基于空频双域特征融合的高迁移性对抗样本生成方法

张世辉[1,2],赵鹏宇[1*],张 尧[1],韩少杰[1]

(1. 燕山大学人工智能学院,河北秦皇岛 066000;2. 河北省计算机虚拟技术与系统集成重点实验室,河北秦皇岛 066000)

摘 要: 尽管深度神经网络在许多领域中均表现出卓越的性能,但对抗样本的存在暴露出其在安全方面的显著缺陷。现有黑盒攻击方法通常仅在单一域中进行对抗攻击,忽视了多域特征协同扰动在提升对抗样本迁移性中的重要作用,且多存在损失函数功能单一问题,难以兼顾目标类别导向与梯度稳定。鉴于此,本文提出了一种基于空频双域特征融合的高迁移性对抗样本生成方法(Spatial-Frequency Dual-domain Feature Fusion,SFDFF)。首先,使用离散余弦变换将输入样本从空间域转换至频率域,区域级融合输入样本与原始样本的频率域特征;其次,利用逆离散余弦变换将输入样本还原至空间域,并向其注入基于原始样本统计特征的噪声;然后,通道级融合输入样本与原始样本的空间域特征;最后,设计了一种兼具目标引导与稳定梯度的双导向损失以进一步提高攻击性能。在ImageNet-Compatible与CIFAR-10数据集上的大量实验验证了所提方法的性能。例如,在ImageNet-Compatible数据集上,当从adv-RN-50模型迁移至LeViT模型时,所提SFDFF方法的攻击成功率较当前最优方法提升了2.5%。本文代码见https://github.com/ipkpkpk/SFDFF。

关键词: 对抗样本;特征融合;频率域;空间域;黑盒攻击;迁移性

基金项目: 国家自然科学基金(No.62476235);河北省自然科学基金(No.F2023203012)

中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0125-16

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250521

A Highly Transferable Adversarial Example Generation Method via Spatial-Frequency Dual-Domain Feature Fusion

ZHANG Shihui[1,2], ZHAO Pengyu[1*], ZHANG Yao[1], HAN Shaojie[1]

(1. *School of Artificial Intelligence*, *Yanshan University*, *Qinhuangdao*, *Hebei* 066000, *China*;
2. *Hebei Key Laboratory of Computer Virtual Technology and System Integration*, *Qinhuangdao*, *Hebei* 066000, *China*)

Abstract: Despite the remarkable performance of deep neural networks across various fields, the existence of adversarial examples reveals significant security vulnerabilities. Existing black-box attack methods typically operate within a single domain, overlooking the importance of multi-domain feature co-perturbation in enhancing the transferability of adversarial examples. Moreover, many methods suffer from a single-purpose loss function, making it difficult to balance target class guidance and gradient stability. To address these issues, this paper proposes a high-transferability adversarial examples generation method based on spatial-frequency dual-domain feature fusion (SFDFF). Specifically, the input examples are first transformed from the spatial domain to the frequency domain using the discrete cosine transform, and region-level feature fusion is performed between the input and clean examples in the frequency domain. Then, the input examples are restored to the spatial domain via the inverse discrete cosine transform, and noise based on the statistical characteristics of the original examples are injected. Next, channel-level fusion of spatial features between the input and clean examples are conducted. Finally, a dual-guidance loss function is designed to simultaneously enhance target class directionality and gradient stability. Extensive experiments on ImageNet-Compatible and CIFAR-10 datasets demonstrate the performance of the proposed method. For instance, the attack success rate of the proposed SFDFF increases by 2.5% compared to the state-of-the-art method when transferred from the adv-RN-50 to LeViT model on ImageNet-Compatible dataset. The code is available at https://github.com/ipkpkpk/SFDFF.

Keywords: adversarial examples; feature fusion; frequency domain; spatial domain; black-box attack; transferability

Foundation Item(s): National Natural Science Foundation of China (No.62476235); Natural Science Foundation of Hebei Province (No.F2023203012)

收稿日期:2025-06-19;录用日期:2025-12-30;责任编辑:覃怀银

*通信作者:赵鹏宇

0 引言

近年来，深度神经网络（Deep Neural Networks，DNNs）凭借其强大的特征提取与非线性建模能力，在图像分类[1-2]、目标检测[3-4]等多种计算机视觉任务中取得突破性进展，并逐步应用于自动驾驶、医疗影像分析等高可靠性需求场景[5]。但已有研究表明，在正常样本上添加微小扰动得到对抗样本可诱导DNNs输出高置信度的错误预测，进而引发严重事故[6]。这种生成对抗样本以及利用生成的对抗样本误导模型的行为称为对抗攻击。为增强模型在面对这些潜在风险时的鲁棒性，研究新颖且高效的对抗样本生成及防御方法就显得尤为重要。

在对抗攻击领域，根据攻击者对目标模型的可访问性及攻击策略不同，攻击可分为黑盒攻击与白盒攻击。白盒攻击指攻击者完全掌握模型结构与参数信息，可直接获取模型梯度并据此构造扰动；黑盒攻击指攻击者无法访问模型内部任何信息，仅能通过模型输入与输出间的关系进行推断，从而设计攻击策略，因而更贴近实际应用场景。在黑盒场景中，迁移攻击是一种被广泛使用的攻击策略，其核心思想是在结构已知的源模型上生成对抗样本，并期望这些样本可成功攻击结构未知的黑盒模型。这种对抗样本在不同模型间的有效性被称为迁移性，是评估对抗攻击实际威胁程度的关键指标。迁移性越强，对抗样本越可能在实际应用中突破系统防护，给DNNs带来实质性风险。因此，如何提高对抗样本的迁移性已经成为当前对抗攻击领域的研究重点。

此外，根据攻击目标不同，对抗攻击可进一步划分为非定向攻击与定向攻击。非定向攻击仅要求模型输出非真实标签，而定向攻击则需诱导模型将输入误分类为攻击者指定的目标类别。相比之下，定向攻击任务更具挑战性。特别是在黑盒场景下，定向攻击需生成具有更强导向性的对抗样本，同时需确保其在未知模型上的迁移性，这对扰动方向的控制与扰动的泛化能力提出了更高要求。

然而，现有定向攻击方法在对抗样本优化过程中常受限于梯度饱和问题，导致所生成的对抗样本导向性不足，不能有效误导模型输出既定目标类别。同时，已有的数据增强策略虽可提升扰动多样性，但多局限于在空间域中进行操作，未能充分挖掘频率域特征在增强对抗样本迁移性方面的潜力，导致生成的扰动模式单一，限制了对抗样本跨模型迁移能力。针对上述问题，本文提出了一种定向攻击场景下基于空频双域特征融合的高迁移性对抗样本生成方法（Spatial-Frequency Dual-domain Feature Fusion，SFDFF）。该方法从两方面出发：一方面，在空间域与频率域内分别使用不同设计的策略进行扰动，将数据增强空间由传统空间域扩展到空频双域，并引入了精心设计的基于原始样本统计特征的噪声（Original Sample Statistical Noise，OSSN），引导对抗样本充分挖掘多域特征，生成更具泛化能力的扰动，从而提高对抗样本的迁移性；另一方面，提出一种结合梯度稳定与类别区分作用的双导向损失（Dual-Direction Loss，DDLoss），既可缓解梯度饱和问题，又能更稳定地引导对抗样本向目标类别靠拢，从而显著提高定向攻击场景下对抗样本迁移性。本文的主要贡献如下：

（1）提出一种面向定向攻击的数据增强方法。该方法扩展了数据增强空间，在频率域和空间域上采用差异化增强策略，充分发挥双域特征的互补特性，有效增强了扰动的多样性和鲁棒性，进而提高了所生成对抗样本的迁移性。

（2）设计了一种基于原始样本统计特征的噪声。相较于传统高斯噪声以及均匀分布噪声，该噪声具备从真实图像中提取的统计属性，保留了一定程度的自然图像结构，在避免噪声与模型先验分布偏移问题的同时实现了输入多样化。

（3）提出一种简单有效的双导向损失。该损失在有效推动对抗样本向目标类别靠拢的同时缓解了传统损失中易出现的梯度消失问题，提升了扰动方向的稳定性与跨模型泛化能力。

（4）在ImageNet-Compatible及CIFAR-10数据集上进行了广泛的性能评估。实验结果表明，SFDFF在黑盒场景下的平均攻击成功率表现出较强的竞争力，为增强对抗样本迁移性提供了新的视角。

1 相关工作

在黑盒场景中，根据攻击目标的不同，可将任务划分为非定向攻击与定向攻击。本节将简要介绍非定向攻击与定向攻击两类任务的相关工作。

1.1 非定向攻击

在非定向攻击任务中，研究者提出多种对抗样本生成方法，旨在误导模型输出除真实类别以外的任意类别。快速梯度符号法（Fast Gradient Sign Method，FGSM）[7]是最基本的方法之一，它采用单步更新来优化扰动，方法如下：

$$x^{\mathrm{adv}}=x+\epsilon\cdot\mathrm{sign}\left(\nabla_x\mathcal{L}\left(f(x),y\right)\right) \tag{1}$$

其中，x^{adv}表示对抗样本，通过ℓ_∞范数约束限制其扰动幅度；$\mathcal{L}$通常为交叉熵损失；y表示该样本真实类别。

迭代快速梯度符号方法（Iterative Fast Gradient Sign Method，I-FGSM）[8]进一步扩展了FGSM，其在梯度方向上以较小的步长迭代更新图像，公式如下：

$$x_0^{\mathrm{adv}}=x,\ x_{i+1}^{\mathrm{adv}}=x_i^{\mathrm{adv}}+\alpha\cdot\mathrm{sign}\left(\nabla_{x_i^{\mathrm{adv}}}\mathcal{L}\left(f(x_i^{\mathrm{adv}}),y\right)\right) \tag{2}$$

