基于多视角置信度融合的对抗样本迁移性提升方法

赵畅菲[1],邓鑫洋[1],蒋 雯[1*],朱金彪[2],耿 杰[1]
(1. 西北工业大学电子信息学院,陕西西安 710129;2. 中国科学院空天信息创新研究院,北京 100094)

摘 要: 对抗攻击揭示了深度学习模型的脆弱性,有效的对抗攻击方法有助于发现模型的潜在漏洞。现有的梯度对抗攻击方法过度拟合受攻击的白盒模型特性,对黑盒模型的迁移攻击性能较差。针对黑盒模型开展迁移对抗攻击研究,提出了一种基于多视角置信度融合的对抗样本迁移性提升方法,并作为通用模块嵌入到基于梯度的对抗攻击过程,以提升对抗样本的迁移性。具体而言,设计了基于双像素空间的多视角变换策略,引导模型在不同通道与空间尺度下感知图像信息,从而扩充模型的重点关注区域,针对图像形成差异化的关注分布,实现对图像信息的多视角感知;为建模视角间的冲突与不确定性,利用证据理论框架提出了基于冲突感知的置信度融合方法,从模型多个视角下的输出置信度提取预测的共性信息,避免视角特异性带来的决策干扰,有效提升模型多视角决策融合的可靠性;设计了一种双向损失优化函数,优化对抗样本偏离正确的模型决策边界,引导其处于跨视角、跨模型共享的脆弱区域,从而提升对黑盒模型的迁移攻击能力。实验表明,本文方法在跨模型架构攻击场景下能够有效提升对抗样本的迁移性,现有梯度对抗攻击组合多视角置信度融合方法后,对常规训练的卷积神经网络(Convolutional Neural Network,CNN)和Transformer架构模型的迁移攻击成功率平均提升了21.15%和13.02%,对防御模型的迁移攻击成功率平均提升了13.84%,对集成模型的迁移攻击成功率平均提升了16.14%。

关键词: 深度学习;对抗攻击;图像分类;迁移性;置信度
基金项目: 国家自然科学基金(No.62573350);陕西省重点研发计划(No.2024QY2-GJHX-05)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2026)03-1062-16
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20260011

A Method for Enhancing the Transferability of Adversarial Examples Based on Multi-Perspective Confidence Fusion

ZHAO Changfei[1], DENG Xinyang[1], JIANG Wen[1*], ZHU Jinbiao[2], GENG Jie[1]
(1. *School of Electronics and Information*, *Northwestern Polytechnical University*, *Xi'an*, *Shaanxi* 710129, *China*;
2. *Aerospace Information Research Institute*, *Chinese Academy of Sciences*, *Beijing* 100094, *China*)

Abstract: Adversarial attacks expose the vulnerabilities of deep learning models, and effective adversarial attack methods aid in uncovering potential weaknesses. Existing gradient-based adversarial attacks overfit the characteristics of attacked white-box models, resulting in poor transferability for black-box models. This paper investigates transferable adversarial attacks for black-box models, proposing a multi-perspective confidence fusion-based method to enhance the transferability of adversarial examples. This approach is integrated as a universal component into gradient-based adversarial attack processes to improve transferability. Specifically, a multi-perspective transformation strategy based on dual-pixel space is designed to guide the model in perceiving image information across different channels and spatial scales, thereby expanding the model's areas of focus, generating a differentiated attention distribution for the image, and enabling multi-view perception of image information. To model conflicts and uncertainties across different perspectives, a conflict-aware confidence fusion method based on the evidence theory framework is proposed. This method extracts common predictive information from the confidence outputs of the model across multiple perspectives, thereby avoiding decision-making interference caused by perspective-specific biases and effectively enhancing the reliability of multi-perspective decision fusion. A bidirectional loss optimization function is designed to optimize the deviation of adversarial examples from the correct model decision boundary, guiding them to lie in the shared vulnerable regions across different views and models, thereby improving the transfer attack performance against black-box models. Experiments show that the proposed method can effectively improve the transferability of adversarial examples in cross-model architecture attack scenarios. After integrating the existing gradient-based adversarial attacks with the multi-perspective confidence fusion method, the transfer attack success rate is improved by an average of 21.15% and 13.02% for conventionally trained convolutional neural networks (CNNs) and Trans-

收稿日期:2026-02-11;录用日期:2026-03-19;责任编委:唐旭
*通信作者:蒋雯

former models, respectively, by 13.84% for defense models, and by 16.14% for ensemble models.

Keywords: deep learning; adversarial attack; image classification; transferability; confidence score

Foundation Item(s): National Natural Science Foundation of China (No.62573350); Key Research and Development Program of Shaanxi (No.2024QY2-GJHX-05)

0 引言

深度学习模型具备强大的特征学习与复杂问题建模能力,被广泛应用于目标识别[1-2]、语义分割[3]、图像生成[4]等领域,成为赋能各行业智能化转型与升级的技术支撑。然而,研究表明,通过对输入样本施加人类难以察觉的精细扰动,可形成具有攻击能力的对抗样本,使深度学习模型以高置信度产生错误的推理结果[5-6]。凭借对深度模型的误导能力,对抗样本在无人机导航避障、飞行器物理一致性验证[7]等领域得到了广泛应用。一方面,对抗攻击揭示了深度学习模型的脆弱性,使模型给出错误的预测结果。另一方面,有效的对抗攻击方法有助于发现模型的潜在漏洞。从以攻促防的角度来看,研究对抗样本的生成方法有利于模型的安全性评估与鲁棒性提升,对于模型的安全部署与应用具备积极意义。

依据攻击者是否能够获取模型的结构与参数信息,对抗攻击大致可分为白盒攻击[8-10]和黑盒攻击[11]两类。在白盒场景下,攻击者完全掌握受攻击模型的结构与参数信息,可利用模型推理过程中的梯度信息进行对抗样本的生成和优化。而黑盒场景下,攻击者仅能获取受攻击模型的输入与输出,常通过迁移攻击[12]或查询攻击[13]方式实现对黑盒模型的决策误导。其中,迁移攻击是黑盒攻击场景中最具实用性的方案,该技术利用已知的白盒模型生成对抗样本,基于不同模型间决策边界的相似性,实现对黑盒模型的对抗攻击。

当前,迁移攻击大都采用基于梯度的攻击策略,通过计算模型对输入样本的梯度信息,确定最大化模型预测误差的扰动方向,再以该方向为指引生成对抗样本。Dong等人[8]提出了动量迭代法(Momentum Iterative Method,MIM),该算法在生成对抗样本时考虑历史扰动的累积效应,通过最大化目标损失函数,并使用动量策略来更新扰动,缓解了对抗样本对白盒模型的过拟合现象。Xiong等人[14]提出的随机方差缩减集成攻击(Stochastic Variance Reduced Ensemble attack,SVRE)将迭代集成攻击视为一个随机梯度下降优化过程,在外循环计算批次模型的平均梯度,内循环计算随机模型的当前梯度,避免了求解过程的局部最优现象。基于集成梯度在模型间表现出强相似性,Ma等人[15]提出了动量积分梯度(Momentum Integrated Gradients,MIG),以动量方式将不同迭代阶段得到的积分梯度结合进而得到累积梯度,并利用其符号对扰动进行迭代修正。这些方法在一定程度上提升了对抗样本的迁移性,增强了对黑盒目标模型的攻击性能。

然而,上述研究未能引导模型探索潜在判别区域,导致模型对输入的重点关注区域固化,难以感知不同视角下决策边界的差异,从而限制了对抗样本迁移性的有效提升。模型的注意力高度集中于固定区域,对其他潜在判别信息关注不足,这种单一视角下获得的梯度信息仅刻画了模型在输入样本特定特征空间的脆弱性,扰动趋向于拟合该模型的局部决策模式。与之相比,若为模型引入多个关注视角,则有助于多维度探索模型决策边界,挖掘普适的扰动更新方向,使生成的扰动更贴近不同模型共有的对抗子空间[16]。然而,模型以不同的关注视角提取输入信息会导致输出的置信度存在差异,使得预测类别间的冲突难以被精准识别,模型决策的不确定性也无法得到充分表征。因此,需要构建一种高效的置信度融合机制,有效整合不同视角下模型预测的判别差异,从而捕捉跨视角的一致性脆弱特征,提升对抗样本的迁移性。

针对梯度对抗攻击方法存在的视角探索不足、跨模型扰动普适性差等问题,本文提出一种基于多视角置信度融合的对抗样本迁移性提升方法,引导模型感知更具多样性的图像主体区域,从而生成具备跨模型适用性的对抗样本。具体来说,首先利用色相-饱和度-明度(Hue,Saturation,Value,HSV)颜色空间的通道解耦特性以及红绿蓝(Red,Green,Blue,RGB)空间的卷积契合特性,对输入图像进行双像素空间增强变换,提升模型对图像关注范围的多样性。变换后的图像输入到模型生成多个视角的置信度输出。其次,基于视角间的冲突与一致性,对变换后的图像输入到模型进行置信度融合,通过显式建模不同视角间的预测冲突与不确定性,抑制视角特异性,优先保留被多视角共同认可的扰动成分。最后,基于融合后的置信度信息构建双向损失函数,引导对抗样本的更新优化。其中损失函数由两部分组成:一方面,最小化正确类别的输出置信度,引导对抗样本偏离正确的模型决策边界;另一方面,最大化错误类别的输出置信度,增强模型在非真实类别方向上的判别倾向,从而推动扰动逼近跨视角

与跨模型共享的脆弱区域。通过以上设计的多视角置信度融合方法,引导模型全面感知图像信息,提升对抗扰动跨视角、跨模型的适用性,从而增强对抗样本的迁移攻击能力。此外,提出的方法能够作为即插即用模块嵌入现有的梯度攻击框架,具备良好的兼容性。

本文的主要贡献包括以下三个方面:

