计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(3): 752−767, 2026

基于双重信息对齐的多目标生成式对抗攻击

钱亚冠[1] 余芹芹[1] 沙嘉强[2] 鲍琦琦[1] 顾钊铨[3] 纪守领[4] 王 滨[5]

[1](浙江科技大学人工智能与信息工程学院 杭州 310023)

[2](浙江科技大学理学院 杭州 310023)

[3](哈尔滨工业大学(深圳)计算机科学与技术学院 广东深圳 518055)

[4](浙江大学计算机科学与技术学院 杭州 310007)

[5](全省智能物联网络与数据安全重点实验室(杭州海康威视数字技术股份有限公司) 杭州 310052)

(qianyaguan@zust.edu.cn)

Multi-Target Generative Adversarial Attacks Based on Dual-Information Alignment

Qian Yaguan[1], Yu Qinqin[1], Sha Jiaqiang[2], Bao Qiqi[1], Gu Zhaoquan[3], Ji Shouling[4], and Wang Bin[5]

[1] (*School of Artificial Intelligence and Information Engineering*, *Zhejiang University of Science and Technology*, *Hangzhou* 310023)

[2] (*School of Science*, *Zhejiang University of Science and Technology*, *Hangzhou* 310023)

[3] (*School of Computer Science and Technology*, *Harbin Institute of Technology* (*Shenzhen*), *Shenzhen*, *Guangdong* 518055)

[4] (*College of Computer Science and Technology*, *Zhejiang University*, *Hangzhou* 310007)

[5] (*Zhejiang Key Laboratory of Artificial Intelligence of Things Network and Data Security* (*Hikvision Digital Technology Co.*, *Ltd.*), *Hangzhou* 310052)

Abstract Deep neural networks (DNNs) have demonstrated remarkable success in numerous application areas, including computer vision, natural language processing, and speech recognition. However, recent studies have revealed that DNNs are vulnerable to adversarial attacks, particularly targeted attacks capable of precisely manipulating the outputs of unknown models, which poses significant risks to data privacy, model trustworthiness, and system security. Generative attack methods, which can efficiently create adversarial examples, have become a critical tool in advancing targeted attack techniques due to their ability to automate the generation process and reduce manual effort. Despite their potential, most existing generative attack approaches mainly focus on crafting adversarial examples for a single target class. This design results in clear limitations, such as low efficiency, restricted flexibility, and poor scalability in multi-target tasks, making them difficult to apply in complex scenarios where multiple targets must be addressed simultaneously. To address these challenges, we propose a multi-target generative attack based on dual-information (MTGA-DI). MTGA-DI employs a conditional generative model that integrates both semantic and visual information from multiple target classes, enabling efficient and adaptable multi-target attacks. Furthermore, the method significantly improves the transferability and stability of generated adversarial examples across different models and defense settings. Experimental results demonstrate that MTGA-DI surpasses previous methods on standard models as well as models enhanced with robust training and input preprocessing defenses, achieving higher attack success rates and better generalization.

收稿日期: 2025-01-10; **修回日期**: 2025-09-09

基金项目: 国家重点研发计划项目(2022YFB3102100); 国家自然科学基金项目(62476250, 62472335, 62372137); 浙江省自然科学基金重点项目(LZ22F020007)

This work was supported by the National Key Research and Development Program of China (2022YFB3102100), the National Natural Science Foundation of China (62476250, 62472335, 62372137), and the Key Project of Zhejiang Provincial Natural Science Foundation (LZ22F020007).

通信作者: 王滨(wbin2006@gmail.com)

Key words adversarial attack; targeted attack; generative model; semantic information; visual information; deep learning

摘 要 深度神经网络在很多应用领域取得了显著成功。然而，近年来的研究表明，它们容易受到对抗攻击的威胁。尤其是有目标的对抗攻击，能够精确控制未知模型的输出，对数据隐私和系统安全构成严重挑战。生成式攻击方法因其高效生成对抗样本的能力，近年来逐渐应用于有目标攻击的研究中。然而，现有的生成式攻击方法通常针对单一目标类别生成对抗样本，在多目标任务中表现出计算效率低下、灵活性不足和扩展性有限等问题。针对这些不足，提出了一种基于双重信息的多目标生成式攻击（multi-target generative attack based on dual-information, MTGA-DI）方法。该方法通过设计一个条件生成模型，充分融合目标类别的语义和视觉信息，不仅具备多目标攻击能力，还显著提升了对抗样本的迁移性和鲁棒性。实验结果表明，与现有多目标攻击方法相比，MTGA-DI在标准训练模型和鲁棒模型上的性能更优，在应对基于输入预处理的防御模型时也展现出更强的适应能力。

关键词 对抗攻击；有目标攻击；生成模型；语义信息；视觉信息；深度学习

中图法分类号 TP391.41；TP393.08

DOI: 10.7544/issn1000-1239.202550024 **CSTR:** 32373.14.issn1000-1239.202550024

随着人工智能时代的到来，深度神经网络（deep neural networks，DNNs）在多个领域中得到了广泛的应用。但研究表明[1]，它们容易受到对抗攻击的威胁。利用对抗样本的迁移特性，攻击者可以仅针对1个公开模型生成对抗样本，以黑盒方式攻击其他模型。这种对抗攻击方式还可以通过精准地操控未知模型的输出，实现有目标的攻击，严重威胁人工智能系统的安全。随着研究的不断深入，基于生成模型的对抗攻击方法逐渐引起研究者的关注[2-5]。这类方法通过训练专门的生成器直接生成对抗样本，与传统的迭代优化方法相比，显著提高了对抗样本的生成效率。同时，生成的扰动通常不依赖于特定输入，表现出较高的通用性。这一特性使得对抗样本能够更好地学习目标类别的特征，从而实现更加精确的有目标攻击。

尽管如此，目前的大多数有目标生成式攻击方法仅关注单一目标类别的对抗样本生成。这类方法存在以下关键局限性：1）计算效率低。每个目标类别需要独立训练生成器，导致计算资源和时间成本的浪费。2）灵活性不足。生成器一旦完成训练，其攻击目标即被固定，难以动态调整。3）扩展性有限。随着目标类别数量的增加，所需生成器的数量呈线性增长，在大规模攻击场景中难以高效应用。这些问题在需要频繁切换攻击目标或同时针对多个目标类别进行攻击的场景中尤为突出。同时，我们发现现有的多目标生成式攻击方法C-GSP（conditional generative models of se-mantic pattern）[6]，CGNC（CLIP-guided generative network with cross-attention modules）[7]主要集中于目标类别的语义信息学习，这导致了“振幅谱偏离”现象，影响对抗样本的攻击效果。因此，开发一种高效灵活的多目标生成式攻击方法，成为解决上述问题的重要方向。

为此，本文提出了一种基于双重信息的多目标生成式攻击（multi-target generative attack based on dual-information, MTGA-DI）方法以应对现有方法的局限性。1）为提高计算效率，本文采用了一种单一生成器的多目标优化策略，使同一生成器能够适应多个目标类别，避免了针对每个目标类别单独训练生成器的计算开销，从而降低训练和生成成本；2）为提高灵活性，本文设计一种条件生成模型，用映射网络将目标类别标签转为含语义和视觉信息的潜在向量，通过复合损失机制（语义损失与视觉损失）确保生成器全面学习利用目标信息，使其能动态适应不同类别，攻击时可灵活切换目标，无需重训生成器；3）为提高扩展性，本文提出双重信息对齐策略，融合高层语义与低层视觉信息，让对抗样本既符合目标类别特征，又具有跨类别泛化能力，使同一生成器适用于多类别，避免目标类别数量增加时训练多个生成器，以提升适应性。

为了全面验证MTGA-DI的有效性，本文进行了广泛的实验，涵盖标准模型与鲁棒模型上的攻击效果分析、不同扰动预算下的迁移性能评估，以及各方法在计算效率、灵活性和多目标场景扩展性方面的对比分析。实验结果表明，MTGA-DI有效提升了攻击成功率和对抗样本迁移能力，显示了其在多目标攻击任务中的高效性、灵活性和扩展性。

1 预备知识

1.1 生成对抗网络

生成对抗网络(generative adversarial network, GAN)是一种基于博弈论的生成模型, 由 Goodfellow 等人[8]于 2014 年提出。GAN 作为生成模型的一种重要类别, 其主要目标是学习数据的潜在分布, 从而生成与训练数据分布相似的高质量新样本。GAN 的核心机制是通过对抗学习的方式训练生成器和判别器 2 个功能互补且相互竞争的神经网络。

在这一对抗结构中, 生成器负责将随机噪声输入转化为虚拟样本, 其目标是生成足够逼真的样本以欺骗判别器; 而判别器的任务则是根据输入样本(可能是真实数据或生成器生成的虚假样本)判断其来源并输出样本为真实数据的概率值。判别器的目标是尽可能准确地区分真实样本与生成样本。GAN 的训练过程可被视为一个极小极大的博弈问题, 其目标是实现生成器与判别器之间的动态平衡, 从而使生成器生成的样本质量不断提升。其优化目标可被形式化为以下问题:

$$\min_G \max_D V(D,G)=\mathbb{E}_{x\sim p_{\text{data}}(x)}[\lg D(x)]+\mathbb{E}_{z\sim p_z(z)}[\lg(1-D(G(z)))], \tag{1}$$

其中 p_{data} 表示真实数据分布, $p_z(z)$ 表示噪声的先验分布, G 和 D 分别代表生成器和判别器。

1.2 对抗攻击

根据攻击者对目标模型掌握信息的不同, 对抗攻击可分为白盒攻击和黑盒攻击[9]。由于黑盒攻击不需要了解目标模型的内部信息, 更符合实际攻击场景, 因此本文聚焦黑盒攻击。经典的黑盒攻击方法主要分为 2 类: 1)实例相关的攻击, 如 MI-FGSM (momentum iterative fast gradient sign method)[10], DI-FGSM(diverse inputs FGSM)[11], TI-FGSM(translation-invariant FGSM)[12], SU(self-universality)[13]; 2)实例无关的攻击, 如 CDA(cross-domain transferability of adversarial perturbations)[14], C-GSP[6], CGCN[7]。