其中,$\alpha=\epsilon/T$用于控制每步扰动的幅度,从而确保在最多T次迭代内所生成的对抗样本始终满足ℓ_∞范数约束。此外,对抗样本通常具有跨模型迁移能力,即在一个源模型上生成的对抗样本往往能误导其他未知模型。为进一步提升对抗样本迁移性,现有研究主要从两方面展开。

一方面,稳定梯度更新有助于缓解扰动在源模型上的过拟合,防止其陷入局部最优解,从而有效提升对抗样本在黑盒模型上的迁移能力。例如,Dong等人[9]提出了动量迭代快速梯度符号方法(Momentum Iterative FGSM,MI-FGSM),通过在迭代生成扰动的过程中引入动量项以稳定梯度方向。此外,Lin等人[10]提出了尺度不变方法(Scale Invariant,SI),通过在多尺度图像上计算梯度以生成扰动,避免了扰动在单一尺度上的过拟合。类似地,Wang等人[11]进一步提出方差调整方法(Variance Tuning,VT),通过利用梯度方差信息调整当前梯度,从而增强迭代优化过程的稳定性。

另一方面,输入级数据增强策略可有效提升扰动的泛化能力。例如,Xie等人[12]提出的多样化输入方法(Diverse Inputs,DI),通过在每次迭代中对输入图像施加随机变换,多样化输入,以避免扰动在特定输入模式上的过拟合。在此基础上,Zou等人[13]提出缩放多样性输入方法(Resized Diverse Inputs,RDI),通过将经DI变换后的图像缩放回原始尺寸,进一步提升了对抗样本在不同模型间的迁移能力。针对对抗样本在防御模型上迁移性较差的问题,Dong等人[14]提出了平移不变性方法(Translation Invariant,TI),通过对原始图像及其多个平移版本求加权平均梯度,令对抗样本具备一定程度的平移不变性,从而提升对抗样本在防御模型上的迁移能力。此外,Wang等人[15]提出了Admix方法,通过融合不同类别图像,使得融合后的图像更能接近决策边界,从而生成迁移性更强的对抗样本。Long等人[16]则从频率域角度出发,提出频谱模拟攻击方法(Spectrum Simulation Attack,SSA),通过对输入样本应用频谱变换,提高输入多样性,以防止扰动陷入局部最优。

除上述两方面外,Li等人[17]从特征的角度出发,提出特征分布感知攻击方法(Feature Distribution-Aware Attack,FDAA),通过对图像不同区域实施差异化策略,以提升对抗样本迁移性。

需明确,尽管上述方法均基于非定向攻击设定提出,但其中部分方法常作为提升定向攻击场景下对抗样本迁移性的基础模块。然而,由于定向攻击需使样本被误分类为特定目标类别,因此其在损失函数构造、扰动方向收敛性及特征利用方式上存在更高要求。仅靠直接将上述非定向攻击方法套用在定向攻击中,常导致扰动陷入局部最优、攻击失败以及迁移性能严重下降的问题,难以在定向攻击场景中取得理想的迁移性能。

1.2　定向攻击

定向攻击要求将对抗样本误分类为攻击者指定的目标类别,相较于非定向攻击,定向攻击对扰动方向的控制更为严格,同时也更容易陷入局部最优,迁移难度显著提升。在黑盒场景下,定向攻击方法除了可借助1.1节中的基础机制外,还需结合特定的方法以进一步增强其攻击性能。现有提高定向攻击下对抗样本迁移性的方法主要集中在基于损失与基于特征两个方向。

基于损失的方向主要侧重于优化损失函数以缓解梯度饱和问题,从而增强定向攻击下对抗样本的迁移性。例如,Zhao等人[18]提出的Logit损失,通过显式增加目标类别的Logit值,保证目标类别梯度稳定,以缓解梯度饱和问题,同时提高攻击过程中的梯度导向性。Weng等人[19]针对交叉熵损失在训练过程中易导致Logit边距迅速饱和的问题提出了三种Logit校准方法(基于温度、基于边距与基于角度),用以优化Logit边距,稳定梯度。尽管上述方法在一定程度上提升了定向攻击场景下对抗样本的迁移性,但各自仍存在关键性局限:Logit损失虽然能够直接放大目标类别的Logit值,但无法保证其相对优势,忽视了对目标类别分类置信度的有效优化;而引入边距校准的交叉熵损失虽然增强了目标类别与非目标类别之间的边距,但其在高置信度区域中梯度易饱和。针对上述问题,本文提出一种既能显式增强目标类别置信度又能缓解梯度饱和问题的双导向损失,可进一步增强对抗样本在多种异构黑盒模型上的迁移能力。

基于特征的方向则是利用了对抗样本的特征来进行特征级数据增强,进而提高所生成对抗样本的定向攻击能力。Inkawhich等人[20]提出了特征分布攻击方法(Feature Distribution Attack,FDA),通过训练特定类别的辅助分类器,对各层特征分布进行建模以提高对抗扰动的迁移效果。此外,Byun等人[21]提出了干净特征混合方法(Clean Feature Mixup,CFM),通过在空间域内混合当前样本特征与原始样本特征以提高对抗样本的迁移能力。为进一步缓解扰动在源模型上过拟合,Weng等人[22]提出了截断特征混合方法(Truncated Feature Mixing,TFM),通过将当前样本特征与移除Rank-1特征的原始样本特征混合,减少主导性特征影响,从而提高对抗样本迁移性。类似地,Liang等人[23]提出了特征调整混合方法(Feature Tuning Mixup,FTM),通过在空间域内引入可学习的扰动,对空间域特征进行调整以提高扰动的泛化能力。与上述方法不同,本文所提方法将仅发生在空间域内

的扰动扩展到了空频双域，在频率域与空间域内分别使用不同的特征融合策略将当前样本特征与原始样本特征进行融合，并在其间注入基于原始样本统计特征的噪声，促使在优化过程中生成更加鲁棒的扰动，进一步减少源模型的影响，提高对抗样本的迁移性。

2 基于空频双域特征融合的高迁移性对抗样本生成方法

2.1 方法概述

现有定向攻击方法在生成对抗样本的过程中常面临两个问题：一是生成的扰动模式较为单一，难以在不同模型间保持良好的迁移性；二是在训练后期易出现梯度饱和现象，导致对抗样本向目标类别收敛不稳定。针对上述不足，本文的学术构想如下：空间域表征能够刻画局部纹理和语义，但在建模整体结构和长程依赖时相对不足；频率域表征则更擅长保留全局结构，却难以体现局部语义，在表征能力上二者具有互补性。因此，本文将扰动空间由单一的空间域扩展至空频双域，在两个域中分别引入差异化数据增强策略，以充分发挥其互补优势，从而增强扰动的泛化能力，提高对抗样本迁移性。在双域特征融合基础上，为进一步提高扰动迁移性，提出在对抗样本内注入基于原始样本统计特征的噪声。该噪声既能保持扰动与自然图像分布的一致性，避免过度偏离模型先验，又通过增加输入多样性防止扰动陷入局部最优，从而提升对抗样本的鲁棒性与迁移能力。最后，为缓解优化过程中梯度饱和问题并提高目标类别导向的稳定性，设计了 DDLoss，兼顾稳定梯度与目标引导，使扰动在优化过程中能够持续收敛至预定目标类别。

基于上述构想，本文提出了一种新颖的基于空频双域特征融合的 SFDFF。方法总体结构如图 1 所示。

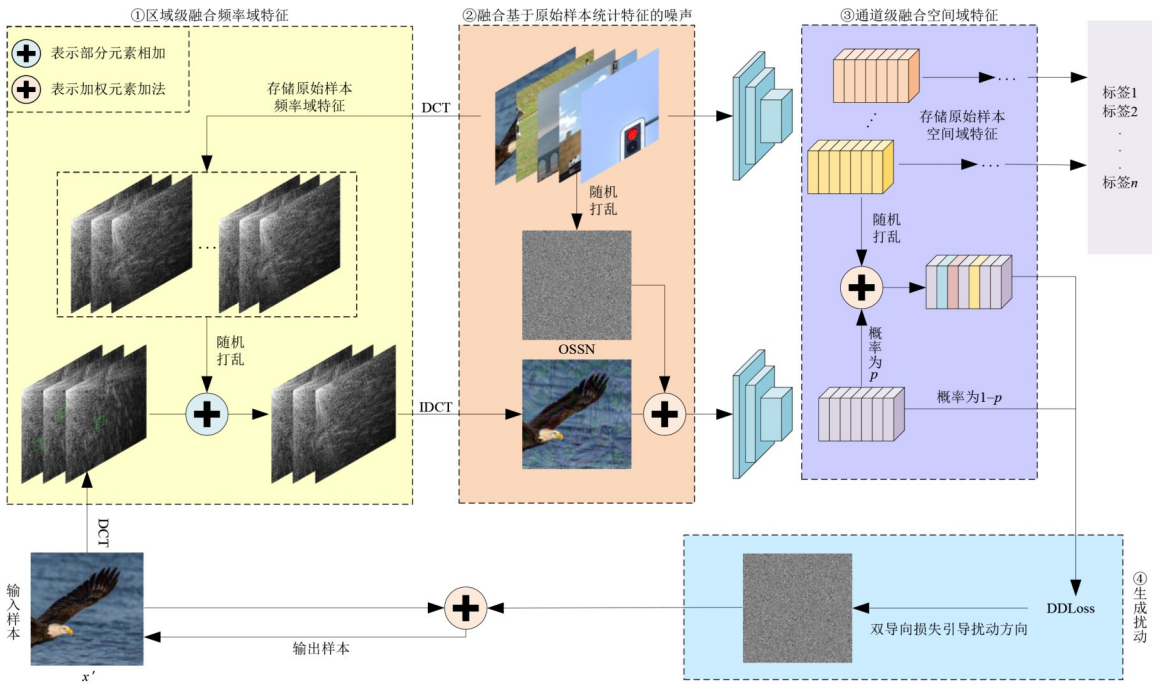

图 1 所提 SFDFF 方法总体结构

Figure 1 Overall architecture of the proposed SFDFF method

由图 1 可知，SFDFF 主要包括四个阶段：首先，使用离散余弦变换（Discrete Cosine Transform，DCT）将输入样本从空间域转换至频率域，并在频率域内随机选取区域，将原始样本与当前样本的频率域特征进行区域级融合，从而削弱模型对特定频率结构的依赖；其次，使用逆离散余弦变换（Inverse Discrete Cosine Transform，IDCT）将输入样本还原回空间域，并向融合后的样本中注入 OSSN，以进一步扰乱模型对关键语义信息的建模过程；然后，将经噪声扰动后的样本与原始样本在空间域内进行通道级特征融合，实现更深层次的扰动信息整合；最后，结合设计合理的 DDLoss，引导扰动方向向目标类别优化，从而生成高迁移性对抗样本。各阶段的具体实现细节详见 2.2 节。