(1)提出基于双像素空间的多视角变换策略,通过对输入图像进行HSV空间通道级全局增强和RGB空间区域差异化增强变换,提升模型对图像目标关注范围的多样性。

(2)提出基于冲突感知的置信度融合方法,通过量化不同变换视角间的置信度冲突,有效处理多视角间的预测不一致性,从而整合多视角共性信息,自适应融合多视角置信度。

(3)针对CNN模型、Transformer模型、防御模型及集成模型进行系统评估,实验结果表明提出的置信度融合方法具备提升现有梯度对抗攻击迁移性的显著效能。

1 相关工作

1.1 白盒对抗攻击

白盒对抗攻击依托对目标模型结构、参数及梯度流的完全知情优势,构建了对抗样本生成的精准优化范式。早期研究聚焦梯度信息的直接利用,Szegedy等人[17]首次通过有限内存布罗伊登-弗莱彻-戈德法布-香农算法(Limited memory Broyden Fletcher Goldfarb Shanno,L-BFGS)优化框架求解带约束的扰动生成问题,从优化理论层面证实深度模型在输入空间存在鲁棒性漏洞。但该方法依赖二次规划求解,计算复杂度随输入维度呈指数增长,难以满足实时性需求。Goodfellow等人[18]提出了快速梯度符号法(Fast Gradient Sign Method,FGSM),通过梯度符号的方向近似替代梯度幅值,显著降低了攻击复杂度,但该单步扰动策略存在梯度离散化误差,存在视觉可感知的像素突变。

针对梯度优化的精度与效率矛盾,后续研究者从优化理论层面提出多类改进方案。Kurakin等人[19]提出了迭代快速梯度符号法(Iterative Fast Gradient Sign Method,I-FGSM),通过小步长多轮迭代实现梯度逼近,但该方法易因梯度方向震荡陷入局部最优。Li等人[20]基于贝叶斯框架推导提出原理性策略,与深度神经网络参数高斯后验近似方法结合实现微调,该从而引导对抗样本的生成。此外,研究发现,同时利用多个白盒模型的梯度信息进行对抗扰动的优化,可以缓解对抗样本对单个模型的过拟合现象。这类基于集成模型的方法通常采用多个不同架构或训练方式的模型作为替代模型,在攻击时综合它们的损失梯度或决策边界生成扰动。Chen等人[21]通过监测多个替代模型对抗目标贡献的差异比率,动态调控其输出融合权重,并引入差异度降低滤波器同步更新方向,实现了多个白盒模型的集成攻击。Liu等人[16]最小化对抗样本在多个模型输出逻辑上的差异来生成扰动,通过优化对抗样本使其在集成模型的预测分布上保持一致。

1.2 黑盒对抗攻击

1.2.1 基于迁移的黑盒对抗攻击

基于迁移的对抗攻击基于对抗样本的跨模型迁移性,通过在可获取的白盒替代模型上生成对抗样本,利用深度模型间特征提取机制的共性,实现对黑盒目标模型的间接攻击,核心逻辑是以白盒模型的梯度信息替代黑盒模型的未知信息。受更平坦的局部极小值可提升迁移性的观察启发,梯度聚合攻击(Gradient Aggregation Attack,GAA)[22]将反映对抗样本邻域内最大损失值的最差意识损失与量化最差意识损失与经验损失差异的替代损失同时纳入优化目标,从而在同步增强区域平坦度的同时在该区域生成对抗样本。强化动量攻击(Enhanced Momentum Attack,EMI)[12]对累加迭代过程中梯度的优化策略进行改进,在前一次迭代的梯度方向上对采样数据点的平均梯度进行累加,以稳定更新方向,避免局部极值。全局动量初始化(Global Momentum Initialization,GMI)[23]首先利用集成梯度计算模型预测对输入的显著性评分,然后通过积分梯度引导扰动的更新过程,避免模型特定噪声所导致的次优解。考虑到具有高迁移性的对抗样本通常携带多类特征,Foolmix[24]提出基于随机像素块的图像混合和基于标签损失的梯度混合策略,并利用基于初始正向方向的更新方法,迫使对抗样本穿透多个类别区域并在隐空间获取多类特征,提升生成对抗样本的可迁移性。

部分迁移对抗攻击方法引入输入变换,在将样本输入模型前对其进行数据增强操作,避免扰动优化陷入局部最优解。块混洗与旋转攻击(Block Shuffle and Rotation attack,BSR)[25]方法首先对输入图像进行分块操作,然后随机地对这些块进行洗牌和旋转,以构造一组新的图像用于梯度计算,通过破坏图像的内在联系进而改变原始图像的注意力热图,达到提高可迁移性的效果。AdMix[26]将输入图像作为主成分、其他图像作为次成分进行混合,同时使用输入的原始标签执行梯度反向传播,并提出利用批次混合的聚合平均梯度来生成可迁移的对抗样本。混合频率输入

(Mixed-Frequency Inputs,MFI)[27]从频域视角提出一种混合输入策略,该策略将高频成分融入梯度迭代过程以抑制过拟合,借助高频信息的累积有效稳定梯度表示,从而引导攻击搜索趋向更优局部极小值,增强对抗样本的跨架构迁移能力。算子扰动随机优化(Operator Perturbation-based Stochastic optimization,OPS)[28]基于模型泛化能力与迁移能力之间的镜像关系,引入了变换算子和随机扰动建立随机优化范式,通过求解该优化问题引导具备泛化性的对抗样本生成。

现有的迁移攻击方法无论是通过优化算法改进梯度搜索路径,还是结合输入变换策略提升样本多样性,其模型输入视角通常仅局限于原始图像尺度或其单一的变形版本。这种单视角的处理方式忽略了在不同观测视角下,模型输出信息之间的内在关联,生成的对抗样本易陷入源模型特定视角下的局部最优解。因此,需要融合多观测视角下的模型决策信息,挖掘不同视角间的内在关联,打破单视角局限,引导对抗样本生成跳出源模型特定视角的局部最优解,提升迁移攻击的泛化性能。

1.2.2　基于查询的黑盒对抗攻击

基于查询的对抗攻击通过与黑盒模型的主动交互,构建模型决策边界的近似估计,再基于近似信息生成针对性对抗样本,核心逻辑是以有限查询代价换取黑盒模型的决策规律认知。Ma等人[29]提出利用有限差分方法构建梯度估计框架,首先对输入图像的像素特征进行局部扰动,依据模型输出反馈分析特征变化与模型输出的关联,进而近似估计模型梯度方向以生成对抗样本。Yin等人[30]在良性样本条件下构建元学习框架,首先利用元生成器基于新任务反馈信息及少量历史攻击数据快速微调,并利用模型级对抗可迁移性在白盒替代模型上训练元生成器,从而针对新的样本生成有效扰动。基于圆几何性质的黑盒攻击(Black box attack based on circular geometric properties,CBA)[31]利用离散余弦变换将攻击位置选在频率空间,基于低频信息在决策边界附近的思想,利用圆的几何性质不断迭代,获取低频空间中的对抗样本,最后经逆离散余弦变换转换回输入空间。该方法避免梯度估计,能够在保证攻击成功率的同时显著降低查询次数。

然而,基于查询的对抗攻击存在查询开销过高的固有缺陷。由于需要通过反复向目标模型发起黑盒查询获取反馈信息、迭代优化对抗样本,此类攻击难以适配实时性需求较高的场景;此外,高频率查询易形成可被识别的异常访问模式,对于配置了访问行为监测或异常检测模块的智能系统而言,此类攻击极易被精准识别并触发防御拦截,大幅降低攻击成功率。

2　本文方法

梯度对抗攻击普遍受限于梯度计算对模型当前决策边界的强依赖性,攻击过程中模型对输入图像的关注区域呈现明显的固化现象。由于模型往往仅聚焦于梯度响应强度较高的局部像素区域,忽略了图像中其他蕴含关键语义信息的全局特征,导致对图像信息的提取缺乏全面性与完整性。这种局部化的信息处理模式,使得生成的对抗样本过度拟合于白盒模型的局部决策空间,难以与其他黑盒模型的决策边界相匹配,导致生成的对抗样本迁移性不足。针对梯度对抗攻击存在的视角探索局限、扰动易拟合模型局部决策模式的问题,本文提出基于多视角置信度融合的对抗迁移性提升方法,通过优化生成具备跨视角、跨模型普适性的对抗扰动,提升对抗样本的迁移性。提出方法的总体流程图如图1所示,主要包括多视角变换模块、置信度融合模块,以及一种双向损失设计。在多视角引导模块中,分别在HSV空间和RGB空间对输入图像进行通道级全局增强和区域差异化增强变换,提升模型对图像目标关注范围的多样性。变换后的图像输入到模型进而获得对应的置信度分布。在置信度融合模块中,每个视角下模型的置信度输出可以看作是一个独立的信息源。为了综合来自不同视角的信息,采用基于冲突感知的置信度融合机制将这些信息源进行合并,根据视角之间的差异性抑制视角特异性,优先保留被多视角共同认可的扰动成分。最后,基于融合置信度设计双向损失并执行梯度反向传播,带有判别倾向地引导对抗样本偏离正确的模型决策边界,增强对抗样本对黑盒模型的攻击效果,提高其在多个模型之间的迁移性。

2.1　基于双像素空间的多视角变换

基于梯度的对抗攻击方法存在明显的迁移性差的现象,这一问题的核心在于,攻击过程中生成的扰动与模型的判别特征高度相关,而模型通常在输入图像中聚焦于有限的主体区域,因此生成的扰动难以覆盖更多潜在的判别信息。已有工作尝试通过输入变换来提升对抗样本的迁移性,但这些方法无法显著改变模型的关注区域,因而对丰富模型提取的特征空间信息的贡献有限。基于此,本文提出一种基于双像素空间的多视角变换方法。该方法利用HSV颜色空间对色彩与亮度的解耦特性以及RGB空间结构与卷积操作的契合特性,通过通道级全局增强及区域差异化增强,引导模型在更广的区域和更丰