1)MI-FGSM 是在每次迭代中利用 FGSM[15]获取梯度, 同时通过引入动量项累计上一步的梯度方向, 以稳定当前迭代步的总梯度更新。其优化过程如式(2)所示。

$$\begin{aligned} g_{i+1}&=\mu\times g_i+\frac{\nabla_x\mathcal{L}\left(f\left(x_i';\theta\right),y_{\text{tar}}\right)}{\|\nabla_x\mathcal{L}\left(f\left(x_i';\theta\right)\right),y_{\text{tar}}\|_1},\\ x_{i+1}'&=x_i'-\alpha\times\text{sgn}(g_{i+1}), \end{aligned} \tag{2}$$

其中 g_i 为梯度累积项, g_i 的初始值为 0, μ 为动量系数, $x_0'=x$ 代表干净样本。DI-FGSM 在梯度迭代过程中融入随机输入变换操作, 以增加对抗样本的多样性, 从而提升攻击的泛化能力。其迭代更新过程可表示为式(3)。

$$x_{i+1}'=Clip_{x,\varepsilon}\left\{x_i'-\alpha\times\text{sgn}\left(\nabla_{x_i'}\mathcal{L}\left(T\left(x_i',p\right),y_{\text{tar}}\right)\right)\right\}, \tag{3}$$

其中 T 为输入变换函数, p 为变换的概率。TI-FGSM 通过在迭代过程中对对抗样本引入平移操作, 利用 DNNs 的平移不变性, 转化为对梯度的平移操作, 从而增强样本迁移性, 如式(4)所示。

$$\begin{aligned} &\arg\min_{x'}\sum_{i,j}\omega_{i,j}\mathcal{L}(T(x'),y_{\text{tar}};\theta),\ \text{s.t.}\|x'-x\|_\infty\leqslant\varepsilon,\\ &x'=x-\varepsilon\times\text{sgn}(\boldsymbol{\varpi}\times\nabla_x\mathcal{L}(x,y_{\text{tar}})), \end{aligned} \tag{4}$$

其中 $\omega_{i,j}$ 为权重矩阵, i,j 分别表示图像在水平、垂直方向上的平移距离, $\boldsymbol{\varpi}$ 为预定义的核矩阵。MI-DI-TI 是结合 MI-FGSM, DI-FGSM, TI-FGSM 这 3 种方法优势的综合攻击框架。SU[13]是一种新的自通用性攻击方法, 它让对抗扰动对图像内部的不同局部区域保持通用性, 即“自通用性”, 其优化目标为

$$\begin{aligned} g_{i+1}=&\nabla_\delta[J(f(x+\delta_i),y_{\text{tar}})+J(f(Loc(x,s)+\delta_i),y_{\text{tar}})-\\ &\lambda\times CS(f_l(x+\delta_i),f_l(Loc(x,s)+\delta_i))], \end{aligned} \tag{5}$$

其中 $Loc(x,s)$ 表示随机裁剪并调整大小后的局部图像, CS 是余弦相似性, 用于对齐全局和局部对抗图像的特征表示, λ 是权重参数。

2)CDA 是由 Naseer 等人[14]提出的一个基于共享对抗模式的框架, 利用 GAN 生成对抗样本。具体优化过程如式(6)所示。

$$\begin{aligned} \theta^*\leftarrow&\arg\min_\theta CrossEntropy\left(D_\psi(x'),\boldsymbol{I}_{y_{\text{tar}}}\right)+\\ &CrossEntropy\left(D_\psi(x'),\boldsymbol{I}_y\right), \end{aligned} \tag{6}$$

其中 θ 表示生成器的可学习参数, $CrossEntropy(\cdot)$ 表示损失, D_ψ 表示判别器, $D_\psi(x')$ 则表示判别器对对抗样本 x' 的分类结果。$\boldsymbol{I}_{y_{\text{tar}}}$, $\boldsymbol{I}_y$ 则分别表示类别 y_{tar} 和 y 的独热编码。C-GSP 是由 Yang 等人[6]提出的一种条件生成模型, 解决传统生成式攻击方法中每个目标类别需要单独训练生成器的问题。C-GSP 的目标函数和优化机制如式(7)(8)所示。

$$\theta^*\leftarrow\arg\min_\theta CrossEntropy\left(D_\psi(x_{\text{sou}}+\mathcal{G}_\theta(x_{\text{sou}},\mathcal{W}(\boldsymbol{I}_c))),c\right), \tag{7}$$

$$\delta=\varepsilon\times\tanh(\boldsymbol{f}), \tag{8}$$

其中 $\mathcal{W}$ 是一个映射网络, 其作用是将指定目标类别 c 的独热编码 $\boldsymbol{I}_c$ 映射成一个潜在向量 $\boldsymbol{q}\in\mathbb{R}^M$ 输入生成网络 $\mathcal{G}_\theta$ 用于后续计算, 其中 $c\in C$, 而 C 表示所有可能的

目标类别，$\boldsymbol{f}\in\mathbb{R}^d$表示生成网络输出的特征图，tanh(·)表示双曲正切函数，δ表示生成的扰动。CGCN[7]是一种具有跨注意层的新型CLIP(contrastive language-image pre-training)[16]引导生成网络，充分利用高级CLIP模型提供的文本知识进行增强的多目标攻击。

$$\boldsymbol{q}^{*}\leftarrow\arg\min_{\boldsymbol{q}}\mathcal{L}(\boldsymbol{f}_{\theta}(x_{\text{sou}}+G_{\boldsymbol{q}}(x_{\text{sou}},\Phi(t_c))),c),\tag{9}$$

其中t_c表示目标类c对应的文本提示，$\mathcal{L}$表示交叉熵损失。

1.3 图像的频率域处理

本文主要基于频域变换中的离散傅里叶变换(discrete Fourier transform, DFT)展开研究。DFT的基本思想是将一个函数分解为其组成的频率成分。对于单通道图像，2维DFT的数学表达式为

$$\mathcal{F}(u,v)=\frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}X(h,w)\mathrm{e}^{-\mathrm{j}2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)}=\mathcal{F}_{\text{real}}(u,v)+\mathrm{j}\mathcal{F}_{\text{img}}(u,v),\tag{10}$$

其中$X(h,w)$表示图像在空间坐标(h,w)处的像素值或强度，h和w分别表示像素的水平位置和垂直位置。$u=0,1,\cdots,H-1$，$v=0,1,\cdots,W-1$，j是虚部单位。另外，$\mathcal{F}_{\text{real}}(u,v)$和$\mathcal{F}_{\text{img}}(u,v)$分别是$\mathcal{F}(u,v)$的实部和虚部。

$$\mathcal{F}_{\text{real}}(u,v)=\frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}X(h,w)\cos 2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right),\tag{11}$$

$$\mathcal{F}_{\text{img}}(u,v)=\frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}X(h,w)\sin 2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)。\tag{12}$$

振幅谱$\mathcal{A}(X)$是频域表示的关键部分，揭示了图像不同频率成分对整体特性的影响，计算公式为

$$\mathcal{A}(X)(u,v)=|\mathcal{F}(u,v)|=\sqrt{\mathcal{F}_{\text{real}}(u,v)^2+\mathcal{F}_{\text{img}}(u,v)^2}。\tag{13}$$

相位谱$\mathcal{P}(X)$是频域的另一重要组成部分，与振幅谱共同完整描述了图像的频域特征，计算公式为

$$\mathcal{P}(X)(u,v)=\arctan\left(\frac{\mathcal{F}_{\text{img}}(u,v)}{\mathcal{F}_{\text{real}}(u,v)}\right)。\tag{14}$$

2 双重信息对齐策略

主流有目标生成式攻击方法多借助预训练分类器获取目标类别信息，通过构建基于目标标签[6,14]或图像分布[17]的损失函数，驱使生成器从分类器提取语义以攻击特定类别。但此类损失函数核心依赖替代模型最后一层输出，使生成器主要捕获目标类别高级语义[18]。由于高级语义在不同模型间差异显著[17]，单纯依此训练生成器，易致使生成的对抗样本迁移性不佳。Zeiler等人[19]的研究表明，高层特征编码高级语义，判别性强却易过拟合特定架构，低层特征捕获颜色、边缘等低级视觉信息，迁移性强但对输出影响小[20]，中间层特征兼具两者优势。基于此，本文提出双重信息对齐策略，其融合低、高级信息的优势引导生成器生成高迁移性对抗样本，提升其跨模型迁移能力。双重信息对齐策略的基本框架如图1所示。