2.2 方法设计

设一个批次中所包含的原始样本集合为 $D=\{x_1,x_2,\cdots,x_n\}$，对应目标类别为 $D_t=\{y_1,y_2,\cdots,y_n\}$，其中 $x_i\in\mathbb{R}^{3\times H\times W}$ 表示第 i 个样本，H 与 W 分别表示图像的高度和宽度。由于算法在第一次迭代时仅涉及存

储频率域与空间域特征,所以当总迭代次数为K时,真正参与对抗样本更新的迭代次数为$K-1$。下文将依次介绍各阶段的具体实现细节与相互作用关系。

2.2.1　区域级频率域特征融合

在扰动优化初期,所提SFDFF方法会将当前样本的频率域特征与原始样本的频率域特征进行区域级融合,以提高对抗样本的结构多样性与区域迁移能力。

对于$x_i \in D$,在第一次迭代过程中,先对x_i应用DCT,得到其在RGB三个通道上的频率域特征,见式(3):

$$F_i=\left\{F_i^r,F_i^g,F_i^b\right\}=\left\{\mathrm{DCT}(x_i^r),\mathrm{DCT}(x_i^g),\mathrm{DCT}(x_i^b)\right\} \tag{3}$$

在提取完D内所有样本频率域特征后,将其组成集合$F_{\mathrm{clean}}=\left\{F_1,F_2,\cdots,F_n\right\}$。随后,对该集合执行图像级别的随机打乱操作$F_{\mathrm{clean}}^s=S(F_{\mathrm{clean}})$,提高融合过程中的随机性,确保在每次迭代过程中不仅可以融合自身原始频率域特征,还可融合其他样本的原始频率域特征。一方面,当与自身原始频率域特征融合时,自身原始频率域特征会引导模型回归原始类别,而对抗扰动则会推动模型朝目标类别分类。两者的竞争关系促使扰动探索更加多样的特征,以克服原始特征带来的干扰,从而提高对抗样本迁移性。另一方面,当与其他图像原始频率域特征融合时,这些特征会引导模型偏向其他类别,而扰动仍试图引导模型输出目标类别。这两类干扰共同作用,一方面促使扰动在优化过程中持续探索不同特征,从而减弱模型对原始特征的依赖,另一方面也增强了扰动在复杂模型上的鲁棒性和跨架构泛化能力。因此,该机制不仅提升了对抗样本迁移性,也为提升黑盒定向攻击的有效性提供了新视角。

在后续迭代过程中,令x_i在上一轮迭代中产生的对抗样本为x_i',对x_i'应用DCT变换,得到其频率域特征$F_i'=\left\{F_i'^{,r},F_i'^{,g},F_i'^{,b}\right\}$。对于$F_i'$内的元素$F_i'^{,c}$,随机选取一个$n\times n$的块状局部区域$R_i^c$,并将其与$F_{\mathrm{clean}}^s$中第$i$个样本$F_{i,\mathrm{clean}}^s$对应区域的频率域特征$F_{i,\mathrm{clean}}^{s,c}(R_i^c)$线性融合,见式(4):

$$F_i'^{,c}(R_i^c)=\alpha\cdot F_i'^{,c}(R_i^c)+(1-\alpha)\cdot F_{i,\mathrm{clean}}^{s,c}(R_i^c),\quad \alpha\in(0,1],\ c\in\left\{r,g,b\right\} \tag{4}$$

最后,将融合后的频率域特征通过IDCT还原回空间域,得到$x_i'=\left\{\mathrm{IDCT}(F_i'^{,r}),\mathrm{IDCT}(F_i'^{,g}),\mathrm{IDCT}(F_i'^{,b})\right\}$。该阶段的融合机制有助于引入跨样本频率域特征,缓解源模型对特定区域的依赖,为后续噪声扰动与空间域特征融合提供了多样性的输入基础。

2.2.2　线性融合基于原始样本统计特征的噪声

在获取经频率域融合后的图像x_i'后,进一步注入基于原始样本统计特征的噪声OSSN。与传统基于全局统计量或高斯分布采样得到的噪声不同,OSSN的生成方式更具针对性,其由随机选取的原始样本通道级统计特征构造而成。这种方式使得噪声能够更贴近自然图像的局部统计规律,从而在保持与真实图像分布一致性的同时,保留一定的结构特征。此外,OSSN不是静态全局噪声,而是动态、依赖随机原始样本的结构化扰动。在每次迭代中,OSSN的来源样本均是随机选取的,因此其统计特征具有较强的随机性。这种样本依赖性使得在优化过程中不断有新的结构化噪声被注入,有助于避免扰动陷入空间域的局部最优解。与传统噪声相比,OSSN既能维持与自然分布的贴合度,减少对模型先验分布的破坏,又能显著提升输入的多样性,使得生成的扰动在不同模型和不同架构上具备更强的迁移性与鲁棒性。

对于经频率域融合后的样本x_i',随机在D内选择一个图像$x_j\in D$,对于x_j的每个通道$c\in\left\{r,g,b\right\}$,分别计算均值与标准差,见式(5)、式(6):

$$\mathrm{mean}_c=\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}x_j[c,h,w] \tag{5}$$

$$\mathrm{std}_c=\sqrt{\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(x_j[c,h,w]-\mathrm{mean}_c\right)^2} \tag{6}$$

使用上述统计量特征构造与x_j同形状的噪声张量$\mathrm{noise}_j\in\mathbb{R}^{3\times H\times W}$,见式(7):

$$\mathrm{noise}_j=N(0,1)\times\mathrm{std}_c+\mathrm{mean}_c,\ c\in\left\{r,g,b\right\} \tag{7}$$

最终,将生成的噪声张量noise_j与x_i'加权融合,完成OSSN的注入,其中$\beta\in(0,1]$控制噪声注入强度,见式(8):

$$x_i'=\beta\cdot x_i'+(1-\beta)\cdot\mathrm{noise}_j \tag{8}$$

该步骤在频率域融合的基础上进一步扰乱图像的统计特征,丰富了扰动分布,避免陷入过拟合的扰动方向,有助于提升对抗样本迁移性。下一步将在空间域深层语义特征层面引入通道级扰动,实现更精细的特征融合。

2.2.3　通道级空间域特征融合

在注入OSSN后,为进一步增加空间域内的扰动表达能力,本文引入了空间域内通道级特征融合机制,融合原始样本与输入样本的空间域特征。此外,为了避免因融合过多低级特征而引入不必要的噪声,该操作仅在输出尺寸显著小于输入尺寸的深层卷积层以及所有全连接层上进行。

在第一次迭代过程中,存储D内样本的原始空间域特征,记为$M_{\mathrm{clean}}=\left\{f_{1,\mathrm{clean}},f_{2,\mathrm{clean}},\cdots,f_{n,\mathrm{clean}}\right\}$,其中$f_{i,\mathrm{clean}}\in\mathbb{R}^{C\times H\times W}$且$i=1,2,\cdots,n$,并对该集合执行图像级别的随机打乱操作,$M_{\mathrm{clean}}^s=S(M_{\mathrm{clean}})$。对于每个样本而言,其特征可能会与同一批次中其他样本的原始特征进行融合,从而增加扰动多样性,促使生成高迁移

性对抗样本。

在后续迭代过程中，为了在保持一定随机性的同时确保在一定比例的层中进行特征融合，以一定概率随机激活特征融合操作。当激活特征融合操作时，将x_i'的空间域特征f_i与M^s_{clean}中的第i个原始空间域特征$f^s_{i,\text{clean}}$进行通道级的随机线性融合。融合后的特征图f_i'计算公式见式(9)：

$$f_i'=(1-\gamma_i)\odot f_i+\gamma_i\odot f^s_{i,\text{clean}} \tag{9}$$

其中，$\odot$表示元素级乘法；γ_i是随机采样的通道级融合比例向量，其形状为$C\times 1\times 1$，并且γ_i中每个元素均从均匀分布$U(0,\gamma_{\max})$中采样，其中$0\leqslant\gamma_{\max}\leqslant 1$。

该机制通过在空间域内引入通道级干扰，从而有效防止对抗样本过度依赖于源模型中的特定特征组合。在此基础上，结合2.2.4节中所提到的双导向损失，在反向传播过程中有效地引导对抗扰动朝更具迁移性的目标类别方向优化，为生成高迁移性对抗样本提供了关键支持。