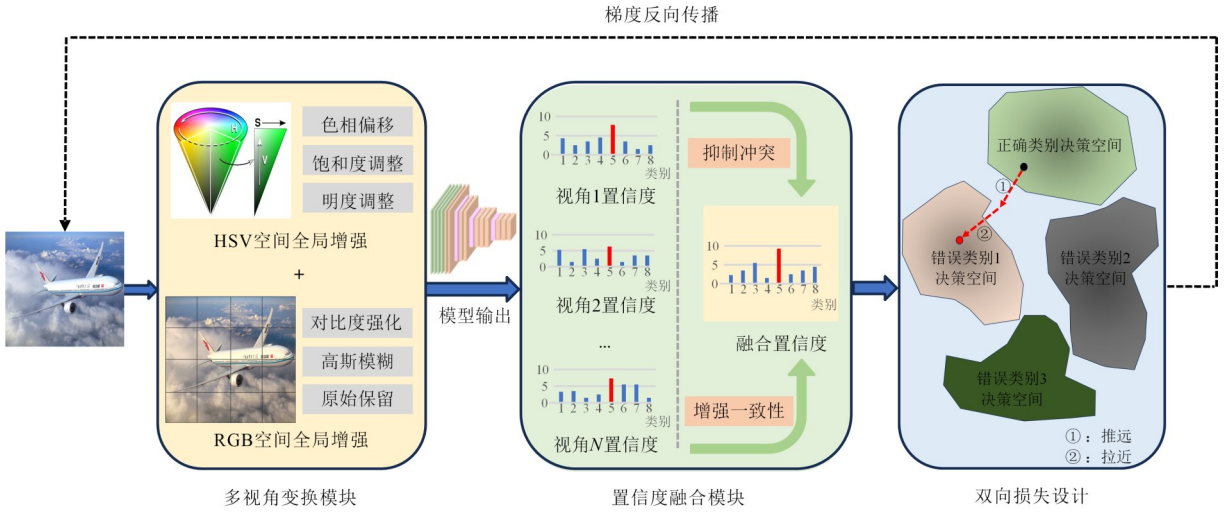


图1 基于多视角置信度融合的对抗迁移性提升方法的总体流程图

Figure 1 Overall flowchart of the method for enhancing the adversarial transferability based on multi-perspective confidence fusion

富的属性上进行感知，从而帮助模型在不同视角下做出决策，缓解对抗样本对白盒模型的过拟合问题，进而有效提升迁移性。

2.1.1 HSV空间通道级全局增强

HSV像素空间由三个分量组成，其中，色调（Hue，H）反映颜色的基本属性，通常以角度表示，范围为0~360°；饱和度（Saturation，S）表示颜色的纯度或鲜艳度，范围为[0, 1]；明度（Value，V）描述颜色的亮度或强度，范围为[0, 1]。HSV空间通过将颜色信息与亮度信息显式分离，使得模型能够对不同维度的变化建立更稳定的响应。Van De Weijer等人[32]通过卷积神经网络的实验指出，网络在处理HSV输入时，能够在H和S通道中提取出更具判别力的边界与区域特征。此外，HSV空间中在特征解耦方面优异，能够将物体的不同属性在特征空间中独立分开，且对不同架构的模型具备通用性。

通过在HSV空间进行通道级的全局增强，可以独立地对色彩偏移、饱和度变化和光照强度进行调整，从而生成具有多样化色彩组合和亮度条件的图像。这种全局一致性的增强既保持了图像语义主体的一致性，又能够在颜色与亮度的维度上引入丰富的变化，使模型关注到更多潜在的特征模式，避免过度依赖单一颜色分布或固定亮度条件，从而提升特征提取的全面性。因此，本文在HSV空间对图像进行通道级的全局增强操作。

对于图像的每个HSV通道，生成统一增强参数，且单个通道内所有像素共享同一参数，不同通道参数独立。增强参数均从调整参数ψ控制的均匀分布中采样，确保增强多样性与合理性：

$$\alpha_H, \alpha_S, \alpha_V \sim U(0, 1+\psi) \tag{1}$$

色相参数α_H控制色相偏移程度，色相偏移后需保持在0°~360°，故对结果取模：

$$H_{\text{aug}}(h,w) = \left(H(h,w) \times \alpha_H\right) \bmod 360^{\circ} \tag{2}$$

饱和度参数α_S和明度参数α_V分别控制饱和度的强弱和亮度的高低，且调整后的数值均需保持在[0, 1]，采用截断函数限制调整后的数值范围：

$$S_{\text{aug}}(h,w) = \text{clip}\left(S(h,w) \times \alpha_S, 0, 1\right) \tag{3}$$

$$V_{\text{aug}}(h,w) = \text{clip}\left(V(h,w) \times \alpha_V, 0, 1\right) \tag{4}$$

其中，clip(x, a, b)为截断函数，当$x<a$时取a，$x>b$时取b，否则取x。增强后得到HSV空间图像$I_{\text{hsv,aug}} = \{H_{\text{aug}}, S_{\text{aug}}, V_{\text{aug}}\}$，将其转回RGB空间得到$I_{\text{rgb,aug}} \in \mathbf{R}^{H \times W \times 3}$，以便后续RGB空间的图像处理。

2.1.2 RGB空间通道级全局增强

RGB空间能够直接保留图像的空间结构特征，每个通道的值对应于原始成像设备采集的强度信号，三通道在空间上严格对齐，因此卷积神经网络能够在局部感受野内直接捕捉邻域之间的纹理、边缘和形状模式。研究表明，神经网络的早期卷积层对RGB输入能够敏锐地捕捉边缘和局部对比度模式[33]，在低层卷积核中往往优先学习边缘和局部结构特征，而这些特征与RGB通道的空间强度梯度密切相关[34]。在实际任务中，RGB空间常被用作局部化操作的基础。例如，基于RGB的局部对比度增强和模糊处理被用于调整模型的注意力分布，帮助其从更广的区域提取判别信息[35]。这些研究共同表明，RGB空间在保留图像空间结构和支持局部特征建模方面具有显著优势。

与HSV空间适合进行通道级全局增强不同，RGB空间更直接地保留了图像的空间结构特征，因而更适

合在空间维度上进行局部化操作。卷积神经网络在处理RGB输入时往往表现出较强的局部依赖性,即模型容易集中关注于目标主体的显著区域,而忽视图像中其他潜在的判别信息。RGB空间直接对应模型卷积层的输入通道,其空间结构与卷积操作高度契合。在该空间进行区域级差异化增强,通过网格化划分图像并在不同区域施加差异化操作,能够有效打破模型对有限局部判别区域的依赖,迫使其在更广的空间范围内重新分配注意力。因此,本文在RGB空间执行区域级的差异化增强操作。

将 $I_{\text{rgb, aug}}$ 均匀划分为 $G\times G$ 个非重叠区域,同时确保每个区域包含足够像素以保留细节。设每个区域的高度和宽度为

$$h_{\text{cell}}=\left\lfloor\frac{H}{G}\right\rfloor,\ w_{\text{cell}}=\left\lfloor\frac{W}{G}\right\rfloor \tag{5}$$

第 i 行第 j 列区域 $(i,j\in[1,G])$ 的像素坐标范围为

$$\begin{aligned} h&\in\left[(i-1)\times h_{\text{cell}}+1,\min\left(i\times h_{\text{cell}},H\right)\right]\\ w&\in\left[(j-1)\times w_{\text{cell}}+1,\min\left(j\times w_{\text{cell}},W\right)\right]\end{aligned} \tag{6}$$

通过上述划分得到区域集合 $R=\{R_{i,j}\mid 1\leqslant i,j\leqslant G\}$,每个区域 $R_{i,j}\subseteq I_{\text{rgb, aug}}$。

从 R 中随机选择三类区域,比例为1∶1∶1,确保增强多样性且不破坏图像完整性。

(1)对比度强化区域 R_{enh}:随机选择 $G^2/3$ 个区域,用于增强细节以引导模型关注。

(2)高斯模糊区域 R_{blur}:从剩余区域中随机选择 $G^2/3$ 个区域,用于降低局部依赖。

(3)原始保留区域 R_{raw}:最后剩余的 $G^2/3$ 个区域保持不变,确保图像信息完整。

针对对比度强化区域 R_{enh},通过区域均值调整增强细节。首先计算区域 $R_{i,j}$ 的RGB通道均值:

$$\mu_{i,j,c}=\frac{1}{\left|R_{i,j}\right|}\sum_{(h,w)\in R_{i,j}} I_{\text{rgb, aug}}(h,w,c),\ c\in\{R,G,B\} \tag{7}$$

其中,$\left|R_{i,j}\right|$ 为区域 $R_{i,j}$ 的像素总数。然后,增强区域像素值:

$$I_{\text{enh}}(h,w,c)=\mu_{i,j,c}+\left(I_{\text{rgb, aug}}(h,w,c)-\mu_{i,j,c}\right)\times e \tag{8}$$

其中,e 为对比度强化系数。最后,对溢出像素进行截断:

$$I_{\text{enh, clip}}(h,w,c)=\text{clip}\left(I_{\text{enh}}(h,w,c),0,1\right) \tag{9}$$

针对高斯模糊区域 R_{blur},通过3×3高斯核平滑区域,降低局部过度依赖。首先定义3×3高斯核:

$$\boldsymbol{Q}=\frac{1}{16}\times\begin{bmatrix}1&2&1\\2&4&2\\1&2&1\end{bmatrix} \tag{10}$$

然后对区域 $R_{i,j}$ 进行卷积模糊:

$$I_{\text{blur}}(h,w,c)=\sum_{p=1}^{3}\sum_{q=1}^{3}\boldsymbol{Q}(p,q)\times I_{\text{rgb, aug}}(h-p+2,w-q+2,c) \tag{11}$$

HSV结合RGB空间的增强实现了互补效应:HSV全局增强从颜色和亮度维度引入多样性,改变了模型对输入的整体表征方式;RGB区域差异化增强从空间维度打破局部依赖,扩展了模型的关注范围。二者结合,使模型能够在不同通道与不同空间尺度下感知更丰富的图像信息,以不同的视角对输入进行决策判断。图2展示了经主流图像增强方法、双像素空间变换后的图像,以及CNN模型对这些图像的预测类别和热力图。可以看到,现有的图像增强方法并不能改变CNN模型对图像目标的关注区域,模型的关注度仍集中在目标的局部关注区域。而本文提出的方法将模型的重点关注区域扩充至目标的更广范围,这意味着双像素空间变换能够引导模型在更广的区域上进行感知,以更加多样化的视角提取图像信息。

