang, & Gao, 2022; Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). Nevertheless, their security vulnerabilities have raised widespread concerns recently — artificially generated adversarial examples by adding imperceptible perturbations to clean data will fool DNNs to output incorrect predictions. As a result, these maliciously crafted data not only present potential threats for digital systems (Vidnerová & Neruda, 2020), but also cause real-world applications in unexpected and dangerous behaviors (Huang et al., 2020).

The discovery of adversarial examples is a double-edged sword. On the one hand, the attackers can create malignant data for fooling remote DNNs and inducing security flaws (Li et al., 2020). On the other hand, adversarial attacks also can be regarded as an essential surrogate to evaluate the robustness of DNNs (Carlini & Wagner, 2017). As a result, the potential value of the adversarial example has triggered an arms race between adversaries and defenders.

Under black-box settings, the attackers exploit the transferability of adversarial examples for fooling unseen remote models with the least cost, collectively as transferable attacks. In this field, current works can be divided into two categories: (a) **gradient-based attacks**, which disturb the final classification space by using momentum (Dong et al., 2018), input transformation (Xie et al., 2019), or translation operation (Dong, Pang, Su, & Zhu, 2019); and (b) **feature-level attacks**, which perturb the intermediate feature space to enhance adversarial examples (Huang et al., 2019; Naseer, Khan, Rahman, & Porikli, 2018). Both of them have high success rates for fooling undefended networks (i.e., normally trained models). For resisting attacks, researchers built a variety of defenses (Guo, Rana, Cisse, & van der Maaten, 2018; Madry, Makelov, Schmidt, Tsipras, & Vladu, 2018; Naseer, Khan, Hayat, Khan, & Porikli, 2020; Xie, Wang, Zhang, Ren, & Yuille, 2018). In consequence, these transferable attacks suffer degraded performance under defenses scenarios. However, we

* Corresponding author at: School of Computer Science and Engineering, Sun Yet-Sen University, China.
E-mail address: liuning2@mail.sysu.edu.cn (N. Liu).

https://doi.org/10.1016/j.neunet.2022.09.009
0893-6080/

observe a crucial phenomenon that some advanced feature-level attacks are *weaker* than most gradient-based methods to circumvent defenders, i.e., their transferability becomes much lower across robustified models. Briefly, although most feature-level attacks illustrate a high performance against undefended models, some of them perform worse to fool robust defenses. This might lead one to wonder: what reason cause the behavior difference between gradient-based methods and feature-level attacks?

Our goal is to provide the explanation for this phenomenon, and go a step further in boosting the power of feature-perturbed adversarial examples for breaking defenses. We first analyze the main reason that caused their degenerated capacity of fooling defenses is the *domain-overfitting* effect. This conception is inspired by the two-domain hypothesis that clean images and adversarial images are drawn from different domains (Xie & Yuille, 2019). Concretely, vanilla models have trained on clean data to learn representations from the benign domain, which are distinguishable with those internal features learned by robustified networks from the adversarial domain (Dong et al., 2019). Therefore, feature-level attacks manipulate the intermediate representations in the benign domain so that they transfer well across normally trained models. Meanwhile, the perturbed benign representations exhibit considerable differences to their counterparts in the adversarial domain, therefore lowering the transferability of feature-perturbed images for fooling adversarially trained models. In contrast, gradient-based attacks merely disrupt the final outputs to generate perturbations, so they are not sensitive to domain diversity. We note that the domain-overfitting is similar to the network-overfitting (Dong et al., 2018) for adversarial examples, while the main difference is that the domain-overfitting effect exhibits more bias on different domains above (see Section 3.2).

To narrow the gap between these two types of methods, we propose **De**coupled **Fe**ature **At**tack (DEFEAT) for mitigating the domain-overfitting effect. Current adversarial attacks often estimate the gradient direction and update adversarial perturbations in a round-robin manner, i.e., one-stage methods. It may induce the overfitting effect of generated adversarial examples. To this end, we design DEFEAT as a two-stage method by *decoupling perturbation generation from the optimization process*. In the first phase (i.e., learning stage), rather than searching the optimal but overfitted adversarial example, DEFEAT adopts the designed optimization strategies into the basic feature disrupting process to learn a lower-dimensional adversarial distribution, which fills with plentiful perturbations regard to relatively high loss. At the second phase (i.e., generation stage), DEFEAT iteratively samples the noises from the learned distribution to construct adversarial perturbations. By performing this attack paradigm, the gap between two domains of feature-perturbed images becomes smaller. Thereby DEFEAT has superior transferability against both benign and adversarial domains compared with state-of-the-art transferable attack methods. On this basis, we further study the relationship between latent features and transferability, including the factors that may affect the adversarial strength and the variations of attention regions, which can helps the community to improve the robustness of models.

In summary, the contributions of our work are four-fold:

- We discover that some feature-level attacks demonstrate poor transferability across from benign domains to adversarial domains and provide a detailed analysis and plausible explanation for this phenomenon i.e., the domain-overfitting effect.
- To mitigate the domain-overfitting effect and boost transferability across domains, we decouple the representation distorting optimization and perturbation generation to form a novel two-stage feature-level adversarial attack method, namely Decoupled Feature Attack (DEFEAT).
- Empirical results suggest that DEFEAT yields much superiority over current feature-level methods across adversarial domains. By integrating gradient-based methods into the proposed framework, DEFEAT further improves transferability and thereby establishes state-of-the-art for transferable attacks under black-box settings.
- We study the correlation between transferability and deep representations, which can help the community understand the inner mechanism of adversarial attacks and provide guidance to improve the robustness of networks.

2. Related work

2.1. Adversarial attacks

Recently, adversarial examples raised widespread attention since they cause the potential security issues (Bai, Yang, & Liu, 2020; Maimon & Rokach, 2022). Even in the real world, maliciously generated objects can fool deep learning models (Huang et al., 2020). Generally, two main types of adversarial attack have been developed based on the attacking setups, i.e., white-box and black-box methods. In this paper, we focus on studying the untargeted black-box attacks for image classifiers within the constraint ℓ_∞ norm-balls under black-box scenarios.

Black-box Attacks. The underlying information of the targeted model is limited, i.e., attackers cannot obtain the gradient information directly. One popular line of the method is transferable adversarial attacks, which exploit the transferability of adversarial examples to fool unknown models, i.e., malicious data is created from an off-the-shelf white-box model and used to mislead the targeted model. However, the adversarial examples generated by gradient-based attacks often suffers from the network-overfitting effect, which may degenerate the transferable strength. To overcome this drawback, Dong et al. introduce the momentum strategy to escape from local optimal solutions (Dong et al., 2018). Xie et al. apply random input transformations in each attacking iteration (Xie et al., 2019). Translation-Invariant method (TIM) optimizes on a substantial amount of translated images to obtain the synthesized gradient (Dong et al., 2019). Moreover, taking into account diverse gradients information from multiple networks also boost the transferable power (Huang, Wei, Gao, & Liu, 2022; Liu, Chen, Liu, & Song, 2017). Due to the insensitivity of domain differences, these gradient-based methods not only fool black-box vanilla models, but also partly break defense networks who trained on adversarial data.

Apart from the above gradient-based attacks, recent studies demonstrate that feature-level methods are efficient for fooling black-box models (Wang et al., 2021). Instead of perturbing the final classification distribution, prior works find that distorting the intermediate representations for a specific network also achieves a high performance to generate adversarial examples. For instance, Mopuri et al. perturb feature activations at multiple layers to generate universal adversarial examples under white-box scenarios (Mopuri, Ganeshan, & Babu, 2018). By distorting the neural representations and maximizing the perceptual distance, adversarial examples can improve transferable strength and exhibit generalizability across tasks, i.e., Neural Representation Distortion (NRD) (Naseer et al., 2018). Based on the linearity of decision boundary, Huang et al. propose Intermediate Level Attack (ILA) by controlling the direction/disturbance trade-off at feature layers, which boost the power of basic methods (Huang et al., 2019). Feature Importance-aware Attack (FIA) creates a batch of masked images for aggregating the gradients as semantically object-aware attentions, and improves its transferability by suppressing important features intensified with attention mechanisms (Wang et al., 2021). Though those feature-level attacks

perform excellently as gradient-based methods on misleading undefended networks, most of them exhibit lower performance for fooling robustified adversarial models, which is largely due to the domain-overfitting effect. To improve the performance of feature-distorted images against defenses, we apply optimization strategies to the feature-perturbing process to propose a novel two-stage method, namely Decoupled Feature Attack (DEFEAT).

Beyond this, the query-based method is another popular line to generate black-box adversarial examples. Logbarrier (Finlay, Pooladian, & Oberman, 2019) sample the vectors from the Bernoulli distribution to estimate the decision boundaries. Square Attack (Andriushchenko, Croce, Flammarion, & Hein, 2020) and Sign-Opt (Cheng et al., 2020) generate the perturbations based on the Uniform distribution. NES (Ilyas, Engstrom, Athalye, & Lin, 2018) draws noises from standard normal distribution at each query to facilitate adversarial attacks. These methods benefit from various distributions with *fixed* parameters. Among them, $\mathcal{N}$Attack (Li, Li, Wang, Zhang, & Gong, 2019) approximates the probability density distribution in a query-feedback manner, which is similar to our method. The major difference between $\mathcal{N}$Attack (Li et al., 2019) and our method is three-fold. First, $\mathcal{N}$Attack spends thousands of queries and feedbacks to fool targeted models, while DEFEAT optimizes the distribution based on white-box models for crafting transferable attacks, and its transferability is much higher than $\mathcal{N}$Attack. Second, $\mathcal{N}$Attack is a one-stage method that approximates the distribution and samples the perturbations simultaneously, while DEFEAT decouples the perturbation generation from the optimization process to improve its transferability (i.e., two-stage). Third, $\mathcal{N}$Attack utilizes the final outputs to update the parameters, while DEFEAT disturbs the intermediate features to approximate a generalizable distribution across domains. In this paper, we focus on searching normal distributions with high adversarial transferability to benefit the first stage, and leave other plausible choices (e.g., Bernoulli or Poisson distributions) as future work.

2.2. Defense methods

Motivated by the threat of adversarial attacks, many parallel algorithms have been proposed to improve the robustness of DNNs, which can be roughly divided into two categories: **(a)** adversarial training, which utilizes malicious examples as augmented data to train the models (Goodfellow, Shlens, & Szegedy, 2014; Zhang, Huang, Zhu, & Liu, 2021); **(b)** gradient obfuscation, which mitigates the adversarial strength by masking the gradient (Xie et al., 2018). For adversarial training, malicious data is regenerated at each iteration to maximize the adversarial effect and they are used to enhance the resistance for trained networks (Borkar, Heide, & Karam, 2020; Kurakin, Goodfellow, & Bengio, 2017; Madry et al., 2018). For instance, Tramr et al. increases the diversity of perturbations during training by leveraging the adversarial examples transferred from other white-box models into an ensemble paradigm (Tramèr et al., 2018). Cohen et al. certify the robustness through random smoothing operation (Cohen, Rosenfeld, & Kolter, 2019). To date, adversarial training is still one of the most effective defenses. Besides, gradient obfuscation is an alternative way to resist adversarial attacks. For example, image modifications (Xie et al., 2018), data compression and reconstruction (Guo et al., 2018; Jia, Wei, Cao, & Foroosh, 2019), guided denoising process (Liao et al., 2018), self-supervised trained neural representation purifier (Naseer et al., 2020) are used to mitigate attack strength. Most of them can be applied as "plug-in" methods combined with adversarial training to form stronger defenders. The defenses above have shown strong resistance against transferable adversarial attacks. To this end, we design a novel feature-level attack to break black-box defenses successfully.

3. Method

In this section, we start by analyzing the reason for the poor transferability of some existing feature-level attacks across from benign to adversarial domains. To this end, we rethink the potential strategies which can boost the transferability in feature-perturbed optimizations. This inspires us to propose a novel two-stage transferable method for mitigating the domain-overfitting effect, referred to as Decoupled Feature Attack (DEFEAT).

3.1. Motivation

Feature-level transferable methods perform excellently to fool undefended black-box networks (i.e., normally trained models), but some of them gain worse results against robustified models (Huang et al., 2019; Naseer et al., 2018). Based on the two-domain hypothesis (Xie & Yuille, 2019), feature-perturbed images are generated by disturbing the deep representations of normally trained models. Therefore they exhibit effectiveness across the benign domains and transfer well to other networks. On the other hand, adversarially trained models have learned features from adversarial domains so that those feature-level attacks can hardly generalize to fool robustified models. This so-called domain-overfitting effect may lead to their low transferability across robust defenses. We note that the domain-overfitting is different from the network-overfitting that it shows *strong bias* on domain differences (see Section 3.2). To push for further advances in this study, we propose to learn a novel and robust feature-level method for mitigating the domain-overfitting effect, namely **De**coupled **Fe**ature **At**tack (DEFEAT). Unlike existing methods that produce adversarial examples in a one-stage paradigm, DEFEAT decouples the optimization process and perturbation generation. It applies optimization strategies to search the adversarial space in the first stage and then repeatedly samples unit noises to construct perturbations in the second stage. Consequently, the feature-perturbed images become less sensitive to domain diversity, thereby achieving higher transferability across benign domains to adversarial domains.

3.2. The domain-overfitting effect

We aim to find the reason that caused the poor performance of some feature-level attacks upon defense models. One plausible explanation is the two-domain hypothesis (Xie & Yuille, 2019), i.e., clean images and adversarial examples are drawn from two different domains. We relate this hypothesis to the weakness of some feature-level attacks: feature-perturbed images crafted on normally trained models overfit benign domains, thus they can hardly fool those robustified models that learn representations from the adversarial domain.

To verify our speculation, we observe the variations of internal representations as the transferable adversarial examples are propagated through the black-box networks. In the evaluations, we consider three powerful gradient-based attacks (i.e., BIM Kurakin et al., 2017, DIM Xie et al., 2019 and TIM Dong et al., 2019) and two state-of-the-art feature-level methods, (i.e., NRD Naseer et al., 2018 and ILA Huang et al., 2019).

We randomly choose 1000 images x from ImageNet dataset and craft their adversarial counterparts x' by attacking three normally trained white-box models, including DenseNet161 (D161), InceptionResNetV2 (IR2) and ResNet152 (R152). Two black-box networks are selected as targeted models, i.e., normally trained InceptionV3 (I3) and adversarially trained InceptionV3$_{ens3}$ (I3$_{ens3}$). These two black-box models share the same architecture but separately learn representations from clean images (benign domain) and adversarial examples (adversarial domain). When an image x

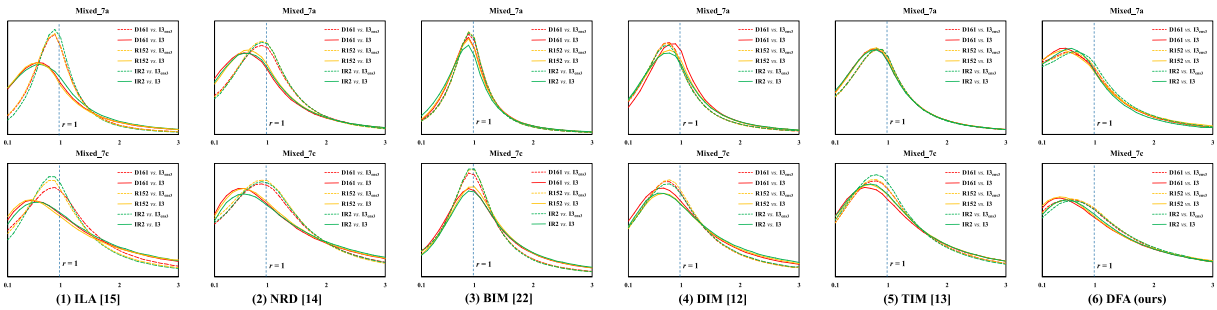


Fig. 1. Variation statistics of deep representations at Mixed_7a (top) and Mixed_7c (bottom) layers. The white-box models include D161 (colored in **red**), I4 (colored in **yellow**), IR2 (colored in **green**), and the black-box models are I3 (solid line) and $I3_{ens3}$ (dotted line). We also plot the ratio $r = 1$ in **blue** dotted line. The figures suggest the feature-level attacks (i.e., ILA Huang et al., 2019, NRD Naseer et al., 2018) leads to huge different statistics for benign domain and adversarial domain. On the contrary, the gradient-based attacks (i.e., BIM Kurakin et al., 2017, DIM Xie et al., 2019, TIM Dong et al., 2019) are not sensitive to the domain difference. This confirms our speculation that the prior feature-level attacks craft on the benign domains can hardly generalize to the adversarial domains, i.e., the domain-overfitting effect. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

passing through a specific layer l of the black-box model $f(\cdot)$, the mean value at ith channel of feature maps is denoted as $f_l^i(x)$. In summary, we can measure the pairwise variation degree of the latent representations between a clean image and its adversarial counterpart under black-box model $f(\cdot)$ as:

$$v = \left\| \frac{f_l^i(x')}{f_l^i(x)} \right\|_1, \tag{1}$$

We report the statistics of variations v within the range of $[0.1, 3.0]$ in Fig. 1. Intuitively, the farther shift distance of the statistic distribution away from $r = 1$, the more representations of clean images x are perturbed to generate adversarial examples x' for the black-box model $f(\cdot)$, meaning the stronger perturbation under the transferable scenarios.

According to Fig. 1 (1, 2), we find that feature-perturbed images generated by using ILA and NRD induce huge different trends under normally trained model I3 and adversarial model $I3_{ens3}$ if they disturb the representations in benign domains (i.e., D161, IR2, and R152). More specifically, ILA and NRD are evidently powerful to disturb the representations across benign domains but also show their inefficacy to generalize to the adversarial domains: the statistic distribution of variations upon I3 is much farther away from $r = 1$ (solid lines), while the statistic distribution for $I3_{ens3}$ is nearly centered around $r = 1$ (dotted lines), which implies their poor capacity to affect the adversarial representations extracted by robustified model $I3_{ens3}$. In contrast, gradient-based methods BIM, DIM, and TIM are not sensitive to domain diversity, so that the gaps between two categories of statistic distribution are relatively small, as illustrated in Fig. 1 (3, 4, 5). Among them, the statistic distributions of DIM and TIM shift farther than others when transferring to $I3_{ens3}$. Therefore, they perform better to fool robustified models than BIM, ILA, and NRD.

This observation reveals that feature-perturbed images crafted on normally trained models usually exhibit overfitting to the benign domain. Besides, we also report the detailed transfer success rates of these methods in Section 4.2. Both qualitative and quantitative results confirm the phenomenon of domain-overfitting effect, which may degrade the transferability of some feature-level methods across from vanilla networks to robustified models.

3.3. Rethinking optimization strategy for transferability

To improve the transferability of BIM, gradient-based methods DIM and TIM introduce diversity input (Xie et al., 2019) and translation operation (Dong et al., 2019) to optimize adversarial examples. Consequently, DIM and TIM induce more considerable variations for latent representations than BIM, and therefore mitigate the network-overfitting and achieve superior performance for fooling black-box models, as shown in Fig. 1. Following this line, we briefly consider three plausible optimization strategies to substantially mitigate the domain-overfitting effect.

Perturbation Augmentation. Input augmentation is widely used for training networks to prevent from overfitting effect. Similarly, it can be utilized in the adversarial learning field, such as derivative-free optimizations in query-based attacks (Ilyas et al., 2018; Li et al., 2019) or adversarially training the robust defenses (Cohen et al., 2019). Rather than input augmentation, we design *perturbation augmentation* for optimizations to improve transferability. In other words, simultaneously optimizing multiple adversarial perturbations over a single input data facilitates to search for diverse updated directions, mitigating the domain-overfitting effect. Intuitively, we can view this optimization as finding an adversarial space assembled the perturbations with high transferability.

Noise Smoothing. Query-based attacks show that estimating gradient at low dimensional space can improve their efficacy (Chen, Zhang, Sharma, Yi, & Hsieh, 2017; Ilyas, Engstrom, & Madry, 2019; Li et al., 2019). Meanwhile, prior works (Dong et al., 2019; Zhou et al., 2018) indicates that some defenses are sensitive to high-frequency noise, including image resizing (Xie et al., 2018) and wavelet denoising (Prakash, Moran, Garber, DiLillo, & Storer, 2018). It motivates us to generate the perturbation at a lower-dimensional space $\mathbb{R}^{\mathcal{L}}$ rather than the original space $\mathbb{R}^{\mathcal{H}}$ ($\mathcal{L} \ll \mathcal{H}$) for smoothing the constructed perturbations.

Neighbor Interpolation. The transferability of adversarial examples is similar to the generalization performance of models. It inspires us to enhance adversarial examples by the concept used for improving the generalizability of networks (i.e., mixup) (Zhang, Cisse, Dauphin, & Lopez-Paz, 2018). Besides, Pang, Xu, and Zhu (2019) have shown that adversarial examples can be correctly classified if they interpolate with clean images. Similarly, we propose neighbor interpolation to diversify the optimized perturbations, i.e., optimizing over surrogate data by interpolating a clean image and its neighbors from the identical distribution in attacks can be regarded as searching the directions that simultaneously push multiple data towards the decision boundary. It is similar to the universal attack (Moosavi-Dezfooli, Fawzi, Fawzi, & Frossard, 2017) that generates a single global perturbation to fool different data with high transferability.

3.4. Decoupled feature attack

Overview. Unlike current one-stage methods that alternately calculate the gradients and update perturbations, DEFEAT decouples

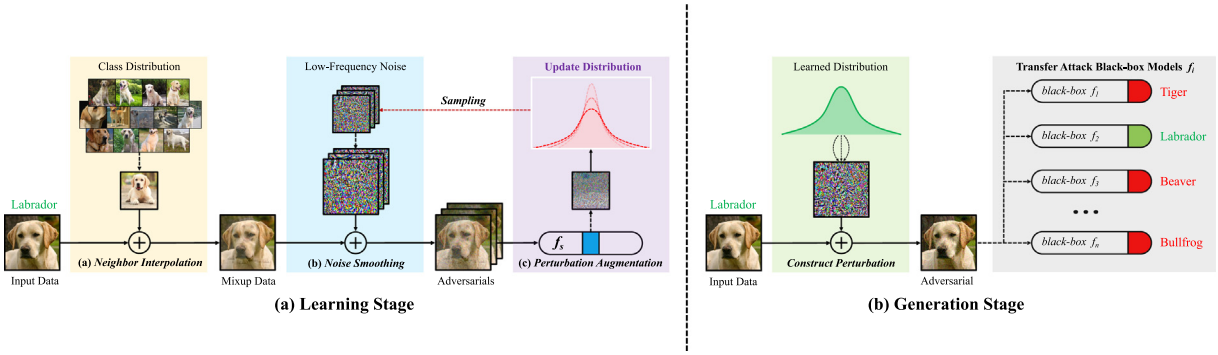


Fig. 2. The pipeline of the proposed DEFEAT framework. DEFEAT decouples the optimization and perturbation generation. (a) Learning Stage. To mitigate domain-overfitting effect, we explore three optimization strategies in the basic feature distortion process to learn a low dimensional adversarial distribution, including: neighbor interpolation (yellow area), noise smoothing (blue area), and perturbation augmentation (purple area). (b) Generation Stage. DEFEAT samples the noise from learned distribution in an iterative manner to generate perturbations (green area). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

perturbation generation from adversarial optimization to form a two-stage method. For every clean image, DEFEAT learns a specific adversarial distribution in the first stage, and craft its adversarial counterpart at the second stage.

As illustrated in Fig. 2(a), DEFEAT performs 3 steps for adversarial optimizations in the first phase (i.e., Learning Stage):

- **Step 1**. Producing the surrogate data by interpolating the input data with a resampled clean image (yellow area).
- **Step 2**. A set of adversarial examples are generated by adding the surrogate data with a bunch of smoothed perturbations, which are drawn from a low-dimensional distribution (blue area).
- **Step 3**. For maximizing the divergence between original images and adversarial images, DEFEAT optimizes over a group of adversarial examples to learn an adversarial distribution (purple area).

In the second phase (i.e., Generation Stage), DEFEAT samples the unit noise in an iterative manner to generate adversarial perturbations (green area), as shown in Fig. 2(b).

According to Fig. 1 (6), the statistics gaps between two domains become minimal, demonstrating the efficacy of DEFEAT for mitigating the domain-overfitting effect. Note that both two-stage paradigm and proposed optimization strategies are important in the DEFEAT framework, which are separately discussed in Section 4.2.