2.2.4 双导向损失

为更有效地引导扰动朝目标类别方向优化，并缓解传统损失在训练后期梯度消失的问题，本文设计了双导向损失DDLoss。

对于迭代过程中所生成的对抗样本集合$D'=\{x_1',x_2',\cdots,x_n'\}$。设$C$为类别总数，将源模型对这些样本的Logit输出记为$\boldsymbol{Z}=\left[z_1^{\mathrm{T}}\ z_2^{\mathrm{T}}\cdots z_n^{\mathrm{T}}\right]^{\mathrm{T}}\in\mathbb{R}^{N\times C}$，针对$x_i'$所输出的Logit向量为$z_i=\left[z_i^1,z_i^2,\cdots,z_i^C\right]\in\mathbb{R}^C$，设$x_i'$的目标类别为$y_i$，则有其在$y_i$上的Logit值$z_i^{y_i}$，在$z_i$的非目标类别中，选取Logit最大的前$m$个元素组成的集合$\Lambda_i=\text{TOP}_m\{1,2,\cdots,C\}\setminus\{y_i\}$，令权重系数$\varphi_i^j=\left(m-\text{rank}(j)+1\right)\Big/\sum_{k=1}^{m}k$，其中$j\in\Lambda_i$，$\text{rank}(j)$表示类别$j$在$\Lambda_i$内按Logit排序后的名次，排名越靠前的非目标类别越容易与目标类别混淆，其惩罚权重越大。将z_i中最大与次大元素分别记作k_1与k_2，定义其差值为$\Delta_i=k_1-k_2$，将z_i按差值归一化，得到$\tilde{z}_i=z_i/\Delta_i$。双导向损失的定义见式(10)：

$$\mathcal{L}=\lambda\cdot\sum_{i=1}^{N}\left(z_i^{y_i}-\sum_{j\in\Lambda_i}\varphi_i^j z_i^j\right)+(1-\lambda)\cdot\sum_{i=1}^{N}\log\left(\frac{\mathrm{e}^{\tilde{z}_i^{y_i}}}{\sum_{j=1}^{C}\mathrm{e}^{\tilde{z}_i^j}}\right) \tag{10}$$

该损失由两项组成：前项直接最大化目标类别的Logit值并抑制最具威胁的非目标类别，从而在增强对抗样本对目标类别响应的同时拉大目标类别与非目标类别的间隔；后项为基于归一化Logit向量构建的交叉熵损失，利用边距缩放缓解梯度消失问题。两项损失相辅相成，既增强了目标导向性，又保持了梯度的稳定传播，进一步促进对抗扰动朝目标类别方向优化，从而显著提升了定向攻击下所生成对抗样本的迁移能力。

2.2.5 算法描述

基于空频双域特征融合的高迁移性对抗样本生成方法首先在频率域内融合原始样本与当前样本的特征，削弱模型对特定频率域特征的依赖；再还原至空间域并注入OSSN，扰乱语义建模；随后在空间域内融合原始样本与当前样本的特征，强化扰动表达；最后通过DDLoss优化扰动方向，生成高迁移性的对抗样本。具体步骤见算法1。

算法1 基于空频双域特征融合的高迁移性对抗样本生成算法

输入： 数据集内某一批次样本集合$D=\{x_1,x_2,\cdots,x_n\}$，目标类别$D_t=\{y_1,y_2,\cdots,y_n\}$；最大迭代次数$K$；步长$\epsilon$；特征融合比例为$\alpha$、$\beta$、$\gamma$

输出： 对抗样本集合$D'=\{x_1',x_2',\cdots,x_n'\}$

1. IF $k=1$:
2. 提取D内样本的频率域特征F_{clean}，按图像级别打乱得到F^s_{clean}；
3. 提取D内样本的空间域特征M_{clean}，按图像级别打乱得到M^s_{clean}；
4. END IF;
5. $D'=D$;
6. FOR k TO K DO:
7. FOR x_i' IN D':
8. 提取x_i'的频率域特征：$F_i'=\{F_i'^{,r},F_i'^{,g},F_i'^{,b}\}$；
9. 随机选择每个通道的一个局部块：R_i^c，$c\in\{r,g,b\}$；
10. 将F_i'与M^s_{clean}中第i个样本的频率域特征$F^s_{i,\text{clean}}$区域级融合：$F_i'^{,c}(R_i^c)=\alpha\cdot F_i'^{,c}(R_i^c)+(1-\alpha)\cdot F^{s,c}_{i,\text{clean}}(R_i^c)$；
11. 通过IDCT还原回空间域：$x_i'=\text{IDCT}(F_i')$；
12. 从D中随机选取图像x_j，生成基于其统计特征的噪声：$\text{noise}_j=N(0,1)\times\text{std}_c+\text{mean}_c$，$c\in\{r,g,b\}$；
13. 图像加噪：$x_i'=\beta\cdot x_i'+(1-\beta)\cdot\text{noise}_j$；
14. 源模型提取x_i'的空间域特征f_i'；
15. 将f_i'与M^s_{clean}中第i个样本的空间域特征$f^s_{i,\text{clean}}$通道级融合：$f_i'=(1-\gamma_i)\odot f_i+\gamma_i\odot f^s_{i,\text{clean}}$；
16. 前向传播，获得其Logit输出$z_i=\left[z_i^1,z_i^2,\cdots,z_i^C\right]\in\mathbb{R}^C$，其中$z_i^{y_i}$为其目标类别的Logit值；
17. 得到z_i中最大与次大元素：k_1、k_2；
18. 在z_i的非目标类别中，选取Logit最大的前m个类别，形成集合Λ_i；
19. 动态分配权重：$\varphi_i^j=\left(m-\text{rank}(j)+1\right)\Big/\sum_{k=1}^{m}k$；
20. Logit校准：$\tilde{z}_i=z_{ij}/\Delta_i$
21. 损失：$\mathcal{L}_i=\lambda\cdot\left(z_i^{y_i}-\sum_{j\in\Lambda_i}\varphi_i^j z_i^j\right)+(1-\lambda)\cdot\log\left(\frac{\mathrm{e}^{\tilde{z}_i^{y_i}}}{\sum_{j=1}^{C}\mathrm{e}^{\tilde{z}_i^j}}\right)$；
22. END FOR;
23. 计算总损失与梯度：$\mathcal{L}=\sum_{i=1}^{n}\mathcal{L}_i$、$\nabla_{D'}\mathcal{L}$；
24. 更新对抗样本：$D'=D'+\epsilon\cdot\text{sign}\left(\nabla_{D'}\mathcal{L}\right)$；
25. END FOR;
26. RETURN D'.

3 实验及分析

3.1 实验设置

数据集:本文采用对抗攻击领域常用的ImageNet-Compatible数据集及CIFAR-10数据集评估各方法的性能。其中,ImageNet-Compatible数据集包含1 000张尺寸为299 × 299的图像。CIFAR-10数据集包含60 000张32 × 32的图像。为确保测试数据的类别均衡性及评估的代表性,从CIFAR-10测试集中采用分层采样方法,选取1 000张图像,每个类别均包含100张图像。

实验模型。为验证本文方法的有效性,在ImageNet-Compatible数据集上采用三类模型对本文方法进行测试:第一类为基于卷积神经网络架构的模型,包括VGG-16(VGG16)[24]、ResNet-18(RN-18)[1]、ResNet-50(RN-50)[1]、DenseNet-121(DN-121)[2]、Xception(Xcep)、MobileNet-v2(MB-v2)[25]、EfficientNet-B0(EF-B0)[26]、Inception ResNet-v2(IR-v2)[27]、Inception-v3(Inc-v3)[28]和Inception-v4(Inc-v4)[27];第二类为基于Transformer架构的模型,包括ViT[29]、LeViT[30]、ConViT[31]、Twins[32]和PiT[33];第三类为经对抗训练增强的ResNet-50模型(adv-RN-50),该模型是在满足ℓ_2范数约束($\|\delta\|_2 \leqslant 0.1$)的对抗样本上进行对抗性训练得到的鲁棒模型。此外,在CIFAR-10数据集上也进行了实验,采用由三个ResNet-20[1]所组成的集成模型(ens3-RN-20),并在Baseline、ADP[34]与GAL[35]三种防御设置下进行评估,以全面验证所提方法的泛化能力。

基线攻击。为了评估所提SFDFF方法的具体效果,选择13种现有技术的各种组合组成基线攻击:DI[12]、RDI[13]、MI[9]、TI[14]、SI[10]、VT[11]、Admix[15]、ODI[36]、CFM[21]、TFM[22]、TFM+NCE[22]、FTM[23]与FTM-E[23]。本文将MI[9]与TI[14]技术与实验中所有攻击方法相结合,故在后续中省略"MI-TI"的表示。鉴于在固定输入尺寸下迭代优化对抗样本时扰动容易在源模型上过拟合,本文在除DI和ODI之外的所有实验中选用具备输入多样化机制的RDI方法作为辅助基线攻击(因DI与RDI功能重叠,ODI与RDI设计逻辑冲突,所以实验中未将二者与RDI相结合),以更公平、全面地评估各方法所生成对抗样本迁移性的差异。此外,本文所有实验均使用NVIDIA GeForce RTX 4090完成。

3.2 对比实验

3.2.1 定量实验

首先,本文选取RN-50、adv-RN-50与DN-121三个基础模型作为源模型以生成对抗样本,在ImageNet-Compatible数据集上与现有先进方法进行对比实验,以此来全面验证本文方法所生成对抗样本的迁移性。具体实验结果如表1所示。

表1　在ImageNet-Compatible数据集上针对10个目标模型的攻击情况　　单位:%

Table 1　Targeted attack success rates against ten target models on the ImageNet-Compatible dataset　　unit: %