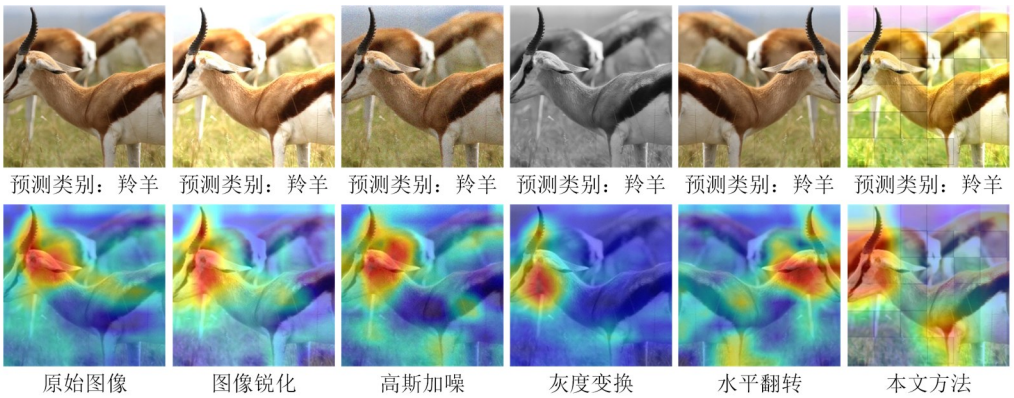


图2　经数据增强后的图像(第一行)及对应的模型热力图(第二行)

Figure 2　Data-enhanced images (first row) and corresponding model heatmaps (second row)

2.2 基于冲突感知的置信度融合

上节提出的基于双像素空间的多视角变换方法使得模型能够从不同角度捕捉图像中的关键信息，从而针对变换后的图像输出具有差异化视角的置信度。然而，不同视角下的置信度存在差异甚至冲突，部分视角特异性的预测结果会干扰最终决策的准确性。为有效解决多视角置信度的冲突问题，实现合理的置信度融合，本文引入证据理论作为融合工具。证据理论作为一种不确定性推理方法，能够有效处理具有未知性、不确定性和冲突性的信息，通过将每个视角的置信度视为一条独立证据，利用基本概率分配、证据组合规则等核心机制，对多源冲突证据进行有效融合，为多视角置信度融合提供实现路径。

具体来说，本文通过证据建模生成与Dempster组合规则融合等过程，提出一种基于冲突感知的置信度融合方法，从多视角置信度输出中提取模型预测的共性信息，避免视角特异性带来的决策干扰，确保预测结果的可靠性。

对于同一张输入图像，重复多次双像素空间增强变换，每次变换生成一个新的视角。这些变换增加了图像的多样性，相当于在输入空间的邻域采样。设输入图像为x_0，应用N次视角变换，生成N张图像：

$$x_i = T_i(x_0) \tag{12}$$

其中，$i \in \{1,2,\cdots,N\}$代表视角的次数；$T(\cdot)$代表双像素空间增强变换。生成的图像及原始图像均作为模型的输入。

模型对于输入图像的预测可以看作一个响应。假设有N个变换视角及原视角生成的输出，每个模型的logits层输出对应一个置信函数$\mathrm{CS}(\theta_j)$，表示对于类别θ_j的置信度，且$j \in \{1,2,\cdots,K\}$，K为总类别数。x_i输入到模型进行预测，得到类别概率分布$P(\theta_j|x_i)$，为满足基本概率分配的非负性，进行以下操作：

$$CS_i(\theta_j) = \mathrm{ReLU}\left(P(\theta_j|x_i)\right) \tag{13}$$

其中，ReLU函数对置信度进行截断处理。其次，计算归一化因子并构建单个类别上的基本概率分配：

$$S_i = \sum_j \left(\mathrm{CS}_i(\theta_j) + 1\right) \tag{14}$$

$$\mathrm{CS}_i(\theta_j) = \frac{\mathrm{CS}_i(\theta_j)}{S_i} \tag{15}$$

同时，引入全集来表征该视角下的全局不确定性，从而实现完整的证据建模生成。在这里，该视角的不确定性定义为

$$u_i = \frac{K}{S_i} \tag{16}$$

在进行多源置信度融合时，设有两个来源的证据$\left(\mathrm{CS}_a(\theta_j), u_a\right)$和$\left(\mathrm{CS}_b(\theta_j), u_b\right)$，则类别$\theta_j$的融合置信度为

$$\mathrm{CS}_{\mathrm{ab}}(\theta_j) = \frac{\mathrm{CS}_a(\theta_j)\mathrm{CS}_b(\theta_j) + \mathrm{CS}_a(\theta_j)u_b + \mathrm{CS}_b(\theta_j)u_a}{1 - C_{\mathrm{ab}}} \tag{17}$$

其中冲突度为

$$C_{\mathrm{ab}} = \sum_{i \neq j} \mathrm{CS}_a(\theta_i)\mathrm{CS}_b(\theta_j) \tag{18}$$

对应的融合后全局不确定性为

$$u_{\mathrm{ab}} = \frac{u_a u_b}{1 - C_{\mathrm{ab}}} \tag{19}$$

若进一步引入第三个来源$\left(\mathrm{CS}_c(\theta_j), u_c\right)$，则融合结果为

$$\mathrm{CS}_{\mathrm{abc}}(\theta_j) = \frac{\mathrm{CS}_{\mathrm{ab}}(\theta_j)\mathrm{CS}_c(\theta_j) + \mathrm{CS}_{\mathrm{ab}}(\theta_j)u_c + \mathrm{CS}_c(\theta_j)u_{\mathrm{ab}}}{1 - C_{\mathrm{abc}}} \tag{20}$$

其中，

$$C_{\mathrm{abc}} = \sum_{i \neq j} \mathrm{CS}_{\mathrm{ab}}(\theta_i)\mathrm{CS}_c(\theta_j) \tag{21}$$

全局不确定性为

$$u_{\mathrm{abc}} = \frac{u_{\mathrm{ab}} u_c}{1 - C_{\mathrm{abc}}} \tag{22}$$

依此类推，对于N个视角，最终类别的融合置信度可以递归定义为

$$\mathrm{CS}_{\mathrm{fused}}\left(\theta_j\right) = \mathrm{CS}_{1,2,\cdots,N}\left(\theta_j\right) \tag{23}$$

$\left(\mathrm{CS}_{1,2,\cdots,n}, u_{1,2,\cdots,n}\right)$表示前$n$个视角融合的中间结果，通过上述两两融合规则迭代计算得到。

通过以上的置信度建模与融合过程，一方面，当不同视角在同一类别上的预测结果趋于一致时，该类别的支持力度会得到放大。这意味着视角间的一致性将被充分利用，模型在融合后的结果中能够更强烈地体现共同认可的类别，从而提升决策的稳定性与可靠性。另一方面，当不同视角的预测结果存在分歧时，冲突量会显著增加，从而削弱直接相互矛盾的置信度贡献。这种机制有效避免了单一视角的偏差对最终结果的主导作用，使得模型在面对不同视角时能够给出更为稳健的决策，从而利用处理得到的共性信息引导生成具备视角普适性的对抗样本。

2.3 双向损失函数设计

得到多视角融合置信度之后，需要利用模型损失执行梯度反向传播，从而对对抗样本进行优化。现有梯度对抗攻击普遍采用交叉熵损失，该损失依赖于单一样本输入的模型预测，缺乏针对多个视角决策边界的综合感知，从而限制了对抗扰动的普适性。针对交

叉熵损失在多视角融合置信度中的局限性,本文设计一种双向损失函数,通过最大化错误类别的置信度并最小化正确类别的置信度,引导对抗样本偏离正确的模型决策边界,来生成更具攻击性的对抗样本。具体来说,降低正确类别的置信度以减少模型对正确类别的信心,通过错误类别置信度的最大化,使对抗样本偏向于具备错误类别的数据属性,使模型产生错误的预测的同时,引导对抗样本处于跨视角、跨模型共享的脆弱区域,从而提高对抗样本的迁移性。

首先,针对每个视角下的输出置信度,提取正确类别的置信度。给定融合后的置信度 $\mathrm{CS}_{\mathrm{fused}}(\theta_j)$,正确类别的置信度可以表示为

$$\mathrm{CS}_{\mathrm{correct}}=\mathrm{CS}_{\mathrm{fused}}\left(\theta_j\right),\ \theta_j=\mathrm{gt} \tag{24}$$

其中,gt是输入图像的真实标签,表示正确类别。

为了计算错误类别的置信度,首先将正确类别的置信度在融合置信度中“屏蔽”掉,得到错误类别的置信度:

$$\mathrm{CS}_{\mathrm{masked}}=\mathrm{CS}_{\mathrm{fused}}\left(\theta_j\right),\ \theta_j\neq\mathrm{gt} \tag{25}$$

通过取最大值操作计算错误类别中置信度最大的类别,定义为

$$\mathrm{CS}_{\mathrm{wrong}}=\max(\mathrm{CS}_{\mathrm{masked}}) \tag{26}$$

这一过程确保可以找到模型在错误类别上的最大置信度,从而引导对抗样本针对错误类别进行优化更新。

为了引导对抗样本生成,本文设计了一个双向损失函数,目标是最大化错误类别的置信度,同时最小化正确类别的置信度:

$$\mathcal{L}_{\mathrm{total}}=\lambda\cdot\mathrm{CS}_{\mathrm{wrong}}-\mathrm{CS}_{\mathrm{correct}} \tag{27}$$

其中,λ是置信度损失权重,控制错误类别置信度和正确类别置信度之间的权重关系。该优化目标以输出置信度最大的错误类别为引导,增强模型在非真实类别方向上的判别倾向。该损失具有一定倾向性的将对抗样本引导至确定的错误类别,使扰动方向倾向于拟合跨模型共享的脆弱区域,避免梯度过度震荡导致不稳定,从而增强对抗扰动的普适性,提升对抗样本的迁移性。

2.4　与梯度攻击方法的兼容性设计

本文提出的多视角置信度融合方法能够作为即插即用模块,嵌入到基于梯度的对抗攻击方法,提升对抗样本的迁移性。本节以MIM对抗攻击方法为例,说明MIM方法与本文方法的兼容性设计流程。