Learning Stage. We apply the optimization strategies for the feature distortion process.

• *Perturbation Augmentation*. Instead of directly searching for the optimal feature-perturbed image with the highest loss value, DEFEAT optimizes a bunch of sub-optimal perturbations to mitigate the domain-overfitting effect. Therefore, we can apply these perturbations to the clean image to form an adversarial space, i.e., the adversarial example cluster with relatively high loss. Following this idea, we design perturbation augmentation to find an adversarial space within a vicinity region over a probability density distribution $\mathcal{F}$, such that every sampled perturbation can make the feature divergence between the original image and its corresponding adversarial image large. The objective loss is defined as:

$$\begin{aligned}\underset{\theta}{\operatorname{argmax}}\ \widetilde{J}(\theta) &= \mathbb{E}\big[J(x, x')\big] \\ &= \mathbb{E}\big[J\big(x, \phi(x, u)\big)\big], \quad u \sim \mathcal{F}(\theta) \\ &\approx \frac{1}{m}\sum_{i=1}^{m} J\big(x, \phi(x, u_i)\big), \quad u_i \sim \mathcal{F}(\theta)\end{aligned} \tag{2}$$

where the adversarial example x' is generated by the attacking procedure $\phi(\cdot)$, i.e., $x' = \phi(x, u)$ (detailed in Eq. (4)), and u denotes the unit noise of perturbation drawn from the distribution $\mathcal{F}(\theta)$ whose parameter θ is to be optimized. For better optimization, we can increase the size m to reduce the randomness. By maximizing Eq (2), the loss $\widetilde{J}(\cdot)$ over the distribution $\mathcal{F}(\theta)$ will become large in expectation. Intuitively, the larger the value of $\widetilde{J}(\cdot)$, the higher chance an adversarial perturbation will be generated by the unit noise u_i sampled from $\mathcal{F}(\theta)$.

• *Noise Smoothing*. We consider two aspects of sampling unit noise from the adversarial distribution $\mathcal{F}(\theta)$. First, the low-frequency perturbations benefit the transferability of adversarial examples, motivating us to parameterize the distribution $\mathcal{F}(\theta)$ in a lower-dimensional space $\mathbb{R}^{\mathcal{L}}$ rather than the original space $\mathbb{R}^{\mathcal{H}}$. Second, an appropriate distribution makes the objective function $\widetilde{J}(\cdot)$ to be smooth that we can easily optimize the parameter θ. Based on the analysis, we sample the unit noise u_i from $\mathcal{F}(\theta)$ by transformation as:

$$u_i = \frac{2}{\pi} \cdot \operatorname{atan}\big(\mathrm{S}(z_i, \beta)\big), \quad z_i \sim \mathcal{N}(\mu, \sigma^2), \tag{3}$$

where u_i is transformed from a seed z_i, which is drawn from a low-dimensional distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2. Following Li et al. (2019), we compute z_i by a transformation $z_i = \mu + z_0\sigma$, where z_0 is sampled from the standard normal distribution, i.e., $z_0 \sim \mathcal{N}(0, 1)$, $z_0 \in \mathbb{R}^{\mathcal{L}}$. Here we use atan function to control the unit noise u_i within ℓ_∞ norm constraint $[-1, 1]$. $\mathrm{S}(\cdot)$ denotes the upsampling operation to keep the consistency with the dimensions of the input data, and we set the factor as $\beta = l/h$, where the lower dimensional space and the original space are defined as $\mathcal{L} = l \times l \times c$ and $\mathcal{H} = h \times h \times c$ with c channels, respectively.

• *Neighbor Interpolation*. To learn a more generalizable distribution, we integrate neighbor interpolation process into the attacking procedure $\phi(\cdot)$ to generate adversarial example x':

$$\phi(x, u_i) = \big((1 - \lambda) \cdot x + \lambda \cdot \hat{x}\big) + \epsilon \cdot u_i, \quad \hat{x} \sim \mathcal{D}(y) \tag{4}$$

where λ denotes the ratio to control the linearity of interpolated data; $\hat{x}$ is a clean image resampled from the identical distribution of original data x, i.e., $\mathcal{D}(y)$, the unit noise u_i is calculated by Eq. (3), and ϵ is used to control the perturbation magnitude under ℓ_∞ norm $[-\epsilon, \epsilon]$. Note that if $\lambda = 0$ this optimization is equal to $\phi(x, g) = x + \epsilon \cdot u_i$, which is similar to the gradient-based attack FGSM (Goodfellow et al., 2014), while we remove its sign$(\cdot)$ operation for better optimization.

Notice, we do not follow data augmentation (e.g., mixup Zhang et al., 2018) to sample the coefficient λ from the Beta distribution.

Conversely, we choose a small fixed coefficient for interpolation to improve the perturbation diversity. There are two reasons for this behavior. First, we aim to construct transferable perturbation for a specific image x that we select smaller λ for preserving the majority features of it. Second, we empirically find that the transferability of adversarial examples generated by sampling from Beta distribution is slightly worse than their counterparts based on the fixed coefficient, even if the expectation and variance of Beta distribution are small (see supplemental material).

• *Overall Objective.* Instead of Euclidean Distance defined in Naseer et al. (2018), we select an another basic loss $J(\cdot)$ (i.e., Cosine similarity) to maximize the diversity at layer l between clean images and feature-perturbed images in Eq. (2) as:

$$J(x, x') = -\text{cosine}\big(\text{N}(f_l(x)), \text{N}(f_l(x'))\big), \tag{5}$$

where $\text{N}(\cdot)$ is the normalization function, which used to make the activation at layer l contribute equally. We have verified these two objectives and the experimental results demonstrate that Cosine Similarity is better to achieve a higher performance (see supplemental material).

According to the discussion above, the objective function $\widetilde{J}$ in Eq. (2) is reformulated as:

$$\underset{\mu,\sigma}{\text{argmax}}\ \widetilde{J} = \frac{1}{n}\sum_{i=1}^{n} J\Big(x, \phi\big(x, \frac{2}{\pi}\cdot\text{atan}(\text{S}(z_i,\beta))\big)\Big), \quad z_i \sim \mathcal{N}(\mu,\sigma^2) \tag{6}$$

where we aim to learn the parameters of an adversarial distribution (mean μ and variance σ^2) for a specific input image x.

Generation Stage. After learning the parameters, the noise drawn from the distribution $\mathcal{N}(\mu,\sigma^2)$ have an adversarial effect in expectation. We freeze the approximated distribution and sample the unit noises in an iterative manner to assemble transferable adversarial examples as:

$$x'_{t+1} = x'_t + \alpha\cdot\text{sign}\Big(\frac{2}{\pi}\cdot\text{atan}\big(\text{S}(z_t,\beta)\big)\Big), \quad z_t \sim \mathcal{N}(\mu,\sigma^2), \tag{7}$$

where α denotes the step size, μ and σ are freezed parameters, and z_t denotes a seed sampled at iteration t. The detailed framework of DEFEAT is illustrated in Algorithm 1.

Combination Variants. DEFEAT is a generalizable framework that we can naturally combine existing transferable methods with DEFEAT to build its improved variants. For example, we denote the integration of diversity input transformations (Xie et al., 2019) and the objective function Eq. (2) as DI-DEFEAT:

$$\underset{\theta}{\text{argmax}}\ \widetilde{J}(\theta) = \mathbb{E}\Big[J\big(\text{R}(x,p), \text{R}(x',p)\big)\Big], \tag{8}$$

where $\text{R}(\cdot)$ is random transformation operations, and p is transformation probability in DIM (Xie et al., 2019).

Similar, we can combine translation invariance (Dong et al., 2019) with the unit noise transformation Eq. (3) as TI-DEFEAT:

$$u_i = \frac{2}{\pi}\cdot\text{atan}\big(W * \text{S}(z_i,\beta)\big), \quad z_i \sim \mathcal{N}(\mu,\sigma^2), \tag{9}$$

where W denotes the guassian kernel matrix in TIM (Dong et al., 2019).

Algorithm 1 Decoupled Feature Attack

Input: Input images x; true label y; feature-based loss function $J(\cdot)$; maximum magnitude of perturbation ϵ, learning and generation iterations T_1, T_2; perturbation augmentation size m; upsampling factor β; interpolation ratio λ;
Output: Adversarial Example x'_{T_2}

Initialize μ, σ, $t \leftarrow 0$
for $t = 0$ to T_1 **do** # ***Learning Stage***
  $\widetilde{J} \leftarrow 0$,
  $\widetilde{x} = (1-\lambda)\cdot x' + \lambda\cdot\hat{x}, \quad \hat{x} \sim \mathcal{D}(y)$
  for $i = 0$ to m **do**
    $u_i = 2/\pi\cdot\text{atan}(\text{S}(z_i,\beta)),\ z_i \sim \mathcal{N}(\mu,\sigma^2), 0<\beta<1$
    $x'_i \leftarrow \widetilde{x} + \epsilon\cdot u_i$
    $\widetilde{J} \leftarrow \widetilde{J} + \frac{1}{m}J(x, x'_i)$
  end for
  Optimize μ and σ by maximizing the objective $\widetilde{J}$
end for
$t \leftarrow 0$, $\alpha \leftarrow \epsilon/T_2$, $x'_0 \leftarrow x$
for $t = 0$ to T_2 **do** # ***Generation Stage***
  Freeze parameters μ and σ
  $\hat{u}_t = 2/\pi\cdot\text{atan}(\text{S}(z_t,\beta)), \quad z_t \sim \mathcal{N}(\mu,\sigma^2), 0<\beta<1$
  $x'_{t+1} = x'_t + \alpha\cdot\text{sign}(\hat{u}_t)$
end for

4. Experiment

In this section, we provide experimental results to demonstrate the efficacy of our proposed DEFEAT framework. We start by providing experimental setups in Section 4.1. Then we conduct extensive evaluations about our proposed DEFEAT in Section 4.2. Based on the discussion above, we further compare the proposed method and baselines under combination scenarios in Sections 4.3–4.5. Finally, we provide insights into the relationship between transferability and internal representations in Section 5.

4.1. Experimental setup

Models and Defenders. Following Dong et al. (2019), Huang et al. (2019) and Wang et al. (2021), we select four normally trained models, include DenseNet161 (D161) (Huang, Liu, Van Der Maaten, & Weinberger, 2017), InceptionV3 (I3) (Szegedy et al., 2016), InceptionResNetV2 (IR2) (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017), ResNet152 (R152) (He, Zhang, Ren, & Sun, 2016). These four vanilla models are used as white-box models to generate transferable adversarial examples.

For defenders, we consider a variety of adversarial robustified models:

- $\text{InceptionV3}_{\text{ens3}}$ (I3_{ens3}), $\text{InceptionV3}_{\text{ens4}}$ (I3_{ens4}), Adversarial InceptionV3 (AdvI3), and $\text{InceptionResNetV2}_{\text{ens}}$ (IR2_{ens}) (Tramèr et al., 2018).
- High-level Representation Guided Denoiser (HGD) (Liao et al., 2018).
- Input transformation defense by bit depth reduction operations (BDR) (Guo et al., 2018).
- The combination of pixel deflection and total variance minimization (PDT) (Prakash et al., 2018).
- Image transformations including random resizing and padding (R&P) (Xie et al., 2018).
- Comdefend: denoising by compression and reconstruction models (COM) (Jia et al., 2019).
- A classifier trained by randomized smoothing with tight robustness guarantee (RS) (Cohen et al., 2019).
- Pre-trained selective feature regeneration defense (SFR) (Borkar et al., 2020).
- Self-supervised trained neural representation purifier (NRP) (Naseer et al., 2020).

For test-time "plug-in" defenses (e.g., BDR, R&P, etc.), we use adversarial I3_{ens3} as underlying model to form a more robust defense.

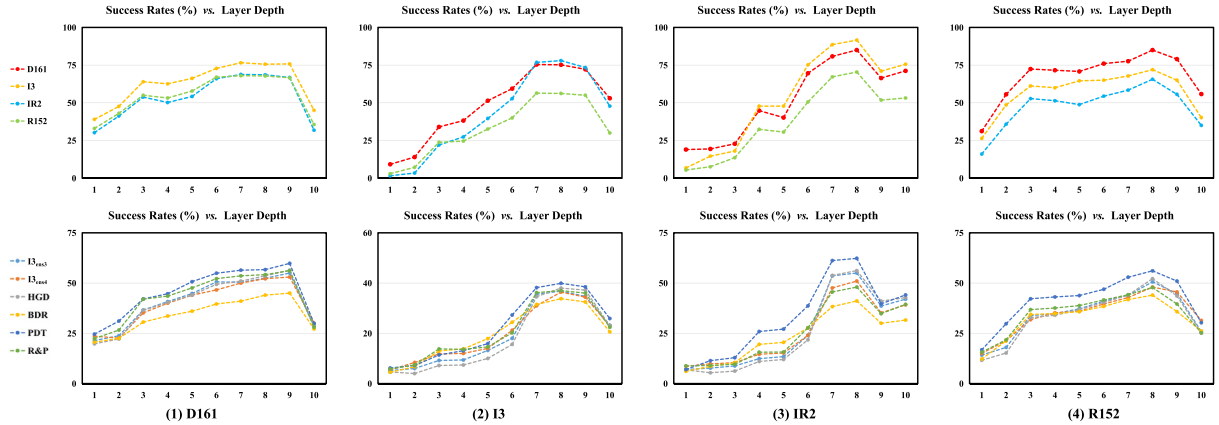


Fig. 3. The success rates (%) *vs.* layer depth of the proposed DEFEAT against four normally trained models (top row) and six defenses (bottom row). In the top row, the adversarial examples are crafted by attacking four vanilla models, including (1) D161 (colored in **red**), (2) I3 (colored in **yellow**), (3) IR2 (colored in **blue**), and (4) R152 (colored in **green**). Specially, the result under white-box attacks is labeled as symbol "×". In the bottom row, we also plot the transfer results of DEFEAT under each defense in different colors. Better viewed with zoom-in. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Baseline Attacks. Our experiments include three state-of-the-art feature-level baselines, including Neural Representation Distortion (NRD) (Naseer et al., 2018), Intermediate Level Attack (ILA) (Huang et al., 2019), and Feature Importance-aware Attack (FIA) (Wang et al., 2021). Beyond feature-level attacks, we also consider two gradient-based baselines, including DIM (Xie et al., 2019) and TIM (Dong et al., 2019). In our experiments, we select the *strongest combination* versions of gradient-based attacks reported in the literature as default setups (e.g., MI-DIM for DIM, TI-MI-DIM for TIM, etc.).

Dataset. We randomly select 1000 images from ImageNet validation dataset. All images can be correctly classified by vanilla models, and thereby treated as the standard benchmark to be collected in the SACP2019 competition (Tianchi Security AI Challenger Program Competition).

Parameters Setup. For the proposed DEFEAT, we utilize Adam optimizor (Kingma & Ba, 2014) and learning rate 0.2 to search the distribution parameters. Following Wang et al. (2021) and Xie et al. (2019), we set the maximum magnitude of perturbation $\epsilon = 16$ (i.e., equivalent to the magnitude 0.062 within the normalized range $[0, 1]$). We choose the size of perturbation augmentation $m = 3$, scale factor $\beta = 0.5$, and upsamling ratio $\lambda = 0.1$. We follow the suggestion in Zhao, Liu, and Larson (2021) to make the comparisons under convergence scenarios, e.g., the learning iteration $T_1 = 30$ and the generation iteration $T_2 = 10$. Those hyper-parameters are detailed discussed in the supplemental material. As for the improved variants DI-DEFEAT and TI-DEFEAT, we set the probability $p = 0.7$ and kernel size 11×11.

To conduct fair comparisons, we set total iteration number $T = 40$ for all baselines, which equals to the sum of learning and attacking steps in DEFEAT. Similarly, we adopt the transformation probability $p = 0.7$ for DIM (Xie et al., 2019), and select the kernel size 11×11 for TIM (Dong et al., 2019). As for feature-level attacks NRD (Naseer et al., 2018), ILA (Huang et al., 2019) and FIA (Wang et al., 2021), we report their optimal layers in supplementary, and set the same parameters as DEFEAT.

Layer Notations. As mentioned in Section 3.4, feature-level methods disturb the latent representations of specific layer l for constructing adversarial examples (see Eq. (5)). Without loss of generality, we choose 10 layers with the same interval step across the whole depth to study the transferability for each white-box model. The notations of the internals are denoted with their corresponding layer depth. For example, the first noted layer of model I3 ($I3_{L1}$) is near the input layer (i.e., *Conv_1a*), and the last layer ($I3_{L10}$) is closer to the output layer (i.e., *Logits*). We follow this rule to set notations for four normally trained white-box models, and detailed notations are shown in the supplemental material.

4.2. Experiments under standalone scenarios

In this section, we evaluate the capability of DEFEAT against different black-box models under *standalone* settings (w/o combining with existing methods). We first empirically analyze the effect of layer depth. Then we compare the performance of DEFEAT with three feature-level baselines. Finally, we provide extensive quantitative ablation studies about DEFEAT.

The Effect of Layer Depth. The adversarial examples are crafted on a total of 10 layers for each white-box model to transfer attack against black-box models. We report the results in Fig. 3, where the performance of DEFEAT against excluded normally trained models (i.e., D161, I3, IR2, and R152) and six defenses (i.e., $I3_{ens3}$, $I3_{ens4}$, HGD, BDR, PDT, and R&P) are plotted in the top and bottom row, separately. The primary axes are attack success rates and layer depth, respectively. The dots with different colors represent the results of adversarial examples generated by attacking various layers against corresponding models.

According to Fig. 3, the implications are three-fold. First, the transferability is *white-box model dependent*. In other words, the curves of success rates from a given model against both benign and adversarial domains exhibit highly resembled trends. It is similar to the conclusion in Inkawhich, Wen, Li, and Chen (2019), while they only includes the black-box normally trained models in studies (*e.g.,* VGG19 Simonyan & Zisserman, 2015, R50 He et al., 2016, etc.). We further provide evidence that this phenomenon has extended into robustified models. Second, we find that the transferability of DEFEAT method is *layer dependent*. Generally, the strength of adversarial examples generated by attacking deeper layers is much stronger than those crafted on shallower features. The success rates under all black-box scenarios increase continually until layer depth exceeds 8 (or 9), and then drop dramatically at the last layer. We relate it to the observations in Zeiler and Fergus (2014), i.e., the shallow layers extract low-level features and deeper internals learn high-level semantic representations. It implies that corrupting high-level features may generate stronger adversarial examples. Third, transferability tendencies are *attacker dependent*. By comparing between Fig. 3, Fig. 14, and Fig. 15 (see supplemental material), we can see the trend gaps between different feature-level methods are huge.

Table 1
The success rates (%) of standalone feature-level attacks against four black-box normally trained models.

(a) Performances of attacking benign model D161 (%)

Model	Attacker	D161	I3	IR2	R152	Avg
D161	NRD (Naseer et al., 2018)	98.9*	68.4	57.9	61.8	71.8
	ILA (Huang et al., 2019)	99.1*	69.7	62.6	67.9	74.8
	FIA (Wang et al., 2021)	98.9*	73.4	65.9	64.4	75.7
	DEFEAT	**99.8***	**75.8**	**66.8**	**66.6**	**77.3**

(b) Performances of attacking benign model I3 (%)

Model	Attacker	D161	I3	IR2	R152	Avg
I3	NRD (Naseer et al., 2018)	43.8	99.6*	62.4	47.7	63.4
	ILA (Huang et al., 2019)	31.8	98.9*	45.9	40.5	54.3
	FIA (Wang et al., 2021)	54.8	99.6*	67.3	56.7	69.6
	DEFEAT	**75.2**	**100.0***	**78.0**	**56.2**	**77.4**

(c) Performances of attacking benign model IR2 (%)

Model	Attacker	D161	I3	IR2	R152	Avg
IR2	NRD (Naseer et al., 2018)	50.6	79.9	96.8*	60.2	71.9
	ILA (Huang et al., 2019)	42.6	68.4	98.0*	52.3	65.3
	FIA (Wang et al., 2021)	54.6	79.3	90.8*	68.2	73.2
	DEFEAT	**85.0**	**91.6**	**100.0***	**70.4**	**86.8**

(d) Performances of attacking benign model R152 (%)

Model	Attacker	D161	I3	IR2	R152	Avg
R152	NRD (Naseer et al., 2018)	59.4	69.0	59.3	98.4*	71.5
	ILA (Huang et al., 2019)	52.0	63.4	54.2	99.0*	67.2
	FIA (Wang et al., 2021)	67.4	72.0	64.3	**99.4***	75.8
	DEFEAT	**85.0**	**72.0**	**65.6**	83.6*	**76.6**

That is to say, transferability also has shown its sensitivity to the attacking mechanism. Therefore, it is difficult to search for a universal optimal layer depth to be attacked for all feature-level attacks.

In summary, these intriguing properties suggest we find the optimal layer based on pairwise white-box models and then generate adversarial examples to fool remote black-box DNNs regardless of their architecture or parameters. Accordingly, we adopt the layers which achieve the best performance in the following comparisons, i.e., $D161_{l9}$, $I3_{l8}$, $IR2_{l8}$, $R152_{l8}$. The optimal layers for feature-level baselines are discussed in the supplemental material.

Comparison with Feature-level Baselines. This section reports the performance of the proposed DEFEAT and three state-of-the-art feature-level baselines, i.e., NRD (Naseer et al., 2018), ILA (Huang et al., 2019), and FIA (Wang et al., 2021). Concretely, the comparisons include two parts: (a) the transfer success rates under four normally trained models; and (b) the performance of attacks against eleven adversarial defenses. In addition, some adversarial examples are visualized in the supplementary.

(1) *The Result Under Benign Domains*. We use the crafted adversarial examples to attack normally trained models (i.e., D161, I3, IR2, and R152, respectively), and compare the success rates between baseline attacks and DEFEAT in the Table 1, where the symbol "$*$" indicates white-box attacks, and others results are conducted under black-box scenarios.

From Table 1, we observe that DEFEAT can achieve better results than baselines under white-box scenarios except for R152. The main reason is that attack methods exhibit different relations between model complexity and transferability. Form this perspective, DEFEAT has to approximate the adversarial distribution thereby it may perform worse to find global optimum from deeper models with large search space (i.e., R152), as discussed in Wang et al. (2021). Moreover, we pay more attention to the black-box experiments that our proposed DEFEAT consistently outperform the baselines by 5%∼20% for all normally trained models under black-box settings (e.g., the average improvements of DEFEAT are 5.5%, 14.0%, 14.9% and 5.1% for D161, I3, IR2 and R152 compared with NRD, respectively). It is interesting that the lower performance for white-box models does not necessarily mean worse transferability for cheating black-box models, as exemplified in R152. In summary, the experiment results confirm the advancement of DEFEAT for benign domains under both white-box and black-box setups.

(2) *The Result Under Adversarial Domains*. The generated adversarial examples are used to attacks against a total of eleven black-box defenses. The performances of various feature-level attacks are reported in the Table 2, which manifest some essential properties of feature-level attacks. On the one side, we observe that two arts feature-level attacks (i.e., NRD and ILA) can achieve high success rates for fooling normally trained models, as illustrated in Table 1. On the other side, NRD and ILA also demonstrate their ineffectiveness to circumvent adversarially trained models that their transferability significantly degrade under defense scenarios (e.g., For NRD: the decreased performance are 47.8%, 53.3%, 55.8%, and 54.4% for D161, I3, IR2, and R152, respectively). These quantitative results confirm the domain-overfitting dilemma of prior feature-level attacks, as discussed in Section 3.2.

The proposed DEFEAT yields much superior performance than current feature-level baselines under all black-box defense scenarios by mitigating the domain-overfitting effect. For example, DEFEAT achieves an average success rate of 47.8% by attacking IR2, outperforming baselines NRD, ILA, and FIA by a large margin of 31.7%, 30.2% and 14.5%, respectively. More significantly, we observe that *standalone* DEFEAT also performs better transferability across defenders than state-of-the-art gradient-based attacks DIM and TIM even they are selected as *combination* versions (see Table 5). Therefore, we establish state-of-the-art for feature-level method against normally trained models and robust defense mechanisms.

The Influence of Two-Stage Paradigm. Recall that DEFEAT is designed as a two-stage framework, which is different from most of the current adversarial attacks. For example, both of feature-level NRD, ILA, FIA, and gradient-based DIM and TIM are one-stage methods that perform optimization and update perturbations

Table 2
The success rates (%) of standalone feature-level attacks against eleven black-box defense models.