源模型:RN-50	目标模型										
攻击方法	VGG16	RN-18	RN-50	DN-121	Xcep	MB-v2	EF-B0	IR-v2	Inc-v3	Inc-v4	Avg.
DI	62.5	56.6	98.9	72.3	5.7	28.2	29.3	4.5	9.2	9.9	37.7
RDI	65.4	71.8	98.0	81.3	13.1	46.6	46.6	16.8	30.7	23.9	49.4
SI-RDI	70.5	79.8	98.8	88.9	29.5	56.2	66.2	37.9	56.4	43.6	62.8
VT-RDI	68.8	78.7	98.2	82.5	27.9	54.5	56.1	32.8	45.8	37.9	58.3
Admix-RDI	74.2	80.7	98.7	86.8	20.9	59.4	56.1	26.7	42.7	34.1	58.0
ODI	78.3	77.1	97.6	87.0	43.8	67.3	70.0	49.5	65.9	55.4	69.2
CFM-RDI	84.7	88.4	98.4	90.3	51.1	81.5	78.8	48.0	65.5	59.3	74.6
TFM-RDI	87.1	88.5	98.7	90.6	50.7	83.1	79.7	47.7	66.4	61.9	75.4
TFM-RDI+NCE	91.0	93.8	**100.0**	95.9	56.1	89.9	85.5	52.9	72.5	68.3	80.6
FTM-RDI	86.3	87.5	97.9	89.7	56.1	83.3	81.2	54.7	70.8	66.6	77.4
FTM-RDI-E	88.1	88.6	98.3	91.8	59.4	85.4	84.3	56.9	73.4	69.3	79.6
SFDFF-RDI	92.6	95.4	**100.0**	97.6	56.1	90.1	86.7	53.1	72.4	65.7	81.0
SFDFF-RDI-E	**93.7**	**97.4**	**100.0**	**98.2**	**61.3**	**92.5**	**90.9**	**59.0**	**76.6**	**71.9**	**84.2**
源模型:adv-RN-50	目标模型										
攻击方法	VGG16	RN-18	RN-50	DN-121	Xcep	MB-v2	EF-B0	IR-v2	Inc-v3	Inc-v4	Avg.
DI	65.3	81.5	91.5	87.0	32.6	62.5	68.8	36.9	55.3	42.2	62.4
RDI	59.7	83.5	90.7	85.9	39.7	67.0	68.8	44.2	62.4	45.1	64.7
SI-RDI	53.9	79.4	87.1	83.8	46.6	66.5	69.5	52.0	69.1	52.2	66.0

续表

源模型:adv-RN-50	目标模型										
攻击方法	VGG16	RN-18	RN-50	DN-121	Xcep	MB-v2	EF-B0	IR-v2	Inc-v3	Inc-v4	Avg.
VT-RDI	54.0	76.8	84.7	81.2	38.5	60.3	58.7	42.7	56.1	44.9	59.8
Admix-RDI	62.7	83.0	90.3	86.6	46.9	71.8	72.4	48.8	66.3	53.0	68.2
ODI	62.0	77.6	84.3	85.0	56.3	66.9	73.0	61.1	71.9	60.0	69.8
CFM-RDI	76.7	86.3	90.9	87.6	67.1	82.4	83.4	64.7	77.1	67.4	78.4
TFM-RDI	79.8	87.9	92.4	89.8	67.5	84.6	85.3	67.4	79.7	69.9	80.4
TFM-RDI+NCE	85.7	93.2	95.4	94.4	73.3	89.3	89.7	70.9	83.5	75.3	85.1
FTM-RDI*	75.7	83.9	87.6	85.6	65.1	79.9	79.3	62.4	74.2	66.0	76.0
FTM-RDI-E*	78.7	85.8	88.8	86.8	67.6	83.2	82.1	68.0	78.3	70.3	79.0
SFDFF-RDI	85.0	**95.8**	97.1	95.5	73.6	91.4	91.5	71.5	86.0	75.1	86.3
SFDFF-RDI-E	**87.2**	95.6	**97.4**	**95.7**	**74.6**	**92.4**	**92.1**	**73.5**	**87.1**	**76.2**	**87.2**
源模型:DN-121	目标模型										
攻击方法	VGG16	RN-18	RN-50	DN-121	Xcep	MB-v2	EF-B0	IR-v2	Inc-v3	Inc-v4	Avg.
DI	37.4	28.7	44.4	98.7	5.2	13.1	18.7	4.3	7.1	8.3	26.6
RDI	42.1	48.8	55.7	98.5	10.1	21.0	29.0	12.8	20.8	18.8	35.8
SI-RDI	45.4	53.0	60.1	98.6	16.1	27.8	37.3	22.0	34.3	25.8	42.0
VT-RDI	47.7	56.7	62.1	98.6	20.3	28.7	36.9	25.4	31.5	27.2	43.5
Admix-RDI	49.6	60.4	65.3	98.6	21.6	34.8	43.5	28.9	41.0	34.3	47.8
ODI	64.2	64.2	71.7	98.0	31.4	45.9	56.1	39.8	52.8	45.9	57.0
CFM-RDI	76.2	79.0	83.9	97.8	41.1	62.5	68.6	43.6	56.1	53.8	66.3
TFM-RDI	76.2	80.7	84.3	98.1	42.4	64.2	69.1	44.6	59.9	53.8	67.3
TFM-RDI+NCE	**85.7**	89.6	92.5	**100.0**	**48.2**	73.4	**78.0**	**50.6**	67.4	62.5	74.8
FTM-RDI*	79.3	81.6	85.4	97.8	47.6	67.6	73.3	48.0	62.8	58.7	70.2
FTM-RDI-E*	81.6	84.4	86.4	97.7	47.9	68.8	75.0	49.6	65.5	61.6	71.9
SFDFF-RDI	83.3	87.2	93.1	**100.0**	44.2	70.0	76.1	48.2	65.1	58.9	72.6
SFDFF-RDI-E	85.6	**90.8**	**93.8**	**100.0**	47.9	**73.5**	77.0	49.1	**68.5**	**62.6**	**74.9**

注:最优实验结果用加粗表示,“E”表示集成攻击(即集成两个相同源模型,对损失求和以生成对抗样本),“*”表示在与原论文设置相同条件下复现的方法。

由表1实验结果可知,本文所提SFDFF在大多数目标模型上表现出较好的攻击性能,整体水平优于现有基线方法。但当以DN-121作为源模型生成对抗样本攻击目标模型时,SFDFF的平均攻击成功率比TFM+NCE低2.2%。其主要原因在于DN-121的密集连接结构导致Rank-1特征主导性更强,TFM+NCE通过截断干净样本高层Rank-1特征直接缓解对抗扰动对DN-121的过拟合,提高扰动在不同目标模型上的迁移性。尽管如此,SFDFF在其他源模型上的表现仍具有明显优势,且其衍生的集成攻击方法SFDFF-E的平均攻击成功率超过了所有对比方法,达到最优,充分体现出本文所提SFDFF在多数场景下的通用性与有效性。

为进一步验证所提方法的有效性,本文在ImageNet-Compatible数据集上构建了跨架构攻击实验,继续使用前述的三个基础模型作为源模型,对五个基于Transformer架构的模型以及adv-RN-50模型进行攻击测试,实验结果如表2所示。

表2 在ImageNet-Compatible数据集上针对鲁棒模型与5个基于Transformer模型的攻击情况 单位:%

Table 2 Targeted attack success rates against a robust model and five Transformer-based classifiers with the ImageNet-Compatible dataset unit: %

源模型:RN-50	目标模型						
攻击方法	adv-RN-50	ViT	LeViT	ConViT	Twins	PiT	Avg.
DI	10.9	0.1	3.6	0.3	1.3	1.5	3.0
RDI	34.8	0.7	13.1	1.9	5.9	6.8	10.5
SI-RDI	59.9	2.9	29.4	6.3	15.5	17.9	22.0
VT-RDI	64.2	2.9	28.1	5.2	15.0	14.0	21.6

续表

源模型:RN-50	目标模型						
攻击方法	adv-RN-50	ViT	LeViT	ConViT	Twins	PiT	Avg.
Admix-RDI	52.4	1.3	22.5	2.5	8.5	8.4	15.9
ODI	64.7	5.1	37.0	10.7	20.1	29.1	27.8
CFM-RDI	75.5	4.3	46.1	8.9	25.2	24.7	30.8
TFM-RDI	75.9	4.4	47.0	8.8	25.1	27.6	31.5
TFM-RDI+NCE	82.1	4.5	52.4	11.1	26.9	30.6	34.6
FTM-RDI*	78.1	5.9	52.9	10.8	32.4	31.5	35.3
FTM-RDI-E *	81.3	**6.8**	**58.6**	**13.6**	**35.2**	**34.9**	**38.4**
SFDFF-RDI	82.4	4.5	51.1	10.6	28.5	29.6	34.5
SFDFF-RDI-E	**85.9**	6.3	56.8	12.7	32.6	34.2	38.1
源模型:adv-RN-50	目标模型						
攻击方法	adv-RN-50	ViT	LeViT	ConViT	Twins	PiT	Avg.
DI	98.9	5.7	36.9	10.1	19.2	20.5	31.9
RDI	98.8	10.8	49.5	19.9	29.4	35.8	40.7
SI-RDI	98.7	19.4	57.6	35.3	35.2	52.1	49.7
VT-RDI	98.5	10.6	46.3	20.0	27.1	34.4	39.5
Admix-RDI	98.9	12.1	55.5	23.1	32.4	38.9	43.5
ODI	97.3	22.2	57.7	38.8	40.0	54.9	51.8
CFM-RDI	98.3	29.5	69.8	41.8	52.7	59.8	58.7
TFM-RDI	98.4	30.0	70.8	45.4	54.1	63.4	60.4
TFM-RDI+NCE	99.6	32.6	76.7	47.5	59.4	68.1	64.0
FTM-RDI*	97.8	31.0	70.0	45.6	52.8	60.9	59.7
FTM-RDI-E*	97.9	**35.4**	71.7	47.6	57.1	64.6	62.4
SFDFF-RDI	**100.0**	31.5	79.2	47.7	57.0	68.0	63.9
SFDFF-RDI-E	**100.0**	33.2	**80.7**	**48.7**	**60.8**	**69.8**	**65.5**
源模型:DN-121	目标模型						
攻击方法	adv-RN-50	ViT	LeViT	ConViT	Twins	PiT	Avg.
DI	3.2	0.2	3.0	0.4	1.0	1.1	1.5
RDI	10.1	0.8	8.5	1.3	3.7	4.5	4.8
SI-RDI	19.2	2.0	16.1	2.4	8.2	11.7	9.9
VT-RDI	26.6	2.2	19.2	3.5	8.3	11.7	11.9
Admix-RDI	19.2	1.0	14.7	1.7	6.8	7.4	8.5
ODI	35.6	3.3	26.9	7.4	14.7	21.9	18.3
CFM-RDI	43.2	3.6	32.8	6.4	17.3	21.1	20.7
TFM-RDI	43.5	2.7	34.8	6.1	19.3	21.4	21.3
TFM-RDI+NCE	50.7	3.1	41.8	7.2	22.3	23.7	24.8
FTM-RDI*	47.7	4.1	41.1	**8.8**	22.9	25.4	25.0
FTM-RDI-E*	**52.8**	**4.7**	**44.1**	8.1	**26.1**	**28.3**	**27.4**
SFDFF-RDI	48.8	4.0	38.1	7.3	21.2	23.7	23.9
SFDFF-RDI-E	52.3	4.5	42.5	7.8	23.0	27.8	26.3