MIM对抗攻击方法将样本x_t^{adv}输入到深度学习模型后得到对应输出,将模型输出与真实标签之间的交叉熵作为损失函数,通过梯度反向传播实现对抗样本的优化更新。嵌入本文提出的多视角引导融合方法,需要进行以下两点调整。

(1)对输入到深度学习模型的样本x_t^{adv}执行本文提出的基于双像素空间的多视角变换操作;

(2)将用于执行梯度反向传播的交叉熵损失修改为本文设计的双向损失函数。

综上,对抗样本的更新流程为

$$\begin{aligned}x_t^{\mathrm{adv}}&=T_i(x_t^{\mathrm{adv}})\\ g_{t+1}&=\mu\cdot g_t+\frac{\nabla_x\mathcal{L}_{\mathrm{total}}}{\left\|\nabla_x\mathcal{L}_{\mathrm{total}}\right\|_1}\\ x_{t+1}^{\mathrm{adv}}&=\mathrm{Clip}_x\left\{x_t^{\mathrm{adv}}+\alpha\cdot\mathrm{sign}\left(g_{t+1}\right)\right\}\end{aligned} \tag{28}$$

3　实验结果与分析

3.1　参数测试

(1)数据集。ImageNet large scale visual recognition challenge (ILSVRC) 2012竞赛数据集作为ImageNet数据集的轻量版[36],涵盖了从自然场景到人工制品的广泛领域,是计算机视觉领域的重要基准。本文从ILSVRC 2012验证集随机选取了1 000张图像用于生成对抗样本,这1 000张图像分属于1 000个类别,图像尺寸均为3 × 299 × 299。

(2)深度学习分类模型。为充分验证方法的有效性,本文选取以下模型开展对抗攻击方法的性能评估:

CNN分类模型:Inception-v3(Inc-v3)[37]、Inception-v4(Inc-v4)、Inception-Resnet-v2(IncRes-v2)[38]、Resnet-v2-50(Res-50)、Resnet-v2-101(Res-101)和Resnet-v2-152(Res-152)[39]。

防御模型:包括3个对抗训练模型:Inc-v3ens3、Inc-v3ens4和IncRes-v2ens[40],以及4种基于输入变换的防御策略:Feature Distillation (FD)[41]、Bit-Reduction (Bit-Red)[42]、Joint Photographic Experts Group(JPEG)[43]和Random Resize and Padding (RP)[44]。对抗样本经输入变换后被送入Inc-v3ens4以给出最终预测。

Transformer模型:为验证提出方法在跨模型架构方面的有效性,6个Transformer架构的模型被选取作为黑盒目标模型:ViT、ViT-L[45]、Swin-ViT、Swin-ViT-S[46]、Swin-ViT-V2[47]以及Max-ViT[48]。

(3)对比算法。本文选取了MIM[8]、DIM[9]、TIM[10]、GMI[23]、BSR[25]及MIG[15]作为基线方法,通过比较基线方法与其嵌入多视角置信度融合方法后的攻击性能验证本文方法的有效性,并分别将嵌入了多视角置信度融合方法后的攻击算法记为MIM-MG、DIM-MG、TIM-MG、GMI-MG、BSR-MG及MIG-MG。

(4)参数设置。为保证实验的公平性,对于所有攻击方法,设置最大扰动$\xi=16$,迭代次数$T=10$,步长$\alpha=\xi/T=1.6$,衰减因子$\mu=1.0$。对于DIM,设置转

换概率$P=0.5$。对于TIM,设置内核长度$k=7$。对于BSR,设置最大旋转角度$\tau=36°$,图像分割块为2×2。对于本文提出的融合引导模块,设置调整参数$\psi=0.2$,视角变换次数$N=3$,置信度损失权重$\lambda=0.2$,区域划分数$G=5$。

3.2 实验一:针对CNN模型的攻击测试

本节将选取的六种基线方法分别与多视角置信度融合方法集成,以Inc-v3模型和IncRes-v2模型为白盒模型生成对抗样本,测试对抗样本对6个正常训练模型的攻击成功率。如表1所示,集成引导方法后生成的对抗样本的白盒攻击成功率与集成前几乎相同,但对于黑盒模型的迁移攻击成功率显示出明显的提升。例如,以IncRes-v2模型为白盒模型,MIM方法对Inc-v4模型的迁移攻击成功率仅为53.40%,而MIM-MG方法能达到84.30%。尤其值得注意的是,以IncRes-v2模型为白盒模型,MIG-MG方法对其余5种黑盒模型的迁移攻击成功率均超80.00%,平均达到了84.44%。这表明基于视角引导方法可以作为一个有效的融合模块来提高对抗样本的可迁移性。

表1 针对CNN模型的攻击成功率 单位:%

Table 1 Attack success rate against CNN models unit: %

白盒模型	算法	Inc-v3	Inc-v4	IncRes-v2	Res-50	Res-101	Res-152	平均值
Inc-v3	MIM	100.00	45.00	41.20	40.00	36.30	35.00	49.58
	MIM-MG	100.00	71.70	71.60	65.80	62.30	61.90	72.22
	DIM	99.90	67.70	63.00	58.40	55.20	52.70	66.15
	DIM-MG	100.00	85.40	83.60	79.00	75.60	72.50	82.68
	TIM	100.00	47.90	40.90	39.80	38.80	36.30	50.62
	TIM-MG	100.00	75.60	73.50	69.50	63.80	65.00	74.57
	GMI	100.00	49.60	47.00	43.10	36.60	34.70	51.83
	GMI-MG	100.00	73.50	73.60	69.50	63.20	61.10	73.48
	BSR	98.80	75.60	69.50	69.40	62.80	59.70	72.63
	BSR-MG	99.70	87.40	84.10	79.80	77.90	73.50	83.73
	MIG	98.70	69.10	65.40	61.20	57.80	55.50	67.95
	MIG-MG	99.90	81.90	79.50	78.60	75.10	75.80	81.80
IncRes-v2	MIM	61.40	53.40	99.40	50.50	46.60	43.60	59.15
	MIM-MG	89.50	84.30	99.80	78.90	78.40	76.60	84.58
	DIM	74.00	68.40	97.60	62.70	60.00	59.40	70.35
	DIM-MG	93.50	91.90	99.40	88.80	88.00	86.70	91.38
	TIM	64.50	57.70	98.70	54.50	50.20	45.90	61.92
	TIM-MG	90.80	86.80	99.60	80.90	79.80	79.40	86.22
	GMI	64.80	56.80	99.90	47.00	42.90	40.00	58.57
	GMI-MG	89.50	83.10	100.00	76.20	76.00	71.60	82.73
	BSR	81.20	78.30	94.40	73.70	68.60	67.00	77.20
	BSR-MG	94.50	93.20	99.50	89.20	86.10	85.60	91.35
	MIG	74.60	71.50	92.30	65.20	66.20	65.30	72.52
	MIG-MG	90.50	86.00	99.20	83.20	81.90	80.60	86.90

3.3 实验二:针对防御模型的攻击测试

对防御模型进行对抗攻击,不仅能评估防御模型的抗攻击能力,还能验证攻击方法在面对防御机制时的攻击鲁棒性。本节针对防御模型展开攻击测试,表2展示了基线攻击方法与集成多视角引导模块后的攻击方法在面对不同防御模型时的攻击效果对比。通过数值可以看出,加入多视角引导方法后的攻击普遍表现出更高的迁移攻击成功率。例如,以Inc-v3模型为白盒模型,JPEG为黑盒受害模型,GMI攻击加入多视角引导模块后,迁移攻击成功率从25.20%提高至46.30%,提升了21.10%;此外,以IncRes-v2模型为白盒模型,TIM-MG方法对7种黑盒防御模型的迁移攻击成功率均超52.00%,展现出良好的鲁棒性。这表明,加入融合方法后的攻击方法能显著突破防御,提升攻击成功率。

3.4 实验三:针对Transformer模型的攻击测试

依赖局部卷积的CNN与基于全局自注意力机制的Transformer核心架构差异显著,这种结构差异导致

表2　针对防御模型的攻击成功率　　单位:%

Table 2　Attack success rate against defence models　　unit: %

白盒模型	算法	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens	JPEG	FD	Bit-Red	RP
Inc-v3	MIM	18.70	16.60	8.70	29.90	37.00	16.90	29.70
	MIM-MG	29.40	27.40	15.30	51.50	50.70	26.30	48.30
	DIM	25.50	24.80	12.30	42.10	43.40	22.00	43.30
	DIM-MG	34.90	34.40	18.00	61.00	56.20	31.70	59.60
	TIM	26.40	26.70	15.90	33.10	44.90	22.10	36.20
	TIM-MG	52.50	48.60	33.00	61.80	62.20	42.60	63.70
	GMI	15.50	13.70	7.30	25.20	37.80	15.00	26.00
	GMI-MG	19.00	17.40	8.80	46.30	46.80	19.50	36.40
	BSR	28.40	27.50	14.20	40.80	49.70	25.40	50.60
	BSR-MG	35.50	34.20	18.30	62.50	60.10	32.20	64.30
	MIG	37.70	36.30	20.30	58.00	54.60	35.40	49.40
	MIG-MG	42.30	41.90	23.80	70.40	66.20	44.40	61.00
IncRes-v2	MIM	21.20	21.60	13.80	30.20	39.50	18.60	32.50
	MIM-MG	42.20	35.80	28.70	61.80	56.80	33.70	58.30
	DIM	32.20	30.60	22.20	44.60	48.30	25.30	48.20
	DIM-MG	52.20	46.50	35.80	72.80	65.20	42.10	72.90
	TIM	33.60	31.30	27.70	41.70	48.80	27.30	42.10
	TIM-MG	65.70	59.90	58.00	71.20	71.30	52.70	74.50
	GMI	15.10	13.40	8.10	27.30	39.70	15.10	27.80
	GMI-MG	24.70	22.10	14.40	54.90	52.30	24.70	43.80
	BSR	42.10	39.70	26.90	55.00	55.20	31.80	62.70
	BSR-MG	50.30	44.50	30.90	72.40	66.70	40.20	75.30
	MIG	47.40	45.60	38.90	72.20	68.00	50.40	69.10
	MIG-MG	52.50	48.50	39.90	75.30	70.20	50.30	67.90