Model	Attacker	$I3_{ens3}$	$I3_{ens4}$	$IR2_{ens}$	HGD	BDR	PDT	R&P	COM	RS	SFR	NRP	Avg
D161	NRD (Naseer et al., 2018)	26.8	24.6	10.7	27.6	16.8	33.5	30.2	37.2	12.2	35.2	7.8	23.9
	ILA (Huang et al., 2019)	36.0	35.5	24.0	37.0	18.8	30.1	39.5	34.5	9.3	39.1	9.9	28.5
	FIA (Wang et al., 2021)	42.1	40.9	29.5	39.7	26.1	45.7	40.9	48.5	18.5	47.1	15.2	35.8
	DEFEAT	**55.0**	**53.0**	**36.6**	**56.4**	**45.0**	**59.8**	**56.2**	**62.4**	**39.2**	**62.3**	**22.2**	**49.8**
I3	NRD (Naseer et al., 2018)	6.3	10.0	1.3	5.4	13.2	19.6	7.9	18.3	4.9	17.7	6.4	10.1
	ILA (Huang et al., 2019)	9.2	10.5	2.5	9.4	12.2	16.6	10.1	13.2	2.9	19.1	6.7	10.2
	FIA (Wang et al., 2021)	29.8	30.5	14.4	27.9	23.7	35.1	28.4	39.8	15.4	36.4	**14.8**	26.9
	DEFEAT	**37.8**	**36.4**	**18.8**	**38.0**	**33.8**	**39.9**	**37.0**	**44.2**	**15.8**	**43.1**	13.6	**32.6**
IR2	NRD (Naseer et al., 2018)	19.5	14.7	4.6	17.5	14.5	21.6	20.1	24.7	4.6	30.3	5.5	16.1
	ILA (Huang et al., 2019)	20.2	20.3	11.7	20.3	16.3	24.0	22.1	19.8	4.2	28.3	6.7	17.6
	FIA (Wang et al., 2021)	36.8	31.9	29.3	38.2	30.1	39.5	38.5	41.9	19.5	46.3	14.8	33.3
	DEFEAT	**55.0**	**51.0**	**38.2**	**56.2**	**41.1**	**62.3**	**48.0**	**68.2**	**23.6**	**63.6**	**19.0**	**47.8**
R152	NRD (Naseer et al., 2018)	16.8	16.8	5.0	19.1	14.6	26.8	19.8	25.1	6.3	30.9	6.5	17.1
	ILA (Huang et al., 2019)	19.5	19.0	10.7	19.9	15.9	21.8	21.3	16.6	4.3	26.6	7.3	16.6
	FIA (Wang et al., 2021)	33.4	33.3	22.0	36.8	29.3	43.8	36.1	42.8	21.9	40.3	15.5	32.3
	DEFEAT	**50.6**	**48.0**	**35.4**	**52.2**	**44.0**	**56.1**	**47.8**	**59.7**	**44.8**	**57.4**	**22.8**	**47.2**

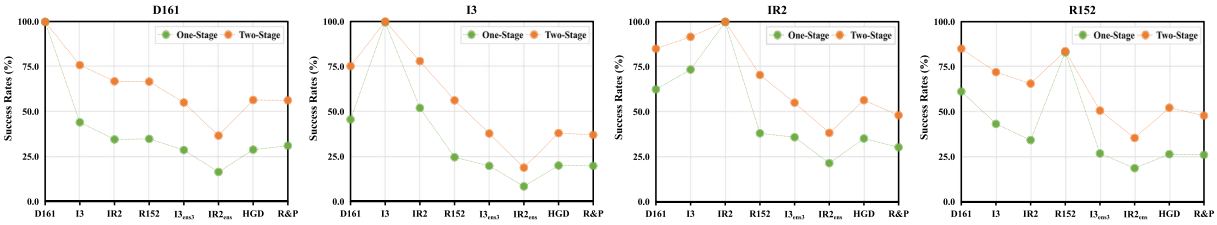


Fig. 4. The performance of different stage strategies. We find that two-stage policy can largely reduce the overfitting effect, thus achieves better results.

in a round-robin manner. However, we speculate this type of attack manner may induce the overfitting effect, thus lower the transferability. To overcome this shortcoming, we decouple the adversarial optimization and generation process: DEFEAT aims to search the parameters of adversarial distribution at the first stage and then generates transferable adversarial examples at the second stage.

To verify the efficacy of the two-stage paradigm, we follow prior one-stage methods (Huang et al., 2019; Naseer et al., 2018) to choose the best adversarial example in the first stage as the final solution, i.e., the optimal x' at the 40-steps learning stage corresponding to the maximum loss in Eq. (2), namely one-stage DEFEAT. By comparison, we select the adversarial results produced at the generation stage as two-stage DEFEAT. A total of eight black-box models are tested in the evaluations, including normally trained D161, I3, IR2, R152 (benign domains), and four defense models $I3_{ens3}$, $IR2_{ens}$, HGD, and R&P (adversarial domains).

As can be seen from Fig. 4, one-stage DEFEAT presents similar trends under all white-box models. Particularly, it achieves comparable performance with two-stage DEFEAT under white-box scenarios. However, its transferability is noticeably inferior to the two-stage DEFEAT when fooling other black-box models, including normally trained networks and defenses. This phenomenon provides evidence that one-stage DEFEAT usually suffers from the overfitting, and using the two-stage paradigm can largely mitigate this negative effect.

The Effect of Optimization Strategies. We provide quantitative studies about each optimization strategy adopted in the DEFEAT framework, including perturbation augmentation (PA), noise smoothing (NS), and neighbor interpolation (NI). Specifically, we consider nine black-box models in the evaluations, including four normally trained models and five defenses (i.e., $I3_{ens3}$, $IR2_{ens}$, HGD, and R&P). The attacks are crafted on I3 and IR2 by following settings (see Table 3): **(1)** Basic feature distortion, i.e., merely distorting the representations of the given model based on the Eq. (5) (1st row); **(2)** Single Strategy, i.e., combining the basic feature distortion with each strategy. Notice, introducing PA means to learn the distribution instead of generating a single optimal perturbation (2nd row); applying NS or NI indicates that calculating the gradient at lower dimensional spaces or interpolating the input with other clean images in optimizations (3rd $\sim$ 4th rows); **(3)** Double Strategies, i.e., removing a single optimization strategy from the proposed DEFEAT framework (5th $\sim$ 7th rows); **(4)** Full version of DEFEAT framework (8th row).

Table 3 reveals several intriguing conclusions. First, we observe that each optimization strategy improves the transferability upon the basic feature distortion process against black-box models. Generally, the more adopted strategies, the higher performances. In particular, the domain-overfitting effect is remarkably reduced when a total of three strategies are applied in attacks, i.e., the average improvements of success rates are 28.8%, 34.2%, 38.8%, and 22.5% for D161, I3, IR2, and R152 compared with the basic attack manner. Second, these optimization strategies exhibit domain biases. As illustrated in Single Strategy setups, PA facilitates feature-level attacks for both benign and adversarial domains (2nd rows). By contrast, NS shows its efficacy to enhance the attacking ability across adversarial domains, while it slightly degrades the transferability in few benign domains (e.g., see 3rd rows in IR2). Moreover, NI substantially boosts the transferability within benign domains regardless of the attacked white-box models (4th rows). It is mainly because NI samples various clean data from identical distribution for interpolation and produces perturbations with a large diversity. Third, simultaneously using PA and NS remarkably boosts the strength for transferring to adversarial domains, even outperforms the sum improvements of single PA and NS (see 5th row). It confirms that *there indeed exists an adversarial space in the lower dimensional space for every input image*, which fills with the adversarial examples with high transferability across from benign to adversarial domains.

Table 3
Experiments For Each Transferable Transform, including perturbation augmentation (PA), Noise Smoothing (NS), and neighbor interpolation (NI).

Model	PA	NS	NI	D161	I3	IR2	R152	$I3_{ens3}$	$IR2_{ens}$	HGD	BDR	R&P	Avg
I3				30.0	99.8	35.0	20.0	6.0	1.6	3.2	15.4	6.0	24.1
	✓			37.0	100.0	51.2	30.6	6.8	2.0	4.4	16.0	6.8	28.3
		✓		49.4	99.0	36.2	26.8	9.2	2.4	4.4	17.0	10.6	28.3
			✓	45.0	99.8	62.0	34.8	6.2	1.6	3.4	16.2	6.4	30.6
	✓	✓		68.2	100.0	65.2	49.0	33.2	15.4	32.2	25.4	31.2	46.6
	✓		✓	50.0	100.0	70.0	46.0	7.5	2.4	4.6	14.8	6.6	33.5
		✓	✓	61.4	99.8	53.8	37.0	10.6	4.4	6.0	16.8	13.0	33.6
	✓	✓	✓	**75.2**	**100.0**	**78.0**	**56.2**	**37.8**	**18.8**	**38.0**	**33.8**	**37.0**	**52.8**
IR2				45.0	73.6	99.6	39.2	5.2	1.2	1.6	14.8	7.2	31.9
	✓			57.2	84.4	99.6	55.0	8.4	2.8	4.8	16.2	8.2	37.4
		✓		64.8	71.8	99.6	43.0	15.2	5.0	5.6	21.8	16.4	38.1
			✓	64.2	91.2	100.0	60.0	6.4	1.4	2.6	16.4	8.4	39.0
	✓	✓		78.6	88.4	99.4	64.8	49.2	32.1	49.4	34.4	42.2	59.8
	✓		✓	70.4	90.6	99.8	68.4	12.2	6.2	10.7	17.0	11.2	42.9
		✓	✓	76.2	82.6	99.8	55.2	18.0	9.2	11.4	25.0	19.1	44.1
	✓	✓	✓	**85.0**	**91.6**	**100.0**	**70.4**	**55.0**	**38.2**	**56.2**	**41.1**	**48.0**	**65.1**

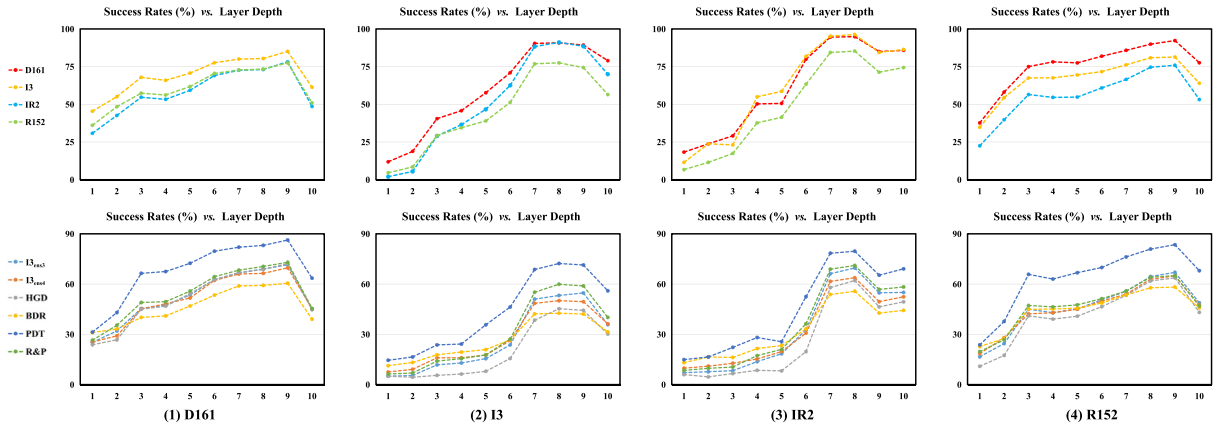


Fig. 5. The success rates (%) *vs.* layer depth of combination **DI-DEFEAT** against four normally trained models (top row) and six defenses (bottom row).

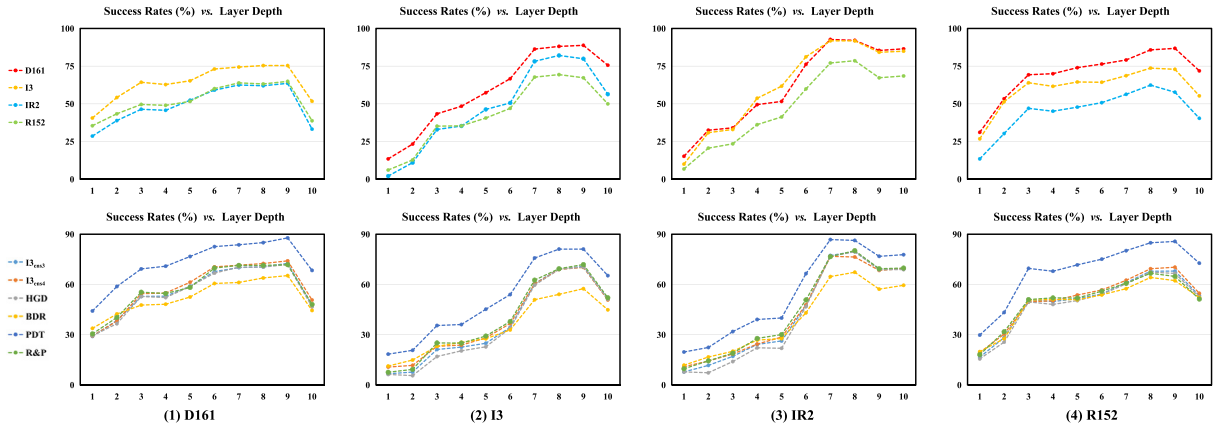


Fig. 6. The success rates (%) *vs.* layer depth of combination **TI-DEFEAT** against four normally trained models (top row) and six defenses (bottom row).

4.3. Experiments under combination scenarios

As mentioned in Section 3.2, we can combine gradient-based methods into the DEFEAT framework to form stronger variants. In this paper, we consider two variant attacks: (1) DI-DEFEAT, which introduces diversity transformations (Xie et al., 2019) into DEFEAT; and (2) TI-DEFEAT, which integrates DEFEAT with diversity transformations and translation invariance (Dong et al., 2019).

The Transferability of Combination Variants. Following the same settings in Section 4.2, we perturb a total of 10 layers to generate adversarial examples by using combination versions DI-DEFEAT and TI-DEFEAT. We report their performance in Figs. 5 and 6, separately. Similarly, the targeted black-box models include four normally trained networks (top row) and six robust models (bottom row). Two interesting behaviors that catch our attention.

The first one is that the transferability tendencies exhibit the transformation-agnostic property. We can see that the trends of success rates are surprisingly similar across the DEFEAT framework and its combination methods DI-DEFEAT and TI-DEFEAT. It indicates that two additional combined methods do not change

Table 4
The success rates (%) of combination attacks against black-box normally trained models.

(a) Performances for undefended models (%)

Model	Attacker	D161	I3	IR2	R152	Avg
D161	DIM (Xie et al., 2019)	98.8	72.3	66.4	69.0	76.6
	DI-NRD (Naseer et al., 2018)	96.9	74.3	66.8	69.7	76.9
	DI-ILA (Huang et al., 2019)	99.5	80.5	75.6	**78.4**	83.5
	DI-FIA (Wang et al., 2021)	98.9	85.7	72.5	70.8	82.0
	DI-DEFEAT	99.7	**85.0**	**78.2**	**77.3**	**85.1**
	TIM (Dong et al., 2019)	98.5	67.3	59.1	59.0	71.0
	TI-NRD (Naseer et al., 2018)	92.2	54.7	40.5	47.6	58.8
	TI-ILA (Huang et al., 2019)	99.3	75.5	70.7	73.1	79.7
	TI-FIA (Wang et al., 2021)	98.4	71.4	64.5	64.1	74.6
	TI-DEFEAT	**99.9**	75.2	63.5	64.8	75.9

(b) Performances for undefended models (%)

Model	Attacker	D161	I3	IR2	R152	Avg
I3	DIM (Xie et al., 2019)	45.1	98.7	54.9	41.6	60.1
	DI-NRD (Naseer et al., 2018)	71.3	99.2	77.7	68.4	79.2
	DI-ILA (Huang et al., 2019)	45.8	98.7	59.4	52.3	64.1
	DI-FIA (Wang et al., 2021)	63.9	99.8	78.0	62.4	76.0
	DI-DEFEAT	**90.7**	99.5	**91.0**	**77.5**	**89.7**
	TIM (Dong et al., 2019)	51.5	98.7	45.8	37.1	58.3
	TI-NRD (Naseer et al., 2018)	65.6	95.5	55.2	45.5	65.5
	TI-ILA (Huang et al., 2019)	46.8	97.6	48.6	39.4	58.1
	TI-FIA (Wang et al., 2021)	56.5	99.6	65.3	58.7	70.0
	TI-DEFEAT	88.0	**100.0**	81.8	69.2	84.8

(c) Performances for undefended models (%)

Model	Attacker	D161	I3	IR2	R152	Avg
IR2	DIM (Xie et al., 2019)	55.8	72.2	98.5	53.4	70.0
	DI-NRD (Naseer et al., 2018)	58.0	78.3	88.1	60.2	71.2
	DI-ILA (Huang et al., 2019)	59.5	79.7	96.8	64.3	75.1
	DI-FIA (Wang et al., 2021)	60.1	87.2	91.9	74.2	78.3
	DI-DEFEAT	**94.8**	**96.1**	99.5	**85.1**	**94.0**
	TIM (Dong et al., 2019)	62.1	63.9	96.4	50.5	68.2
	TI-NRD (Naseer et al., 2018)	56.6	59.6	64.1	42.2	55.6
	TI-ILA (Huang et al., 2019)	56.9	70.5	94.2	54.4	69.0
	TI-FIA (Wang et al., 2021)	60.4	76.3	89.8	69.1	74.2
	TI-DEFEAT	92.0	91.5	**99.8**	78.4	90.4

(d) Performances for undefended models (%)

Model	Attacker	D161	I3	IR2	R152	Avg
R152	DIM (Xie et al., 2019)	73.0	77.7	73.1	79.4	75.8
	DI-NRD (Naseer et al., 2018)	82.0	**85.0**	**77.8**	97.0	**85.5**
	DI-ILA (Huang et al., 2019)	71.9	79.3	74.4	**99.3**	81.2
	DI-FIA (Wang et al., 2021)	75.1	79.2	70.7	**99.5**	81.1
	DI-DEFEAT	**89.8**	80.7	74.4	87.1	83.0
	TIM (Dong et al., 2019)	75.2	65.3	62.9	98.5	75.5
	TI-NRD (Naseer et al., 2018)	72.5	64.0	50.7	87.1	68.6
	TI-ILA (Huang et al., 2019)	67.9	74.2	69.3	99.2	77.7
	TI-FIA (Wang et al., 2021)	64.4	71.6	64.5	99.1	74.9
	TI-DEFEAT	85.6	73.6	62.3	81.3	75.7

the transferability tendency of the basic DEFEAT. It is a favorable character that we can easily extend the transferability trends from DEFEAT to its improved versions once the optimal layer depth of any white-box model is discovered. To this end, we adopt the same layers for DI-DEFEAT and TI-DEFEAT in the following testing, i.e., $D161_{l9}$, $I3_{l8}$, $IR2_{l8}$, $R152_{l8}$, respectively.

Moreover, though both diversity transformations (Xie et al., 2019) and translation invariance (Dong et al., 2019) boost the transferability under all black-box models, they also show their different domain bias properties, similar to the proposed noise smoothing and neighbor interpolation (see Table 3). Specifically, diverse input transformations demonstrate more power to enhance adversarial examples across benign domains, i.e., the performance of DI-DEFEAT is better than TI-DEFEAT to fool normally trained models (see top row). Conversely, translation invariance noticeably improves the strength of DEFEAT across adversarial domains, especially when attacking I3 and IR2 to craft adversarial examples against black-box defenses, as shown in Fig. 6 (2) & (3).

Comparison with Combination Baselines. To fairly compare the transferability between gradient-based and feature-level methods, we similarly introduce DIM and TIM into NRD, ILA and FIA, denoted as DI-NRD, TI-NRD, DI-ILA, TI-ILA, DI-FIA, and TI-FIA, respectively. Total of ten attacks are included in the tests (see Tables 4 & 5):

(1) Gradient-based method DIM, and four combined feature-level attacks who applied diversity transformations, i.e., DI-NRD, DI-ILA, DI-FIA and the proposed DI-DEFEAT.

(2) Gradient-based method TIM, and feature-level variants who adopt translation invariance, i.e., TI-NRD, TI-ILA, TI-FIA and the proposed TI-DEFEAT.

Table 5
The success rates (%) of combination attacks against black-box defense models.

Model	Attacker	$I3_{ens3}$	$I3_{ens4}$	$IR2_{ens}$	HGD	BDR	PDT	R&P	COM	RS	SFR	NRP	Avg
D161	DIM (Xie et al., 2019)	41.6	39.6	23.5	44.1	27.6	41.5	48.7	49.1	23.6	47.3	11.3	36.2
	DI-NRD (Naseer et al., 2018)	44.0	43.7	26.2	44.9	26.4	45.4	47.4	52.0	23.6	53.3	10.8	38.0
	DI-ILA (Huang et al., 2019)	52.1	49.7	36.2	52.7	25.0	37.4	53.4	47.3	13.2	54.3	9.9	39.2
	DI-FIA (Wang et al., 2021)	48.9	47.6	34.4	46.2	30.6	53.3	47.6	56.3	30.2	58.4	22.9	43.3
	DI-DEFEAT	71.5	69.7	52.5	**72.2**	60.5	86.3	**73.0**	**87.0**	63.4	80.6	28.3	67.7
	TIM (Dong et al., 2019)	53.2	54.9	44.5	53.1	36.5	51.4	56.5	55.8	44.1	55.2	23.6	48.1
	TI-NRD (Naseer et al., 2018)	45.8	50.6	39.6	45.9	39.6	50.7	48.1	54.2	50.7	49.7	23.5	45.3
	TI-ILA (Huang et al., 2019)	51.9	49.2	34.8	52.2	25.9	39.1	53.5	50.2	18.7	55.7	11.4	40.2
	TI-FIA (Wang et al., 2021)	56.6	55.1	41.9	53.6	46.6	64.1	54.1	63.6	34.1	62.5	31.1	51.2
	TI-DEFEAT	**72.6**	**74.1**	**60.0**	71.9	**65.3**	**87.8**	71.8	82.8	**71.1**	**82.3**	**47.3**	**71.5**
I3	DIM (Xie et al., 2019)	15.7	17.4	6.4	15.8	16.2	25.2	18.0	22.1	7.7	21.9	9.9	16.0
	DI-NRD (Naseer et al., 2018)	21.0	18.9	3.4	17.9	14.9	28.9	26.0	42.4	6.5	38.7	6.5	20.5
	DI-ILA (Huang et al., 2019)	12.3	14.9	3.9	12.0	14.1	19.1	15.2	16.2	3.0	16.9	6.6	12.2
	DI-FIA (Wang et al., 2021)	33.1	34.1	22.8	30.9	28.5	39.1	31.7	44.8	23.9	44.2	23.0	32.4
	DI-DEFEAT	54.7	49.5	26.6	44.3	42.1	71.4	58.9	79.8	35.9	70.4	16.1	50.0
	TIM (Dong et al., 2019)	36.7	39.7	26.4	37.2	26.2	40.3	37.1	39.2	24.9	39.4	16.8	33.1
	TI-NRD (Naseer et al., 2018)	45.3	43.7	28.7	45.6	27.7	50.2	46.2	52.7	22.6	51.4	11.8	38.7
	TI-ILA (Huang et al., 2019)	12.1	16.4	5.4	12.3	14.4	23.8	17.2	20.8	6.8	20.1	7.3	14.2
	TI-FIA (Wang et al., 2021)	45.0	47.8	37.7	45.6	41.8	53.6	43.3	51.5	32.7	52.4	27.7	43.6
	TI-DEFEAT	**70.8**	**70.2**	**52.8**	**70.4**	**57.5**	**81.1**	**71.9**	**80.1**	**56.1**	**80.3**	**31.7**	**65.7**
IR2	DIM (Xie et al., 2019)	29.3	26.9	16.2	31.0	20.5	31.5	34.1	30.5	11.5	34.0	14.8	25.5
	DI-NRD (Naseer et al., 2018)	30.8	25.4	11.2	28.9	17.8	27.0	34.1	37.6	5.6	44.4	6.4	24.5
	DI-ILA (Huang et al., 2019)	28.4	24.7	18.9	29.2	16.4	23.6	31.1	23.6	4.5	35.6	7.1	22.1
	DI-FIA (Wang et al., 2021)	36.1	30.6	27.5	37.9	30.8	39.6	38.4	42.2	23.5	50.7	22.8	34.6
	DI-DEFEAT	69.6	63.8	48.8	62.1	55.7	79.6	71.1	**88.3**	43.5	81.4	20.9	62.3
	TIM (Dong et al., 2019)	47.3	45.4	44.4	47.4	31.7	44.5	47.3	47.4	27.9	50.1	18.9	41.1
	TI-NRD (Naseer et al., 2018)	38.0	35.8	25.1	37.3	29.6	46.1	39.3	45.5	25.8	46.5	11.3	34.6
	TI-ILA (Huang et al., 2019)	31.0	28.5	21.1	32.9	19.8	30.1	33.3	31.9	9.8	38.3	8.1	25.9
	TI-FIA (Wang et al., 2021)	45.5	44.4	40.1	44.2	36.6	53.8	44.6	52.8	34.6	52.1	28.4	43.4
	TI-DEFEAT	**79.7**	**76.4**	**73.5**	**79.5**	**67.3**	**86.3**	**80.2**	87.5	**61.8**	**86.4**	**44.1**	**74.8**
R152	DIM (Xie et al., 2019)	37.7	35.3	20.0	39.5	25.2	36.9	43.3	37.4	14.9	45.1	14.7	31.8
	DI-NRD (Naseer et al., 2018)	39.5	32.7	16.7	40.2	21.0	39.2	43.9	46.0	9.9	50.6	7.3	31.5
	DI-ILA (Huang et al., 2019)	32.9	30.5	21.0	34.4	17.9	25.9	36.8	22.6	5.7	34.3	8.6	24.6
	DI-FIA (Wang et al., 2021)	40.0	40.1	26.9	44.0	37.7	44.2	43.4	51.1	33.3	50.8	24.1	39.6
	DI-DEFEAT	64.5	62.9	49.0	61.9	57.8	80.9	64.1	**84.0**	55.0	79.4	24.7	62.2
	TIM (Dong et al., 2019)	56.0	55.0	48.0	56.7	37.7	55.8	57.8	56.1	37.7	59.2	26.5	49.7
	TI-NRD (Naseer et al., 2018)	50.0	48.7	38.7	50.1	38.9	58.3	50.6	60.3	37.0	59.9	22.6	46.8
	TI-ILA (Huang et al., 2019)	33.9	33.7	22.3	35.4	20.4	27.9	38.0	28.6	7.8	38.6	10.1	27.0
	TI-FIA (Wang et al., 2021)	56.0	54.8	44.8	57.3	50.3	61.1	56.3	64.7	42.1	61.6	31.1	52.7
	TI-DEFEAT	**67.9**	**69.4**	**59.1**	**67.4**	**64.2**	**84.9**	**66.7**	81.3	**65.8**	**76.7**	**45.3**	**68.1**