注:最优实验结果用加粗表示,“E”表示集成攻击(即集成两个相同源模型,对损失求和以生成对抗样本),“*”表示在与原论文设置相同条件下复现的方法。

由表2可知,SFDFF的平均攻击成功率显著优于大部分基线方法,且以adv-RN-50为源模型时,其集成攻击方法SFDFF-E的平均攻击成功率相比当前最优的集成攻击方法FTM-E提高了3.1%,达到了最优。但当以RN-50和DN-121作为源模型生成对抗样本攻击目标模型时,SFDFF的平均攻击成功率略低于FTM。其主要原因在于,FTM通过在空间域内引入干净特征与可学习扰动,对当前样本特征进行调整,从

而在一定程度上缓解了扰动对源模型的过拟合。但据FTM原论文所述,启用干净特征后,随着更新概率p从0.1增大到1.0,FTM的性能会出现显著下降[23]。总体而言,本文方法在攻击鲁棒模型以及基于Transformer架构的目标模型时,表现出较强竞争力。

为充分检验本文方法的性能,本文进一步在CIFAR-10数据集上展开攻击实验。实验选取RN-50作为源模型,对五种常规模型与三种经对抗训练的集成模型进行攻击测试。详细实验结果如表3所示。

表3 在CIFAR-10数据集上针对8个目标模型的攻击情况 单位:%

Table 3 Targeted attack success rates against eight target models on the CIFAR-10 dataset unit: %

源模型:RN-50	目标模型								
攻击方法	VGG16	RN-18	MB-v2	Inc-v3	DN-121	ens3-RN-20			Avg.
						Baseline	ADP	GAL	
DI	66.4	71.5	62.7	71.1	84.2	77.9	56.5	14.3	63.1
RDI	66.4	70.9	64.1	73.4	82.8	76.3	55.8	13.5	62.9
SI-RDI	72.9	76.3	77.1	77.0	84.7	81.2	65.5	20.0	69.3
VT-RDI	89.8	87.1	92.6	92.9	93.7	94.4	82.3	24.3	82.1
Admix-RDI	74.2	78.8	76.2	82.7	89.2	85.2	66.4	17.3	71.3
CFM-RDI	98.3	97.7	99.0	99.0	99.2	98.8	97.2	54.9	93.0
TFM-RDI	98.3	97.9	99.2	98.7	99.2	98.3	97.4	61.2	**93.8**
TFM-RDI+NCE	98.2	97.5	98.4	98.8	99.2	98.5	97.9	**62.0**	**93.8**
SFDFF-RDI	**99.1**	**99.0**	**99.5**	**99.5**	**99.6**	**99.4**	**98.1**	55.8	**93.8**

注:表中最优实验结果用加粗表示。

由表3可知,SFDFF在CIFAR-10数据集上的表现良好,在除GAL之外的所有目标模型上的攻击成功率均超过了当前最优的TFM+NCE方法,且SFDFF的平均攻击成功率达到了最优。这一结果不仅进一步验证了所提方法在小尺寸、结构紧凑图像任务中的有效性,也侧面印证了本文所提SFDFF方法在不同任务场景下的通用性,体现出该方法具有较强的泛化能力与实用潜力。

为进一步评估不同方法生成的对抗样本图像质量及对抗扰动的抗检测能力,本文引入四个常用图像质量评价指标:均方误差(Mean Squared Error,MSE)、峰值信噪比(Peak Signal-to-Noise Ratio,PSNR)、结构相似性(Structural Similarity Index Measure,SSIM)以及学习感知图像块相似度(Learned Perceptual Image Patch Similarity,LPIPS),以RN-50作为源模型,在ImageNet-Compatible数据集上展开实验。在实验过程中,本文首先采用多种攻击方法针对数据集内全体样本生成所对应的对抗样本。随后,以每对原始样本及其对抗样本为单位计算四项指标,并将全体样本的结果取算术平均,从而获得不同方法的评估指标。详细实验结果如表4所示。此外,表4中所有方法的数据均为在与其原论文相同环境下复现出来的结果。

表4 图像质量评估实验 单位:%

Table 4 Image quality evaluation experiments unit: %

源模型:RN-50	图像质量评估指标			
攻击方法	SSIM ↑	MSE ↓	PSNR ↑	LPIPS ↓
DI	0.670 4	0.003 6	24.479 5	0.369 9
RDI	0.670 7	0.003 6	24.450 1	0.378 9
SI-RDI	0.678 1	0.003 6	24.435 0	0.365 9
VT-RDI	0.686 8	0.003 6	24.400 0	0.364 4
Admix-RDI	0.679 3	0.003 6	24.454 6	0.368 4
ODI	0.665 6	0.003 6	24.443 4	0.392 2
CFM-RDI	0.670 9	0.003 6	24.443 0	0.379 0
FTM-RDI	0.670 9	0.003 6	24.437 5	0.383 2
FTM-RDI-E	0.673 5	0.003 6	24.447 8	0.377 8
SFDFF-RDI	0.669 3	0.003 6	24.441 8	0.382 5
SFDFF-RDI-E	0.671 7	0.003 6	24.448 6	0.375 4

注:↑表示越大越好,↓表示越小越好。

根据表4实验结果可知,SFDFF方法生成的对抗样本在整体表现上较为优异。其中,SFDFF在MSE指标上取得最优,在SSIM、PSNR与LPIPS指标上与最优指标接近,从而充分体现出本文方法所生成的对抗样本与原始样本具有较高的相似性以及对抗扰动具有较强的抗检测能力。综合各项指标分析,SFDFF在提升对抗样本迁移性的同时,能够兼顾并保障样本质量,表现出较为均衡的优势。

3.2.2　定性实验

在定性实验中,本文选用adv-RN-50作为源模型,随机选取两个原始样本,并基于所选样本分别采用六种先进方法生成对抗样本。为更直观地展示各方法生成对抗样本的差异及其攻击特性,进一步生成对抗样本在RN-50模型上的热力图。同时,为客观评估各方法的攻击效果,本文统计了每种方法生成的对抗样本在十六个异构目标模型上的平均攻击成功率,以衡量其在黑盒场景下的迁移能力。不同对抗样本的可视化结果如图2所示。图2中,第一行、第二行图像分别表示不同方法所生成的对抗样本及其热力图。

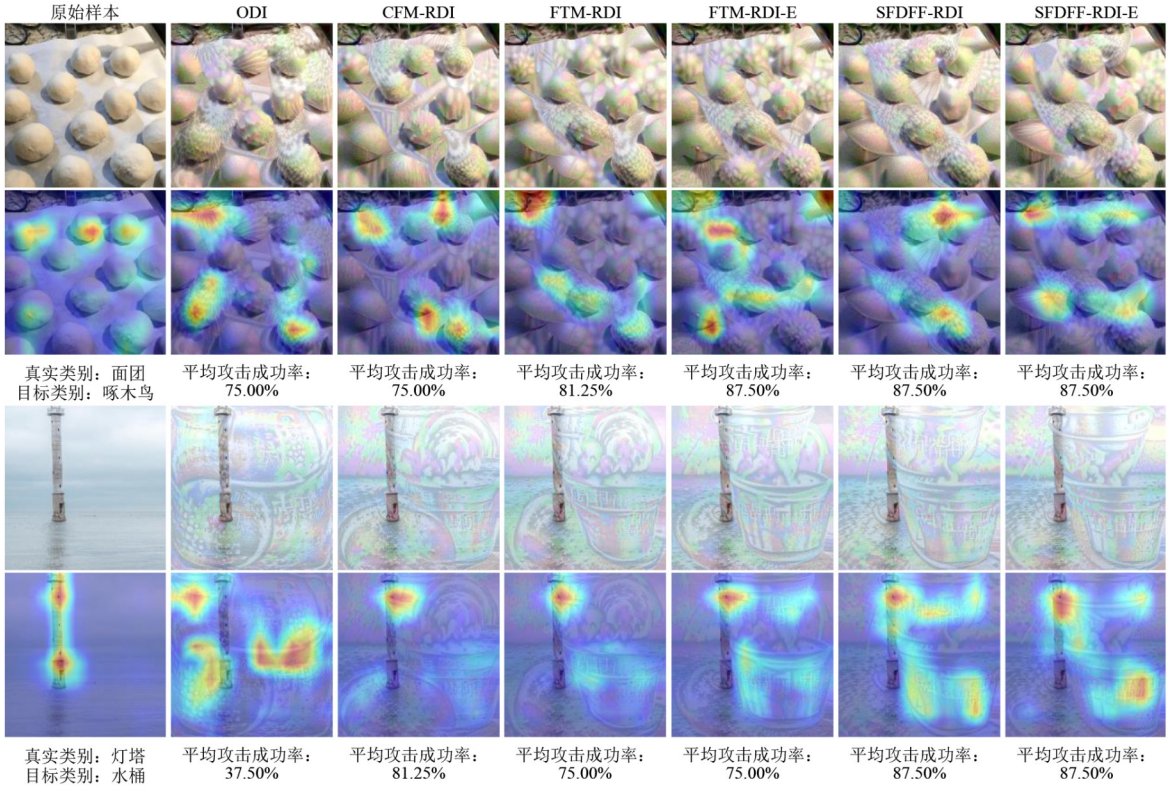

图2　多种攻击方式下的对抗样本可视化

Figure 2　Adversarial example visualization with different attack methods

由图2可知,SFDFF与SFDFF-E所生成对抗样本的热力图高亮区域分布杂乱,对关键部位的关注减弱,异常区域突出,且平均攻击成功率显著优于其他方法,充分说明了SFDFF与SFDFF-E在提升对抗样本迁移性方面的有效性。