传统CNN生成的对抗样本对Transformer模型的迁移性普遍偏低。跨模型结构攻击能验证实际场景中攻击者难以获取目标模型具体架构的情况下对抗攻击的实用性。本节以CNN为白盒模型生成对抗样本,针对基于Transformer的目标模型展开攻击测试。表3中的结果显示,引入多视角融合方法后,对抗样本攻击成功率得到显著提升。例如,以Inc-v3模型为白盒模型,Swin-ViT为黑盒受害模型,GMI攻击加入多视角引导模块后,迁移攻击成功率从22.20%提高至34.70%;以IncRes-v2模型为白盒模型时,对各类Transformer模型的平均攻击成功率提升了12.26%~20.32%。综上可见,多视角融合方法能够有效突破CNN与Transformer之间的结构壁垒,大幅提升对抗样本的迁移性,验证了其在梯度攻击中的兼容性。

3.5　实验四:针对集成模型的攻击测试

针对集成模型的对抗攻击通过捕获稳定欺骗多个白盒模型的扰动方向,使得生成的对抗样本易于偏离模型共享的正确类别决策边界。针对集成模型开展攻击测试,一方面可以评估攻击方法对构成集成模型的多个基础模型的对抗有效性,另一方面也能验证生成的对抗扰动对黑盒模型的普适性。本节通过攻击集成模型来验证提出的多视角置信度融合方法的性能。具体来说,分别使用基线攻击方法与集成多视角置信度融合方法后的攻击方法攻击具有相等集成权重的正常训练模型(包括Inc-v3、Inc-v4和IncRes-v2)的集成。如表4所示,融合多视角引导方法后,针对集成模型的白盒攻击成功率与基线方法基本持平,但对于其余黑盒模型的迁移攻击成功率得到了明显的提升。具体来说,MIM方法对于剩余3个正常训练的黑盒模型的迁移攻击成功率均值仅为64.53%,而MIM-MG达到了91.43%,实现了26.90%的提升;除此以外,TIM-MG方法对于3个对抗训练防御模型的攻击成功率均在70%以上。这些结果表明本文方法对于提升针对集成模型的对抗样本可迁移性具有良好的效果。

3.6　对抗样本热力图可视化分析

为揭示所提方法提升对抗样本迁移性的成因,本节基于模型注意力热力图开展可视化分析。图3展示了干净样本、基线方法及其融合多视角引导模块后生成的对抗样本,以及模型对这些样本的热力图。可

表3 针对Transformer模型的攻击成功率 单位:%

Table 3 Attack success rate against Transformer models unit: %

白盒模型	算法	ViT	ViT-L	Swin-ViT	Swin-ViT-S	Swin-ViT-V2	Max-ViT	平均值
Inc-v3	MIM	17.10	17.00	24.50	19.70	23.20	15.90	19.57
	MIM-MG	24.80	23.10	36.80	30.70	34.90	29.60	29.98
	DIM	20.50	20.10	33.40	27.10	30.90	23.60	25.93
	DIM-MG	29.90	28.00	47.10	37.20	43.90	38.40	37.42
	TIM	23.40	21.10	27.20	22.20	26.10	16.90	22.82
	TIM-MG	34.90	33.40	41.70	35.30	40.90	29.30	35.92
	GMI	16.00	15.30	22.20	16.50	20.00	13.90	17.32
	GMI-MG	21.30	20.30	34.70	26.60	33.20	24.50	26.77
	BSR	25.60	23.90	39.50	30.50	37.70	27.70	30.82
	BSR-MG	31.80	29.10	47.90	37.30	46.60	37.40	38.35
	MIG	30.60	25.80	40.40	34.10	37.00	31.50	33.23
	MIG-MG	41.30	35.00	52.40	45.40	50.20	45.10	44.90
IncRes-v2	MIM	20.60	20.20	27.80	23.30	26.70	21.70	23.38
	MIM-MG	33.40	29.60	47.10	39.70	45.10	40.00	39.15
	DIM	26.10	23.90	37.30	32.30	36.20	32.70	31.42
	DIM-MG	38.90	36.00	59.20	52.60	56.50	55.60	49.80
	TIM	27.90	26.20	31.60	26.80	31.10	22.90	27.75
	TIM-MG	46.90	42.50	53.90	47.80	53.30	44.00	48.07
	GMI	17.60	18.60	23.70	20.20	22.20	15.40	19.62
	GMI-MG	28.50	25.40	38.90	32.00	36.30	30.20	31.88
	BSR	30.20	29.20	46.10	37.30	43.70	37.00	37.25
	BSR-MG	39.70	36.20	59.50	53.20	56.30	53.60	49.75
	MIG	35.80	30.80	46.20	40.50	43.30	40.90	39.58
	MIG-MG	47.30	41.70	61.30	54.50	58.10	54.90	52.97

表4 针对集成模型的攻击成功率 单位:%

Table 4 Attack success rate against integrated models unit: %

算法	Inc-v3	Inc-v4	IncRes-v2	Res-50	Res-101	Res-152	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
MIM	100.00	99.30	97.60	65.80	65.30	62.50	33.20	33.00	20.60
MIM-MG	100.00	99.80	99.60	91.60	91.10	91.60	59.40	53.30	40.30
DIM	99.70	98.50	96.80	84.10	82.60	80.80	50.80	49.20	34.10
DIM-MG	100.00	99.90	99.60	95.30	95.20	95.00	67.60	62.80	49.10
TIM	99.90	98.70	97.10	67.40	65.40	64.70	52.80	50.50	40.00
TIM-MG	100.00	99.80	99.60	92.80	92.20	91.50	85.90	82.70	72.20
GMI	100.00	100.00	99.70	72.00	69.40	68.30	25.60	21.90	13.20
GMI-MG	100.00	100.00	100.00	94.80	94.20	93.40	34.70	30.80	20.40
BSR	98.70	97.30	95.70	88.40	86.30	85.40	57.40	54.00	37.10
BSR-MG	99.70	99.70	99.40	95.10	94.60	94.50	64.80	59.10	44.30
MIG	98.60	95.40	91.10	82.50	81.40	82.70	66.40	65.20	52.80
MIG-MG	99.60	98.80	98.00	91.60	91.10	90.40	70.40	65.00	55.20

见基线方法与融合模块后生成的对抗样本均难以察觉,融合多视角引导模块对可迁移性的提升并未以牺牲对抗样本的视觉感知性为代价。此外,热力图直观显示模型对融合多视角引导模块生成的对抗样本的关注重心已脱离目标主体区域,转而聚焦于背景等非关键区域。本文提出的双像素空间增强变换,结合置信度融合与梯度反向传播机制,核心通过拓宽模型关注区域提升对抗样本迁移性。双像素空间增强变换打破了模型对图像主体区域的固有关注模式,引导模型拓宽感知范围,这种拓宽效应使得白盒模型的关注

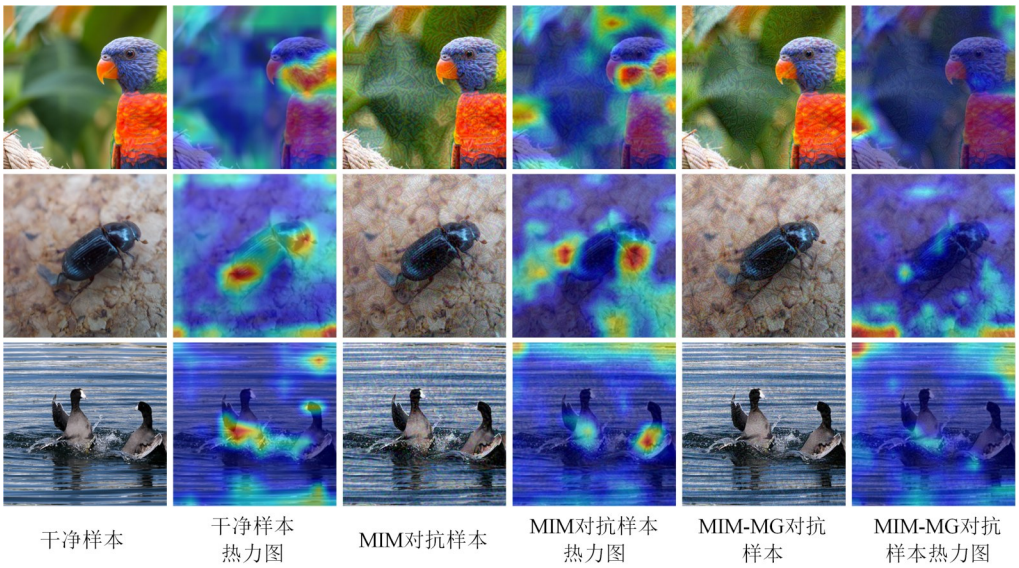


图3　干净样本、对抗样本以及模型的热力图可视化

Figure 3　Visualization of clean examples, adversarial examples, and model heatmaps

区域覆盖了黑盒模型可能存在的关注模式。在此基础上,对模型输出进行置信度融合,可整合多视角下的模型决策特征,使梯度反向传播过程精准聚焦于模型拓宽后的关注区域,针对性施加扰动,从而引导对抗样本干扰模型对主体区域的聚焦能力。最终生成的对抗样本使得模型无法从主体区域提取有效判别信息,从而显著提升对不同黑盒模型的迁移攻击性能。

3.7　像素空间增强变换组合的影响

为验证本文提出的RGB与HSV双像素空间增强变换相较于其他组合方案在提升对抗样本迁移性方面的优势,我们选取YCbCr像素空间的增强变换方法[49],构建了三类双像素空间增强变换组合方案:RGB+YCbCr、HSV+YCbCr及本文提出的RGB+HSV,进而对比不同双像素空间增强变换设置下生成的对抗样本的攻击效果。实验结果如图4所示,RGB+HSV组合生成的对抗样本对黑盒模型的迁移攻击成功率最优,相较于RGB+YCbCr、HSV+YCbCr组合,平均攻击成功率分别提升了10.80%、7.38%,充分证明该像素空间增强变换组合在提升迁移性方面的优越性。

