E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4; TP18　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)01-0154-13
论文引用格式:Mi J X, Zeng H Q, Zhou L F, Chen T and Cheng X. 2026. Adaptive frequency interference: generator-driven transferable adversarial attack. Journal of Image and Graphics, 31(1):0154-0166(米建勋, 曾鸿泉, 周丽芳, 陈涛, 程晓. 2026. 自适应频率干扰:生成器驱动的可迁移对抗样本攻击. 中国图象图形学报, 31(1):0154-0166)[DOI:10.11834/jig.250075]

自适应频率干扰:生成器驱动的可迁移对抗样本攻击

米建勋[1,2*],曾鸿泉[1,2],周丽芳[1,2],陈涛[3],程晓[3]
1. 重庆邮电大学计算机科学与技术学院,重庆　400065; 2. 重庆邮电大学图像认知重庆市重点实验室,重庆　400065;
3. 国网重庆市电力公司电力科学研究院,重庆　401123

摘　要: 目的　对抗样本作为人工智能安全领域的基础性问题,其生成方法的研究具有重要意义。当前的全局扰动优化策略生成的对抗样本存在源模型过拟合问题,这显著降低了跨模型迁移攻击的效果。为了解决这一问题,一种有效的思路是将图像转换到频率域,提取对决策相关的频率信息后进行针对性干扰以生成对抗样本。但现有频率域攻击方法过度依赖低频信息,导致对抗攻击的可迁移性受限。**方法**　为了突破传统攻击方法在迁移性与生成效率方面的双重局限,本文提出一种基于生成器架构的频率域对抗样本生成方法。该方法能够在频率域中自适应地保留对分类决策有益的高频和低频信息,并通过干扰这些信息生成对抗样本。通过利用生成器架构,本文方法无需真实标签,一旦生成器训练完成,即可直接将输入图像送入生成器生成对抗样本,从而显著提高生成效率,同时保持了较高的迁移攻击性能。**结果**　在ImageNet上进行的大量实验证明,本方法在卷积神经网络(convolutional neural network, CNN)和视觉变换器(vision Transformer, ViT)上均优于现有方法。此外,本文还利用了动物数据集以及微软通用物体检测数据集(Microsoft common objects in context, MS-COCO)分别在跨数据集图像分类任务以及目标检测任务上验证了攻击方法的有效性。**结论**　本文提出的基于生成器架构的频率域对抗样本生成方法,通过自适应选择并干扰对分类决策有益的频率成分,显著提升了对抗样本的迁移攻击能力,在生成效率和攻击性能上均优于对比方法,为对抗样本生成领域提供了新的解决方案。
关键词: 深度学习;人工智能安全;计算机视觉;图像分类对抗样本;生成式攻击算法

Adaptive frequency interference: generator-driven transferable adversarial attack

Mi Jianxun[1,2*], Zeng Hongquan[1,2], Zhou Lifang[1,2], Chen Tao[3], Cheng Xiao[3]
1. School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing　400065, China;
2. Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing　400065, China;
3. State Grid Chongqing Electric Power Company Electric Power Research Institute, Chongqing　401123, China

Abstract: Objective The goal of digital image adversarial examples is to add subtle, nearly imperceptible noise to an image to deceive computer vision systems into making incorrect decisions. Most existing adversarial example generation methods aim to maximize the loss function, pushing the image away from the true label. These methods typically involve

收稿日期:2025-03-12;**修回日期**:2025-07-21;**预印本日期**:2025-07-28
*通信作者:米建勋　mijianxun@gmail.com
基金项目:国家自然科学基金项目(62276039)
Supported by: National Natural Science Foundation of China (62276039)