3.3　消融实验

为充分验证所提SFDFF方法各组成部分的有效性,本文针对其内部的频域融合、OSSN与DDLoss模块进行了消融实验,详细实验结果如表5所示。实验选用adv-RN-50作为源模型,默认仅在空间域内进行扰动(对应表5中第一行结果)。

由表5实验结果可知,当未引入任何模块,即仅在空间域内进行扰动时,其在9个代表性目标模型上的平均攻击成功率为58.3%。在此基础上进一步引入频域融合,即在空频双域内进行扰动时,平均攻击成功率提高了0.4%,说明了在空频双域内进行协同扰动的有效性。此外,当分别单独引入OSSN与DDLoss模块时,平均攻击成功率相对于表5中第一行的默认结果也分别提升了0.3%与6.0%,说明了OSSN与DDLoss模块的有效性。进一步分析表5结果可知,同时引入频域融合、OSSN与DDLoss模块时获得了最高的平均攻击成功率,充分说明了当本文提出的这三个模块协同作用时可有效提高对抗样本迁移性。

为研究在频率域内所选融合区域大小对实验结

表5 针对SFDFF内部模块的消融实验 单位:%

Table 5 Ablation experiments on the internal modules of SFDFF unit: %

消融项			目标模型									
频域融合	OSSN	DDLoss	Xcep	IR-v2	Inc-v3	Inc-v4	ViT	LeViT	ConViT	Twins	PiT	Avg.
			65.4	63.2	76.4	67.1	29.5	69.6	42.3	51.4	60.0	58.3
√			66.7	64.1	76.2	68.6	29.6	70.3	42.0	51.7	59.3	58.7
	√		66.5	65.3	76.8	68.3	28.2	69.4	41.8	50.8	60.0	58.6
		√	71.2	72.2	84.6	74.1	30.1	77.5	46.9	56.0	66.1	64.3
√	√		65.2	64.0	76.6	67.6	29.1	70.0	42.3	51.8	59.3	58.4
√		√	71.1	71.6	84.8	74.3	30.5	76.7	45.8	57.0	66.8	64.3
	√	√	71.3	70.4	86.5	75.0	31.0	78.2	47.7	56.9	67.1	64.9
√	√	√	73.6	71.5	86.0	75.1	31.5	79.2	47.7	57.0	68.0	65.5

注:当未引入DDLoss时默认使用Logit损失,反之则使用DDLoss。

果的影响,本文选用adv-RN-50作为源模型攻击九个目标模型。具体实验结果如表6所示。由表6实验结果可知,当所选频率域融合区域大小为8×8时,平均攻击成功率最高。

表6 对频率域内使用不同区域大小进行融合的实验结果 单位:%

Table 6 Experimental results of fusion with different region sizes in the frequency domain unit: %

源模型:adv-RN-50	目标模型									
融合区域大小	Xcep	IR-v2	Inc-v3	Inc-v4	ViT	LeViT	ConViT	Twins	PiT	Avg.
4×4	73.2	70.8	85.3	75.1	29.3	78.0	47.5	55.4	65.5	64.5
8×8(默认)	73.6	71.5	86.0	75.1	31.5	79.2	47.7	57.0	68.0	65.5
12×12	72.9	70.6	85.0	75.5	30.7	78.6	46.3	57.4	65.4	64.7
16×16	72.4	70.2	84.8	75.0	32.1	76.4	45.6	55.4	66.5	64.3
20×20	73.1	70.3	84.9	76.0	31.4	78.0	47.0	55.3	65.7	64.6
24×24	71.2	70.3	85.0	75.1	31.0	78.8	46.4	56.2	67.0	64.6

为研究在频率域特征融合过程中应用区域级融合与随机打乱对攻击效果的影响,本文选用RN-50作为源模型进行消融实验。具体实验结果如表7所示。在表7中,当引入区域级融合时,频率域特征将在区域范围内进行融合,反之则在全局范围内进行融合。当引入随机打乱时,将在输入样本内融合随机样本原始频率域特征并引入竞争机制,反之则融合自身原始频率域特征且未引入竞争机制。

表7 针对区域级频率域特征融合的消融实验 单位:%

Table 7 Ablation experiments on region-level frequency domain feature fusion unit: %

消融项		目标模型									
区域级融合	随机打乱	Xcep	IR-v2	Inc-v3	Inc-v4	ViT	LeViT	ConViT	Twins	PiT	Avg.
		55.2	53.2	72.3	65.5	4.8	50.2	9.2	27.5	28.8	40.7
√		56.3	53.6	72.1	64.8	4.8	50.8	9.5	28.4	28.5	41.0
	√	56.7	52.8	71.1	65.9	4.7	50.2	10.1	27.9	28.5	40.9
√	√	56.1	53.1	72.4	65.7	4.5	51.1	10.6	28.5	29.6	41.3

由表7结果可知,当引入区域级融合时,平均攻击成功率提高了0.3%。在此基础上进一步引入随机打乱时,平均攻击成功率再次提高了0.3%,达到了最优。这一结果表明,在频率域特征融合过程中,同时引入区域级融合与随机打乱可有效提高对抗样本迁移性。

此外,为探究丢弃不同比例频率域扰动高频分量对实验结果的影响,本文对丢弃频率域扰动高频分量的比例进行消融实验。实验采用随机丢弃法,在频率域内融合扰动之前,随机丢弃频率域扰动中不同比例的高频分量。详细实验结果如表8所示。

由表8实验结果可知,丢弃频率域扰动高频分量仅能轻微影响所生成对抗样本的攻击效果,且不同丢弃比例下的平均攻击成功率方差仅为0.068,验证了SFDFF在频率域扰动净化场景下具有较强的鲁棒性。需明确,本文方法默认未丢弃频率域扰动高频分量,且由表8实验结果可知,该情况下的平均攻击成功率最高(对应表8中第一行的结果)。

表8　随机丢弃频率域扰动高频分量比例的实验结果　　单位:%

Table 8　Experimental results of random dropping of high-frequency component proportion in frequency domain perturbations　　unit: %

源模型:RN-50	目标模型									
丢弃高频分量比例	Xcep	IR-v2	Inc-v3	Inc-v4	ViT	LeViT	ConViT	Twins	PiT	Avg.
0%(默认)	56.1	53.1	72.4	65.7	4.5	51.1	10.6	28.5	29.6	41.3
20%	55.9	54.0	72.3	65.4	4.6	50.7	10.2	27.6	29.0	41.1
40%	56.1	52.4	73.0	64.4	3.9	51.6	9.3	27.4	28.7	40.8
60%	55.6	51.9	72.7	65.0	4.9	50.2	11.0	26.0	28.3	40.6
80%	56.0	53.6	71.4	65.8	5.0	50.3	9.1	27.0	28.1	40.7

为进一步研究双导向损失(DDLoss)中λ的取值对实验结果的影响,本文选用RN-50作为源模型攻击九个目标模型。详细实验结果如表9所示。

在优化过程中,DDLoss的前项与后项表现出互补作用。前项能够显著推动扰动向目标类别靠拢,但当其权重过大时,可能导致优化过程过于激进,从而出现梯度震荡与过拟合现象;后项则通过概率分布层面的平滑约束,为梯度提供稳定信号,保证扰动的泛化能力。然而,若单独依赖后项,则目标导向性不足,难以充分提升攻击成功率。通过实验对不同权重比例进行对比发现,当$\lambda=0.2$时整体性能最佳,此时前项提供了适度的目标类别增强信号,有效避免陷入局部最优;后项则在主导作用下维持了梯度方向的稳定性,从而在跨模型迁移性与攻击成功率方面均取得最优表现。

同时,为探究DDLoss中m的取值对实验结果的影响,本文选用RN-50作为源模型攻击九个目标模型。实验结果如表10所示。

表9　损失函数中使用不同λ的实验结果　　单位:%

Table 9　Experimental results with different λ values in the loss function　　unit: %

源模型:RN-50	目标模型									
λ取值	Xcep	IR-v2	Inc-v3	Inc-v4	ViT	LeViT	ConViT	Twins	PiT	Avg.
0.0	54.0	53.1	73.1	64.2	4.6	49.1	9.1	27.2	26.8	40.1
0.1	55.8	53.4	72.5	66.0	4.9	51.3	9.7	28.1	29.4	41.2
0.2(默认)	56.1	53.1	72.4	65.7	4.5	51.1	10.6	28.5	29.6	41.3
0.4	54.1	52.7	72.0	63.3	5.0	49.0	10.3	27.8	27.6	40.2
0.6	54.3	51.0	71.5	65.2	4.2	49.3	9.8	26.0	27.3	39.8
0.8	54.4	52.9	71.8	62.8	4.2	47.5	9.7	25.7	26.3	39.5

表10　损失函数中使用不同m的实验结果　　单位:%

Table 10　Experimental results with different m values in the loss function　　unit: %

源模型:RN-50	目标模型									
m取值	Xcep	IR-v2	Inc-v3	Inc-v4	ViT	LeViT	ConViT	Twins	PiT	Avg.
2	55.0	53.2	71.3	64.2	4.6	48.6	9.1	27.9	27.7	40.2
5(默认)	56.1	53.1	72.4	65.7	4.5	51.1	10.6	28.5	29.6	41.3
10	55.7	52.8	72.2	66.3	5.0	49.8	9.1	27.4	30.3	41.0

由表10实验结果可知,当$m=5$时对抗样本平均攻击成功率最高。这一取值既能覆盖最具威胁的若干非目标类别,又能保持梯度稳定性,从而提高对抗扰动在跨模型攻击中的泛化能力。