To better illustrate, we separately report the DI-combined attacks and TI-combined methods in the top and bottom of each given white-box model in Tables 4 & 5. It is no surprise to observe that integrating two gradient-based methods (i.e., diversity transformations and translation invariance) into current feature-level attacks can also promote their transferability across benign and adversarial domains. For example, compared with the basic NRD, the average improvements of DI-NRD against defenses are 14.1%, 10.4%, 8.4%, 14.4% for D161, I3, IR2, and R152, respectively (see Tables 1 & 4). Moreover, these two strategies also impose their domain bias to feature attacks. For undefended models (i.e., benign domains), DI-based methods achieve higher fooling rates than the basic versions, while TI-based attacks slightly degrade their capabilities. Under adversarial domains (i.e., defense settings), all TI-combined versions demonstrate more substantial power than their counterparts. However, these improved feature-level attacks are still worse than gradient-based baselines in some cases, indicating their unstable generalizability. By making the comparisons with them, DI-DEFEAT and TI-DEFEAT further narrow down the performance gaps between two domains, thereby yielding a distinct improvement under benign and achieving the strongest performance under adversarial domains. These quantitative results validate the superior transferability of the proposed method.

4.4. Ensemble-model experiments

We provide the performance of ensemble attacks. The ensemble of four vanilla models (i.e., D161, I3, IR2, and R152) is attacked by different methods, including six basic attacks (i.e., NRD Naseer et al., 2018, ILA Huang et al., 2019, FIA Wang et al., 2021, DIM Xie et al., 2019, TIM (Dong et al., 2019), and DEFEAT) and eight combination variants. Specifically, gradient-based attacks perturb the fusion of logits activations (Dong et al., 2018), and feature-level methods disturb the optimal layers of multiple models simultaneously. We summarize the success rates of fourteen ensemble attacks in Table 6.

Among the basic versions, we find the feature-level NRD and ILA can improve the transferability by introducing ensemble policy, while they only consider the single-model setups in the literature (Huang et al., 2019; Inkawhich et al., 2019). However, these two feature-level attacks continue to perform worse than gradient-based DIM and TIM under ensemble-model scenarios due to the domain-overfitting effect. On the other hand, the basic DEFEAT consistently demonstrates much stronger transferability than baseline methods. It achieves an average success rate of 73.1% against defenses, outperforming the current state-of-the-art TIM and FIA by 2.8% and 11.9%. Therefore, DEFEAT can synchronously reduce the domain-overfitting for multiple models.

Table 6
The success rates (%) of ensemble attacks against defense models..

Model	Method	$I3_{ens3}$	$I3_{ens4}$	$IR2_{ens}$	HGD	BDR	PDT	R&P	COM	RS	SFR	NRP	Average
Basic versions	NRD (Naseer et al., 2018)	34.9	29.6	11.0	32.5	18.5	33.3	39.0	46.6	7.9	43.9	8.0	27.7
	ILA (Huang et al., 2019)	55.9	53.8	42.8	56.1	23.9	36.7	56.9	45.9	13.2	58.4	12.3	41.4
	FIA (Wang et al., 2021)	71.2	70.2	59.6	72.7	51.6	63.7	72.9	68.1	44.5	70.4	28.4	61.2
	DIM (Xie et al., 2019)	67.0	62.0	42.8	67.8	35.3	56.7	71.3	68.0	24.6	69.2	17.9	53.0
	TIM (Dong et al., 2019)	81.2	79.2	**76.6**	79.9	56.0	72.7	80.9	78.6	**51.5**	79.1	37.1	70.3
	DEFEAT	**81.5**	**81.2**	70.2	**83.3**	**67.8**	**84.2**	**81.0**	**84.1**	49.9	**83.7**	**37.7**	**73.1**
Combination versions	DI-NRD (Naseer et al., 2018)	56.2	50.1	26.1	56.7	30.3	48.9	60.1	61.7	12.7	66.9	9.0	43.5
	TI-NRD (Naseer et al., 2018)	60.8	57.7	46.4	60.2	46.4	63.9	59.6	66.3	43.3	68.4	25.0	54.4
	DI-ILA (Huang et al., 2019)	70.2	68.8	59.9	71.3	33.4	47.3	73.4	63.3	17.2	72.1	14.4	53.8
	TI-ILA (Huang et al., 2019)	72.3	68.6	59.4	72.3	36.8	49.6	73.1	64.1	22.4	73.4	18.0	55.5
	DI-FIA (Wang et al., 2021)	80.5	82.1	72.4	79.4	55.8	71.4	80.1	76.7	52.4	81.3	32.7	69.5
	TI-FIA (Wang et al., 2021)	87.0	86.1	78.4	86.4	68.2	80.3	87.1	84.0	59.8	85.7	42.6	76.9
	DI-DEFEAT	90.9	88.3	79.3	88.6	78.4	93.7	90.6	94.7	67.7	91.4	45.2	82.6
	TI-DEFEAT	**92.4**	**91.3**	**87.0**	**92.4**	**83.7**	**95.6**	**93.0**	**95.1**	**80.0**	**92.2**	**69.7**	**88.4**

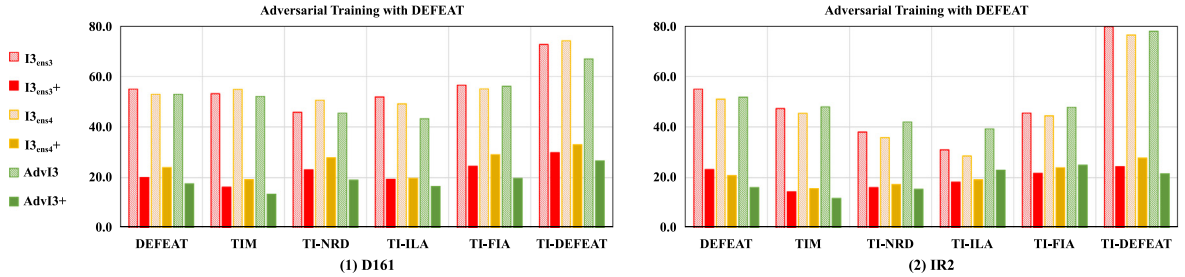


Fig. 7. The performance of adversarially trained models using DEFEAT. The adversarial examples are generated by using different methods to attack D161 and IR2 models, and we record the transferable success rates against pre-trained models (e.g., $I3_{ens3}$, $I3_{ens4}$, etc.) as well as fine-tuned defenses (e.g., $I3_{ens3}$+, $I3_{ens4}$+, etc.). The lower transferability indicates more adversarial examples can be classified correctly, implying higher robustness of trained models.

As for combination variants, it is natural to summarize the higher fooling rates compared with their basic versions. For example, DI-NRD and TI-NRD have average success rates of 43.5% and 54.4%, which exceed the basic NRD by 15.8% and 26.7%, respectively. On average, our proposed two combinations provide superior capabilities to all baselines, among which TI-DEFEAT achieves the highest average result of 88.4% and surpasses the strongest baseline TI-FIA by a large margin of 11.5%, which not only verifies the effectiveness and generalizability of our proposed DEFEAT, but also remind the security issues of the current defenses.

4.5. Adversarial training experiment

To validate the efficacy of DEFEAT for training the robust defenses, we use DEFEAT to generate adversarial examples and augment the training data during the model optimization. Since it requires high computational costs to train from scratch, we introduce the adversarial fine-tuning strategy to optimize the pre-trained robustified models (e.g., $I3_{ens3}$, $I3_{ens4}$, and AdvI3) (Tramèr et al., 2018). Concretely, we attack the images from ImageNet test dataset to generate their corresponding adversarial examples by using the proposed DEFEAT and DI-DEFEAT on the vanilla I3 model. Then the produced images are treated as additional data to fine-tune the weights for 10 epochs, and the new obtained models are denoted as $I3_{ens3}$+, $I3_{ens4}$+, and AdvI3+, respectively. Without loss of generality, we select six strong attackers to generate malicious images for testing the robustness of fine-tuned models. The experimental setup is the same as Section 4.3 and we illustrate the comparison results in Fig. 7, where the primary axes are attack methods (e.g., DEFEAT, TIM, etc.) and their transferable success rates (the lower the better). We can observe that the robustness of newly trained models is largely boosted by optimizing over data generated from DEFEAT method. Compared with the performance of pre-trained adversarial models (dashed bar), the transferability of each attacker against our fine-tuned models becomes lower (solid bar). For instance, the fooling rates of TI-DEFEAT degenerate from 72.6%/74.1%/66.9% to 29.9%/33.0%/26.6% on $I3_{ens3}$+, $I3_{ens4}$+, and AdvI3+, respectively (Fig. 7 (1)). Moreover, it is surprising that the models trained with DEFEAT generation also demonstrate higher robustness against other attacks, including TIM, TI-NRD, TI-ILA, etc. These results verify the effectiveness and generalizability of the proposed DEFEAT in the adversarial training paradigm.

5. Further analysis of transferability

Feature Similarity. One line to study transferability is analyzing the variation of representations between the input image and its adversarial example. Concretely, we explore the correlation as an indication to learn their intrinsic properties. For evaluating the feature similarity, we randomly choose 1000 images from ImageNet validation dataset, and average the cosine of features from layer 1 to 10 (column-wise) based on clean images and their adversarial counterparts generated at different depths (row-wise). The similarity matrix $\mathcal{M}$ of four source models are recorded in Fig. 8, where $\mathcal{M}_{ij}$ indicates the average cosine between jth features extracted from original x and its adversarial x'_{Li} generated by perturbing layer i.

As row-wise observed from the similarity matrices (Fig. 8), we can roughly classify four white-box networks as two types: (1) **Mutational models**, which keep the adjacent cosine values in stable at shallow layers and decrease sharply at specific deep layers, like I3 and IR2; (2) **Gradational models**, which progressively decrease the pairwise cosines from shallow to deep representations, including D161 and R152.

These matrices provide some hints about the trendlines of transferability. For mutational models (*i.e.,* I3 and IR2), the adversarial examples constructed from shallow layers ($x'_{L1} \sim x'_{L5}$)

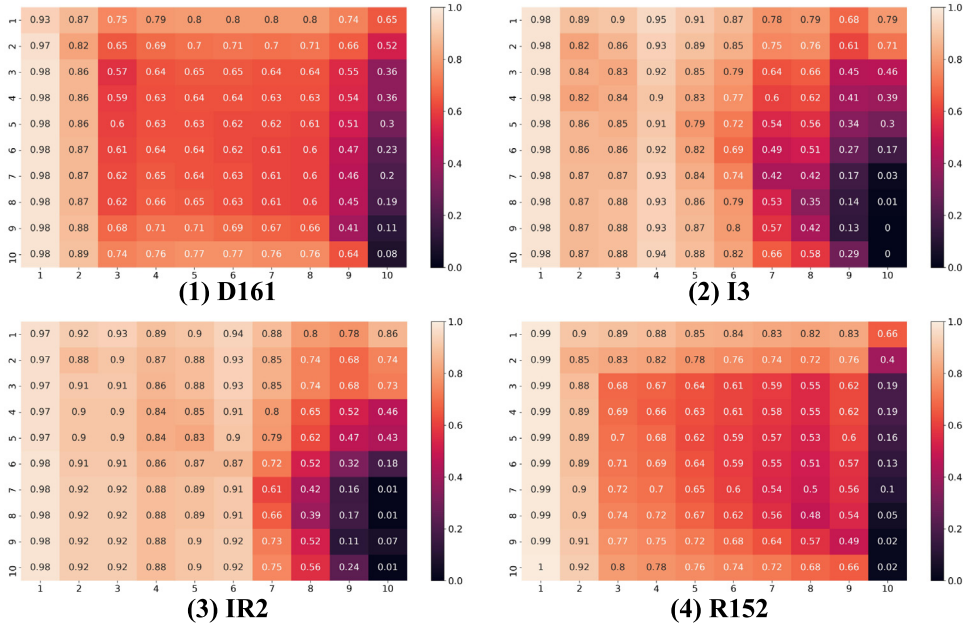


Fig. 8. The correlation matrices of features between adversarial examples and clean images. The darker the color, the less similarity. Each entry shows the cosine of latent representations at column-wise layers between the clean images and the adversarial examples which generated at the row-wise layers.

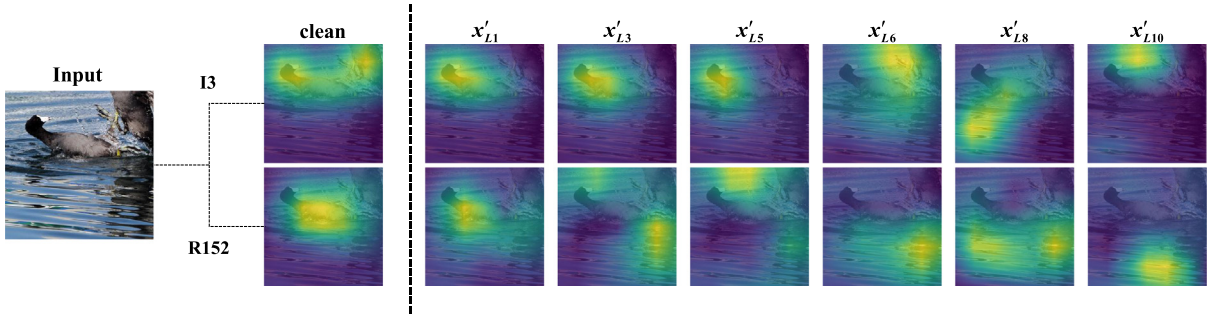


Fig. 9. Visualization of the highlighted discriminative regions on I3 (top) and R152 (bottom). The adversarial images constructed by perturbing the shallow features ($L1, L3, L5$) as well as the deep features ($L6, L8, L10$).

exhibit highly-correlated with their input data, while the deeper instances ($x'_{L6} \sim x'_{L9}$) demonstrate huge variant with the clean images (see bottom right corners). This observation obviously can be related to the curve shapes of I3 and IR2 in Fig. 3 that the transferability is lower before layer $L5$ and then dramatically increases from layer $L6$ to $L9$, like the inverse of correlation. As for gradational models D161 and R152, we find the similarities at the first two layers are high and continue to decrease at deep layers with a slow and steady rate. This tendency is also converse to the curve shape in Fig. 3 that the transfer rates increase shapely from $L1$ to $L3$, and then grow slowly until the layer depth exceeds $L9$. In short, lower correlation indicates the greater dissimilarity between clean data and adversarial ones, thus induces higher transferability.

Attention Mechanism. We provide more qualitative investigations about transferability and attention maps. Given a clean image and a series of adversarial examples generated by attacking different latent layers, we use vanilla model I3 (top row) and R152 (bottom row) as an example to visualize the discriminative regions by using grad-CAM (Selvaraju et al., 2017) in Fig. 9 (more examples are demonstrated in supplemental material). Concretely, the visualization includes the adversarial images generated by disrupting shallow features ($L1, L3, L5$) and deep layers ($L6, L8, L10$). The clean image and its heatmaps are illustrated on the left side to reflect the changes of attentions.

We notice that the variation of attention maps also hints at the transferability trends (Fig. 3) and the correlation matrices (Fig. 8). For I3, it can be seen that the original image shares a similar attention area with shallow adversarial examples (*i.e.*, $x'_{L1}, x'_{L3}, x'_{L5}$), while a distinct different heatmap occurs after disrupting the features at layer 6. This observation further confirms the property of the mutational model that layer 6 is an inflection point that turns the high correlation to little similarity, thus improving the transferability at deeper layers. Similarly, the discriminative region of gradational model R152 progressively moves from the main content area to the background area if the perturbed features is located after 3 layer. Similar behaviors are observed in both transfer trendlines (Fig. 3) and correlation matrices (Fig. 8). Consequently, the variation of heatmaps not only implies the differences between mutational models and gradational models, but also provides a complementary explanation about transferability.

6. Conclusion

In this paper, we focus on the problem of black-box adversarial attacks. Especially, we find that the domain-overfitting effect may lower the transferability of feature-level adversarial examples. To improve their capabilities against both benign to adversarial domains, we propose a novel feature-level framework, referred to as **De**coupled **Fe**ature **At**tack (DEFEAT). Concretely, DEFEAT

decouples the classical one-stage procedure into two phases: it adopts explored optimization strategies to learn an adversarial distribution with relatively high losses in the first stage, and then samples the noises from learned space to construct adversarial perturbations in the second stage. Extensive experiments demonstrate that DEFEAT largely mitigates the domain-overfitting effect, thus achieving better transferability than both feature-level and gradient-based methods. It also indicates that the current defenses are not real security. Finally, we further provide insights into the relationship between transferability and internal representations, which may help the community enhance the robustness of deep learning models.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61972433) and Guangdong Basic and Applied Basic Research Foundation (2021A1515012242).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neunet.2022.09.009.

References

Andriushchenko, M., Croce, F., Flammarion, N., & Hein, M. (2020). Square attack: A query-efficient black-box adversarial attack via random search. In *European conference on computer vision* (pp. 484–501). Springer.

Bai, X., Yang, M., & Liu, Z. (2020). On the robustness of skeleton detection against adversarial attacks. *Neural Networks*, *132*, 416–427.

Borkar, T., Heide, F., & Karam, L. (2020). Defending against universal attacks through selective feature regeneration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 709–719).

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy* (pp. 39–57).

Chen, P. -Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. -J. (2017). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security* (pp. 15–26).

Cheng, M., Singh, S., Chen, P., Chen, P. -Y., Liu, S., & Hsieh, C. -J. (2020). Sign-opt: A query-efficient hard-label adversarial attack. In *International conference on learning representation*.

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In *International conference on machine learning* (pp. 1310–1320).

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., et al. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9185–9193).

Dong, Y., Pang, T., Su, H., & Zhu, J. (2019). Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4312–4321).

Finlay, C., Pooladian, A. -A., & Oberman, A. (2019). The logbarrier adversarial attack: Making effective use of decision boundary information. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4862–4870).

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Guo, C., Rana, M., Cisse, M., & van der Maaten, L. (2018). Countering adversarial images using input transformations. In *International conference on learning representations*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, L., Gao, C., Zhou, Y., Xie, C., Yuille, A. L., Zou, C., et al. (2020). Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 720–729).

Huang, Q., Katsman, I., He, H., Gu, Z., Belongie, S., & Lim, S. -N. (2019). Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4733–4742).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Huang, L., Wei, S., Gao, C., & Liu, N. (2022). Cyclical adversarial attack pierces black-box deep neural networks. *Pattern Recognition*, Article 108831.

Ilyas, A., Engstrom, L., Athalye, A., & Lin, J. (2018). Black-box adversarial attacks with limited queries and information. In *International conference on machine learning* (pp. 2137–2146). PMLR.

Ilyas, A., Engstrom, L., & Madry, A. (2019). Prior convictions: Black-box adversarial attacks with bandits and priors. In *International conference on learning representations*.

Inkawhich, N., Wen, W., Li, H. H., & Chen, Y. (2019). Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7066–7074).

Ito, R., Nakae, K., Hata, J., Okano, H., & Ishii, S. (2019). Semi-supervised deep learning of brain tissue segmentation. *Neural Networks*, *116*, 25–34.

Jia, X., Wei, X., Cao, X., & Foroosh, H. (2019). Comdefend: An efficient image compression model to defend adversarial examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6084–6092).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial machine learning at scale. In *International conference on learning representations*.

Li, Y., Bai, S., Zhou, Y., Xie, C., Zhang, Z., & Yuille, A. L. (2020). Learning transferable adversarial examples via ghost networks. In *AAAI* (pp. 11458–11465).

Li, L., Li, Z., Liu, Y., & Hong, Q. (2021). Deep joint learning for language recognition. *Neural Networks*, *141*, 72–86.

Li, Y., Li, L., Wang, L., Zhang, T., & Gong, B. (2019). Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In *International conference on machine learning* (pp. 3866–3876).

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., & Zhu, J. (2018). Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1778–1787).

Lin, Z., Jia, J., Huang, F., & Gao, W. (2022). Feature correlation-steered capsule network for object detection. *Neural Networks*, *147*, 25–41.

Liu, Y., Chen, X., Liu, C., & Song, D. (2017). Delving into transferable adversarial examples and black-box attacks. In *International conference on learning representations*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International conference on learning representations*.

Maimon, G., & Rokach, L. (2022). A universal adversarial policy for text classifiers. *Neural Networks*, *153*, 282–291.

Moosavi-Dezfooli, S. -M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1765–1773).

Mopuri, K. R., Ganeshan, A., & Babu, R. V. (2018). Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(10), 2452–2465.

Naseer, M., Khan, S., Hayat, M., Khan, F. S., & Porikli, F. (2020). A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 262–271).

Naseer, M., Khan, S. H., Rahman, S., & Porikli, F. (2018). Task-generalizable adversarial attack based on perceptual metric. arXiv preprint arXiv:1811.09020.

Pang, T., Xu, K., & Zhu, J. (2019). Mixup inference: Better exploiting mixup to defend adversarial attacks. In *International conference on learning representations*.

Prakash, A., Moran, N., Garber, S., DiLillo, A., & Storer, J. (2018). Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8571–8580).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations*.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence* (pp. 4278–4284).
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).
Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In *International conference on learning representations*.
Vidnerová, P., & Neruda, R. (2020). Vulnerability of classifiers to evolutionary generated adversarial examples. *Neural Networks*, *127*, 168–181.
Wang, Z., Guo, H., Zhang, Z., Liu, W., Qin, Z., & Ren, K. (2021). Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7639–7648).
Xie, C., Wang, J., Zhang, Z., Ren, Z., & Yuille, A. (2018). Mitigating adversarial effects through randomization. In *International conference on learning representations*.
Xie, C., & Yuille, A. (2019). Intriguing properties of adversarial training at scale. In *International conference on learning representations*.
Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., et al. (2019). Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2730–2739).
Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833).
Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. In *International conference on learning representations*.
Zhang, S., Huang, K., Zhu, J., & Liu, Y. (2021). Manifold adversarial training for supervised and semi-supervised learning. *Neural Networks*, *140*, 282–293.
Zhao, Z., Liu, Z., & Larson, M. (2021). On success and simplicity: A second look at transferable targeted attacks. *Advances in Neural Information Processing Systems*, *34*.
Zhou, W., Hou, X., Chen, Y., Tang, M., Huang, X., Gan, X., et al. (2018). Transferable adversarial perturbations. In *Proceedings of the european conference on computer vision* (pp. 452–467).