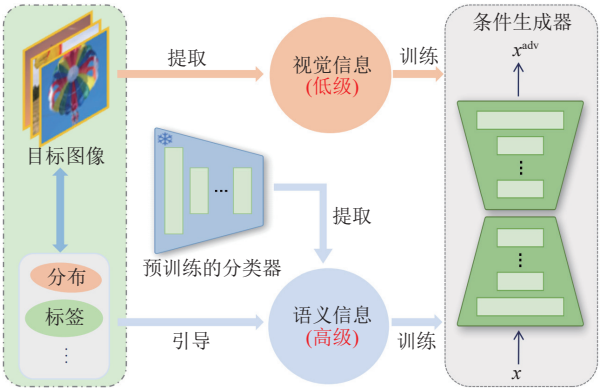


Fig. 1 Basic framework of the dual-information alignment strategy

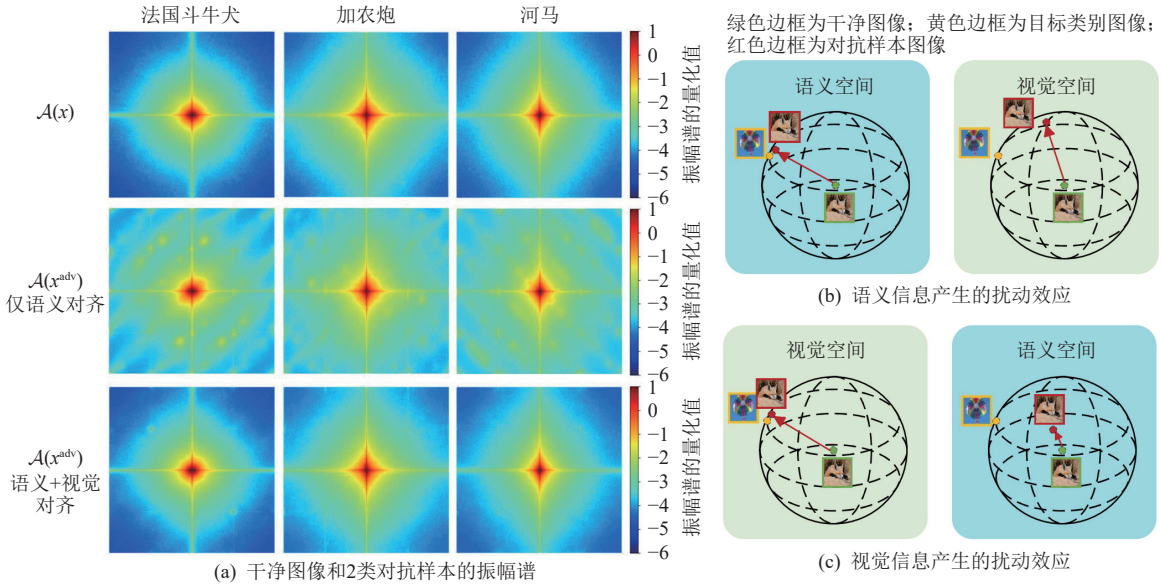
图1 双重信息对齐策略的基本框架

2.1 语义信息对齐

语义信息不仅反映了图像的高层次语义特征，还包含类别的统计分布和类内特征结构等，体现了图像的本质特征。在有目标生成式攻击中，语义信息对齐是引导生成器生成符合目标类别特征的对抗样本[18-21]。早期研究[14,21]主要通过学习替代模型的决策边界，使对抗样本符合目标类别的判别特征。然而，不同模型对类别边界的表达存在显著差异，基于单一模型生成的对抗样本通常难以在其他模型上保持相同的攻击效果，从而限制了对抗样本的跨模型迁移性。为解决这一问题，现有研究[17]引入分布匹配策略，以缩小对抗样本与目标类别真实样本的特征分布差距。具体而言，分布匹配方法不再仅依赖类别边界信息，而是将对抗样本的深度特征对齐至目标类别的整体统计分布，使其在多个模型上的迁移能力得以提升。然而，本文进一步研究发现，单纯依赖语义信息对齐存在明显局限性，特别是在频域特征上容易引发“振幅谱偏离”现象。具体而言，我们观察到，仅基于语义信息生成的对抗样本，其振幅谱与原始图像或目标类别图像的振幅谱之间存在显著差异。这种偏离可能破坏对抗样本在跨域迁移场景中的有效性。因此，本文在语义信息对齐的基础上，引入低级视觉信息作为补充，以形成更全面的双重信息对齐策略，从而增强生成器在多模型、多场景

下的适用性，如图 1 所示。

1)振幅谱偏离。为系统分析振幅谱偏离现象，本文对原始图像及其对应的对抗样本进行了频谱分析，图 2(a)直观展示了其振幅谱的可视化结果。实验表明，仅基于语义信息对齐生成的对抗样本在振幅谱高频区域(图像边缘部分)中表现出明显的扰动和不规则模式。这种杂乱的分布与原始图像中相对平滑的高频特征形成了鲜明对比。虽然中心低频区域(图像中心部分)在结构上相对稳定，但依然可以观察到细微的变化。总体而言，仅语义信息对齐生成的对抗样本的振幅谱呈现出更复杂且不规则的模式，缺乏原始图像振幅谱所表现出的结构性和规律性。



(a) 干净图像和2类对抗样本的振幅谱

(b) 语义信息产生的扰动效应

(c) 视觉信息产生的扰动效应

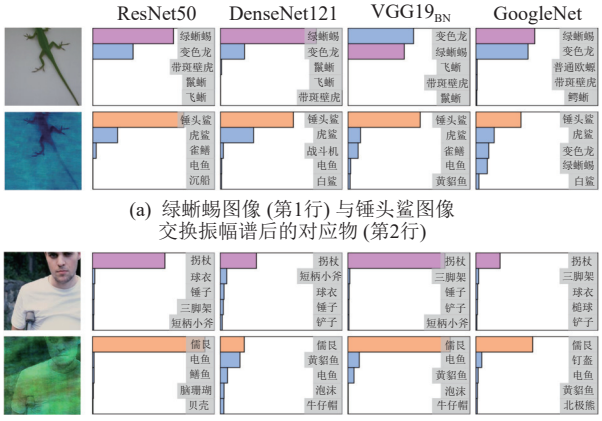
Fig. 2 Analysis of the characteristics of clean images and adversarial samples in the frequency domain and information space

图 2 干净图像与对抗样本在频域与信息空间中的特性分析

这种振幅谱偏离现象在不同类别的图像中均有体现，包括法国斗牛犬、加农炮和河马等，说明这一问题并非局限于特定类别，而是具有普遍性。更为重要的是，即使在将对抗样本攻击为其原始类别的实验中，振幅谱偏离现象依然存在，进一步证明了这一问题的普遍性和严重性。

2)卷积神经网络(CNN)对振幅谱的敏感性。已有研究[22]在 CIFAR-10[23]和 SVHN[24]数据集上发现，CNN 对图像振幅谱的变化极为敏感。为验证和扩展这一发现，本文在频率信息更丰富的 ImageNet[25]数据集上进行了更深入的实验。实验结果进一步证实了这一现象，同时揭示了这种偏离在不同网络架构中的普遍性和一致性。如图 3 所示，在 4 种不同的网络模型中，图像的振幅谱被交换后，模型将其错误分类为与振幅谱匹配的标签，而非原始图像类别。这些结果表明，CNN 对振幅信息的高度敏感性。

基于对抗样本生成过程中出现的振幅谱偏离现象可以推断，这种偏离可能显著影响 CNN 对有目标对抗样本的预测，从而降低攻击的成功率。特别是在目标类别的高频特征未能有效匹配时，这种现象尤为明显。



(a) 绿蜥蜴图像 (第1行) 与锤头鲨图像交换振幅谱后的对应物 (第2行)

(b) 拐杖图像 (第1行) 与儒艮图像交换振幅谱后的对应物 (第2行)

Fig. 3 Illustration of the sensitivity of different CNN architectures to the image amplitude spectrum

图 3 不同 CNN 网络架构对图像振幅谱敏感性的示意图

3)振幅谱偏离的影响。振幅谱偏离现象不仅导致对抗样本在视觉空间中偏离目标类别的真实分布，还可能偏离自然图像的一般特征。图 2(b)展示了这一现象的具体表现：生成的对抗样本的振幅谱不仅未能有效体现目标类别的频率特征，反而出现了与目标类别不相关的频率特性。这种过度偏离可能干

扰 CNN 的预测，使其难以被正确分类为目标类别。更值得注意的是，这种振幅谱偏离为对抗样本的检测提供了潜在的突破口。由于对抗样本的频谱显著偏离了自然图像的频谱分布，防御者可以利用这一差异设计检测机制，从而提高对抗样本的识别能力。

2.2 视觉信息对齐

在探讨了语义信息对齐及其局限性之后，本文提出了一种新的对齐策略——视觉信息对齐，以解决语义信息对齐方法可能导致的振幅谱偏离问题，生成更自然、更具迁移性的有目标对抗样本。

振幅谱是图像的关键频域特征，它编码了诸如颜色、边缘和纹理等低级视觉信息，并在很大程度上决定了图像的整体视觉特性。2.1 节中描述的振幅谱偏离现象表明，基于纯语义对齐生成的对抗样本在这些基础视觉元素上与目标类别图像存在显著差异。因此，本文在语义信息对齐的基础上，通过对抗样本与目标类别图像的振幅谱对齐，进一步优化对抗样本的视觉特性。

通过傅里叶变换，可以将图像从空间域转换到频率域，得到相应的振幅谱和相位谱。振幅谱编码了图像的低级视觉信息，而相位谱则携带了图像的高级语义信息[26-27]。然而，从图 4 中可以观察到，相位谱表现出类似随机噪声的复杂性和无序性，这使得从中提取有用信息极具挑战性。相比之下，振幅谱展现出更为结构化和规律化的模式，为有效捕捉视觉信息提供了可行路径。与模型的低层特征相比，振幅谱具有两大优势：1）与模型无关性。图像的振幅谱与替代模型无关，因此可以显著降低生成器过拟合于特定模型的风险。2）全局性质。振幅谱携带了图像的全局频率信息（如式（8）所示），能够更全面地捕捉图像的低层视觉特性。基于这些优势，本文选择利用振幅谱实现低级视觉信息的对齐，从而弥补语义信息对齐策略的不足。

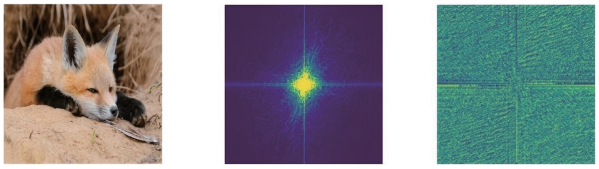


Fig. 4 Amplitude and phase spectra of an image

图 4 图像的振幅谱和相位谱

为了引导生成器从目标类别的振幅谱中学习低级视觉信息，本文提出了振幅谱对齐（amplitude spectrum alignment, ASA）损失。该损失的核心思想是最小化生成器输出与目标类别图像在振幅谱上的差异，从而鼓励生成器$\mathcal{G}_\theta$捕捉并重现目标类别的视觉特性。ASA 损失的定义为：

$$\mathcal{L}_{\mathrm{ASA}}=\frac{1}{N}\sum_{i=1}^{N}\|\mathcal{A}(\mathcal{G}_\theta(x^{(i)}))-\mathcal{A}(x_{\mathrm{tar}}^{(i)})\|_1, \tag{15}$$

其中$\|\cdot\|_1$表示l_1范数。已有研究表明，l_1范数是度量频域距离的有效标准[28-29]。通过最小化 ASA 损失，生成器$\mathcal{G}_\theta$可以学习生成具有目标类别振幅特性的对抗样本。经过训练的生成器$\mathcal{G}_\theta$不仅能够捕捉目标类别的视觉信息，还能合成在视觉空间上更接近目标类别的对抗样本。

为进一步提升生成器生成对抗样本的质量，本文还引入了增广损失，通过结合 ASA 损失与其他损失项（如语义对齐损失），在语义与视觉信息对齐之间实现平衡。增广损失的具体形式为

$$\mathcal{L}_{\mathrm{ASA}}^{\mathrm{aug}}=\frac{1}{N}\sum_{i=1}^{N}\|\mathcal{A}(\mathcal{G}_\theta(\tilde{x}^{(i)}))-\mathcal{A}(x_{\mathrm{tar}}^{(i)})\|_1。 \tag{16}$$