本文提出的RGB与HSV双像素空间增强方法在提升迁移性方面具有明显的性能优势,核心原因在于两类像素空间特性互补且增强策略适配。HSV像素空间实现色调、饱和度与明度解耦,通道级全局增强让全图各区域的色光属性均产生变化,能够拓宽模型对色彩亮度的感知范围。RGB像素空间贴合模型的原生输入,区域级分块增强消除了局部特征的规律性,使得模型需动态关注全图各块的特征变化,引导模型关注全图特征。两种像素空间对应图像的不同视觉属性,增强操作针对的感知维度独立,可协同引导模型拓宽感知范围,进而引导对抗样本干扰模型对主体区域的聚焦,提升对抗样本的迁移性。

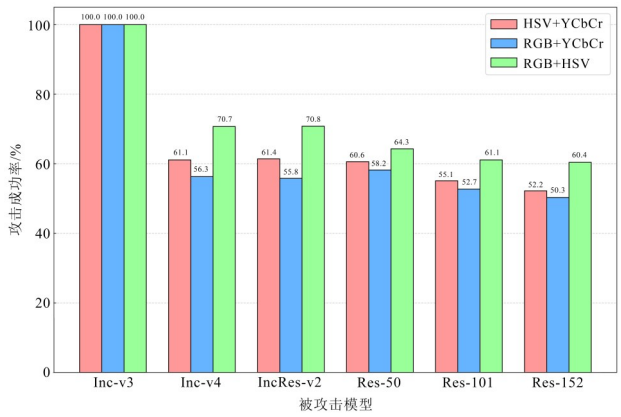


图4　针对像素空间增强变换组合的对照实验,白盒模型为Inc-v3

Figure 4　The controlled experiment on pixel-space enhancement transformation combinations, the white-box model used is Inc-v3

3.8　超参数分析

3.8.1　迭代次数

在基于多视角置信度融合的对抗样本生成中,迭代次数T用于控制对抗样本循环更新优化的次数。针对该超参数的实验结果如图5所示,对抗样本的迁移性随迭代次数的增大呈现提升趋势。这是因为更多的迭代次数能够基于每次迭代的新梯度逐步调整扰动方向与大小,避免陷入局部最优,并逼近使模型损失最大化的最优扰动。然而,迭代次数的增大直接导致计算消耗显著增加。过大的迭代次数会明显降低攻击效率,尤其在处理大规模数据集时,过高的计

算成本使其难以满足实际应用需求。综合迁移性提升与计算成本的权衡，本文设置适中的迭代次数值 $T=10$，此时迁移性已达到较为理想的水平，且计算消耗处于可接受范围，能够在攻击性能与效率间取得较优平衡。

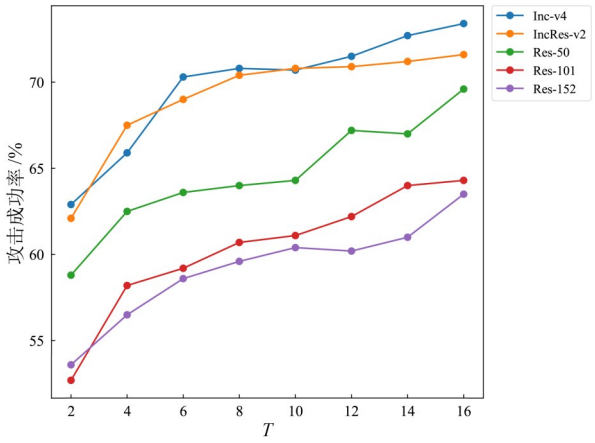


图5 针对迭代次数的超参数实验，白盒模型为Inc-v3

Figure 5 Hyperparameter experiments on the number of iterations, the white-box model is Inc-v3

3.8.2 置信度损失权重

设计双向损失函数时，λ用于控制错误类别置信度 CS_{wrong} 和正确类别置信度 $-CS_{correct}$ 之间的权重关系。针对该超参数的实验结果如图6所示，对抗样本的迁移性随错误类比置信度比例的增大呈现先提升后下降的趋势，并普遍在取值0.2时达到最优性能。超参数λ用于平衡定向诱导与原始特征抑制之间的强度配比。当λ取值较小时，损失函数由 $-CS_{correct}$ 主导。此时优化过程倾向于寻找距离原始样本最近的决策边界进行跨越。然而，此类边界往往受源模型特有的局部特征的影响，并未触及跨模型共享的泛化脆弱区域，因此生成的对抗样本对其他模型的迁移性较差。当λ取值过大时，优化重心偏移至最大化 CS_{wrong}，虽然这能显著提升特定错误类别的置信度，但削弱了对 $-CS_{correct}$ 的惩罚力度，从而导致对抗样本在特征空间中并未彻底远离正确类别的流形，保留了过多的原始类别语义信息。当攻击转移至鲁棒性更强或其他架构的模型时，这些残留的原始特征容易被识别，导致攻击失效。因此，最佳的λ选择应遵循抑制优先，诱导为辅的原则。本文设置置信度损失权重λ取值为0.2，以确保在首要破坏正确类别置信度的基础上，将样本推向跨模型共享的错误区域，从而实现最优的迁移攻击性能。

3.8.3 区域划分数

在执行RGB空间区域差异化增强时，区域划分

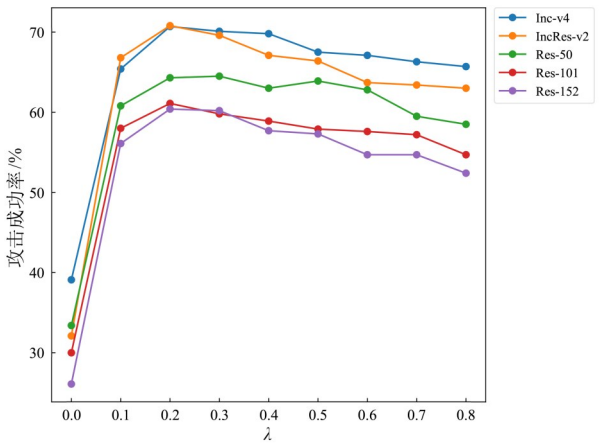


图6 针对置信度损失权重的超参数实验，白盒模型为Inc-v3

Figure 6 Hyperparameter experiments on the confidence loss weights, the white-box model is Inc-v3

数G控制划分网格的数量。图7针对该参数的实验结果表明，对抗样本的迁移性随划分数量的增大呈现出先提升后下降的趋势，并在取值$G=5$时达到最优性能。具体而言，当G取值过小时，划分区域面积过大，导致增强变换趋近于图像层面的全局操作。这种粗粒度的变换难以在空间上实现特征的有效解耦，模型依然能够捕捉到大面积连贯的局部特征，导致产生的梯度方向无法打破模型对特定显著区域的过拟合依赖。反之，当G取值过大时，网格区域变得过于细碎。一方面，由于深度卷积神经网络的深层单元具有较大的感受野，这种非结构化的细微干扰极易被网络作为噪声过滤；另一方面，过细的切分过度破坏了图像的局部邻域相关性，导致原本具有判别意义的结构化特征发生冗余，使得梯度优化过程失去了明确的攻击导向。为保障最优的对抗迁移性能，本文选定G取

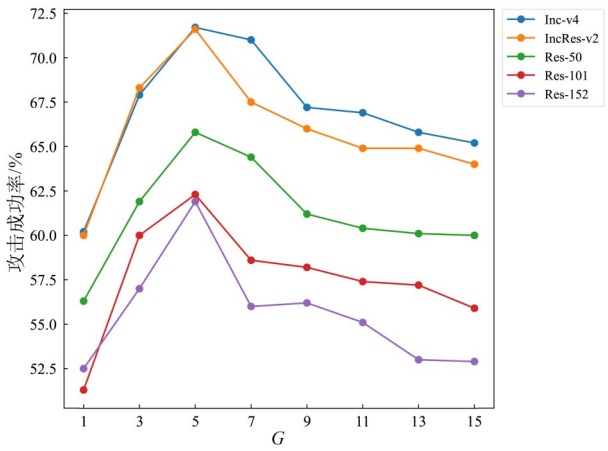


图7 针对区域划分数的超参数实验，白盒模型为Inc-v3

Figure 7 Hyperparameter experiments on the number of regions, the white-box model is Inc-v3

值为5,此时的分块尺度能够最大程度地提升对抗样本对不同模型的迁移性能。

4 结束语

面向深度学习模型的对抗攻击研究中,对抗样本的迁移性不足是制约其实际应用效能的核心瓶颈。现有梯度对抗攻击方法对输入图像的重点关注视角单一,生成的对抗样本过度依赖白盒模型的局部梯度信息,难以适用于黑盒模型的决策机制。针对这一问题,本文提出了一种基多视角置信度融合的对抗样本迁移性提升方法。通过对输入图像进行双像素空间的增强变换,提升模型对图像主体关注范围的多样性,从而促进模型的信息全面感知能力。此外,制定了基于冲突感知的置信度融合策略,处理不同视角下模型输出的冲突和不确定性,有效避免了单一视角的模型决策对扰动方向的主导作用。在损失函数设计上,设计了一种双向损失函数引导对抗样本处于跨视角、跨模型共享的脆弱区域,从而提高对抗样本的迁移性。实验表明,本文方法能够在多类架构模型上促进现有对抗攻击方法的迁移性。

由于本文方法采用了双像素空间变换策略,并对多个视角下模型的输出进行融合,导致对抗样本的生成需要较大的计算开销。为平衡对抗样本的生成效率与攻击性能,未来研究可从筛选关键变换视角、采取轻量化变换操作出发,以满足实时应用场景的需求。此外,由于目标检测与语义分割均以类别判别为基础任务,本文方法可拓展至此类视觉场景。然而,检测与分割任务输出维度更高,损失函数由多分支构成,难以在分类误导、定位精度与区域一致性之间实现有效平衡。后续研究可通过设计面向特定任务的空间感知多视角变换机制,构建分类置信度与定位和分割质量协同的自适应融合策略,为在更复杂视觉任务中的落地提供可行方案。

参考文献

[1] 吴亚军, 刘礼文. 一种基于深度学习水下高速航行器的目标识别方法研究[J]. 指挥控制与仿真, 2025, 47(2): 87-94.
Wu Yajun, Liu Liwen. Research on an underwater high-speed vehicle target recognition method based on deep learning[J]. Command Control & Simulation, 2025, 47(2): 87-94. (in Chinese)

[2] 王浩添, 冀振元, 化青龙, 等. 基于多分支多信息多深度复值特征融合网络的SAR舰船目标识别方法[J]. 电子学报, 2025, 53(10): 3759-3772.
Wang Haotian, Ji Zhenyuan, Hua Qinglong, et al. Recognition method of ship targets for SAR based on M3Net[J]. Acta Electronica Sinica, 2025, 53(10): 3759-3772. (in Chinese)

[3] Han Wenqi, Jiang Wen, Geng Jie, et al. Difference-complementary learning and label reassignment for multimodal semi-supervised Semantic segmentation of remote sensing images[J]. IEEE Transactions on Image Processing, 2025, 34: 566-580.