AUTE: Peer-Alignment and Self-Unlearning Boost Adversarial Robustness for Training Ensemble Models

Lifeng Huang[1], Tian Su[2], Chengying Gao[2], Ning Liu[2], Qiong Huang[1] *

[1]College of Mathematics and Informatics, South China Agricultural University
[2]School of Computer Science and Engineering, Sun Yat-sen University
{huanglf6, qhuang}@scau.edu.cn, {mcsgcy, liuning2}@mail.sysu.edu.cn

Abstract

Adversarial attacks poses a significant threat to the security of AI-based systems. To counteract these attacks, adversarial training (AT) and ensemble learning (EL) have emerged as widely adopted methods for enhancing model robustness. However, a counter-intuitive phenomenon arises where the simple combination of these approaches may potentially compromising adversarial robustness of ensemble models. In this paper, we propose a novel method called ***A**lignment and **U**nlearning for **T**raining **E**nsembles* (AUTE), aiming to effectively integrate AT and EL to maximize their benefits. Specifically, AUTE incorporates two key components. Firstly, AUTE divides the ensemble into a big peer model and a single member in a loop manner, aligning their outputs for boosting robustness of each member. Secondly, AUTE introduces the concept of unlearning, actively forgetting specific data with over-confident properties to preserve model capacity to learn more robust features. Extensive experiments across various datasets and networks illustrate that AUTE achieves superior performance compared to baselines. For instance, a 5-member AUTE with ResNet-20 networks outperforms state-of-the-art method by 2.1% and 3.2% in classifying clean and adversarial data. Additionally, AUTE can easily extend to non-adversarial training paradigm, surpassing current standard ensemble learning methods by a large margin.

Code — https://github.com/mesunhlf/AUTE

Introduction

Deep neural networks (DNNs) have been widely employed in essential systems, including classification (He et al. 2016), recognition (Qiao et al. 2021) and translation (Ouyang et al. 2022). Despite their excellent performance, DNNs are not robust to adversarial examples: adding human-imperceptible perturbations to the clean data can deceive DNNs into outputting unexpected predictions (Wu et al. 2020; Huang et al. 2020; Doan et al. 2022).

There has emerged a number of research on defenses. For example, adversarial training (AT) and ensemble learning (EL) are two promising methods to enhance adversarial robustness. Specially, AT trains DNNs on generated adversarial examples, while most of AT methods merely focus on the defensive capabilities of individual models (Zhang et al. 2019; Wang et al. 2019; Rade and Moosavi-Dezfooli 2021b; Qizhang Li 2023). In contrast, EL methods optimize multiple models on the clean data and then combine them together for joint predictions (Pang et al. 2019; Yang et al. 2020; Deng and Mu 2024; Zhuang et al. 2024; Huang et al. 2023b). In general, EL approaches exhibit desirable robustness against black-box attacks. Given the observations that EL is benefit to constitute stronger standard trained models, a natural question arises: *Can AT and EL be integrated compatibly to further boost accuracy and robustness?*

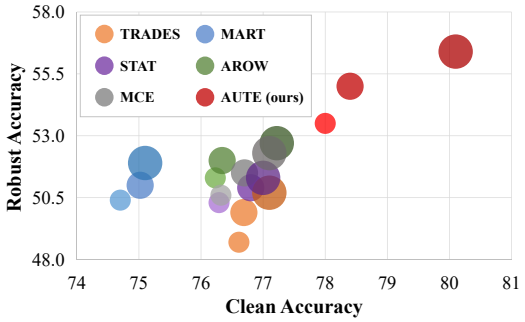


Figure 1: The comparison results between different ensemble models. The primary axes represent clean accuracy and adversarial robustness against the AA attack. The size of each circle (ensemble method) indicates the robustness level of ensemble models with 3, 5, and 8 members, ranging from small to large. Our proposed AUTE ensembles exhibit significantly better performance compared to existing methods.

Empirically, we observe a counter-intuitive phenomenon where the simple combination of AT and EL approaches sometimes offers marginal robustness improvements, while at other times, it inadvertently introduces side effects that can compromise the adversarial robustness. This phenomenon suggests that this simple combination may struggle to learn balanced features from both clean and adversarial data, which aligns the findings in related works (Xie and Yuille 2019). In light of these results, our forthcoming objective is to answer the question: *How can we maximize the benefits from EL for further boosting AT ensembles?*

*Corresponding author.

In this paper, we present a novel method aimed at establishing robust ensemble models, referred to as **A**lignment and **U**nlearning for **T**raining **E**nsembles (AUTE). AUTE consists of two crucial components: the Peer-Alignment scheme and the Self-Unlearning optimization, respectively.

- **Peer-Alignment**: Given that an ensemble of multiple AT models often demonstrates an improved defensive capacity than a single model, we partition the ensemble into two parts—a larger peer model characterized by higher robustness and a single member with weaker resistance. By aligning the robust features of the member with those of the peer model, its robustness is enhanced, and it is then integrated into the peer model to create a more robust ensemble. Since the alignment process between single member and its peer is carried out in a loop manner, the robustness of each ensemble member is gradually enhanced, leading to improved overall defensive capability from an ensemble learning perspective.
- **Self-Unlearning**: Most AT methods usually require network to capture robust features from all adversarial examples. However, it is challenging for models to entirely learn all adversarial data, which arises from the need for the consumption of a significant portion of the model capacity (Zhang et al. 2020). Particularly, we identify that certain adversarial examples are predicted with high confidence, resulting in over-confident predictions. This behavior can inadvertently have a negative impact on model performance (Müller, Kornblith, and Hinton 2019). Therefore, this category of data is designated for *forgetting* to free up model capacity and subsequently relearned by the AT model. As training progresses, the AT model can correctly classify more adversarial data, while simultaneously mitigating the over-confidence dilemma.

We conduct extensive experiments on various datasets and networks under a range of attack scenarios. The empirical results suggest that the proposed AUTE method perform signicantly higher clean accuracy as well as adversarial robustness compared to SOTA methods (see Fig. 1). Furthermore, AUTE exhibits well scalability, allowing it to further boost the performance as the ensemble size increases. For instance, training ResNet-20 networks on CIFAR-10 dataset, when AUTE trains an ensemble with 3 members, it improves average robustness from the SOTA 51.6% to 54.6%, and after increasing the group size to 8 members, AUTE boosts robustness from the SOTA 53.0% to 56.8%. In summary, the contributions of our work are three-fold:

- We introduce Peer-Alignment training scheme, a novel strategy designed to enhance the robustness of ensemble models. It guides each individual model within the ensemble to iteratively align with stronger peers, ultimately strengthening the overall robustness of the ensemble.
- We propose the Self-Unlearning optimization, which force ensembles to learn more robust features. Unlike most methods that continuously perform the learning process, we forces ensemble members to intentionally forget certain adversarial examples with high confidence and then attempt to relearn this type of data.
- Extensive experiments across various datasets and networks demonstrate that AUTE not only achieves the highest performance, but also showcases strong scalability and generalizability to the standard training paradigm.

Related Work

Adversarial Attacks. Given the susceptibility of DNNs to adversarial examples, there has been a significant surge in interest surrounding the development of attack techniques. Specially, an adversarial example is created by adding an imperceptible perturbation to the clean data, which can mislead the model to flip its label to a wrong prediction. A variety of attack methods has been developed recently, including Momentum-based Iterative Method (MIM) (Dong et al. 2018), Projecting Gradient Descent (PGD) (Madry et al. 2017), Parameters-freed Auto-Attack (AA) (Croce and Hein 2020), *etc.* These attacks have demonstrated their capability to achieve a high success rate in misleading DNNs, even when subjected to defensive mechanisms (Prakash et al. 2018; Athalye, Carlini, and Wagner 2018). They also exhibit ability to generalize adversarial effect across different networks and datasets under black-box scenarios (Huang, Gao, and Liu 2023; Huang et al. 2022). This poses a significant threat to the security of AI-controlled systems.

Adversarial Defenses. Adversarial Training (AT) treats adversarial examples as a form of augmentation data to train models (Madry et al. 2017). For example, TRADES (Zhang et al. 2019) is engineered to achieve better balance between clean accuracy and adversarial robustness. MART (Wang et al. 2019) distinguishes between misclassified and correctly classified data during optimization, aiming to improve overall model performance. HAT (Rade and Moosavi-Dezfooli 2021b) introduces helper examples to confine the excessive margin of decision boundaries. STAT (Qizhang Li 2023) creates collaborative examples (instead of adversarial examples) during training, fortifying its defensive capacity. AROW (Yang, Kong, and Kim 2023) pays more attention to less robust data, going a step further in boosting resistance. Although these methods perform well in single-model scenarios, effectively combining multiple AT models to enhance robustness presents an ongoing challenge.

Ensemble Learning (EL) was initially conceived to enhance overall performance in the context of classifying out-of-distribution data (Schapire 2013) or uncertainty estimation (Lakshminarayanan, Pritzel, and Blundell 2017). Recent studies show that EL methods can improve adversarial robustness, particularly in the settings of defending against black-box attacks (Pang et al. 2019; Yang et al. 2020). Several methods focus on training ensembles from an optimization standpoint, such as ADP (Pang et al. 2019) and GAL (Kariyappa and Qureshi 2019). Another way to strengthen the ensemble is employing augmentation and smoothing techniques, including DVERGE (Yang et al. 2020) and TRS (Yang et al. 2021b). However, their robustness is significantly degenerated under white-box attacks. Beyond standard EL methods, some advanced approaches, such as MCE (Zhang et al. 2022) and DRT (Yang et al. 2021a), incorporate both AT and EL concepts to build robust ensemble models, though their improvements are limited.

Methodology

Simple Combinations of AT and EL

We aim to answer two questions in the field: **(1)** Can adversarial training (AT) and ensemble learning (EL) be combined to improve adversarial robustness; and **(2)** If such a harmonious integration is feasible, how can we maximize the benefits derived from both EL and AT to further enhance clean accuracy and robustness?

To answer the first question, we start by building two simple combinations of AT and EL approaches. The first one directly incorporates the principles of AT into EL methods (AT2EL), which utilize adversarial examples for training ensemble models. The second approach involves the implementation of EL concepts into AT methods (EL2AT) that multiple AT models are optimized and subsequently consolidated into a large ensemble group. We test the robustness of these two combinations by using six methods, and detailed experimental results are shown in the Appendix. A.

We draw two key insights from the empirical results: **(1)** combining multiple AT models together to form an ensemble indeed exhibit higher adversarial robustness compared to a single AT model; and **(2)** Diversification regularization (Pang et al. 2019; Yang et al. 2020) sometimes unintentionally degrade the adversarial robustness of ensemble models. Building on these findings, we introduce a novel method to further enhance ensemble performance, termed Alignment and Unlearning for Training Ensembles (AUTE), addressing the second question. This method notably improves performance from both ensemble learning and adversarial training perspectives: Peer-Alignment and Self-Unlearning (Fig. 2).

Peer-Alignment Training Scheme

We introduce the Peer-Alignment (PA) scheme from the perspective of ensemble learning. Two crucial discoveries serve as the basis for it: **(1)** The experimental results above substantiate the conclusion that simply adding AT models in the ensemble can strengthen them to become a stronger defender. This finding is rooted in the intuition that stacking multiple members together to create a larger network provides increased capacity for learning more robust features from adversarial examples; and **(2)** Recent works (Zhao et al. 2022; Zhou et al. 2021; Huang et al. 2023a) have demonstrated that distillation is a valuable technique for training smaller robust networks by transferring defensive knowledge from an existing adversarially trained model. Thus, we can instruct each ensemble member to align features from its partners with higher robustness. These two observations motivate us to include three key steps in Peer-Alignment: separation, alignment, and reallocation. The intuitive mechanism of PA is illustrated in Fig. 2.

Separation, which creates adversarial examples and divides the entire ensemble into two groups. Specifically, given an ensemble model F comprising a total of n members $F = \{f_1, f_2, ..., f_n\}$, it is partitioned into two non-overlapping parts: one is a single member, denoted as f_i, which is presently the subject of optimization; another part is a *freezed* subset ensemble, namely the peer model of P_i, which consists of the rest of $n-1$ members.

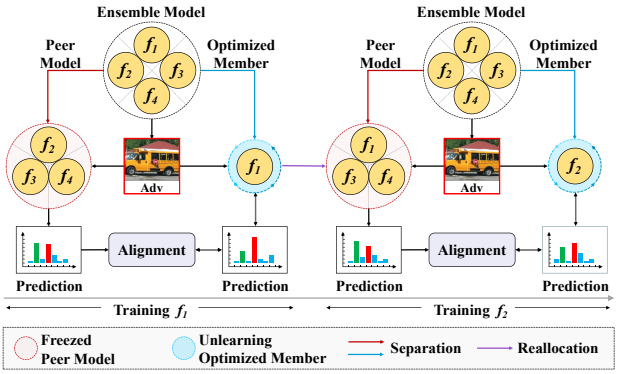


Figure 2: The pipeline of the proposed AUTE. At each iteration, the entire ensemble is divided into a large freezed peer model and an small optimized member. The member undergoes self-unlearning optimization (blue area) as well as aligns with the robust peer model. Afterward, the optimized member is reassigned within the peer model in preparation for alignment in the next iteration. These steps are executed in a round-robin manner.

Alignment, which instructs a optimized member to align with its peer model. Specifically, we use the Kullback-Leibler (KL) divergence to measure and minimize the differences between the two networks, guiding the alignment process, which is defined as:

$$\mathcal{L}_{i,\text{DIS}} = \text{KL}(f_i(x^{\text{adv}}), P_i(x^{\text{adv}})), \tag{1}$$

where x^{adv} is the adversarial example. It is evident that the capacity of the peer model become larger than this of the optimized member when the ensemble size $n \geq 3$.

Due to the increased model capacity, the peer model can effectively achieve higher robustness during training. However, we speculate that this does not necessarily make the peer model a suitable anchor for all adversarial examples. To test our hypothesis, we trained several ensembles using different AT methods and evaluated the robustness of individual members and their peer model (see Appendix. B). Surprisingly, we found that some clean data and adversarial examples were correctly classified by the smaller member but misclassified by the larger peer model, suggesting that the member may have acquired comprehensive knowledge than its peer on these specific data. This finding aligns with recent studies (Zhu et al. 2021). Thus, we employ a *selective* manner to instruct the member in alignment process. The optimization objective is reformulated Eq. (1) as:

$$\mathcal{L}_{i,\text{PA}} = (1 - f_i(x^{\text{adv}})) \cdot \text{KL}(f_i(x^{\text{adv}}), P_i(x^{\text{adv}})), \tag{2}$$

where the term $1 - f_i(x^{\text{adv}})$ is treated as an learning weight for encouraging the member to selectively align with the peer model based on the predicted confidence for the data.

Reallocation, which involves incorporating the optimized member into the peer model for making it stronger, while simultaneously excluding another member from the peer model. The excluded member then prepares for optimizing in the next iteration.

These steps are performed in a round-robin manner to progressively enhance the robustness of the ensemble. Intuitively, the peer model *dynamically* gains strength as each robustified member re-engages with the group, leading to improved accuracy on adversarial examples.

Self-Unlearning Optimization

While most methods typically treat all data equally during the optimization, recent studies emphasize the potential benefits of introducing instance-weighting to enhance robustness (Yang, Kong, and Kim 2023). Nevertheless, to the best of our knowledge, existing AT methods still force the model to *continuously* capture features from the dataset, including both adversarial and clean data. However, we discover this training paradigm may induce larger margin distance, which may harm the robustness of ensembles. To this end, we introduce the Self-Unlearning (SU) to optimize the model.

Rethink of Margin Distance. Most AT methods usually aims to achieve the maximum margin distance $\mathcal{D}$ between the groundtruth and others classes, which is defined as:

$$\mathcal{D}\big(f(x), y\big) = f(x)_y - \max_{k \neq y} f(x)_k. \tag{3}$$

Generally, $\mathcal{D}$ is employed to measure the gap between the data point and the nearest decision boundary in the latent space: the larger $\mathcal{D}\big(f(x), y\big)$, the farther away the data x is from the boundary, and vice versus. This concept is widely adopted in training robust models (Ding et al. 2019) or developing strong attacks (Carlini and Wagner 2017).

However, we observe a perplexing phenomenon where models like MART and AROW, despite having smaller margins on both clean and adversarial data, unexpectedly exhibit stronger defensive capacity compared to models with larger margins, such as SAT and TRADES (see Appendix. B). Two types of research shed light on explaining this behavior. Firstly, large margin distance is often linked to the overconfident property, potentially degrading the generalization of models (Müller, Kornblith, and Hinton 2019). Secondly, an excessive margin from the data to the decision boundary may impose a burden on the model capacity (Rade and Moosavi-Dezfooli 2021a). These observations inspire us to consider minimizing the margin distance, thereby conserving model capacity and improving adversarial robustness.

Forgetting Fewer for Learning More. Recent advanced studies have introduced regularization techniques in the training of AT models, which implicitly penalize overconfident data (Wang et al. 2019; Yang, Kong, and Kim 2023). To further reduce the margin between adversarial examples and the decision boundary, we incorporate the idea of machine unlearning (Sekhari et al. 2021) into the optimization of robust models, referred to as Self-Unlearning (SU).

In contrast to existing machine unlearning techniques that entirely eliminate the features of specific data, our approach emphasizes encouraging models to retain a subset of features from instances demonstrating overconfidence. Specifically, SU comprises two main components: firstly, the model continuously learns features from adversarial examples until they can be accurately classified, and secondly, the model progressively forgets data locates in low-loss regions with high confidence. Therefore, we reformulate standard Cross-Entropy by introducing an unlearning weight as $\mathcal{L}_{\text{UCE}}$:

$$\mathcal{L}_{\text{UCE}}(f(x), y) = -w_{\text{SU}} \cdot \sum_{k=1}^{C} y_k \cdot f(x)_k \tag{4}$$

$$w_{\text{SU}} = \begin{cases} 1 - \mathcal{D}\big(f(x), y\big) & \mathcal{D}\big(f(x), y\big) < \mathcal{M} \\ -\gamma \cdot \mathcal{D}\big(f(x), y\big) & \text{Otherwise} \end{cases} \tag{5}$$

where $\mathcal{M}$ is the threshold of margin distance, γ is a small constant. Intuitively, data with a greater margin has larger weight during the unlearning process, experiencing a faster displacement. As a consequence, both adversarial examples and clean data tend to move away from the low-loss regions.

AUTE Optimization

The proposed AUTE performs the Peer-Alignment (PA) and the Self-Unlearning (SU) for training the ensemble (Fig. 2). Specifically, we follow (Yang et al. 2020; Pang et al. 2019) to optimize members sequentially and then combing them to form a robust ensemble. In this process, each member simultaneously align with their peer model (PA) and captures (or forgets) the features from adversarial examples (SU). Therefore, the overall objective for a single member f_i is

$$\min_{\theta_i} \mathop{\mathbb{E}}_{(x,y)\sim D} \Big[\mathcal{L}_{i,\text{UCE}}(f_i(x^{\text{adv}}), y) + \beta \cdot \mathcal{L}_{i,\text{PA}}\Big], \tag{6}$$

where β is the balance weight, $\mathcal{L}_{i,\text{UCE}}$ and $\mathcal{L}_{i,\text{PA}}$ are objectives defined in SU and PA. Detailed pseudo-code for training an AUTE ensemble is shown in Appendix. C.

Experiments

Experimental Settings

Dataset. We mainly evaluate the ensemble models using the CIFAR-10 dataset. To illustrate the generalizability of the proposed method, we show that AUTE can consistently achieves superior performance across varying dataset scales—specifically, on smaller datasets like MNIST as well as complex datasets such as CIFAR-100 and Tiny-ImageNet.
Network. We align our ensemble setups with those of previous literature (Pang et al. 2019; Kariyappa and Qureshi 2019; Yang et al. 2020) during evaluations. Specifically, we utilize the light-weighting ResNet-20 architecture to develop robust ensemble models. Furthermore, we extend the experiments to include deeper and wider DNNs, such as VggNet-16, ResNet-18 and WideResNet-34. To demonstrate the scalability of AUTE, we build ensemble models with 3, 5, and 8 members together, respectively.
Baselines. We consider several ensemble versions of AT methods in comparisons: TRADES (Zhang et al. 2019), which aims to minimize the empirical risk and the robustness regularization. MART (Wang et al. 2019), which assigns larger weight to adversarial examples where the corresponding clean counterparts are wrongly predicted. HAT (Rade and Moosavi-Dezfooli 2021a) introduces a standard neural network as the helper to handle the overly perturbed adversarial images. STAT (Qizhang Li 2023) creates collaborative examples during the training of robust models. AROW (Yang, Kong, and Kim 2023) applies increased

Method	3 members						5 members						8 members					
	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**
TRADES	76.6	49.2	49.7	48.7	48.7	49.1	76.7	50.4	50.9	50.0	49.9	50.3	78.0	51.3	51.7	50.7	50.7	51.1
MART	74.7	51.1	51.4	50.6	50.4	50.9	75.0	51.6	52.0	51.2	51.0	51.5	75.0	52.6	52.8	52.1	51.9	52.4
HAT	77.2	50.0	50.6	49.3	49.2	49.8	77.9	50.5	51.3	49.9	49.8	50.4	79.2	51.2	51.7	51.0	50.7	51.2
STAT	76.3	50.6	50.9	50.3	50.3	50.5	76.8	51.2	51.5	51.0	50.9	51.2	77.0	51.7	51.9	51.4	51.3	51.6
MCE	76.3	51.1	51.5	51.0	50.6	51.1	76.7	51.9	52.2	51.8	51.5	51.9	77.1	52.5	52.9	52.5	52.3	52.6
AROW	76.2	51.7	52.0	51.3	51.3	51.6	76.3	52.4	52.7	52.1	52.0	52.3	77.4	53.1	53.5	52.8	52.7	53.0
AUTE	**78.0**	**54.8**	**56.1**	**54.1**	**53.5**	**54.6**	**78.4**	**55.7**	**57.0**	**54.4**	**55.0**	**55.5**	**80.1**	**57.1**	**58.1**	**55.6**	**56.4**	**56.8**

Table 1: The robust accuracy (%) of adversarially trained ensembles on CIFAR-10 dataset with ResNet-20 networks.

Method	VggNet-16						Resnet-18						WideResNet-34-10					
	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**
TRADES	82.9	52.9	54.0	52.3	52.1	52.8	81.6	55.7	56.4	55.5	55.2	55.7	84.4	56.7	56.3	56.2	56.0	56.3
MART	82.5	51.5	53.1	50.6	50.1	51.3	81.9	56.6	57.8	56.2	55.8	56.6	86.2	60.9	58.6	58.4	58.2	59.0
HAT	83.7	51.0	52.8	49.8	49.4	50.8	**84.8**	53.7	55.0	53.1	52.6	53.6	**86.3**	58.5	60.3	59.1	57.3	58.8
STAT	82.7	54.1	55.0	53.5	53.4	54.0	83.1	56.8	57.7	56.5	56.2	56.8	86.2	59.8	60.7	59.4	59.1	59.8
MCE	83.1	55.2	55.0	54.1	53.5	54.5	82.2	56.5	57.6	56.1	56.0	56.6	85.3	60.1	61.0	59.4	59.0	59.9
AROW	82.8	54.9	55.7	54.5	54.5	54.9	82.4	57.9	58.3	57.6	57.5	57.8	85.8	59.7	60.5	59.5	59.2	59.7
AUTE	**84.8**	**59.5**	**61.6**	**58.2**	**55.8**	**58.8**	82.1	**64.2**	**65.0**	**63.9**	**61.3**	**63.6**	85.1	**64.6**	**66.1**	**63.9**	**62.6**	**64.3**

Table 2: The robust accuracy (%) of adversarially trained ensembles on CIFAR-10 dataset with different structures.

regularization to data susceptible to adversarial attacks. MCE (Zhang et al. 2022), which is an advanced method for learning an ensemble with maximum margin.