该过程的示意图如图 5 所示，进一步直观地展示了振幅谱对齐和增广 ASA 损失的具体操作流程。

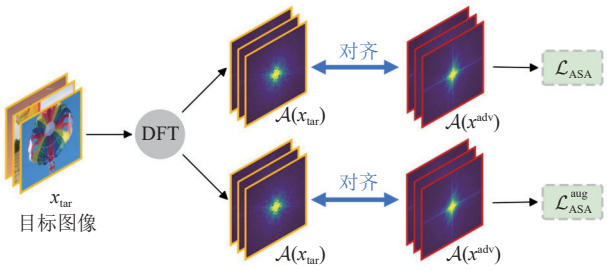


Fig. 5 Illustration of amplitude spectrum alignment

图 5 振幅谱对齐示意图

3 基于双重信息对齐的条件生成模型

本节介绍了一种基于双重信息对齐的条件生成模型，通过映射网络将目标类别的语义与视觉信息编码至统一潜在空间，并设计复合损失函数（语义损失与视觉损失）实现多目标对抗样本的联合优化。该模型在保留目标类别高层语义特征的同时，通过频域视觉信息约束减少扰动偏离，显著提升了对抗样本的跨模型迁移性和物理合理性，突破了传统单信息对齐方法的局限性。工作原理如图 6 所示。

3.1 映射网络

映射网络的主要目标是将目标类别的标签信息转换为包含丰富类别特征的潜在向量。其结构包括输入层、多层感知机（MLP）和归一化层，具体如下：

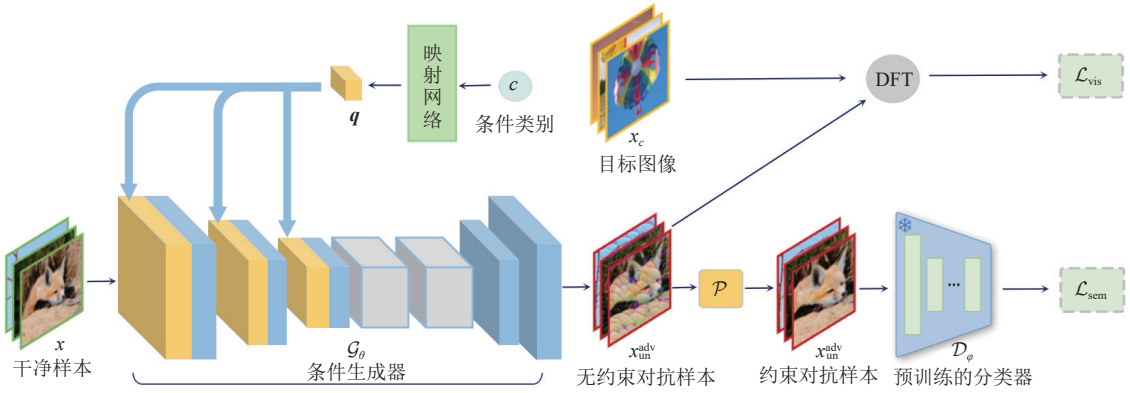


Fig. 6 Illustration of the multi-target generative attack algorithm based on dual information

图 6 基于双重信息的多目标生成式攻击算法的示意图

1)输入层。接收目标类别c的独热编码表示$\boldsymbol{I}_c \in \mathbb{R}^K$, K是类别总数。

2)多层感知机。由多个全连接层组成,每层后接非线性激活函数。其作用是将输入的类别编码转换为高维特征表示,从而捕捉目标类别的复杂语义信息。数学表示为

$$h = MLP(\boldsymbol{I}_c) = f_n(\cdots f_2(f_1(\boldsymbol{I}_c))), \tag{17}$$

其中f_i表示对第i层的操作,包括线性变换和非线性激活。激活函数选用的 ReLU,通过引入非线性,能够有效提高网络的表达能力。

3)归一化层。为了稳定训练并提高网络的泛化能力,在 MLP 输出上应用归一化操作。本文采用层归一化,其定义为

$$\boldsymbol{q} = LayerNorm(h) = \gamma \odot \frac{h-\mu}{\sqrt{\sigma^2+\varepsilon}} + \beta, \tag{18}$$

其中μ和σ分别是h的均值和标准差,γ和β是可学习的缩放和偏移参数,ε是一个小常数以确保数值稳定性。

最终,该映射网络的输出为一个潜在向量$\boldsymbol{q} \in \mathbb{R}^M$, M为预定义的维度。映射过程可概为

$$\boldsymbol{q} = \mathcal{W}(\boldsymbol{I}_c) = LayerNorm(MLP(\boldsymbol{I}_c)) \text{ 。} \tag{19}$$

3.2 双重信息编码机制

为了在多目标生成式攻击中同时捕获目标类别的语义和视觉信息,本文设计了一种双重信息编码机制。该机制通过语义损失和视觉损失的联合优化,确保生成的潜在向量$\boldsymbol{q}$同时包含高级语义信息和低级视觉信息。

1)语义损失函数。语义损失旨在保证潜在向量$\boldsymbol{q}$准确表示目标类别的语义信息。本文采用交叉熵损失来实现这一目标,数学表示为

$$\mathcal{L}_{sem} = \mathbb{E}_{x\sim X, c\sim C}\left[\mathcal{L}_{CE}\left(D_\psi(\mathcal{G}_\theta(x, \mathcal{W}(\boldsymbol{I}_c))), c\right)\right], \tag{20}$$

其中$\mathcal{L}_{CE}$代表交叉熵损失, $\mathcal{G}_\theta$表示条件生成模型, D_ψ表示预训练的分类器, c代表目标类别。通过最小化语义损失,映射网络被鼓励学习能够准确表示目标类别语义特征的潜在向量。

2)视觉损失函数。为了捕捉目标类别的低级视觉特征,本文基于 ASA 的思想设计了视觉损失函数,其形式为

$$\mathcal{L}_{vis} = \mathbb{E}_{x\sim X, c\sim C}\left[\left\|\mathcal{A}\left(\mathcal{D}_\psi(\mathcal{G}_\theta(x, \mathcal{W}(\boldsymbol{I}_c)))\right) - \mathcal{A}(x_c)\right\|_1\right] \text{ 。} \tag{21}$$

通过最小化视觉损失,生成器被引导生成在视觉空间上接近目标类别的对抗样本。

3)复合损失函数。为了同时学习语义和视觉信息,本文将语义损失与视觉损失结合,形成复合损失函数,数学表示为

$$\mathcal{L}_{total} = \mathcal{L}_{sem} + \alpha \times \mathcal{L}_{vis}, \tag{22}$$

其中α为超参数,取值为5×10^{-4},用于平衡损失的影响。通过最小化复合损失,映射网络被优化为能够学习包含语义和视觉信息的潜在向量,从而提升生成器生成对抗样本的迁移能力和攻击成功率。

基于双重信息的条件生成模型通过引入映射网络和双重信息编码机制,实现了语义信息和视觉信息的联合学习。复合损失函数的设计进一步确保了生成器能够生成既具有目标类别高级语义特征又具有低级视觉特性的对抗样本。

3.3 算法实现

基于上述分析,本文提出了一种基于双重信息的多目标生成式攻击方法 MTGA-DI。其训练过程主要包括 5 个步骤:1)特征向量计算。对于给定输入图像x,条件生成模型$\mathcal{G}_\theta$中的编码器部分首先计算特征向量$\boldsymbol{F} \in \mathbb{R}^{N\times H\times W}$,其中 N, H, W 分别为特征映射的通道数、高度和宽度。2)潜在向量生成。针对目标类别c,映射网络$\mathcal{W}$将其转换为潜在向量$\boldsymbol{q} \in \mathbb{R}^{N\times H\times W}$,

使其不再是简单的类别标签，而是蕴含高级语义信息的潜在表示。3）特征融合。将上述 2 个向量沿通道连接，得到 $\boldsymbol{F}' \in \mathbb{R}^{(N+M)\times H\times W}$ 混合特征，此过程不仅引入目标类别信息，还动态调制图像特征，使生成器学习适应性更强的特征映射，而非固定扰动。4）对抗样本生成。将混合特征映射馈送到生成器 $\mathcal{G}_\theta$，生成器将输入图像 x 和目标标签 c 转换为输出图像 $\mathcal{G}_\theta(x,c)$，使 $\mathcal{G}_\theta$ 能在同一生成器上针对不同目标类别生成高效的对抗样本 x_i^{adv}，从而具备真正的多目标攻击能力。5）对抗样本约束。为了确保生成的对抗扰动在约束下既符合优化目标又能避免高频噪声的影响，我们引入平滑投影机制，具体投影操作定义为

$$x' = Clip\left(\min(x_i^{\text{adv}}+\varepsilon, \max(\mathcal{W}\times\mathcal{G}_\theta(x_i^{\text{adv}},c), x_i^{\text{adv}}-\varepsilon))\right), \tag{23}$$

其中 $\mathcal{W}$ 是一个固定权重的平滑算子，用于抑制高频信息，并确保扰动的平稳性。$Clip$ 操作用于确保扰动不会超出给定的约束范围。整个生成过程如图 6 所示。

在整个生成过程中，映射网络起到了关键作用。其通过与预训练分类器 D_ψ 的深度交互，捕获目标类别的丰富语义信息。在训练过程中，随机采样不同的目标类别 $c\in C$，映射网络逐渐学习到一个广泛而有效的潜在表示空间。这个潜在空间为后续对抗样本的生成提供了丰富的条件信息，使生成器能够生成更具迁移性和鲁棒性的多目标对抗样本。算法实现的具体流程见算法 1。

算法 1. 基于双重信息的多目标生成式攻击。

输入：训练数据 $\mathcal{X}$，预训练好的分类器 D_ψ，训练轮次 T，扰动预算 ε；

输出：训练好的多目标生成器 $\mathcal{G}_\theta$。

① 随机初始化生成器 $\mathcal{G}_\theta$；

② for $t=0\rightarrow T-1$ do

③ 从 $\mathcal{X}$ 中随机采样小批量数据 $\{x_i\}_{i=1}^B$ 和随机选择 B 个目标类别 $\{c_i\}_{i=1}^B$，并分别采样；