[4] Xu Mai, Sun Xiancheng, Li Shengxi, et al. Spherical patch generative adversarial net for unconditional panoramic image generation[J]. IEEE Transactions on Image Processing, 2025, 34: 3833-3848.

[5] 刘文钊, 郭凯威. 面向深度神经网络视觉模型对抗鲁棒性的攻击与防御方法研究综述[J]. 网络安全技术与应用, 2025(1): 42-48.
Liu Wenzhao, Guo Kaiwei. A review of attack and defense methods targeting the adversarial robustness of deep neural network vision models[J]. Network Security Technology & Application, 2025(1): 42-48. (in Chinese)

[6] 刘洁怡, 李明哲, 杨曜铭, 等. 基于频域多目标优化的SAR图像对抗样本生成方法[J]. 电子学报, 2025, 53(6): 1958-1968.
Liu Jieyi, Li Mingzhe, Yang Yaoming, et al. A multi-objective optimization method in the frequency domain for SAR image adversarial sample generation[J]. Acta Electronica Sinica, 2025, 53(6): 1958-1968. (in Chinese)

[7] Braiek H B, Reid T, Khomh F. Physics-guided adversarial machine learning for aircraft systems simulation[J]. IEEE Transactions on Reliability, 2023, 72(3): 1161-1175.

[8] Dong Yinpeng, Liao Fangzhou, Pang Tianyu, et al. Boosting adversarial attacks with momentum[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 9185-9193.

[9] Xie Cihang, Zhang Zhishuai, Zhou Yuyin, et al. Improving transferability of adversarial examples with input diversity[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 2725-2734.

[10] Dong Yinpeng, Pang Tianyu, Su Hang, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 4307-4316.

[11] 冯卫栋, 余东, 张淳杰, 等. 基于扰动响应的自适应集成黑盒对抗攻击算法[J]. 自动化学报, 2025, 51(8): 1788-1799.
Feng Weidong, Yu Dong, Zhang Chunjie, et al. Perturbation response-based adaptive ensemble black-box adversarial attack algorithm[J]. Acta Automatica Sinica, 2025, 51(8): 1788-1799. (in Chinese)

[12] Wang Xiaosen, Lin Jiadong, Hu Han, et al. Boosting adversarial transferability through enhanced momentum[C]//32nd British Machine Vision Conference. Durham: BMVA Press, 2021.

[13] Gao Yue, Shumailov I, Fawaz K. SEA: Shareable and ex-

plainable attribution for query-based black-box attacks[C]// 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). Piscataway: IEEE, 2025: 439-458.

[14] Xiong Yifeng, Lin Jiadong, Zhang Min, et al. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 14963-14972.

[15] Ma Wenshuo, Li Yidong, Jia Xiaofeng, et al. Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 4607-4616.

[16] Liu Yanpei, Chen Xinyun, Liu Chang, et al. Delving into transferable adversarial examples and black-box attacks[C]// 5th International Conference on Learning Representations. Toulon: ICLR, 2017: 2235-2248.

[17] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[C]//2nd International Conference on Learning Representations. Banff: ICLR, 2014: 1-10.

[18] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]//3rd International Conference on Learning Representations. San Diego: ICLR, 2015: 1-11.

[19] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world[M]//Yampolskiy R V. Artificial intelligence safety and security. New York: Chapman and Hall/CRC, 2018: 99-112.

[20] Li Qizhang, Guo Yiwen, Zuo Wangmeng, et al. Making substitute models more Bayesian can enhance transferability of adversarial examples[C]//11th International Conference on Learning Representations. Kigali: ICLR, 2023: 37295-37310.

[21] Chen Bin, Yin Jiali, Chen Shukai, et al. An adaptive model ensemble adversarial attack for boosting adversarial transferability[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 4466-4475.

[22] Gan Fuquan, Yan Wo. Boosting the transferability of adversarial examples through gradient aggregation[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 5563-5576.

[23] Wang Jiafeng, Chen Zhaoyu, Jiang Kaixun, et al. Boosting the transferability of adversarial attacks with global momentum initialization[J]. Expert Systems with Applications, 2024, 255: 124757.

[24] Li Zhankai, Wang Weiping, Li Jie, et al. Foolmix: Strengthen the transferability of adversarial examples by dual-blending and direction update strategy[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 5286-5300.

[25] Wang Kunyu, He Xuanran, Wang Wenxuan, et al. Boosting adversarial transferability by block shuffle and rotation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 24336-24346.

[26] Wang Xiaosen, He Xuanran, Wang Jingdong, et al. Admix: Enhancing the transferability of adversarial attacks[C]// 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 16138-16147.

[27] Qian Yaguan, Chen Kecheng, Wang Bin, et al. Enhancing transferability of adversarial examples through mixed-frequency inputs[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 7633-7645.

[28] Guo Yu, Liu Weiquan, Xu Qingshan, et al. Boosting adversarial transferability through augmentation in hypothesis space[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 19175-19185.

[29] Ma Chen, Chen Li, Yong Junhai. Simulating unknown target models for query-efficient black-box attacks[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 11830-11839.

[30] Yin Fei, Zhang Yong, Wu Baoyuan, et al. Generalizable black-box adversarial attack with meta learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(3): 1804-1818.

[31] 郑德生, 郑舜天, 李晓瑜, 等 . CBA: 基于圆几何性质的黑盒攻击方法[J/OL]. 计算机工程, 2025-03-26. https://doi.org/10.19678/j.issn.1000-3428.0070476.
Zheng Desheng, Zheng Shuntian, Li Xiaoyu, et al. CBA: Black box attack based on circular geometric properties[J/OL]. Computer Engineering, 2025-03-26. https://doi. org/10.19678/j.issn.1000-3428.0070476.(in Chinese)

[32] Van De Weijer J, Gevers T, Gijsenij A. Edge-based color constancy[J]. IEEE Transactions on Image Processing, 2007, 16(9): 2207-2214.

[33] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//13th European Conference on Computer Vision. Heidelberg: Springer, 2014: 818-833.

[34] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[35] Hu Tingyu, Yin Haibing, Wang Hongkui, et al. Pixel-domain just noticeable difference modeling with heterogeneous color features[J]. Sensors, 2023, 23(4): 1788.

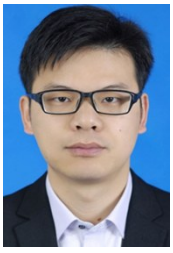
[36] Russakovsky O, Deng Jia, Su Hao, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[37] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the

inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2818-2826.
[38] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1): 4278-4284.
[39] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.
[40] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses[C]//6th International Conference on Learning Representations. Vancouver: ICLR, 2018: 1894-1913.
[41] Liu Zihao, Liu Qi, Liu Tao, et al. Feature distillation: DNN-oriented JPEG compression against adversarial examples[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 860-868.
[42] Xu Weilin, Evans D, Qi Yanjun. Feature squeezing: Detecting adversarial examples in deep neural networks[C]//25th Annual Network and Distributed System Security Symposium. Rosten: The Internet Society, 2018: 1-16.
[43] Guo Chuan, Rana M, Cissé M, et al. Countering adversarial images using input transformations[C]//6th International Conference on Learning Representations. Vancouver: ICLR, 2018: 4914-4925.
[44] Xie Cihang, Wang Jianyu, Zhang Zhishuai, et al. Mitigating adversarial effects through randomization[C]//6th International Conference on Learning Representations. Vancouver: ICLR, 2018: 960-975.
[45] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]//9th International Conference on Learning Representations. Vienna: ICLR, 2021: 611-631.
[46] Liu Ze, Lin Yutong, Cao Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 9992-10002.
[47] Liu Ze, Hu Han, Lin Yutong, et al. Swin transformer V2: Scaling up capacity and resolution[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 11999-12009.
[48] Tu Zhengzhong, Talebi H, Zhang Han, et al. MaxViT: Multi-axis vision transformer[C]//17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 459-479.
[49] Chen Pei, Feng Zhiyong, Xing Meng, et al. Exploring imperceptible adversarial examples in YC_bC_r color space[C]//30th International Conference on Multimedia Modeling. Heidelberg: Springer, 2024: 242-256.

作者简介

赵畅菲　男,1999年8月出生于河南省漯河市。现为西北工业大学电子信息学院博士研究生。主要研究方向为智能算法安全。
E-mail: cfzhao@mail.nwpu.edu.cn



邓鑫洋　男,1988年8月出生于四川省广安市。现为西北工业大学电子信息学院副教授。主要研究方向为多源信息融合、Dempster-Shafer证据理论、不确定信息建模和处理、智能算法安全。
E-mail: xinyang.deng@nwpu.edu.cn

蒋　雯　女,1974年3月出生于陕西省西安市。现为西北工业大学电子信息学院教授。主要研究方向为信息融合、人工智能、遥感图像处理、智能算法安全。中国电子学会会员编号:E190020409S。
E-mail: jiangwen@nwpu.edu.cn

朱金彪　男,1977年12月出生于山东省德州市。现为中国科学院空天信息创新研究院正高级工程师。主要研究方向为智能遥感系统、微波透视探测技术。
E-mail: zhujb@aircas.ac.cn

耿　杰　男,1990年12月出生于山西省晋中市。现为西北工业大学电子信息学院副教授。主要研究方向为SAR图像处理、遥感图像分类、小样本学习。
E-mail: gengjie@nwpu.edu.cn