Attack Models and Metrics. We test the robustness of ensembles under different attack scenarios. Particularly, we include four white-box attacks: **1)** 50-step Momentum-based Iterative attack Method (MIM) (Dong et al. 2018) with step size $\epsilon/5$; **2)** 10-step and 100-step PGD (Madry et al. 2017), denoted as PGD-10 and PGD-100, respectively; and **3)** Auto-Attack (AA) (Croce and Hein 2020), which is an ensemble of parameter-free attacks to fool classifiers. We evaluate the accuracy of ensembles on clean data (clean accuracy) and robustness against adversarial examples generated with a perturbation magnitude of $\epsilon = 8/255$. More experimental results of ensemble voting, smaller and larger perturbations are reported in Appendix.

Experimental Results of AUTE

We mainly assess the performance of ensemble models on the light-weighting ResNet-20 structure using CIFAR-10 dataset. We also extend our experiments to different datasets (*i.e.*, MNIST, CIFAR-100, and Tiny-ImageNet). We also explore the combination of various network architectures (*i.e.*, VggNet-16, ResNet-18, and WideResNet-34).

(1) Performance on CIFAR-10 Dataset. We demonstrate the clean accuracy and adversarial robustness under different white-box attacks in Tab. 1. In particular, we include the adversarially trained ensembles with various group size (*i.e.*, 3, 5, and 8 members) in the evaluations.

Referring to Table 1, it is consistent with our observations in Section that most ensembles exhibit similar trends. As the number of members in the ensemble models increases, their performance gradually improves. Among baselines, HAT consistently demonstrates superior performance in classifying clean data among the baselines. Conversely, MART emerges as a robust defender against adversarial examples, albeit at the cost of recording the lowest accuracy on clean data, thereby limiting its practical applicability. Similarly, methods like TRADES and STAT also exhibit varying degrees of bias towards either clean accuracy or robustness. The experimental results illustrate the challenge of balancing these two metrics simultaneously.

In comparison to baseline methods, AUTE demonstrates significantly enhanced performance in classifying both clean and adversarial examples. Notably, a 3-member AUTE accurately identifies 54.6% of adversarial examples on average, surpassesing the runner-up AROW ensemble by 3.5%. Moreover, the scalability of AUTE is remarkable, as evidenced by its favorable outcomes with larger ensemble groups. By training with more members, AUTE consistently enhances its performance. Particularly, an 8-member AUTE surpasses the nearest competitors AROW by a considerable margin of 3.9% in robustness against PGD-100 attacks. This improvement can be attributed to the fact that AUTE combines the alignment process with an unlearning paradigm to optimize ensemble members, thereby demonstrating better trade-off between robustness and accuracy.

(2) Performance on Complicated Structures. Instead of merely using the light-weighting ResNet-20 network to form ensemble models, we consider following settings: **1)** ResNet family but with different network depths and widths, *i.e.*, ResNet-18 and WideResNet-34-10; and **2)** the VggNet network family, *i.e.*, VggNet-16. The performance of these variants on CIFAR-10 dataset is reported in Tab. 2.

According to Tab. 2, there are three implications. Firstly, it is evident that incorporating deeper ResNets into ensemble models results in significant enhancements across both clean accuracy and adversarial robustness metrics. This observa-

Method	MNIST						CIFAR-100						Tiny-ImageNet					
	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**	Clean	MIM	PGD-10	PGD-100	AA	**Avg**
TRADES	98.2	93.7	93.7	93.3	93.1	93.4	58.0	30.9	31.7	30.6	30.2	30.9	51.5	25.1	25.3	24.8	24.8	25.0
MART	99.1	93.3	93.6	92.3	89.7	92.2	56.1	31.0	32.1	30.3	30.0	30.9	50.1	24.2	24.6	23.9	23.7	24.1
HAT	**99.2**	93.6	93.9	93.2	93.4	93.5	58.2	31.4	31.3	30.8	30.3	31.0	50.3	24.4	25.0	24.1	24.0	24.4
STAT	98.7	93.6	94.3	93.5	90.3	92.9	57.3	31.8	32.0	31.6	31.3	31.7	50.2	24.8	25.0	24.8	24.8	24.9
MCE	98.5	93.1	94.0	93.0	92.9	93.3	58.1	31.1	32.3	32.0	31.6	31.8	50.0	25.0	25.6	25.0	24.8	25.1
AROW	97.8	93.3	93.3	93.1	92.8	93.1	58.4	32.1	32.4	31.9	31.0	31.9	50.6	25.2	25.6	25.1	25.0	25.2
AUTE	98.9	**94.9**	**94.9**	**94.3**	**94.2**	**94.6**	**59.3**	**32.6**	**33.2**	**32.3**	**32.0**	**32.5**	**52.5**	**25.9**	**26.3**	**25.4**	**25.3**	**25.7**

Table 3: The robust accuracy (%) of adversarially trained ensembles on different datasets.

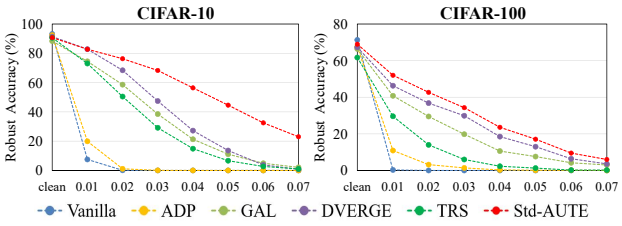


Figure 3: Robustness of standard (non-adversarial) trained ensembles on CIFAR-10 and CIFAR-100 datasets.

Settings	PA	SU	Clean	MIM	PGD-10	AA	Avg
AT	×	×	76.7 (-)	49.5	49.9	48.9	49.4 (-)
w/o PA	×	✓	78.4 (+1.7)	**55.2**	56.1	52.0	54.4 (+5.0)
w/o SU	✓	×	79.0 (+2.3)	50.4	51.2	49.4	50.3 (+0.9)
w/o Weight	⋆	✓	78.5 (+1.8)	54.9	55.9	52.5	54.4 (+5.0)
w/ ADP	⋆	✓	71.3 (-5.4)	45.0	45.4	44.6	45.0 (-4.4)
AUTE-LS (0.1)	✓	⋆	**79.2 (+2.5)**	48.6	49.3	47.6	48.5 (-0.9)
AUTE-LS (0.3)	✓	⋆	79.1 (+2.4)	49.0	49.8	48.3	49.0 (-0.4)
AUTE-LS (0.5)	✓	⋆	78.6 (+1.9)	49.1	50.0	48.4	49.2 (-0.2)
AUTE (**ours**)	✓	✓	78.0 (+1.3)	54.8	**56.1**	**53.5**	**54.8 (+5.4)**

Table 4: The robust accuracy (%) of adversarially trained ensembles trained by using different setups.

tion is consistent with prior studies that networks with more parameters usually brings larger model capacity, which supports models in achieving higher performance against adversarial attacks. Secondly, the choice of network family emerges as a crucial factor influencing learning preferences during model optimization. Specifically, VggNet ensembles significantly enhances clean data classification capabilities, *e.g.*, the improvements are 5.9% and 7.8% for TRADES and MART, respectively. However, the improvements in robustness are comparatively less pronounced. This fact reveals the accuracy and robustness bias related to the network structures. Thirdly, the proposed AUTE still demonstrate superior performance compared to current methods. Specially, leveraging both alignment and unlearning mechanisms, AUTE achieves significant advantages from larger model capacities. The average enhancements of these three complicated ensembles are 3.2%, 9.0% and 9.7% in adversarial robustness, surpassing those of baseline methods. This observation highlights the potential and effectiveness of AUTE.

(3) Performance on Different Datasets. We report the experimental results of various ensemble models with 3-member setup on MNIST, CIFAR-100 and Tiny-ImageNet datasets in Tab. 3. Specially, MNIST has resolutions of 28×28, which is smaller than the dimension of CIFAR-10 dataset. CIFAR-100 consists of 100 classes, and Tiny ImageNet comprises 200 classes with a resolution of 64×64. We train ResNet-20 ensembles on MNIST and ResNet-18 ensembles on CIFAR-100 and Tiny-ImageNet, respectively.

In Table 3, a noteworthy deviations are observed in the behaviors of MART, STAT, and HAT compared to their performances on the CIFAR-10 datasets. Specifically, the defensive capabilities of MART and STAT diminish significantly when confronted with different attacks. On the contrary, HAT exhibits an noticeable improvement in both clean accuracy and robustness on MNIST compared to its performance on CIFAR-10 (see Table. 1). We hypothesize that this success can largely be attributed to the fact that its helper model, trained using the standard paradigm, can achieve nearly 100% accuracy on these datasets. Moreover, it is clearly indicates that AUTE maintains state-of-the-art performance in defending against various attacks. Specifically, AUTE stands out as the frontrunner, outperforming the strongest competitor, AROW, by 0.9% in clean accuracy and 0.6% in average robustness on the CIFAR-100 dataset. Similarly, on the Tiny ImageNet dataset, AUTE achieves an average improvement of 1.9% in clean accuracy and 0.5% in robustness. This outcome provides compelling evidence to support the efficacy of employing AUTE for training more complex datasets.

(4) Performance on Standard Training. We explore the potential of AUTE by training ensembles exclusively on natural data, without incorporating any adversarial examples. We consider five standard ensemble methods in evaluations: Vanilla, ADP, GAL, DVERGE, and TRS. We note that these methods aim to achieve dual objectives: maintaining high clean accuracy while simultaneously enhancing adversarial robustness. More details are introduced in the Appendix. E.

We plot the black-box robustness of different methods in Fig. 3 and report detailed white-box robustness in Appendix. E. We can see that the proposed Std-AUTE also exhibits remarkable robustness against black-box attack with large perturbation compared to baselines. A similar tendency is observed under white-box attacks that AUTE surpasses the second-place baseline by a substantial margin.

The Effect of Alignment and Unlearning

We study the influence of Peer-Alignment (PA) and Self-Unlearning (SU) in AUTE. Concretely, we investigate each component in 3-member AUTE ensembles as following:

- Removing either PA or SU from the AUTE ensemble is denoted as *w/o PA* or *w/o SU*, respectively.
- Alignment process without a sample-selective manner in PA, where the weight in Eq. (2) is set equally for all adversarial data, denoted as *w/o Weight*.
- Replace the PA with a diversification ADP (Pang et al. 2019) to regularize the ensemble, denoted as *w/ ADP*.
- Replacing the SU with the Label Smoothing (LS) with a cofficient λ, denoted as *AUTE-LS* (λ).
- The full version of the proposed method, *i.e.*, *AUTE*.

The comparison results for the aforementioned settings are presented in Table 4. The symbol $\star$ denotes the substitution of PA or SU with other mechanisms.

Observing the results in Table 4, both PA and SU contribute significantly to enhancing the performance of ensemble models (rows 2 and 3). Specifically, PA notably improves clean accuracy by 2.3%, while AU demonstrates a greater tendency to enhance adversarial robustness, with an average increase of 4.8%. Furthermore, removing the learning weights of PA results in a slight boost in clean accuracy, but at the expense of degraded robustness, particularly against attackers with large perturbations (row 4). Besides, the diversification regularizer may compromise the robustness, aligning with our discussions in Methodology (row 5). As for the label smoothing strategy, we observe that it indeed marginally improves clean accuracy. However, its defensive capacity diminishes significantly. An interesting observation is that as the label smoothing increases (rows 6-8), the model's robustness improves while its clean accuracy decreases. This aligns with findings from a related study (Yang et al. 2021b), suggesting that smoothing the model could be a viable defense against attacks. Consequently, the combination of PA and SU achieves desirable clean accuracy and the highest robustness compared to other setups (row 9).

Ablation Studies of AUTE

We train ensembles with 5-member on CIFAR-10 dataset, where quantitative results are presented in the Appendix.

Group Size of Peer Models. We consider to selectively combine fewer robust members to form the peer model. A common trend is observed: both clean accuracy and robustness gradually increase as the peer model becomes larger. This supports the conclusion that a larger capacity helps the model capture more robust features. Thus, we select all partner members within the peer model.

Threshold of Margin Distance. The threshold $\mathcal{M}$ determines unlearning behaviors of ensembles (Eq. (5)). We evaluate thresholds ranging from 0.01 to 1.0, where a smaller threshold indicates that adversarial data are positioned closer to the decision boundary. We find that variations in the threshold do not significantly affect clean accuracy, maintaining a stable performance of 78.0$\pm$0.5%. However, there is a decline in robustness with increasing thresholds.

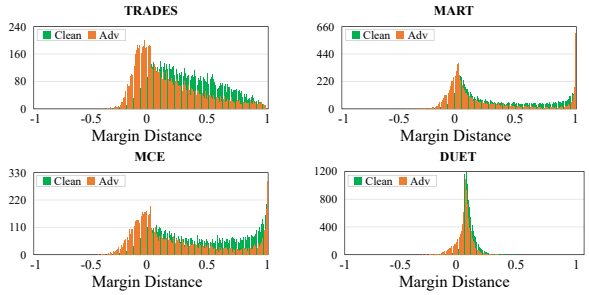


Figure 4: Statistic of margin distance on different ensembles.

Balance Weight. We train the AUTE with different weights β (Eq. (6)). We can see that clean accuracy and robustness initially improve and then decline after reaching the peaks. This suggests that emphasizing the learning from the peer model too much may lead to inverse consequences. Therefore, we select a medium weight to strike a better balance between these competing objectives.

Statistic of Margin Distance. We conducted a statistical analysis on the margin distances of 10000 clean data and adversarial examples (Fig. 4). We observe that MART and MCE ensembles tend to memorize all data, resulting in a phenomenon where a portion of data are classified with absolute confidence. Conversely, TRADES exhibited a relatively more uniform distribution trend. In comparison to baselines, AUTE showcased different statistical patterns that most of data points are concentrated within a small region, which is largely attributed to the unlearning process.

Conclusion

In this paper, we focused on enhancing the robustness of ensemble models. We introduced a novel learning method, termed AUTE, aimed at further improving ensemble robustness. AUTE comprises Peer-Alignment (PA) and Self-Unlearning (SU), which enhance performance from the perspectives of ensemble learning and adversarial training, respectively. Specifically, PA employing a selective alignment process to fortify the ensemble member in a iterative manner, and SU facilitates the ensemble in forgetting adversarial examples with overconfidence property. Extensive experiments show that AUTE not only achieves higher accuracy and robustness across different scenarios, including large datasets, complicated structures, and challenging attacks, but also showcases scalability, enabling extension to larger group and compatibility with standard training paradigms.

Acknowledgments

This work was supported by the National Key Research and Development Plan in China (2023YFC3306100), the National Natural Science Foundation of China (62472182), the Guangdong Basic and Applied Basic Research Foundation (2023A1515110075, 2024A1515010950), the Science and Technology Program of Guangzhou (2024A04J6542), and Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang) (ZJW-2023-04).

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, 274–283. PMLR.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2206–2216. PMLR.

Deng, Y.; and Mu, T. 2024. Understanding and improving ensemble adversarial defense. *Advances in Neural Information Processing Systems*, 36.

Ding, G. W.; Sharma, Y.; Lui, K. Y. C.; and Huang, R. 2019. MMA Training: Direct Input Space Margin Maximization through Adversarial Training. In *International Conference on Learning Representations*.

Doan, B. G.; Xue, M.; Ma, S.; Abbasnejad, E.; and Ranasinghe, D. C. 2022. Tnt attacks! universal naturalistic adversarial patches against deep neural network systems. *IEEE Transactions on Information Forensics and Security*, 17: 3816–3830.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, B.; Chen, M.; Wang, Y.; Lu, J.; Cheng, M.; and Wang, W. 2023a. Boosting Accuracy and Robustness of Student Models via Adaptive Adversarial Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24668–24677.

Huang, L.; Gao, C.; and Liu, N. 2023. Erosion Attack: Harnessing Corruption To Improve Adversarial Examples. *IEEE Transactions on Image Processing*.

Huang, L.; Gao, C.; Zhou, Y.; Xie, C.; Yuille, A. L.; Zou, C.; and Liu, N. 2020. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 720–729.

Huang, L.; Huang, Q.; Qiu, P.; Wei, S.; and Gao, C. 2023b. FASTEN: Fast Ensemble Learning For Improved Adversarial Robustness. *IEEE Transactions on Information Forensics and Security*.

Huang, L.; Wei, S.; Gao, C.; and Liu, N. 2022. Cyclical Adversarial Attack Pierces Black-box Deep Neural Networks. *Pattern Recognition*, 108831.

Kariyappa, S.; and Qureshi, M. K. 2019. Improving adversarial robustness of ensembles with diversity training. *arXiv preprint arXiv:1901.09981*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Pang, T.; Xu, K.; Du, C.; Chen, N.; and Zhu, J. 2019. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, 4970–4979. PMLR.

Prakash, A.; Moran, N.; Garber, S.; DiLillo, A.; and Storer, J. 2018. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8571–8580.

Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; and Zhang, C. 2021. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8681–8690.

Qizhang Li, W. Z. H. C., Yiwen Guo. 2023. Squeeze Training for Adversarial Robustness. In *International Conference on Learning Representations*.

Rade, R.; and Moosavi-Dezfooli, S.-M. 2021a. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*.

Rade, R.; and Moosavi-Dezfooli, S.-M. 2021b. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*.

Schapire, R. E. 2013. Explaining adaboost. In *Empirical inference*, 37–52. Springer.

Sekhari, A.; Acharya, J.; Kamath, G.; and Suresh, A. T. 2021. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34: 18075–18086.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2019. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*.

Wu, J.; Chen, B.; Luo, W.; and Fang, Y. 2020. Audio steganography based on iterative adversarial attacks against convolutional neural networks. *IEEE transactions on information forensics and security*, 15: 2282–2294.

Xie, C.; and Yuille, A. 2019. Intriguing Properties of Adversarial Training at Scale. In *International Conference on Learning Representations*.

Yang, D.; Kong, I.; and Kim, Y. 2023. Improving Adversarial Robustness by Putting More Regularizations on Less Robust Samples. In *International conference on machine learning*. PMLR.

Yang, H.; Zhang, J.; Dong, H.; Inkawhich, N.; Gardner, A.; Touchet, A.; Wilkes, W.; Berry, H.; and Li, H. 2020. DVERGE: diversifying vulnerabilities for enhanced robust generation of ensembles. *arXiv preprint arXiv:2009.14720*.

Yang, Z.; Li, L.; Xu, X.; Kailkhura, B.; Xie, T.; and Li, B. 2021a. On the certified robustness for ensemble models and beyond. *arXiv preprint arXiv:2107.10873*.

Yang, Z.; Li, L.; Xu, X.; Zuo, S.; Chen, Q.; Rubinstein, B.; Zhang, C.; and Li, B. 2021b. Trs: Transferability reduced ensemble via encouraging gradient diversity and model smoothness. *arXiv preprint arXiv:2104.00671*.

Zhang, D.; Zhang, H.; Courville, A.; Bengio, Y.; Ravikumar, P.; and Suggala, A. S. 2022. Building robust ensembles via margin boosting. In *International Conference on Machine Learning*, 26669–26692. PMLR.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.

Zhang, J.; Zhu, J.; Niu, G.; Han, B.; Sugiyama, M.; and Kankanhalli, M. 2020. Geometry-aware Instance-reweighted Adversarial Training. In *International Conference on Learning Representations*.

Zhao, S.; Yu, J.; Sun, Z.; Zhang, B.; and Wei, X. 2022. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *European Conference on Computer Vision*, 585–602. Springer.

Zhou, S.; Wang, Y.; Chen, D.; Chen, J.; Wang, X.; Wang, C.; and Bu, J. 2021. Distilling holistic knowledge with graph neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10387–10396.

Zhu, J.; Yao, J.; Han, B.; Zhang, J.; Liu, T.; Niu, G.; Zhou, J.; Xu, J.; and Yang, H. 2021. Reliable Adversarial Distillation with Unreliable Teachers. In *International Conference on Learning Representations*.

Zhuang, W.; Huang, L.; Gao, C.; and Liu, N. 2024. LAFED: Towards robust ensemble models via Latent Feature Diversification. *Pattern Recognition*, 150: 110225.

Boosting Imperceptibility of Adversarial Attacks for Environmental Sound Classification

Shaojian Qiu[1,2], Xiaokang You[1], Wei Rong[1], Lifeng Huang[1,*], Yun Liang[1,2]
[1]South China Agricultural University, Guangzhou, China
[2]Guangzhou Key Laboratory of Intelligent Agriculture, Guangzhou, China

2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI) | 979-8-3315-2723-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICTAI62512.2024.00116

***Abstract*—As artificial intelligence (AI) continues to advance, AI-based audio systems are becoming increasingly vulnerable to adversarial attacks. However, most current studies overlook the scenes of environmental sounds and the imperceptibility of attack. In response to these, we propose a novel frequency-weighted perturbation algorithm for environmental sounds called the Frequency Psychological Attack Algorithm (FPAA). This innovative algorithm incorporates auditory thresholds with psychoacoustic principles during the perturbation generation process to create highly imperceptible adversarial examples. Extensive experiments conducted on two public datasets using multiple models demonstrate that our FPAA algorithm can produce adversarial audio examples that are not only imperceptible to the human ear but also maintain high offensive capability against AI-based audio systems.**

***Index Terms*—Audio adversarial attack, Environmental sound classification, Imperceptibility, Psychoacoustic model**

I. Introduction

The rapid advancement of deep learning has propelled artificial intelligence (AI)-centric systems to achieve remarkable success in various intelligent applications. These achievements are tethered to the utilization of large-scale datasets, with the proficiency of the AI models being honed through the analysis of vast amounts of data [1], [2]. This profound dependence on data puts the credibility of AI models at risk. For example, adversarial attacks on speech could result in voice assistants performing erroneous commands [3], while tampering with facial data might compromise user privacy [4]. In this context, understanding adversarial attacks is crucial for bolstering the robustness of deep learning models and safeguarding them against data vulnerabilities. It is necessary to dive into the generation mechanics of adversarial examples to enhance model robustness and ensure the reliability of AI systems across various applications.

Research into adversarial attacks has seen commendable progress across several critical domains, e.g., computer vision [5], [6], natural language processing [7], and speech recognition [8]. However, exploring these attacks within Environmental Sound Classification (ESC) is markedly scant. In fact, the security of ESC tasks plays an important role in practical artificial intelligence applications. For example, in autonomous driving, tampering with the sound of surrounding vehicle horns may cause the system to mistake them for non-threatening. This misclassification poses serious risks to road safety [9], [10]. Besides, a home security system that relies on visual analysis may be less effective in low visibility conditions, so the system will rely on sound signals for analysis. At this point, the intruder may be able to manipulate environment sound to avoid detection. These scenarios underscore the urgent need for rigorous research in ESC adversarial attacks to ensure the security of AI systems and prevent potentially disastrous consequences.

Some existing adversarial attack algorithms [11]–[13] have a high success rate in perturbing environmental sounds, but the adversarial examples they generate are often poorly concealed. This is because environmental sound data is different from common adversarial attack targets. First, compared to speech data, environmental sounds do not have semantic and grammatical features in linguistics. If only tampering with such features, we cannot effectively attack the ESC system. Second, compared to image data, environmental sounds pay more attention to frequency information rather than visual content. If the spectrogram converted from audio is processed in the traditional image perturbation method, it is easily detected by the human.

To address these limitations, we propose the FreqPsy Attack Algorithm (FPAA), a novel approach that incorporates the auditory threshold and psychoacoustic principles. We evaluate FPAA's effectiveness through comprehensive experiments on two publicly available ESC datasets. The empirical results demonstrably support our claims, showing that FPAA generates adversarial perturbations with superior imperceptibility compared to existing baseline techniques.

The main contributions of this paper are as follows:

- We propose FPAA, a novel framework for crafting adversarial examples in ESC tasks. This framework incorporates auditory threshold and psychoacoustic principles, explicitly targeting the human auditory system's vulnerabilities.

*Corresponding author: Lifeng Huang (huanglf6@scau.edu.cn)
First Author and Second Author contribute equally to this work.
This work was supported by Guangdong Basic and Applied Basic Research Foundation (2022A1515110564, 2023A1515110075), Guangzhou Basic and Applied Basic Research Foundation (Grant No.2024A04J4382), and Special Fund for the Rural Revitalization Strategy of Guangdong (2023TS-3, 2024TS-3).

2375-0197/24/$31.00 ©2024 IEEE
DOI 10.1109/ICTAI62512.2024.00116

- We demonstrate how FPAA integrates with existing gradient-based attack methods. This integration significantly improves the imperceptibility of adversarial examples without sacrificing attack effectiveness.
- We conduct comprehensive experiments on publicly available ESC datasets. The results demonstrably validate the efficacy of FPAA in generating highly imperceptible adversarial examples, representing a significant advancement in ESC security.