④ 每个类别对应的样本 $\{x_{c_i}\}_{i=1}^B$；

⑤ 将 c_i 通过映射网络 $\mathcal{W}$ 前向传播来计算各个目标类别的潜在向量 $\boldsymbol{q}_i$；

⑥ 将 x_i 通过生成器 $\mathcal{G}_\theta$（包含了潜在向量 $\boldsymbol{q}_i$）前向传播以生成无界对抗样本 x_i^{adv}；

⑦ 使用无界对抗样本 x_i^{adv} 和目标类别中的样本 x_{c_i} 来计算视觉损失 $\mathcal{L}_{\text{vis}}$（式（21））；

⑧ 使用平滑投影（式（22））约束 x_i^{adv} 以满足：$\|x_i^{\text{adv}}-x_i\|_\infty \leqslant \varepsilon$；

⑨ 使用约束后的 x_i^{adv} 计算语义损失 $\mathcal{L}_{\text{sem}}$（式（20））；

⑩ 计算总损失 $\mathcal{L}_{\text{total}}$ 并反向传播更新 $\mathcal{G}_\theta$；

⑪ end for

⑫ 返回收敛后的生成器 $\mathcal{G}_\theta$。

4 实　　验

4.1 实验设置

1）数据集。为测试本文提出的多目标生成式攻击方法（MTGA-DI）的有效性，实验主要在 ImageNet 完整训练集（约 1 300 000 张图像）[25] 上进行。为评估跨域迁移性能，补充使用 Comics 数据集（约 40 000 张图像），该数据集具备与 ImageNet 显著不同的视觉特征与风格。所有图像均统一调整为 224×224 像素，以标准化输入，保持关键特征并降低计算复杂度。

2）模型。本文采用 ResNet 架构作为条件生成器的基础网络，选择预训练的 ResNet-50[30] 和 Densenet-121[31] 作为分类器，为生成器提供辅助监督。在实验中，共评估了 9 个标准训练模型和 6 个鲁棒模型作为受害者模型。其中标准训练模型包括 VGG19_{BN}[32]，DenseNet-121（Dense-121）[31]，ResNet-50（Res-50）[30]，ResNet-152（Res-152）[30]，WideResNet-50-2（WRN-50-2）[33]，WideResNet-101-2（WRN-101-2）[33]，MobileNetV2[34]，GoogleNet[35]，ViT（vision Transformer）[36]。鲁棒的受害模型包括对抗训练过的模型 adv-Inception-v3（Adv-Inc-v3）[15]，ens-adv-Inception-ResNet-v2（Ens-Adv-IncRes-v2）[37] 以及几种鲁棒的 ResNet-50 变体[38-39]。

3）对比方法。为验证 MTGA-DI 的有效性，选择了一系列经典的对抗攻击方法作为基准，分为 2 类：实例相关攻击和实例无关攻击。实例相关攻击中我们选择了 MI-FGSM[10]，DI-FGSM[11]，TI-FGSM[12] 的组合版本 MI-DI-TI 和 SU[13]，实例无关攻击包括单目标攻击方法 CDA[14] 与多目标攻击方法 C-GSP[6] 和 CGNC[7]。

4）应用细节。在生成器的训练阶段，预训练分类器参数保持固定，仅更新生成器参数。本文使用 Adam 优化器[40] 训练模型，学习率为 0.000 2，参数设为 β_1=0.5，β_2=0.999，训练轮次 T=20。所有扰动预算 ε 统一设为 16（像素值范围为 0~255）。对于实例相关攻击 MI-DI-TI 和 SU[13]，由于其无需训练生成器，其实验结果与训练集无关，因此直接使用标准设置进行实验；对于单目标生成式攻击 CDA[14]，本文将其训练轮数增加至 20，并采用 ImageNet_50k 子集进行训练。由于该方法在子集上训练即可获得收敛结果，进一步扩展至完整训练集的性能提升有限，因此直接使用子集结果进行对比。对于 C-GSP[6]，CGNC[7] 和本文方法，它们需要使用 ImageNet 的整个训练集。为高

效利用计算资源, 三者的训练轮数统一降低至 10。此外, 由于 ImageNet 数据集足够庞大, 实验未引入额外的数据增强技术。所有实验均在相同硬件配置下完成, 使用 NVIDIA RTX A6000 GPU 进行模型训练和推理。本文基于 PyTorch 框架实现所有模型与算法, 确保计算的高效性和良好的扩展性。

5)评估指标。本文通过攻击成功率评估攻击方法的性能。攻击成功率定义为被误分类为指定目标类别的对抗样本数量占总对抗样本数量的百分比。实验报告了 8 个选定目标类别的平均攻击成功率, 包括: 鹅、法国斗牛犬、河马、加农炮、消防车、T 型车、降落伞和雪地摩托车。

4.2 攻击标准训练模型

为评估 MTGA-DI 的攻击性能和迁移能力, 本文对比了实例相关攻击中表现最强的 MI-DI-TI 和 SU、生成式攻击方法中的单目标攻击方法 CDA[14]、多目标攻击方法 C-GSP[6] 和 CGNC[7], 以及本文提出的 MTGA-DI。值得注意的是, CDA 需要训练 8 个生成器才能实现多目标攻击, 而 C-GSP, CGNC, MTGA-DI 仅需训练 1 个生成器即可完成任务, 大大降低了训练成本。实验采用 Dense-121 和 Res-50 作为替代模型, 评估了 6 种攻击方法在黑盒和白盒场景下的攻击成功率。具体实验结果如表 1 和表 2 所示。

Table 1 Attack Results on the Standard Training Model Using Dense-121 as the Alternative Model

表 1 使用 Dense-121 为替代模型在标准训练模型上的攻击结果 %

标准训练模型	MI-DI-TI	CDA	SU	C-GSP	CGNC	MTGA-DI
$VGG19_{BN}$	9.17	43.33	49.20	53.92	**61.60**	57.45
Dense-121	**99.84***	96.74*	96.25*	95.74*	94.12*	93.91*
Res-50	10.07	60.90	44.30	62.55	61.30	**70.96**
Res-152	6.16	45.39	33.20	45.78	49.96	**50.56**
WRN-50-2	7.58	47.82	49.60	54.41	51.61	**56.43**
WRN-101-2	4.32	39.92	37.72	39.26	43.95	**45.37**
MobileNetV2	4.90	37.41	9.91	**45.55**	37.30	40.66
GoogleNet	4.36	32.85	9.30	32.14	**33.87**	30.22
ViT	0.35	11.73	7.16	11.93	7.75	**12.93**
平均攻击成功率	5.86	39.92	39.42	49.03	49.05	**50.94**

注: "*"表示白盒攻击, 最优结果和次优结果分别用黑体数值和下划线数值表示。

通过观察和分析表 1 和表 2, 可知:

1)在白盒场景下, MTGA-DI 的攻击成功率与 2 种多目标攻击 C-GSP 和 CGNC 相当, 显示出良好的攻击性能。尽管 C-GSP 和 CGNC 在部分场景中略占优势, 但 MTGA-DI 在攻击效率和迁移性上的综合表现更为突出。

Table 2 Attack Results on the Standard Training Model Using Res-50 as the Alternative Model

表 2 使用Res-50为替代模型在标准训练模型上的攻击结果 %

标准训练的模型	MI-DI-TI	CDA	SU	C-GSP	CGNC	MTGA-DI
$VGG19_{BN}$	12.43	67.15	54.74	56.69	**67.36**	59.05
Dense-121	17.84	75.38	48.79	74.88	55.43	**76.59**
Res-50	**97.75***	96.97*	93.17*	94.46*	93.56*	93.39*
Res-152	14.00	**70.05**	50.2	65.44	68.52	63.53
WRN-50-2	15.48	71.03	67.54	69.68	71.13	**72.20**
WRN-101-2	9.84	62.17	59.60	57.63	62.42	**63.27**
MobileNetV2	7.34	52.58	12.78	**55.12**	42.63	46.60
GoogleNet	6.02	**46.25**	30.45	37.77	33.58	34.00
ViT	0.39	7.37	8.96	12.60	10.96	**17.00**
平均攻击成功率	20.12	**60.99**	47.36	58.25	56.18	58.40

注: "*"表示白盒攻击, 最优结果和次优结果分别用黑体数值和下划线数值表示。

2)在 16 种不同的黑盒场景中, MTGA-DI 在 9 种情况下表现出最佳的迁移性能。此外, 在 2 种不同的替代模型下, 其平均攻击成功率也达到较高水平, 整体表现优越, 充分证明了其在不同受害模型间的高迁移性。

3)以 Dense-121 为替代模型时, MTGA-DI 对 8 个标准训练模型的黑盒平均攻击成功率达到 45.57%, 优于 CGNC(43.42%)和 C-GSP(43.19%), 并超越 CDA 和 MI-DI-TI 方法。

4.3 攻击防御模型

虽然许多攻击方法能够轻松欺骗标准训练的模型, 但在攻击防御模型, 尤其是黑盒防御模型时, 其攻击成功率往往较低。为了验证 MTGA-DI 生成的对抗样本在鲁棒模型上的有效性和鲁棒性, 实验评估了 6 种攻击方法 MI-DI-TI, SU, CDA, C-GSP, CGNC, MTGA-DI 对多种鲁棒模型的攻击性能。实验分别以 Dense-121 和 Res-50 作为替代模型, 评估对 Adv-Inc-v3, Ens-Adv-IncRes-v2 以及 4 种鲁棒 ResNet-50 变体等防御模型的攻击成功率。相应的实验结果如表 3 和表 4 所示。

通过观察与分析表 3 和表 4, 可知:

1)在有目标攻击场景下, 攻击防御模型的难度显著增加。特别是对于只使用 1 个生成器来实现多目标攻击的生成式方法而言, 这需要生成器同时学习多个目标类别的判别信息。尽管如此, MTGA-DI

Table 3 Attack Results on the Robust Model Using Dense-121 as the Alternative Model

表 3 使用Dense-121为替代模型在鲁棒模型上的攻击结果 %

鲁棒模型	MI-DI-TI	CDA	SU	C-GSP	CGNC	MTGA-DI（本文）
Adv-Inc-v3	0.55	0.40	0.20	**7.83**	5.62	5.79
Ens-Adv-IncRes-v2	0.51	0.57	0.21	5.93	**6.23**	6.01
Res-50_SIN	0.94	8.90	5.62	**10.72**	8.85	8.07
Res-50_SIN_IN	4.22	45.33	26.32	49.57	49.35	**50.55**
Res-50_SIN_fine_IN	7.16	50.00	30.25	51.98	52.18	**58.85**
Res-50_AugMix	1.60	40.36	22.61	38.32	40.33	**50.57**
平均攻击成功率	2.50	24.26	14.20	27.39	27.09	**29.97**