为研究在样本中线性融合不同噪声对实验结果的影响,本文分别向样本中融合高斯噪声(GN)、均匀分布噪声(UN)、基于训练数据整体统计特征的噪声(SN)与基于原始样本统计特征的噪声(OSSN)进行实验。实验结果如表11所示。表11结果表明,当融合OSSN时,SFDFF的平均攻击成功率最高,说明本文所提OSSN在提升对抗样本迁移性方面具有一定的优势。

最后,为研究不同扰动量ϵ对实验结果的影响,本文选用RN-50作为源模型攻击十六个目标模型,取攻击结果的平均值进行比较。实验结果如表12所示。表12实验结果表明,SFDFF的平均攻击成功率较当前最优的非集成攻击方法TFM+NCE提高了0.11%,且SFDFF-E的平均攻击成功率较当前最优的集成攻击方法FTM-E提高了2.55%,验证了本文所提方法的先进性。

表 11 线性融合不同种类噪声的对比实验 单位:%

Table 11 Comparison experiments on linear fusion of different types of noise unit: %

源模型:RN-50	目标模型									
噪声类型	Xcep	IR-v2	Inc-v3	Inc-v4	ViT	LeViT	ConViT	Twins	PiT	Avg.
GN	56.0	53.5	71.3	66.0	4.5	50.4	9.6	28.9	29.2	41.0
UN	55.2	52.8	72.4	65.6	5.1	49.9	10.2	27.8	27.6	40.7
SN	55.5	53.0	72.8	66.1	4.9	50.2	10.5	27.7	29.6	41.1
OSSN(默认)	56.1	53.1	72.4	65.7	4.5	51.1	10.6	28.5	29.6	41.3

表 12 针对不同扰动量ϵ的对比实验 单位:%

Table 12 Comparison experiments for different perturbation budgets ϵ unit: %

源模型:RN-50	扰动量			
攻击方法	$\epsilon=8$	$\epsilon=16$	$\epsilon=32$	Avg.
DI	14.64	34.00	34.11	27.58
RDI	20.12	34.84	49.32	34.76
SI-RDI	24.55	47.48	65.96	46.00
VT-RDI	24.39	44.54	59.59	42.84
Admix-RDI	25.09	42.24	56.24	41.19
ODI	29.12	53.66	69.43	50.74
CFM-RDI	34.09	58.17	74.44	55.57
TFM-RDI	34.54	59.95	75.63	56.71
TFM-RDI+NCE	37.50	63.34	80.32	60.39
FTM-RDI*	37.04	61.61	76.51	58.39
FTM-RDI-E*	38.78	64.12	79.71	60.87
SFDFF-RDI	37.48	63.53	80.50	60.50
SFDFF-RDI-E	40.27	66.88	83.10	63.42

4 结论

针对现有对抗攻击方法所生成对抗样本迁移性不足的问题,本文提出了基于空频双域特征融合的高迁移性对抗样本生成方法。所提方法通过在频率域与空间域内融合原始样本特征,促使扰动在优化过程中充分利用空频双域内广泛的特征以克服原始特征的干扰,并在两者之间穿插注入基于原始样本统计特征的噪声以多样化特征分布。同时,所提出的双导向损失使得生成的对抗样本能够更好地误导目标模型。基于 ImageNet-Compatible 与 CIFAR-10 数据集的全面实验结果表明,本文所提方法在提高对抗样本迁移性方面表现较出较强的竞争力。

参考文献

[1] 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.

[2] Huang Gao, Liu Zhuang, Van Der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2261-2269.

[3] Zhang Peiyuan, Luo Junwei, Yang Xue, et al. PointOBB-v3: Expanding performance boundaries of single point-supervised oriented object detection[J]. International Journal of Computer Vision, 2025, 133(9): 6108-6128.

[4] Lin Zhiwei, Liu Zhe, Xia Zhongyu, et al. RCBEVDet: Radar-camera fusion in bird's eye view for 3D object detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 14928-14937.

[5] 张世辉, 张晓微, 宋丹丹, 等. 基于逆扰动融合生成对抗网络的对抗样本防御方法[J]. 电子学报, 2023, 51(4): 879-884.
Zhang Shihui, Zhang Xiaowei, Song Dandan, et al. Adversarial example defense method based on inverse perturbation fusing generative adversarial network[J]. Acta Electronica Sinica, 2023, 51(4): 879-884. (in Chinese)

[6] 吴骥, 邵文泽, 葛琦, 等. 一种基于迭代累积梯度的多层特征重要性攻击方法[J]. 电子学报, 2024, 52(11): 3798-3808.
Wu Ji, Shao Wenze, Ge Qi, et al. A multi-layer feature importance attack method based on iterative accumulated gra-

dients[J]. Acta Electronica Sinica, 2024, 52(11): 3798-3808. (in Chinese)

[7] 王硕, 徐茹枝, 关志涛. 基于主特征归因的对抗样本生成方法研究[J]. 电子学报, 2023, 51(11): 3137-3145.
Wang Shuo, Xu Ruzhi, Guan Zhitao. Research on the generation of adversarial samples based on the attribution of principal features[J]. Acta Electronica Sinica, 2023, 51(11): 3137-3145. (in Chinese)

[8] 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.

[9] 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.

[10] Lin Jiadong, Song Chuanbiao, He Kun, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[PP/OL]. V5. arXiv (2020-02-03) [2025-06-20]. https://arXiv.org/abs/1908.06281.

[11] Wang Xiaosen, He Kun. Enhancing the transferability of adversarial attacks through variance tuning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 1924-1933.

[12] 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.

[13] Zou Junhua, Pan Zhisong, Qiu Junyang, et al. Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting[C]// Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 563-579.

[14] 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.

[15] 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, 2022: 16138-16147.

[16] Long Yuyang, Zhang Qilong, Zeng Boheng, et al. Frequency domain model augmentation for adversarial attack[C]// Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 549-566.

[17] Li Jiachun, Hu Yuchao, Yan Cheng. FDAA: A feature distribution-aware transferable adversarial attack method[J]. Neural Networks, 2024, 178: 106467.

[18] Zhao Zhengyu, Liu Zhuoran, Larson M. On success and simplicity: A second look at transferable targeted attacks[PP/OL]. V4. arXiv (2021-10-26) [2025-10-10]. https://arxiv.org/abs/2012.11207.

[19] Weng Juanjuan, Luo Zhiming, Li Shoazi, et al. Logit margin matters: Improving transferable targeted adversarial attack by logit calibration[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 3561-3574.

[20] Inkawhich N, Liang K J, Carin L, et al. Transferable perturbations of deep feature distributions[PP/OL]. V1. arXiv (2020-04-27)[2025-06-20]. https://arXiv.org/abs/2004.12519.

[21] Byun J, Kwon M J, Cho S, et al. Introducing competition to boost the transferability of targeted adversarial examples through clean feature mixup[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 24648-24657.

[22] Weng Juanjuan, Luo Zhiming, Li Shaozi. Improving transferable targeted adversarial attack via normalized logit calibration and truncated feature mixing[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 4595-4609.

[23] Liang Kaisheng, Dai Xuelong, Li Yanjie, et al. Improving transferable targeted attacks with feature tuning Mixup[C]// 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 25802-25811.

[24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[PP/OL]. V6. arXiv (2015-04-10)[2025-06-20]. https://arXiv.org/abs/1409.1556.

[25] Sandler M, Howard A, Zhu Menglong, et al. MobileNetV2: Inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4510-4520.

[26] Tan Mingxing, Le Q V. EfficientNet: Rethinking model scaling for convolutional neural networks[PP/OL]. V5. arXiv (2020-09-11) [2025-10-10]. https://arXiv.org/abs/1905.11946.

[27] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[PP/OL]. V2. arXiv (2016-08-23) [2025-10-10]. https://arxiv.org/abs/1602.07261.

[28] 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.

[29] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[PP/OL]. V2. arXiv (2021-06-03) [2025-06-18]. https://arXiv.org/abs/2010.11929.

[30] Graham B, El-Nouby A, Touvron H, et al. LeViT: A vision transformer in ConvNet's clothing for faster inference[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 12239-12249.

[31] D'Ascoli S, Touvron H, Leavitt M L, et al. ConViT: Improving vision transformers with soft convolutional inductive biases[J]. Journal of Statistical Mechanics: Theory and Experiment, 2022, 2022(11): 114005.

[32] Chu Xiangxiang, Tian Zhi, Wang Yuqing, et al. Twins: Revisiting the design of spatial attention in vision transformers[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: 716.

[33] Heo B, Yun S, Han D, et al. Rethinking spatial dimensions of vision transformers[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 11916-11925.

[34] Pang Tianyu, Xu Kun, Du Chao, et al. Improving adversarial robustness via promoting ensemble diversity[PP/OL]. V3. arXiv (2019-05-29) [2025-10-10]. https://arxiv.org/abs/1901.08846.

[35] Kariyappa S, Qureshi M K. Improving adversarial robustness of ensembles with diversity training[PP/OL]. V1. arXiv (2019-01-28)[2025-06-20]. https://arXiv.org/abs/1901.09981.

[36] Byun J, Cho S, Kwon M J, et al. Improving the transferability of targeted adversarial examples through object-based diverse input[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 15223-15232.

作者简介

张世辉 男，1973年出生于河北省赞皇县。现为燕山大学人工智能学院教授、博士生导师。主要研究方向为计算机视觉、人工智能与模式识别、对抗样本生成与防御等。

E-mail：sshhzz@ysu.edu.cn

张 尧 男，2003年5月出生于江苏省淮安市。现为燕山大学硕士研究生。主要研究方向为计算机视觉。

E-mail：19932810534@163.com

赵鹏宇 男，2002年10月出生于河北省保定市。现为燕山大学硕士研究生。主要研究方向为对抗样本生成和计算机视觉。

E-mail：zhaopengyu200210@163.com

韩少杰 男，2001年9月出生于河北省邯郸市。现为燕山大学硕士研究生。主要研究方向为对抗样本生成和计算机视觉。

E-mail：hshaojie2023@163.com