The rest of this paper is organized as follows: Section II presents related work. Section III presents the technical specifications of our FPAA. Section IV and Section V respectively elaborate on the experimental setup and results. Section VI concludes this paper.

II. Related Work

A. Environmental Sound Classification

Environmental sound classification is the task of identifying and categorizing various non-standardized sounds that are crucial for real-world applications, such as car horns for autonomous vehicles and breaking glass for criminal investigations. Due to recent advances in deep learning for image processing [14], [15], the ESC task has been profoundly impacted. Researchers have successfully adapted image classification techniques to audio data by leveraging parallels between spectrogram features and visual features. This cross-disciplinary approach has led to substantial improvements in ESC accuracy [16], [17]. However, the heavy reliance of ESC models on large datasets raises concerns about data reliability. Consequently, there is an increasing need to investigate adversarial attack techniques in ESC tasks to ensure the robustness and reliability of these models.

B. Adversarial Attacks

Adversarial attacks exploit the vulnerabilities of machine learning models by introducing small, carefully crafted perturbations to input data, causing the model to misclassify the altered input. Consider a model with weights θ and an input x that is correctly classified with the label y. By adding a perturbation δ to x, resulting in a new input $x' = x + \delta$, the model may be tricked into misclassifying it, yielding a different predicted label $\hat{y}$. The whole process is described as $\hat{y} = f(x + \delta; \theta)$.

In recent years, significant progress has been made in the field of adversarial attacks, leading to the development of various sophisticated methods that exploit the inherent vulnerabilities in neural network architectures [11], [18]. Notable techniques include the Fast Gradient Sign Method (FGSM) [18], the Projected Gradient Descent (PGD) [19], the Basic Iterative Method (BIM) [20], Autoattack [21], and Lafeat [22]. These methods help researchers understand the sensitivity of neural networks to carefully designed perturbations and promote research aimed at enhancing model robustness. However, there is currently a lack of research on adversarial attacks on environmental sounds, but this type of data has the potential to be attacked, as shown in Fig. 1.

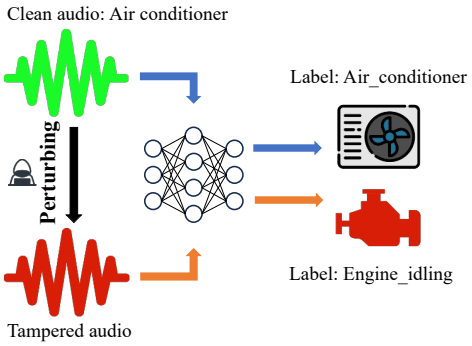


Fig. 1. An example of an audio adversarial attack. The tampered audio resembles its clean counterparts, but classification results from deep learning models are different.

In this paper, we will focus on the unique properties of environmental sounds. We will use the concepts of auditory threshold and psychoacoustic principles to formulate this attack algorithm.

III. FreqPsy Attack Algorithm

Fig. 2 illustrates the overall framework of our proposed method. First, we extract features from the raw audio to generate a spectrogram. Next, we calculate frequency weights using the A-weighting curve from the auditory threshold principle and multiply these weights by the gradient directions obtained through backpropagation with the model, producing initial perturbations. Finally, we further optimize these initial perturbations using a psychoacoustic model to obtain the final perturbations.

In the subsequent sections, we explain the concepts of auditory threshold and psychoacoustics (as shown in the yellow box in Fig. 2 and detailed in Sections III-A and III-B), as well as their roles within the framework (see Sections III-C and III-D).

A. Human Auditory Threshold Curve

The human auditory threshold curve depicts how our hearing sensitivity changes across frequencies, as explained in [23], different curves (A, B, and C) gauge sensitivity at various loudness levels. The A-weighted curve, which best mimics human perception, is crucial for noise standards and regulations.

The A-weighted curve mirrors the human ear's reduced sensitivity to low and high frequencies while emphasizing greater sensitivity in the mid-frequency range. Fig. 3 illustrates this sensitivity pattern. This characteristic makes the A-weighted curve useful in urban noise studies. For example, it helps assess the perceived intensity of urban traffic noise, as highlighted by [24]. Their research underscores the curve's utility in evaluating urban noise's impact on human perception. This provides a robust framework for urban planning and environmental health studies. The A-weighted auditory threshold curve is also crucial in designing audio technology

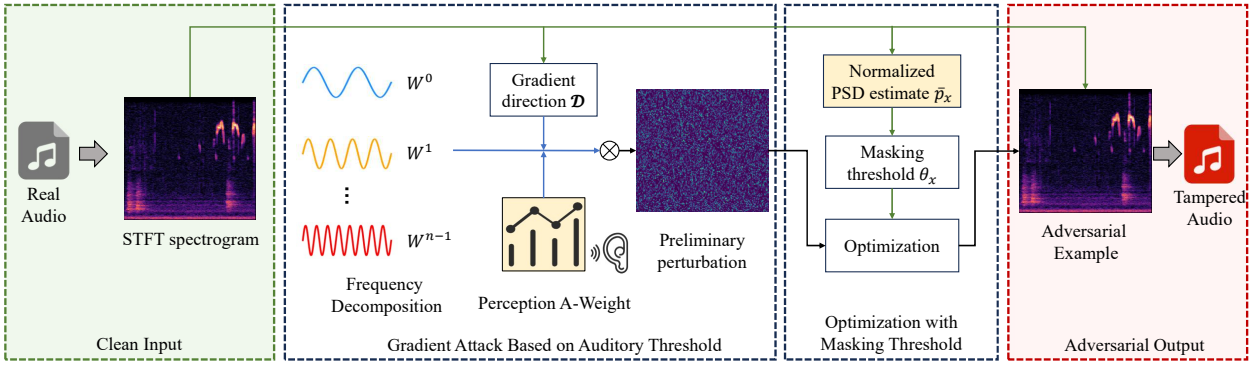


Fig. 2. Framework of our FreqPsy Attack Algorithm (FPAA). It consists of two key components: an acoustics-based optimization module and a human auditory psychology model, enabling the attack to achieve high success rates while maintaining superior imperceptibility.

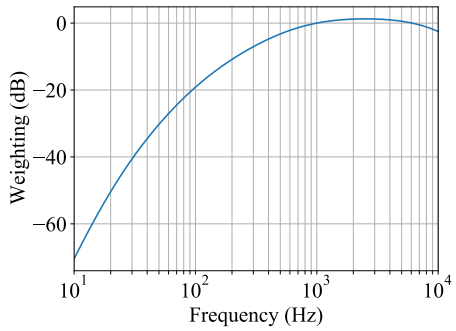


Fig. 3. A-weighted Curve. It reflects the human ear's varying sensitivity across frequencies and is used for noise standards and regulations in our method.

and sound equipment. Manufacturers and audio engineers use this curve to test and evaluate audio devices like speakers, headphones, and amplifiers, aiming to ensure that the audio output matches the perceptual characteristics of human hearing. By leveraging the A-weighted curve, they can optimize audio devices to deliver pleasing and accurate sound. The A-weighted auditory threshold curve is not merely a tool for assessing human auditory sensitivity; it is a cornerstone in designing and evaluating soundscapes and audio technologies. Its applications range from urban noise assessment to optimizing audio equipment, demonstrating its broad utility in aligning our auditory environment with human hearing capabilities.

B. Psychoacoustic Model

As mentioned earlier, subtle alterations within the spectrogram can induce substantial changes in auditory perception. Therefore, we introduce psychoacoustic principles to reduce the components in the spectrogram that can be detected by the human ear. In [25], this principle was first applied to speech, and we extend its application to environmental sounds. By exploiting the principle of frequency masking in psychoacoustics, we generate adversarial samples that remain imperceptible to human ears.

Frequency masking elucidates the phenomenon wherein a louder sound, termed the 'masker,' can obscure other sounds at nearby frequencies, known as the 'maskees.' In practical auditory perception, these maskees become effectively undetectable. The masker sets a 'masking threshold' in the frequency domain, rendering any audio signals below this threshold imperceptible.

First, we compute the log-magnitude Power Spectral Density (PSD) as follows:

$$p_x(k) = 10 \log_{10} |\frac{1}{N} s_x(k)|^2 \tag{1}$$

where $s_x(k)$ denotes the k-th bin of the specturm of frame x, N represents the size of the analysis window. We adopted the normalized PSD estimate $\bar{p}_x(k)$ [26] to compute the masking threshold, which can be expressed as:

$$\bar{p}_x(k) = 96 - \max_k \{p_x(k)\} + p_x(k) \tag{2}$$

The initial step in determining the masking threshold for an audio sample involves pinpointing the primary maskers. These are characterized by their normalized Power Spectral Density (PSD) estimate $\bar{p}_x(k)$. To qualify as maskers, these spectral peaks must fulfill a set of conditions: they should be the most prominent peaks within their spectral neighborhood; their intensity needs to surpass the baseline threshold of audibility; and their dominance in amplitude is required within a 0.5 Bark range, which is a frequency scale attuned to human auditory perception.

A specialized function with a dual slope is utilized to model the masking effect, tailored to mirror the auditory masking phenomenon of each identified masker. The overarching masking threshold, a composite metric, is then derived by combining these individual thresholds with the baseline quiet threshold. In [26], the complex process of computing the masking threshold $t_x(k)$ is explained in detail. The detectability of a perturbation δ in the audio input x is contingent upon the normalized PSD estimate $\bar{p}_\delta(k)$. If this estimate

stays below the original audio's masking threshold $\theta_x(k)$, the initial audio spectrogram effectively masks the perturbation, rendering it imperceptible to human hearing. The method to compute the normalized PSD estimate of the perturbation $\bar{p}_\delta(k)$ is as follows:

$$\bar{p}_\delta(k) = 96 - \max_k \{p_x(k)\} + p_\delta(k) \tag{3}$$

where $p_\delta(k) = 10\log_{10}|\frac{1}{N}s_\delta(k)|^2$ is the PSD estimate of the perturbation noise and $p_x(k) = 10\log_{10}|\frac{1}{N}s_x(k)|^2$ is the original spectrogram input of the audio data.

C. Gradient Attack Based on Auditory Threshold

In this section, we detail the process of computing the initial perturbation by using auditory thresholds and gradient directions in the first stage.

We use the Short-Time Fourier Transform (STFT) of the raw audio as the model's input. First, we employ backpropagation to calculate the gradient direction, denoted as $\mathcal{D}$. This process can be described using the following mathematical expression:

$$\mathcal{D} = \text{sign}\left(\nabla_x \mathcal{L}_{nn}(x, y)\right) \tag{4}$$

where x represents the original spectrogram data, and y signifies the actual label. The function $\mathcal{L}_{nn}(x, y)$ is the computational loss, and the application of the sign function is to determine the orientation of the gradient.

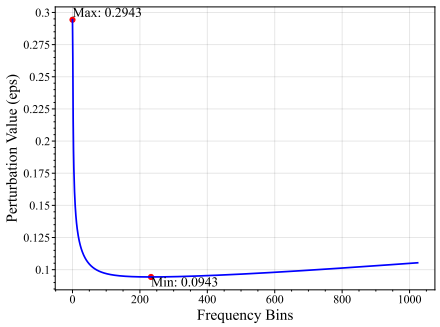


Fig. 4. Distribution of perturbation amplitude.

Then, we use the A-weighted formula of the human auditory threshold to ascertain a set of weights that reflect the human ear's varying sensitivity across frequencies. These weights are mapped to the perturbation range, expressed as $\mathcal{R} = [\epsilon_{\min}, \epsilon_{\max}]$, to determine the corresponding magnitudes. The formula below illustrates this process:

$$\epsilon(k) = M\left(\mathcal{A}(F[k]), \mathcal{R}, d\right), \quad k = 0, 1, 2, \ldots, \left\lfloor \frac{N}{2} \right\rfloor \tag{5}$$

where the $\mathcal{A}$ represents A-weighted function, F is the STFT, k is the index within the spectrogram bin, M signifies the linear mapping function, and d is a dispersion control factor. For example, Fig. 4 illustrates a scenario with an average perturbation size of 0.1 and a control factor of 0.2. In this illustration, the maximum perturbation value indicates the frequency with the least sensitivity, whereas the minimum value corresponds to the most sensitive frequency. To compute the perturbation step $\alpha(k)$ for each frequency in each iteration, we average the perturbations over T iterations. We then perform an element-wise multiplication of $\mathcal{D}$ with ϵ on each row, enabling the calculation of the cumulative perturbation δ:

$$\delta = \{\delta_k | \delta_k = \mathcal{D}(k) \times \frac{\epsilon(k)}{T}, k = 0, 1, 2, ..., \left\lfloor \frac{N}{2} \right\rfloor\} \tag{6}$$

The final step is to add δ to the original spectrogram, as shown in the equation below:

$$x_{t+1} = \Pi_{x+\mathcal{S}}\left(x_t + \delta\right) \tag{7}$$

where t denotes the iteration number, and $\Pi_{x+\mathcal{S}}$ is a projection operation that ensures the adversarial example falls within the range constraint of $x + \mathcal{S}$.

D. Optimization with Masking Threshold

In the second stage, we will further optimize the initial perturbations generated in the first stage, for which the following equation is utilized:

$$\mathcal{L}_\theta(x, \delta) = \frac{1}{\lfloor \frac{N}{2} \rfloor + 1} \sum_{k=0}^{\lfloor \frac{N}{2} \rfloor} \max\{\bar{p}_\delta(k) - \theta_x(k), 0\} \tag{8}$$

In this formula, $\bar{p}_\delta(k)$ is the normalized PSD estimate of the perturbation. The loss function $\mathcal{L}_\theta$ is designed to ensure that $\bar{p}_\delta(k)$ remains below the frequency masking threshold $\theta_x(k)$ of the original spectrogram.

The comprehensive loss function that guides our optimization is composed of two distinct parts:

$$\mathcal{L}(x, y, \delta) = -\mathcal{L}_{nn}(x + \delta, y) + \alpha \cdot \mathcal{L}_\theta(x + \delta, \delta) \tag{9}$$

where α is a parameter that balances the importance of these two loss components, it reconciles the trade-off between the effectiveness of the adversarial perturbation and its imperceptibility under psychoacoustic constraints. During the perturbation optimization process, if the generated adversarial samples are sufficient to classify the model incorrectly, the weight α will increase. Conversely, if the attack fails, the value will decrease. The optimization process iterates a total of I times.

IV. Experimental Setting

A. Datasets

The datasets UrbanSound8K [27] and ESC-50 [28] are commonly used for environmental sound classification. UrbanSound8K contains 8,732 urban sound recordings, each no longer than 4 seconds, organized into ten classes. Its focus on short urban sounds makes it well-suited for machine learning applications. On the other hand, ESC-50 offers a broader range, featuring 2,000 5-second audio samples spread across 50 diverse categories, including natural sounds, human activities, and animal sounds. Our study utilized 50 samples per class from UrbanSound8K and 10 samples per class from ESC-50.

TABLE I

COMPARATIVE ANALYSIS OF STOI AND SNR BETWEEN THE SEVERAL COMMON ADVERSARIAL ATTACK METHODS AND FPAA ACROSS FIVE MODELS ON THE URBANSOUND8K. TO ENSURE FAIRNESS, THE ATTACK SUCCESS RATE OF THE ADVERSARIAL ATTACK ALGORITHM APPLIED TO EACH MODEL IS LIMITED TO A CERTAIN RANGE.

Models	Scope of ASR	FGSM		PGD		LAFEAT		AutoAttack		FPAA	
		STOI	SNR(dB)	STOI	SNR(dB)	STOI	SNR(dB)	STOI	SNR(dB)	STOI	SNR(dB)
VGG13	81.60%-83.00%	0.750	6.01	0.945	30.47	0.826	22.27	0.830	22.35	**0.963**	**44.22**
VGG16	84.60%-85.80%	0.751	6.08	**0.972**	30.71	0.825	19.12	0.832	20.04	0.964	**44.59**
GoogLeNet	98.00%-100.00%	0.953	32.04	0.958	32.07	0.796	3.77	0.819	14.28	**0.967**	**38.40**
DenseNet	90.00%-91.90%	0.950	30.10	**0.956**	32.76	0.814	13.30	0.818	13.11	0.956	**34.07**
SBCNN	56.80%-58.00%	0.923	25.45	0.949	32.05	0.852	51.06	0.855	50.23	**0.959**	**51.69**

TABLE II

COMPARATIVE ANALYSIS OF STOI AND SNR BETWEEN THE SEVERAL COMMON ADVERSARIAL ATTACK METHODS AND FPAA ACROSS FIVE MODELS ON THE ESC-50. TO ENSURE FAIRNESS, THE ATTACK SUCCESS RATE OF THE ADVERSARIAL ATTACK ALGORITHM APPLIED TO EACH MODEL IS LIMITED TO A CERTAIN RANGE.

Models	Scope of ASR	FGSM		PGD		LAFEAT		AutoAttack		FPAA	
		STOI	SNR(dB)	STOI	SNR(dB)	STOI	SNR(dB)	STOI	SNR(dB)	STOI	SNR(dB)
VGG13	78.20%-79.40%	0.768	9.40	0.905	23.88	0.838	26.17	0.838	26.18	**0.947**	**33.52**
VGG16	91.40%-92.80%	0.743	4.31	0.907	23.64	0.788	11.06	0.810	11.99	**0.929**	**27.49**
GoogLeNet	91.40%-92.60%	0.966	34.49	0.980	41.32	0.824	10.39	0.827	11.77	**0.980**	**42.68**
DenseNet	83.60%-84.60%	0.956	33.62	**0.985**	**47.24**	0.837	22.41	0.837	20.51	0.972	37.90
SBCNN	70.20%-71.20%	0.772	7.95	0.882	21.68	0.843	35.03	0.850	35.10	**0.942**	**37.71**

B. Baseline

To demonstrate the effectiveness of our approach, we selected several advanced gradient-based attack algorithms, as well as several adversarial attack algorithms in the speech domain.

In the comparative experiments with the gradient-based adversarial attack algorithms, we have chosen three generic attack algorithms, FGSM and PGD, as foundational elements of our study. Complementing these, we introduce two innovative attack algorithms from the image processing domain: Autoattack [21] and Lafeat [22]. In the comparative experiments with adversarial attack algorithms in the speech domain, we have chosen two classic speech adversarial attack algorithms: Fakebob [29] and Sirenattack [30]. Our experiments focus on two key metrics: signal-to-noise ratio (SNR) and Short-Term Objective Intelligibility (STOI). These metrics are commonly used in current research to gauge both the efficacy and the imperceptibility of adversarial examples [31], [32].

C. Models

In the comparative experiments with gradient-based adversarial attack algorithms, we selected several widely recognized convolutional neural networks: VGG13 [33], VGG16 [33], GoogLeNet [34], DenseNet [35], and SBCNN [17]. The test accuracies of these five target models on the UrbanSound8K dataset are 92.50%, 88.90%, 92.96%, 90.44%, and 90.84%, respectively, while on the ESC-50 dataset, the test accuracies are 92.25%, 90.00%, 63.75%, 60.00%, and 35.25%. For the comparative experiments with adversarial attack algorithms in the speech domain, we selected a one-dimensional CNN [36] to match these attack algorithms.

D. Experimental parameter settings

In our methodology, we carefully select unique initial perturbation values for each model to establish a clear contrast with alternative methods that achieve similar success rates and demonstrate the superiority of FPAA. Specifically, we set the iteration count to 40, adjust the discretization factor d to 0.2, and carry out a Fourier transform process involving 2048 samples (N) to generate the initial perturbation. In the perturbation optimization phase, we start with the initial value of α set to 0.09 and perform 1000 iterations.

V. Experimental results

A. Comparison with gradient-based adversarial attack algorithms

On the UrbanSound8K dataset, FPAA's performance is exemplary across multiple models. For instance, in the VGG16 model, FPAA achieves a remarkable SNR of 44.59 and an STOI of 0.964. This is a significant improvement over other methods, such as FGSM and PGD, which report lower SNR and STOI values, highlighting that FPAA not only ensures the success of adversarial attacks but also does so with minimal degradation to the audio quality. The consistently high SNR and STOI values across all models tested with the UrbanSound8K dataset illustrate FPAA's robustness and effectiveness in preserving the auditory essence of the original signal.

On the ESC-50 dataset, FPAA significantly surpasses other techniques in preserving audio quality. Using the GoogLeNet model, FPAA achieves an SNR of 42.68 dB, outperforming the PGD method's 41.32 dB, and maintains competitive STOI values, showcasing its superior ability to minimize audio distortion. With the DenseNet model, FPAA does not surpass PGD's SNR of 47.24 dB and STOI of 0.985 but still outperforms other adversarial methods like FGSM, LAFEAT, and AutoAttack with an SNR of 37.90 dB and an STOI of 0.972. This demonstrates FPAA's robust capability in preserving high intelligibility under complex auditory conditions, confirming its effectiveness in various attack scenarios where maintaining high audio quality is crucial.

The variation in the performance of FPAA across the UrbanSound8K and ESC-50 datasets on various models can largely be attributed to the intrinsic characteristics of the datasets. UrbanSound8K, with its complex urban noise environments, provides a rich backdrop that may amplify the strengths of FPAA in handling diverse and dynamic sound elements. Conversely, ESC-50's more defined and possibly less complex environmental sounds present a different scenario. This reflects a fundamental aspect of adversarial audio research: the effectiveness of a method is heavily dependent on the dataset's acoustic properties and the specific challenges it poses, necessitating adaptability in approach and execution across different audio contexts.

B. Comparison with speech adversarial attacks

Table. III presents the comparison results between FPAA and speech adversarial attack methods on UrbanSound8K and ESC-50 datasets. We set unified attack success rate ranges of 97.00% to 98.00% and 92.00% to 93.00%.

TABLE III
Comparison with the algorithms of speech adversarial attack

Methods	UrbanSound8K		ESC-50	
	STOI	SNR(dB)	STOI	SNR(dB)
FakeBob	0.852	18.36	0.893	23.65
Sirenattack	0.543	5.97	0.881	22.71
FPAA	0.967	38.43	0.980	42.68

For the UrbanSound8K dataset, FPAA achieves an STOI of 0.967 and an SNR of 38.43 dB, outperforming FakeBob, which has an STOI of 0.852 and an SNR of 18.36 dB, and Siren attack, which achieves an STOI of 0.543 and an SNR of 5.97 dB. Similarly, on the ESC-50 dataset, FPAA records an STOI of 0.980 and an SNR of 42.68 dB, surpassing FakeBob with an STOI of 0.893 and an SNR of 23.65 dB and Sirenattack with an STOI of 0.881 and an SNR of 22.71 dB. These results indicate that the adversarial attack methods commonly used in the speech domain do not perform well on environmental sound data.

C. The effect of Freq and Psy Modules

TABLE IV
Ablation experiments on the UrbanSound8K

Models	FPAA		w/o Freq		w/o Psy		w/o Freq&Psy	
	STOI	SNR	STOI	SNR	STOI	SNR	STOI	SNR
VGG13	**0.963**	**44.22**	0.942	40.75	0.946	30.73	0.945	30.47
VGG16	0.964	**44.59**	0.942	37.79	0.946	30.81	**0.972**	30.71
GoogLeNet	**0.965**	**38.40**	0.964	37.85	0.956	32.44	0.958	32.02
DenseNet	0.956	34.07	**0.962**	**36.90**	0.952	31.23	0.956	32.76
SBCNN	**0.959**	**51.69**	0.956	46.89	0.956	32.30	0.949	32.05

As shown in Table. IV, we conducted ablation experiments on the UrbanSound8K dataset to evaluate the impact of different modules. We compare various models under different conditions: with FPAA, without Freq, without Psy, and without both Freq and Psy. Models like VGG13 and GoogLeNet achieved the highest STOI scores of 0.963 and 0.965, respectively, when using FPAA, demonstrating superior audio signal reconstruction capabilities. However, removing Freq or Psy led to notable performance drops, especially in SNR metrics, where many models dropped below 30 dB without these modules. These results indicate the pivotal role of FPAA in enhancing audio processing tasks, particularly when leveraging frequency and psychoacoustic characteristics to improve adversarial audio quality and clarity.

D. The impact of interaction number

The experiment shows the impact of iterative testing on SNR and STOI for five models using the FPAA method on the UrbanSound8K dataset. Each model was tested from 20 to 200 iterations, capturing results at 20-iteration intervals. The comparison of SNR and STOI is shown in Fig. 5.