注：最优结果和次优结果分别用黑体数值和下划线数值表示。

Table 4 Attack Results on the Robust Model Using Res-50 as the Alternative Model

表 4 使用Res-50为替代模型在鲁棒模型上的攻击结果 %

鲁棒模型	MI-DI-TI	CDA	SU	C-GSP	CGNC	MTGA-DI（本文）
Adv-Inc-v3	0.72	0.99	0.23	4.78	3.77	**4.99**
Ens-Adv-IncRes-v2	0.66	0.96	0.08	4.79	**5.93**	5.17
Res-50_SIN	2.07	15.22	8.34	**27.47**	21.14	23.59
Res-50_SIN_IN	19.17	80.47	54.22	77.06	80.80	**81.68**
Res-50_SIN_fine_IN	37.59	**85.20**	59.65	78.76	79.81	80.63
Res-50_AugMix	3.53	60.34	37.51	59.57	61.29	**61.62**
平均攻击成功率	10.62	40.53	26.66	42.07	42.12	**42.95**

注：最优结果和次优结果分别用黑体数值和下划线数值表示。

方法通过结合语义和视觉信息，在应对多个类别的攻击需求时仍然展现出了较强的泛化能力。

2）在攻击黑盒防御模型的实验中，MTGA-DI 方法在 12 个案例中的 6 个取得了最佳性能。此外，在 2 种不同的替代模型下，其平均攻击成功率均达到最佳，展现了优越的有目标迁移攻击能力。以 Dense-121 作为替代模型时，MTGA-DI 的平均攻击成功率达到了 29.97%，比 C-GSP 和 CGNC 分别高出 2.58 个百分点和 2.88 个百分点，进一步验证了将目标类别的视觉信息纳入生成模型对提升对抗样本鲁棒性的有效性。

4.4 攻击基于输入预处理的防御

除了对抗训练和数据增强策略外，基于输入预处理的防御方法是另一种有效的对抗样本防御手段。这类方法通过在图像输入模型前利用图像处理技术去除对抗性扰动来降低对抗样本的威胁。本文主要评估了 2 种经典的输入预处理防御方法：JPEG 防御[41]和中值平滑防御[42]。JPEG 防御[41]基于图像压缩原理，通过 JPEG 压缩算法的特性来减弱对抗扰动的影响。本文实验中，JPEG 压缩的质量因子设置为 75。中值平滑防御[42]是一种基于非线性中值滤波的防御策略，通过减弱对抗扰动的影响，同时尽可能保持图像的主要结构信息。实验中，滤波内核大小设置为 3。

本文评估了 4 种生成式攻击方法 CDA, C-GSP, CGNC, MTGA-DI 在面对上述 2 种输入预处理防御方法时的攻击效果。此外，实验还评估了不同扰动预算ε下各生成式攻击方法的迁移性，结果如图 7 所示。

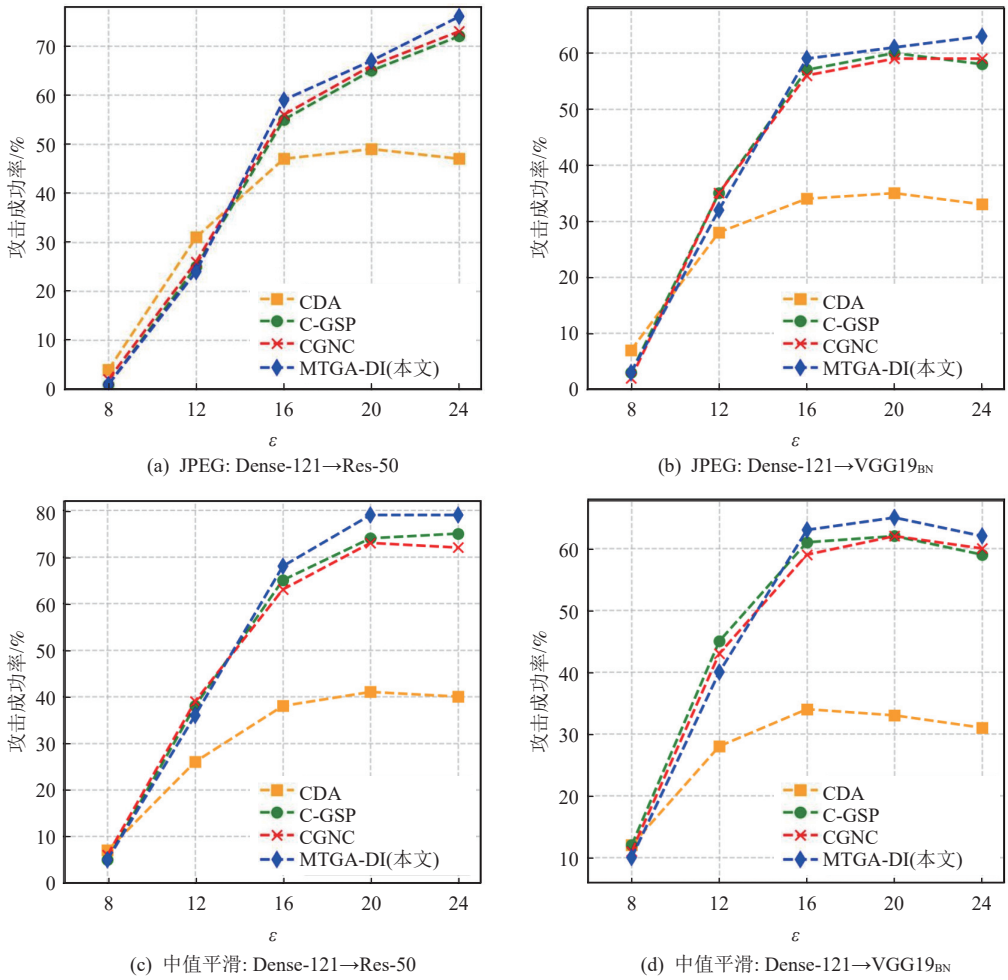
从图 7 中可以观察到，在较低扰动预算下（$\varepsilon \leqslant 12$），无论采用 JPEG 还是中值平滑防御方法，MTGA-DI 的攻击成功率略低于 C-GSP 和 CGNC。这种现象可能是由于 C-GSP 和 CGNC 在小扰动范围内生成的对抗样本更加精细，迁移性稍优，但其攻击成功率的提升幅度较为有限。随着扰动预算的增加，MTGA-DI 的攻击成功率逐渐超越 C-GSP 和 CGNC。例如，当扰动预算为 24，利用 Dense-121 作为替代模型生成对抗样本并攻击目标模型时，MTGA-DI 的攻击成功率为 62.75%，而 CDA, C-GSP, CGNC 的攻击成功率分别为 32.34%, 57.98%, 59.35%，且进一步增加扰动预算并未对攻击成功率带来显著提升。这种优势在目标模型为 Res-50 时也是如此。这表明，通过结合语义和视觉信息，MTGA-DI 能够生成在更广泛扰动范围内具有鲁棒性的对抗样本。

实验结果表明，MTGA-DI 方法在面对基于输入预处理的防御时，具有较高的攻击成功率和鲁棒性。通过将目标类别的视觉信息引入潜在空间的编码过程，MTGA-DI 能够在高扰动预算条件下生成更具迁移性和破坏性的对抗样本。这种能力使其在复杂对抗场景中具有更强的实际应用价值。

4.5 计算效率分析

在对比方法中，实例相关攻击 MI-DI-TI 和 SU 需要执行M步计算梯度来获得对抗样本，因此计算分类器C的前向传播t_C^{FP}和反向传播t_C^{BP}的时间成本可定义为$T^{\mathrm{IS}} = t_C^{\mathrm{FP}} \times M + t_C^{\mathrm{BP}} \times M$。而实例无关攻击仅需计算生成器的推理时间：$T^{\mathrm{IA}} = t_{\mathcal{G}_\theta}^{\mathrm{FP}}$。

实例无关攻击中，我们以单目标生成方法 CDA 作为基线，它需要 8 个模型才能覆盖 8 个类别。相比之下，多目标式生成方法 C-GSP, CGNC 和 MTGA-DI 都仅需训练 1 个模型即可完成推理，有效提升了计算效率，减少了存储开销和训练成本。结合攻击成功率分析，MTGA-DI 仍然优于其他方法。具体各方法的计算效率结果如表 5 所示。



(a) JPEG: Dense-121→Res-50 (b) JPEG: Dense-121→VGG19$_{BN}$

(c) 中值平滑: Dense-121→Res-50 (d) 中值平滑: Dense-121→VGG19$_{BN}$

Fig. 7 Attack success rates of four generative attacks against JPEG and median smoothing under different perturbation budgets

图 7 不同扰动预算下 4 种生成式攻击对 JPEG 和中值平滑的攻击成功率

Table 5 Comparison of Computational Efficiency of Different Attack Methods on Dense-121 and Res-50 Models

表 5 不同攻击方法在 Dense-121 与 Res-50 模型上的计算效率对比

模型	攻击方法	时间/ms	生成器训练数量
Dense-121	MI-DI-TI	约 350	-
	CDA	约 25	8
	C-GSP	约 12	1
	SU	约 220	-
	CGNC	约 13	1
	MTGA-DI	约 12	1
Res-50	MI-DI-TI	约 450	-
	CDA	约 30	8
	C-GSP	约 16	1
	SU	约 300	-
	CGNC	约 16	1
	MTGA-DI	约 15	1

注: "-"表示无需训练生成器。

4.6 灵活性分析

为了评估各方法的灵活性, 我们选取 Dense-121 为替代模型, 迁移到 Res-50 和 WRN-50-2, 测试了 4 种方法(CDA, C-GSP, CGNC, MTGA-DI)在 4 种目标类别(鹅、法国斗牛犬、河马、加农炮)上的攻击成功率。其中 CDA 方法中我们选取了"鹅"类别训练生成器; 而 C-GSP, CGNC, MTGA-DI 方法均无需针对类别单独训练。最终的测试结果如表 6 所示。

实验结果表明 CDA 方法在 Res-50 和 WRN-50-2 上迁移后, 在"鹅"类别上仍保持一定的攻击成功率(Res-50 为 71.32%, WRN-50-2 为 57.78%), 但在其他类别(法国斗牛犬、河马、加农炮)上的攻击成功率极低, 特别是在"加农炮"类别上几乎无效, 说明 CDA 在目标类别变化时无法适应新类别, 必须重新训练, 灵活性最差; C-GSP 和 CGNC 方法在多个目标类别上均能保持较高的攻击成功率, 但仍存在一定波动。例如, CGNC 在"鹅"类别上的表现为: Res-50 为 70.01%; WRN-50-2 为 56.12%。而在"加农炮"类别上的攻击

Table 6 Success Rates Comparison of Different Methods Under Various Models and Attack Settings

表 6 不同模型与攻击设置下不同方法的成功率对比 %

模型	攻击方法	类别				平均值
		鹅	法国斗牛犬	河马	加农炮	
Res-50	CDA	71.32	30.13	15.98	8.83	31.57
	C-GSP	68.45	61.34	63.67	58.23	62.92
	CGNC	70.01	60.56	61.98	60.12	63.17
	MTGA-DI	**71.67**	**69.45**	**71.12**	**68.34**	**70.15**
WRN-50-2	CDA	**57.78**	17.89	8.45	7.67	22.95
	C-GSP	55.89	54.56	54.23	51.78	54.12
	CGNC	56.12	55.90	53.78	52.45	54.56
	MTGA-DI	57.45	**57.23**	**56.89**	**57.34**	**57.23**

注：黑体数值表示最优实验结果。

成功率仅为：Res-50 为 60.12%；WRN-50-2 为 52.45%，均下降约 10 个百分点；而 MTGA-DI 在所有目标类别上均保持较高的攻击成功率，且相比 C-GSP 和 CGNC，其攻击成功率下降幅度最小，能够在目标类别变化时稳定适应新目标类别，具有更好的灵活性。

4.7 消融实验

1）目标类别数量的影响。在多目标生成式攻击中，设置生成模型可同时攻击的目标类别数量。然而，随着目标类别数量的增加，训练的难度也会显著上升。为了研究目标类别数量对攻击成功率的影响，本文进行了相关消融实验，其结果如图 8 所示。

实验观察到无论替代模型是 Dense-121 还是 Res-50，当目标类别数量在 50 个以内时，生成模型的攻击成功率均能保持在较高水平。然而，当目标类别数量超过 50 后，攻击成功率开始显著下降。具体而言：在白盒场景下，当生成模型需要同时攻击 200 个目标类别时，其有目标攻击的成功率相比仅攻击 10 个目标类别时下降了 10~20 个百分点；这一趋势在黑盒场景下同样显著，表明目标类别数量的增加会对生成模型的鲁棒性和迁移性能造成负面影响。

因此，为了在攻击效果与训练效率之间实现平衡，建议单个生成模型的目标类别数量不宜超过 50 个。这一设置可以在保持较高攻击成功率的同时，避免过度复杂的训练过程。

2）超参数α的影响。我们进一步研究了在替代模型 Dense-121 和 Res-50 上，不同 α 值对迁移攻击成功率的影响。从图 9 的结果可以看出：当α超过 1.0×10^{-4}时，视觉损失有效地学习了目标类别的视觉信息，提高了攻击成功率；当$\alpha=5.0\times10^{-4}$时，各模型的攻击成功率均达到了最高值。然而，在攻击 ViT 模型时，无论α取何值，攻击成功率均明显低于其他模型，最高的攻击成功率仅为 12.93% 和 17%。这一现象主要归因于 ViT 采用 Transformer 架构，而其他模型（包括替代模型 Res-50）均基于 CNN 架构，架构上的差异导致了对 ViT 的攻击迁移性较差。此外，当$\alpha=5.0\times10^{-2}$时，所有模型的攻击成功率显著降低，甚至大部分模型的攻击成功率都低于仅使用语义损失（$\alpha=0$）的情形。这表明过度学习视觉信息可能会损害生成器从分类器学习语义信息的能力，从而导致迁移成功率降低。因此在本文中，我们选取了最优参数$\alpha=5.0\times10^{-4}$。

4.8 跨域分析

为了验证 MTGA-DI 的跨域迁移性能，本文在 2 个具有显著差异的数据集 ImageNet[25] 和 Comics 上训

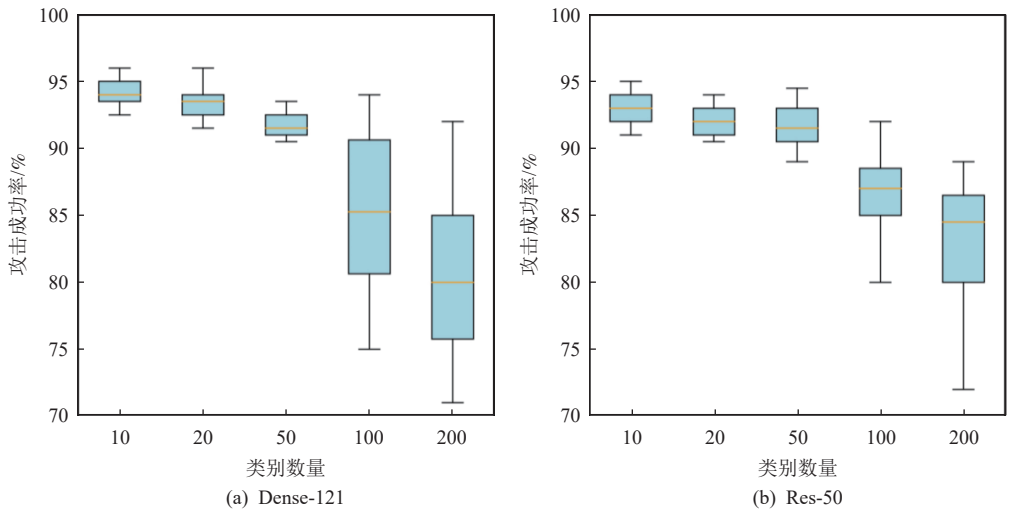
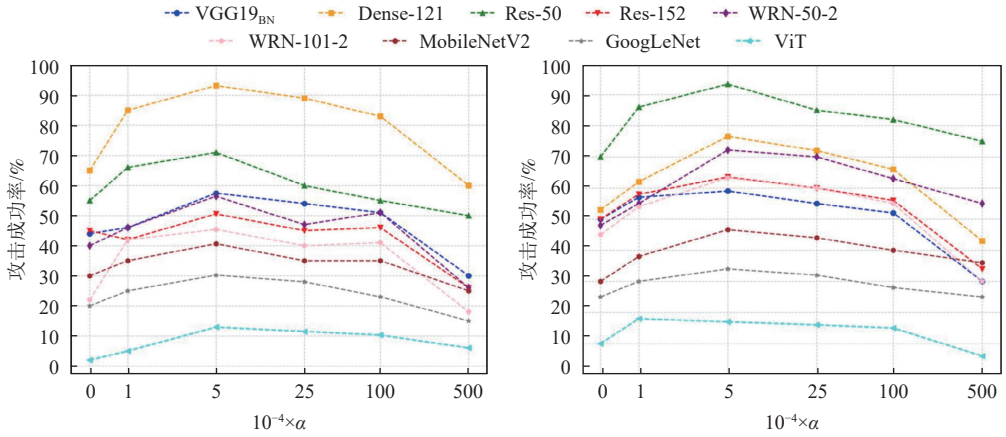


Fig. 8 Boxplot for the relationship between attack success rates and the number of target classes under different models

图 8 不同模型下攻击成功率与目标类别数量关系的箱图



(a) 替代模型: Dense-121 (b) 替代模型: Res-50

Fig. 9 Impact of the values of α on attack success rates under different models

图 9 不同模型下 α 值对攻击成功率的影响

练条件生成模型, 并对生成的对抗样本进行了可视化分析。ImageNet 包含多样化的自然图像, 而 Comics 由风格化的漫画图像构成, 这 2 个数据集在视觉特征和数据分布上具有显著差异。这种差异为评估生成模型在跨数据域上的鲁棒性与泛化性能提供了有效对比。图 10 展示了基于这 2 个数据集训练的条件生成模型生成的有目标对抗样本的示例。

通过观察图 10, 可得到如下结果:

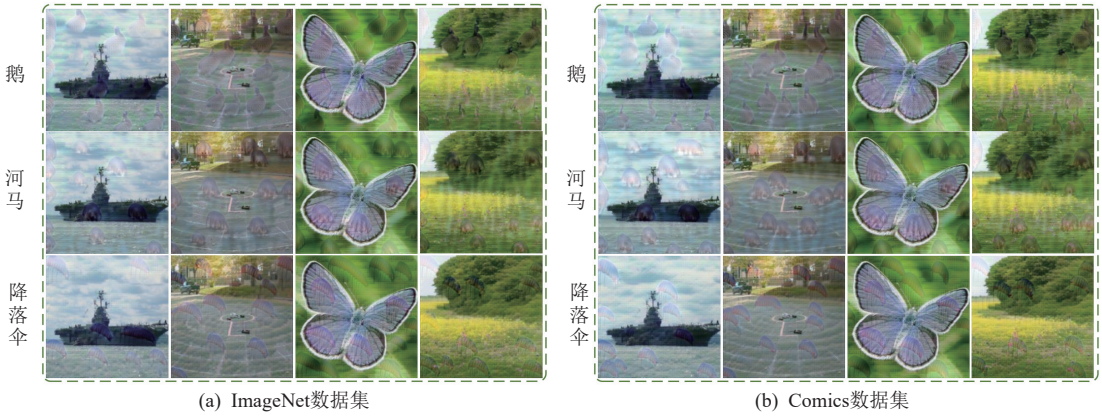
1)视觉模式的相似性。无论目标类别是“鹅”“河马”还是“降落伞”, 基于 ImageNet 和 Comics 训练的条件生成模型所生成的对抗扰动均展现出高度相似的视觉模式和结构特征。这表明, MTGA-DI 能够捕获目标类别的固有语义模式, 并在不同数据分布间保持一致性。

2)跨域鲁棒性。这一现象表明 MTGA-DI 生成的对抗样本对训练数据的分布具有较强的鲁棒性。这是由于 MTGA-DI 在训练阶段通过最小化特定目标类别的复合损失函数, 有效地捕获了目标类别的核心语义特征。