For SNR, the results on SBCNN show a clear upward trend, with the audio signal quality improving as the number of iterations increases. The results on DenseNet show the slightest changes, indicating early saturation or robustness. The results on VGG and GoogLeNet show moderate SNR improvements.

Regarding STOI, all models maintain stability, suggesting that intelligibility is preserved early in the iteration process, regardless of SNR changes. This is crucial for applications

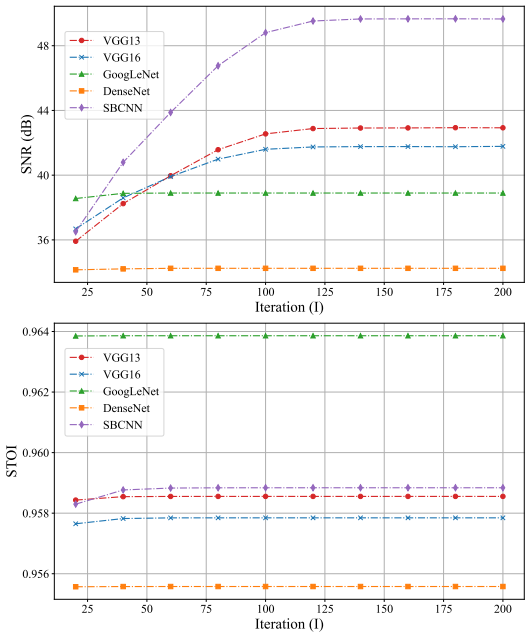


Fig. 5. The impact of optimization iteration times on adversarial examples.

where speech understanding is essential, even under adversarial conditions.

Despite varying SNR trends, the stable STOI values across models suggest that these models effectively preserve intelligibility against iterative adversarial attacks. Enhancing model robustness may require focusing on both maintaining SNR and ensuring intelligibility.

E. The impact of balance weight

In this analysis, we examine the impact of employing dynamically adjusted versus weight α on the SNR (Eq.(9)). As shown in Fig. 6, the results with dynamic adjustment of α show a progressive increase in SNR. It starts at 49.41 dB and ascends to 51.69 dB as the number of iterations increases from 100 to 1000. This suggests that adapting the α to the evolving characteristics of the adversarial sample or the target model's response enhances the stealth and quality of the adversarial examples.

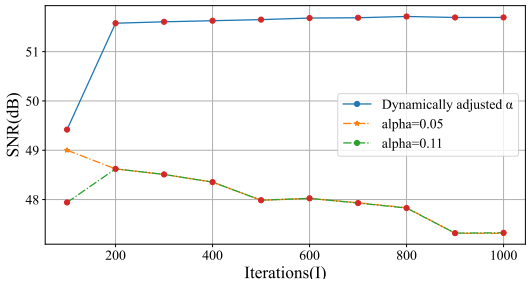


Fig. 6. The impact of dynamically adjusted α on optimization performance.

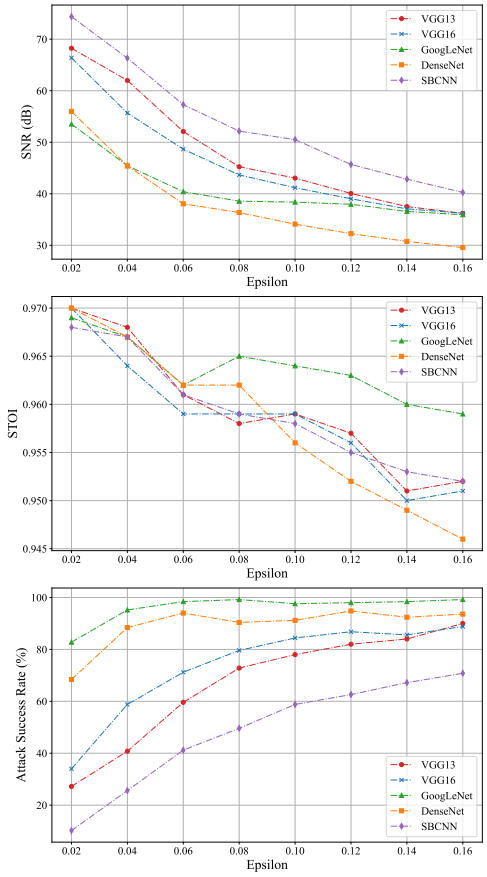


Fig. 7. Performance under different ϵ.

In contrast, using a constant α showed an initial downward trend after 200 iterations, indicating that the quality and clarity of the audio gradually diminished over time. These findings underscore the benefits of a dynamic adjustment strategy in the adversarial process to maintain and enhance the effectiveness and imperceptibility of attacks over extended iterations.

F. The impact of perturbation magnitude

As shown in Fig. 7, for all models, the SNR decreases as ϵ increases from 0.02 to 0.16, indicating that the audio contains more noise as the perturbation level increases. Although the STOI score also decreases, the decline is small, within 0.025. Among all models, GoogLeNet demonstrates the smallest decrease in SNR, suggesting it has the highest robustness against noise. In contrast, SBCNN exhibits the largest decrease in SNR, indicating it is the most susceptible to noise. These results highlight the varying degrees of robustness different models have against adversarial perturbations. The ability of FPAA to maintain audio clarity despite increasing adversarial noise is particularly noteworthy, which is critical to improving the effectiveness of audio adversarial attacks in various complex scenarios.

VI. Conclusion

This work proposes the Frequency Psychological Attack Algorithm (FPAA), which combines human auditory thresholds and psychoacoustic modeling with data perturbation to create robust adversarial examples against environmental sound classification models. By integrating these auditory principles, FPAA ensures that perturbations are highly imperceptible to human listeners while remaining effective against AI-based audio systems. We validated FPAA's effectiveness through extensive experiments on public datasets and various neural network architectures, demonstrating the algorithm's capability to generate adversarial examples that are both difficult for humans to detect and potent in deceiving models.

The FPAA algorithm and its code are publicly available on GitHub at https://github.com/RyunosukeYuu/FPAA.

References

[1] J. A. Kroll and V. Berzins, "Understanding, assessing, and mitigating safety risks in artificial intelligence systems," Report No. NPS-CS-22-001). Naval Postgraduate School, Tech. Rep., 2022.

[2] J. Hu, Z. Zhao, F. Hang, and J. Yin, "Effective application of artificial intelligence techniques in security risk assessment and dependency analysis of open source components," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1.

[3] D. Dai, Z. An, and L. Yang, "Inducing wireless chargers to voice out for inaudible command attacks," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 1789–1806.

[4] S. Jia, C. Ma, T. Yao, B. Yin, S. Ding, and X. Yang, "Exploring frequency adversarial attacks for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4103–4112.

[5] J. Zhang, Y. Huang, W. Wu, and M. R. Lyu, "Transferable adversarial attacks on vision transformers with token gradient regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 415–16 424.

[6] D. Lu, Z. Wang, T. Wang, W. Guan, H. Gao, and F. Zheng, "Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 102–111.

[7] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, 2020.

[8] Q. Liu, T. Zhou, Z. Cai, and Y. Tang, "Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2390–2398.

[9] G. Ciaburro and G. Iannace, "Improving smart cities safety using sound events detection based on deep neural network algorithms," in *Informatics*, vol. 7, no. 3. MDPI, 2020, p. 23.

[10] H. Touqeer, S. Zaman, R. Amin, M. Hussain, F. Al-Turjman, and M. Bilal, "Smart home security: challenges, issues and solutions at different iot layers," *The Journal of Supercomputing*, vol. 77, no. 12, pp. 14 053–14 089, 2021.

[11] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017, pp. 39–57.

[12] Z. Chen, B. Li, S. Wu, K. Jiang, S. Ding, and W. Zhang, "Content-based unrestricted adversarial attack," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] F. Li, R. Liu, Z. He, S. Gao, Y. Dong, and W. Zhou, "Ria: A reversible network-based imperceptible adversarial attack," in *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2022, pp. 996–1001.

[14] K. W. Gunawan, A. A. Hidayat, T. W. Cenggoro, and B. Pardamean, "A transfer learning strategy for owl sound classification by using image classification model with audio spectrogram," *International Journal on Electrical Engineering and Informatics*, vol. 13, no. 3, pp. 546–553, 2021.

[15] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Procedia computer science*, vol. 112, pp. 2048–2056, 2017.

[16] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *arXiv preprint arXiv:1706.07156*, 2017.

[17] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[20] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[21] Y. Liu, Y. Cheng, L. Gao, X. Liu, Q. Zhang, and J. Song, "Practical evaluation of adversarial robustness via adaptive auto attack," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 105–15 114.

[22] Y. Yu, X. Gao, and C.-Z. Xu, "Lafeat: Piercing through adversarial defenses with latent features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5735–5745.

[23] S. Kuwano, S. Namba, and H. Miura, "Advantages and disadvantages of a-weighted sound pressure level in relation to subjective impression of environmental noise," *Noise Control Engineering Journal*, vol. 33, no. 3, pp. 107–115, 1989.

[24] M. E. Nilsson, "A-weighted sound pressure level as an indicator of short-term loudness or annoyance of road-traffic sound," *Journal of Sound and Vibration*, vol. 302, no. 1-2, pp. 197–207, 2007.

[25] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *International conference on machine learning*. PMLR, 2019, pp. 5231–5240.

[26] Y. Lin, W. H. Abdulla, Y. Lin, and W. H. Abdulla, "Principles of psychoacoustics," *Audio Watermark: A Comprehensive Foundation Using MATLAB*, pp. 15–49, 2015.

[27] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[28] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[29] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 694–711.

[30] T. Du, S. Ji, J. Li, Q. Gu, T. Wang, and R. Beyah, "Sirenattack: Generating adversarial audio for end-to-end acoustic systems," in *Proceedings of the 15th ACM Asia conference on computer and communications security*, 2020, pp. 357–369.

[31] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Towards robust speech-to-text adversarial attack," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2869–2873.

[32] C.-H. Yang, J. Qi, P.-Y. Chen, X. Ma, and C.-H. Lee, "Characterizing speech adversarial examples using self-attention u-net enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3107–3111.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[36] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1d convolutional neural network," *Expert Systems with Applications*, vol. 136, pp. 252–263, 2019.

華南農業大學 South China Agricultural University

技术成果申请专利（软件著作权）审批

拟申请软件著作权名　　称	轻量级低成本的鲁棒病虫害检测应用软件		
申请人	华南农业大学	申请类型	软件著作权
联 系 人	黄立峰	手机号码	13929500478
E-Mail	huangIf6@scau.edu.cn		
发明人			
第一发明人	黄立峰	性　别	男
所在单位	数学与信息学院、软件学院	职　称	讲师

姓　　名	性　别	所在单位	职　　称
梁建烨	男	数学与信息学院、软件学院	无
林晖	男	数学与信息学院、软件学院	无
林永鸿	男	数学与信息学院、软件学院	无
罗鑫	男	数学与信息学院、软件学院	无
王凯津	男	数学与信息学院、软件学院	无
吴炜馨	男	数学与信息学院、软件学院	无

支持课题情况				
课题名称	轻量级和低成本的鲁棒病虫害检测技术与Android应用开发			
课题来源及编号	大学生创新创业计划-轻量级和低成本的鲁棒病虫害检测技术与Android应用开发			
申请类别	☐发明专利　☐实用新型专利　☐外观设计专利			
代理机构名称	广州市华学知识产权代理有限公司 代理机构资质查询：http://www.acpaa.cn/view/agencySearch.jhtml			
拟申请软件著作权简　　介	开发目的： 为了帮助农户以低门槛获取精准的病虫害诊断，提高农业生产效率，解决设备成本高、网络依赖强和识别精度低的问题。 面向领域/行业： 应用于农业病虫害检测领域，为个体农户提供技术支持。 软件的主要功能： 1、实现低质量图像的高清修复，通过超分辨率重建和图像修复算法提升图像清晰度。 2、提供鲁棒的病虫害识别模型，能够在复杂环境下准确识别病虫害。 3、开发轻量级的Android离线应用，方便农户在无网络或网络较差的情况下使用。			
发明人承诺	本人将按照学校有关规定，积极配合学校知识产权办的工作，做好专利（软件著作权）申请过程中的答复等相关事宜，并及时将收到的有关通知、材料交学校知识产权办处理。			
	是否为第一发明人	◉是　○否	日　　期	2025-03-21

所在单位审核	本人代表学院对该专利（软件著作权）申请进行了审查。 学院将积极配合学校督促发明人按照学校的有关规定和工作需要做好该专利（软件著作权）的申请、宣传、转化等工作。			
		审核人	夏强	审核时间	2025-03-21

科研院业务科室科员审核	科研院已备存，无需提交纸质版专利（软件著作权）审批表；如有材料需加盖公章，请在办事服务大厅--知识产权类材料用章申请办理。

		审核人	张晓佳	审核时间	2025-03-24

科研院业务科室科长审核	已核。				
		审核人	韩雨辰	审核时间	2025-03-24

科研院分管副处长审核	同意				
		审核人	倪慧群	审核时间	2025-03-24

中华人民共和国国家版权局
计算机软件著作权登记证书

证书号： 软著登字第15519594号

软 件 名 称： 轻量级低成本的鲁棒病虫害检测应用软件
[简称：轻量级病虫害检测软件]
V1.0.0

著 作 权 人： 华南农业大学

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2025SR0863396

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

2025年05月26日

華南農業大學
South China Agricultural University

技术成果申请专利（软件著作权）审批

拟申请软件著作权名　　称	视觉模型稳健性的黑盒自适应评估系统		
申请人	华南农业大学	申请类型	软件著作权
联 系 人	黄立峰	手机号码	13929500478
E-Mail	huanglf6@scau.edu.cn		
发明人			
第一发明人	黄立峰	性　别	男
所在单位	数学与信息学院、软件学院	职　称	讲师

姓　　名	性　别	所在单位	职　　称
高子恒	男	数学与信息学院、软件学院	无
陈思儒	男	数学与信息学院、软件学院	无
陈梓泓	男	数学与信息学院、软件学院	无
何梓峰	男	数学与信息学院、软件学院	无
张喆翔	男	数学与信息学院、软件学院	无
郑汝酬	男	数学与信息学院、软件学院	无

支持课题情况	
课题名称	面向跨域数据和异构模型的迁移场景对抗攻防方法研究
课题来源及编号	粤穗联合基金青年基金 2023A1515110075
申请类别	☐发明专利　☐实用新型专利　☐外观设计专利
代理机构名称	广州市华学知识产权代理有限公司 代理机构资质查询：http://www.acpaa.cn/view/agencySearch.jhtml
拟申请软件著作权简　　介	提供了市面上分类表现较好的基于CNN或者VIT架构的视觉AI分类模型、经典的白盒对抗攻击、表现较好的黑盒对抗攻击，且每种黑盒攻击的侧重点不同、经典的对抗防御算法，且创新性地用自适应评估来方便且系统的对不同算法进行评估，来实验不同侧重点的对抗攻击更容易攻破哪一种模型，方便了研究人员的实验，并能够根据自己的需要进行修改、添加算法。可以部署到服务器。
发明人承诺	本人将按照学校有关规定，积极配合学校知识产权办的工作，做好专利（软件著作权）申请过程中的答复等相关事宜，并及时将收到的有关通知、材料交学校知识产权办处理。

是否为第一发明人	◉是　○否	日　　期	2025-04-25

所在单位审核	本人代表学院对该专利（软件著作权）申请进行了审查。 学院将积极配合学校督促发明人按照学校的有关规定和工作需要做好该专利（软件著作权）的申请、宣传、转化等工作。			
		审核人	夏强	审核时间	2025-04-27

科研院业务科室科员审核	科研院已备存，无需提交纸质版专利（软件著作权）审批表；如有材料需加盖公章，请在办事服务大厅--知识产权类材料用章申请办理。

		审核人	张晓佳	审核时间	2025-04-28

科研院业务科室科长审核	已核。				
		审核人	韩雨辰	审核时间	2025-04-28

科研院分管副处长审核	同意				
		审核人	倪慧群	审核时间	2025-04-28

中华人民共和国国家版权局
计算机软件著作权登记证书

证书号： 软著登字第15757851号

软 件 名 称： 视觉模型稳健性的黑盒自适应评估系统
[简称：视觉模型稳健性评估系统]
V1.0.0

著 作 权 人： 华南农业大学

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2025SR1101653

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

2025年06月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导林永鸿荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛全国总决赛C/C++程序设计大学B组一等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602114801

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年6月23日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导张喆翔荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛全国总决赛C/C++程序设计大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602114933

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年6月23日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导林桂塔荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛全国总决赛Java软件开发大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1605028614

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年6月23日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导马岳骏荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛全国总决赛Java软件开发大学B组优秀奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1605028577

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年6月23日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导陈文叙荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组二等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602058256

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导申昊林荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组二等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602057410

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导房君炫荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602058154

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导伏杨博荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602058152

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导黄悦佳荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602058153

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导赖成文荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602058266

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导廖伟骏荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602057694

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导杨东锦荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602058292

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导林永鸿荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组一等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602057883

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导张喆翔荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区C/C++程序设计大学B组一等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1602057421

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导林桦荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区Java软件开发大学B组二等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1605014786

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导张宝俊荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区Java软件开发大学B组二等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1605014756

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导陈早荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区Java软件开发大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1605014739

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导冯凯禧荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区Java软件开发大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1605014505

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导黄静曦荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区Java软件开发大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1605014444

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导马铭辉荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区Java软件开发大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1605014729

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导张堡焜荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区Java软件开发大学B组三等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1605014732

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导林桂塔荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区Java软件开发大学B组一等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1605014730

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

蓝桥杯大赛

获奖证书

华南农业大学黄立峰：

指导马岳骏荣获第十六届蓝桥杯全国软件和信息技术专业人才大赛广东赛区Java软件开发大学B组一等奖，被评为优秀指导教师。

特发此证，以资鼓励。

证书编号：1605014635

证件号码：431202199002010815

工业和信息化部
人才交流中心

蓝桥杯大赛组委会

2025年5月26日

华南农業大學

本科教学成果奖

获奖证书

获奖成果：国产筑基，数智赋能：卓越软件工程师人才培养模式的探索与实践

获 奖 者：王金凤、黄 琼、孙微微、张丽霞、杨 磊、张 猜、黄立峰

获奖等级：一等奖

证书编号：JXCG24020

2025年9月9日


GDCA
广东省计算机学会
Computer Academy of Guangdong
广东省计算机学会
优秀论文奖
证书
为表彰2025年度广东省计算机学会
优秀论文奖获奖者，特颁发此证书。
项目名称：FASTEN: Fast Ensemble Learning For Improved Adversarial Robustness
奖励等级：二等奖
获奖单位：华南农业大学
获 奖 者：黄立峰 黄琼 邱培超 韦舒心 高成英
项目编号：2025-J2-125
广东省计算机学会
二〇二五年十二月

GDCA

广东省计算机学会
Computer Academy of Guangdong

广东省计算机学会优秀论文奖

证 书

为表彰2024年度广东省计算机学会优秀论文奖获奖者，特颁发此证书。

项目名称：Erosion attack: Harnessing corruption to improve adversarial examples

奖励等级：二等奖

获奖单位：华南农业大学

获 奖 者：黄立峰 高成英 刘宁

项目编号：2024-J2-102

广东省计算机学会
二〇二四年十二月

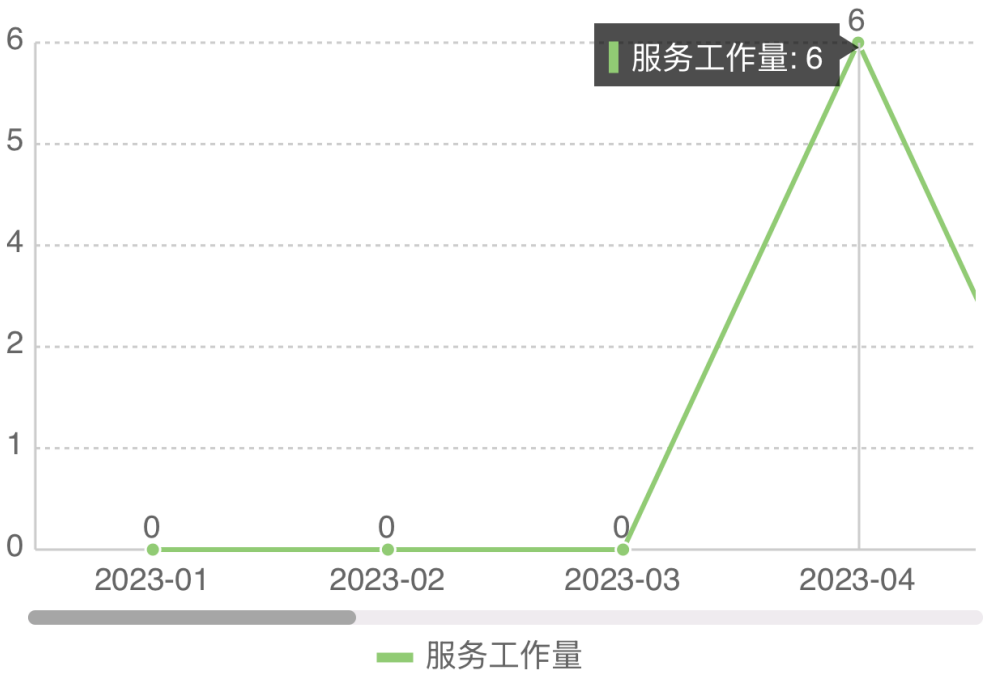

GDCA
广东省计算机学会
Computer Academy of Guangdong
2025年广东省计算机学会
教育教学成果奖
（高等教育）
获奖证书
获奖成果：国产筑基 数智赋能：卓越软件工程师人才培养模式的探索与实践
获 奖 者：王金凤 黄琼 孙微微 张丽霞 杨磊 张猜 黄立峰
获奖等级：一等奖
获奖单位：华南农业大学
证 书 号：GDCA2025GJ1-022
广东省计算机学会
2025年12月

17:35 5G 47

个人绩效统计

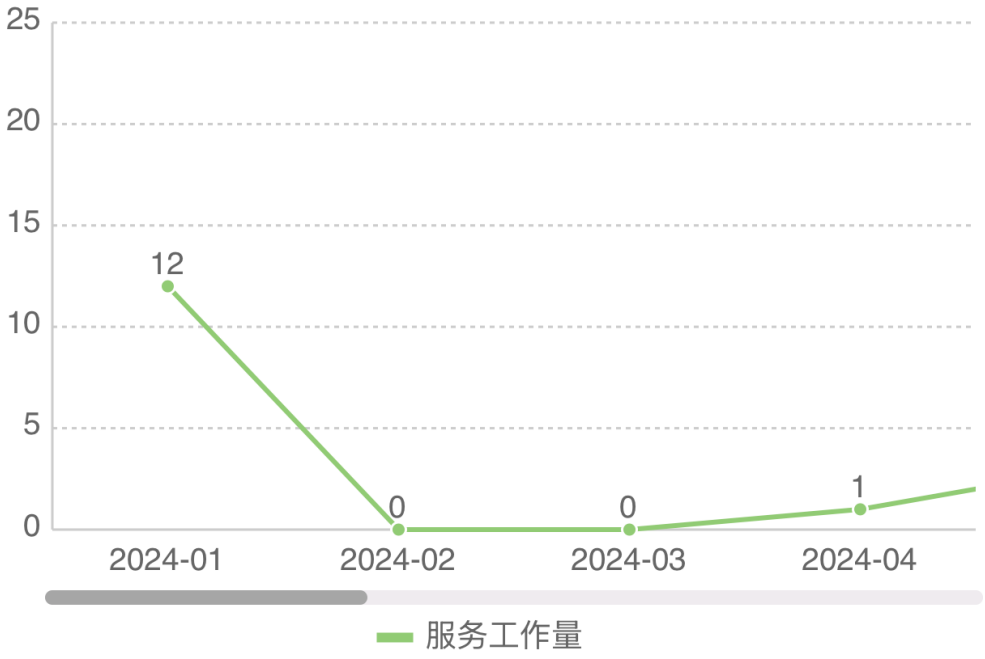
查询月份 2023-01 – 2023-12 确定

6	0	0	0
总工作量	开展培训/讲座	发放技术资料数	培训人数



查询月份 2024-01 – 2024-12 确定

62	6	2	208
总工作量	开展培训/讲座	发放技术资料数	培训人数



查询月份 2025-01 – 2025-12 确定

19	1	0	0
总工作量	开展培训/讲座	发放技术资料数	培训人数