实验结果表明, MTGA-DI 方法不仅能够生成具有较高攻击性能的对抗样本, 还表现出卓越的跨域迁移能力。通过捕获和重现目标类别的语义模式, MTGA-DI 生成的对抗样本在不同数据分布中均能保持一致性。这些观察结果不仅验证了 MTGA-DI 的跨域性能, 还为理解对抗样本的本质特征和迁移机制提供了重要实证依据。

5 结束语

本文研究了语义和视觉双重信息对齐策略, 提出了基于双重信息的多目标生成式攻击方法 MTGA-DI。这一方法将双重信息的思想扩展到多目标场景, 在只训练 1 个生成模型的情况下, 保持对多个目标类别的高攻击成功率。实验结果表明, MTGA-DI 相



(a) ImageNet数据集 (b) Comics数据集

Fig. 10 Targeted adversarial examples generated by conditional generators trained on ImageNet and Comics datasets

图 10 基于 ImageNet 和 Comics 数据集训练的条件生成器所生成的有目标对抗样本

比先前的多目标攻击方法在标准训练模型和鲁棒模型上均获得了更好的实验结果，在面对基于输入预处理防御的模型时也有更好的表现。在将来的工作中，我们将该方法扩展到物理对抗样本，提升深度学习模型在实际应用场景中的鲁棒性和安全性。

作者贡献声明：余芹芹、鲍琦琦负责撰写与修订论文；沙嘉强负责研究设计；钱亚冠、顾钊铨、纪守领、王滨提供研究监督与指导。

参 考 文 献

[1] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint, arXiv: 1312.6199, 2013

[2] Poursaeed O, Katsman I, Gao B, et al. Generative adversarial perturbations[C]//Proc of the 19th Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4422−4431

[3] Zhao Zhenyu, Zhang Hanwei, Li Renjue, et al. Towards good practices in evaluating transfer adversarial attacks[J]. arXiv preprint, arXiv: 2211.09565, 2022

[4] Tian Yu. Research on image adversarial sample generation methods based on GAN[D]. Beijing: Beijing University of Posts and Telecommunications, 2020 (in Chinese)
(田宇. 基于 GAN 的图像对抗样本生成方法研究[D]. 北京：北京邮电大学, 2020)

[5] Pan Wenwen, Wang Xinyu, Song Mingli, et al. A review of adversarial sample generation techniques[J]. Journal of Software, 2020, 31(1): 67−81 (in Chinese)
(潘文雯, 王新宇, 宋明黎, 等. 对抗样本生成技术综述[J]. 软件学报, 2020, 31(1): 67−81)

[6] Yang Xiao, Dong Yinpeng, Pang Tianyu, et al. Boosting transferability of targeted adversarial examples via hierarchical generative networks[C]//Proc of the 12th European Conf on Computer Vision. Berlin: Springer, 2022: 725−742

[7] Fang Hao, Kong Jiawei, Chen Bin, et al. Clip-guided generative networks for transferable targeted adversarial attacks[C]//Proc of the 19th European Conf on Computer Vision. Berlin: Springer, 2024: 1−19

[8] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proc of the 28th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2014: 27−35

[9] Wang Zhibo, Wang Xue, Ma Jingjing, et al. A review of adversarial attacks on computer vision systems[J]. Journal of Computer Research and Development, 2023, 46(2): 436−468 (in Chinese)
(王志波, 王雪, 马菁菁, 等. 面向计算机视觉系统的对抗样本攻击综述[J]. 计算机学报, 2023, 46(2): 436−468)

[10] Dong Yinpeng, Liao Fangzhou, Pang Tianyu, et al. Boosting adversarial attacks with momentum[C]//Proc of the 31st Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 9185−9193

[11] Xie Cihang, Zhang Zhishuai, Zhou Yuyin, et al. Improving transferability of adversarial examples with input diversity[C]//Proc of the 32nd Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 2730−2739

[12] Dong Yinpeng, Pang Tianyu, Su Hang, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//Proc of the 32nd Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4312−4321

[13] Wei Zhipeng, Chen Jingjing, Wu Zuxuan, et al. Enhancing the self-universality for transferable targeted attacks[C]//Proc of the 24th Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 12281−12290

[14] Naseer M, Khan S, Khan M, et al. Cross-domain transferability of adversarial perturbations[C/OL]//Proc of the 33rd Conf on Neural Information Processing Systems. San Diego: NIPS, 2019[2025-01-10]. https://proceedings.neurips.cc/paper_files/paper/2019/file/99cd3843754d20ec3c5885d805db8a32-Paper.pdf

[15] Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint, arXiv: 1412.6572, 2015

[16] Radford A, Kim J, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8748−8763

[17] Naseer M, Khan S, Hayat M, et al. On generating transferable targeted perturbations[C]//Proc of the 18th European Conf on Computer Vision. Berlin: Springer, 2021: 7708−7717

[18] Ganeshan A, BS V, Babu R V. Fda: Feature disruptive attack[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2019: 1−11

[19] Zeiler M, Fergus R. Visualizing and understanding convolutional networks[C]//Proc of the 11th European Conf on Computer Vision. Berlin: Springer, 2014: 818−833

[20] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks[C/OL]//Proc of the 27th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2014[2025-01-10].https://proceedings.neurips.cc/paper_files/paper/2014/hash/532a2f85b6977104bc93f8580abbb330-Abstract.html

[21] Poursaeed O, Katsman I, Gao B, et al. Generative adversarial perturbations[C]//Proc of the 31st Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4422−4431

[22] Chen Guangyao, Peng Peixi, Ma Li, et al. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain[C]//Proc of the 18th European Conf on Computer Vision. Berlin: Springer, 2021: 458−467

[23] Krizhevsky A , Hinton G . Learning multiple layers of features from tiny images[EB/OL]. Toronto, Ontario, Canada: University of Toronto, 2009[2025-01-10]. https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

[24] Netzer Y, Wang Tao, Coates A, et al. Reading digits in natural images with unsupervised feature learning[C/OL]//Proc of the 25th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2011[2025-01-10]. https://static.googleusercontent.com/media/resear-

ch.google.com/zh-CN//pubs/archive/37648.pdf
[25] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]//Proc of the 22nd Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248−255
[26] Yang Yanchao, Lao Dong, Sundaramoorthi G, et al. Phase consistent ecological domain adaptation[C]//Proc of the 34th Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 9011−9020
[27] Liu Quande, Chen Cheng, Qin Jing, et al. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space[C]//Proc of the 35th Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1013−1023
[28] Cai Mu, Zhang Hong, Huang Huijuan, et al. Frequency domain image translation: More photo-realistic, better identity-preserving[C]//Proc of the 18th European Conf on Computer Vision. Berlin: Springer, 2021: 13930−13940
[29] Wang Zhe, Chi Ziqiu, Zhang Yanbing. FreGAN: Exploiting frequency components for training GANs under limited data[C]//Proc of the 36th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 33387−33399
[30] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of the 30th Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778
[31] Huang Gao, Liu Zhuang, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proc of the 31st Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 4700−4708
[32] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint, arXiv: 1409.1556, 2014
[33] Zagoruyko S. Wide residual networks[J]. arXiv preprint, arXiv: 1605.07146, 2016
[34] Sandler M, Howard A, Zhu Menglong, et al. MobileNetV2: Inverted residuals and linear bottlenecks[C]//Proc of the 32nd Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4510−4520
[35] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions[C]//Proc of the 28th Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1−9
[36] Dosovitskiy A. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint, arXiv: 2010.11929, 2021
[37] Hang Jie, Han Keji, Chen Hui, et al. Ensemble adversarial black-box attacks against deep learning systems[J]. Pattern Recognition, 2020, 101: 107184
[38] Geirhos R, Rubisch P, Michaelis C, et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness[C/OL]//Proc of the 7th Int Conf on Learning Representations. Washington: ICLR, 2019[2025-01-10]. https://openreview.net/pdf?id=Bygh9j09KX
[39] Hendrycks D, Mu N, Cubuk E, et al. Augmix: A simple data processing method to improve robustness and uncertainty[J]. arXiv preprint, arXiv: 1912.02781, 2020
[40] Kingma D. Adam: A method for stochastic optimization[J]. arXiv preprint, arXiv: 1412.6980, 2014
[41] Dziugaite G, Ghahramani Z, Roy D. A study of the effect of JPG compression on adversarial images[J]. arXiv preprint, arXiv: 1608.00853, 2016
[42] Ding G W, Wang Luyu, Jin Xiaomeng. AdverTorch v0.1: An adversarial robustness toolbox based on Pytorch[J]. arXiv preprint, arXiv: 1902.07623, 2019



Qian Yaguan, born in 1976. PhD, professor. Member of CCF. His main research interests include AI security, machine learning, pattern recognition, and computer vision.
钱亚冠, 1976 年生。博士, 教授。CCF 会员。主要研究方向为人工智能安全、机器学习、模式识别、计算机视觉。



Yu Qinqin, born in 2000. Master candidate. Her main research interests include computer vision, machine learning, deep learning, and multimodal attacks.
余芹芹, 2000 年生。硕士研究生。主要研究方向为计算机视觉、机器学习、深度学习、多模态攻击。



Sha Jiaqiang, born in 1999. Master. His main research interests include deep learning, machine learning, and computer vision.
沙嘉强, 1999 年生。硕士。主要研究方向为深度学习、机器学习、计算机视觉。



Bao Qiqi, born in 1996. PhD, lecturer. Her main research interests include image restoration, image generation, trustworthy artificial intelligence, and multimodal security.
鲍琦琦, 1996 年生。博士, 讲师。主要研究方向为图像复原、图像生成、可信人工智能、多模态安全。



Gu Zhaoquan, born in 1989. PhD, professor. Member of CCF. His main research interests include cyberspace security, cyber range, big data analysis, and artificial intelligence security.
顾钊铨, 1989 年生。博士, 教授。CCF 会员。主要研究方向为网络空间安全、网络靶场、大数据分析、人工智能安全。



Ji Shouling, born in 1986. PhD, professor. Member of CCF. His main research interests include data-driven security and privacy, artificial intelligence security, and big data analysis.

纪守领, 1986 年生。博士，教授。CCF 会员。主要研究方向为数据驱动的安全与隐私、人工智能安全、大数据分析。



Wang Bin, born in 1978. PhD, professor. His main research interests include IoT security, cryptography, AI security, and new network security architectures.

王　滨, 1978 年生。博士，教授。主要研究方向为物联网安全、密码学、人工智能安全、新型网络安全架构。