multiple searches across the input space to find possible adversarial perturbations. However, this approach often leads to overfitting to the source model, reducing the transferability of adversarial examples, making them minimally effective in attacking other models. Furthermore, performing iterative searches across the entire input space increases the computational cost of generating adversarial samples. To address the above issues, we propose a novel adversarial example generation method based on a generative architecture, which selectively interferes with the frequency components that are crucial for the decision-making process of the model. This method significantly improves the transferability of adversarial attacks. Once the generator is trained, it can rapidly generate adversarial examples by simply feeding in the input, thus enhancing the efficiency of adversarial example generation. Our method not only addresses the challenges of transferability and overfitting but also reduces the computational cost by eliminating the need for exhaustive searches in the input space, enabling fast and effective adversarial attacks. **Method** Specifically, our method begins by feeding the input into a generator architecture, which produces two outputs: Branch 1 generates the initial adversarial noise, and Output 2 provides the initial mask information. The purpose of the initial mask is to retain the feature information that is beneficial for classification in the original image. After obtaining the initial adversarial noise, we add it to the original sample to create the preliminary adversarial example. The next step involves applying a discrete cosine transform (DCT) to the adversarial example, followed by multiplying the DCT-transformed adversarial sample with the mask information to isolate the crucial frequency components that affect the decision-making process. Afterward, an inverse DCT (IDCT) is applied to recover the image in the spatial domain. This processed adversarial image is then fed into an image classifier, and the generator is optimized by maximizing the loss function, which drives the adversarial example to produce the desired misclassification results. The incorporation of the mask ensures that only the most influential frequency components, which directly influence the model's classification, are preserved. This approach not only guides the generator to focus on the critical parts of the image but also enhances the transferability of the adversarial example to different models, given that the manipulated frequency components are less likely to be overfitted to a specific model. By strategically targeting and preserving important frequency information, our method improves the effectiveness and efficiency of generating adversarial examples. **Result** Our method outperforms baseline attack methods in terms of attack effectiveness when applied to convolutional neural networks, vision Transformer models, and adversarially trained models. We also employed our approach to attack Google's commercial API, and the results demonstrated that our method effectively compromised their classifier, causing it to produce incorrect detection outcomes. To further understand the underlying reasons for the success of our method, we conducted a heatmap visualization analysis. This analysis revealed that our method successfully shifted the model's attention away from the foreground and toward the background, thereby manipulating the model's decision-making process. Furthermore, leveraging the generative architecture, we used the trained generator to generate adversarial examples on the Animal dataset and the Microsoft common objects in context dataset. These adversarial samples were then applied to attack the cross-dataset image classification tasks and the downstream object detection tasks. The experimental results indicate that our method maintains its attack effectiveness in cross-dataset image classification tasks and downstream tasks, highlighting its transferability and versatility. Hence, our attack method is not only effective for image classification models but also exhibits strong attack capabilities for more complex tasks such as object detection. These results validate the broad applicability and robustness of our adversarial attack strategy across different types of computer vision models and tasks. **Conclusion** The generator architecture proposed in this paper targets frequency components that are beneficial for classification decisions, significantly improving the transferability of adversarial attacks.

Key words: deep learning; artificial intelligence security; computer vision; adversarial example in image classification; generative attack algorithm

0 引　言

近年来，深度学习技术在计算机视觉领域取得显著进展，例如图像分类（Krizhevsky 等，2012）、目标检测（Huang 等，2017）和语义分割（He 等，2017）。在某些特定场景下，深度学习模型的准确率甚至超越了人类水平。然而，研究表明，基于深度神经网络

的计算机视觉系统容易受到对抗样本攻击。Szegedy等人(2013)发现在图像分类任务中,通过向图像中添加微小的扰动,能够使图像分类系统以极高的置信度做出错误的分类决策。更为令人担忧的是,已有研究发现对抗样本还具有一定的迁移攻击能力(Moosavi-Dezfooli等,2016;Madry等,2017;袁珑等,2022;秦颖鑫等,2024;闫嘉乐等,2022)。

分类器的决策依赖于输入图像的多维度特征,可分为共性特征与个性特征(Wang等,2021;Jin等,2024;Wang和He,2021;王杨等,2022):共性特征是多模型决策的共同依赖依据;个性特征仅对单一模型有效。因此,生成高迁移性对抗样本的核心在于干扰模型的共性特征依赖。

现有迁移攻击方法主要通过梯度信息搜索扰动空间(Dong等,2018;Xie等,2019;Lin等,2019),但存在两方面问题:1)梯度信息反映模型对当前输入的局部反馈,过度依赖会限制跨模型攻击能力,且迭代优化过程效率低下;2)未加约束的全局扰动易破坏模型无关的个性特征,降低扰动对共性特征的干扰强度。

针对问题1),一种主流的方法是采用生成式架构。由于生成器经过大量训练,其中记录了关于当前数据集的丰富信息,通过输入原始图像,能够直接生成其对应的对抗扰动。生成式方法不依赖于梯度信息,也不需要多次迭代,从而提高对抗样本的生成效率以及迁移攻击能力(Xiao等,2018;Isola等,2017;He等,2021;Zhu等,2024;赵恩浩和凌捷,2025;孙曦音等,2019)。

在更集中的特征丰富的频域中,利用频率信息生成对抗样本被证明是有效的。因此,解决问题2)的自然方法是使用离散余弦变换(discrete cosine transform, DCT)将图像从空间域转换为频域。低频信息用于描述整体图像特征。目前,主要的频域攻击方法是通过保留频域中有价值的低频信息,去除高频细节,然后通过逆离散余弦变换(inverse discrete cosine transform, IDCT)将其转换回空间域,得到具有更明显共同特征的图像。接着对该转换后的图像进行攻击,生成梯度和对抗扰动,并将其添加到原始图像中,生成对抗样本(Guo等,2018)。

然而,现有频域攻击方法隐含"模型仅依赖低频特征决策"的假设,忽略了部分模型可能利用高频信息完成分类(Wang等,2020)。因此,研究重点在于通过何种设计能够有效地保留对分类有效的高低频信息。生成器可以通过损失函数约束学习从输入分布中生成图像。同样,利用这些约束,可以引导生成器学习哪些高频和低频信息对分类有益。图1展示了本研究方法与仅保留低频信息方法在保留特征方面的差异。第1行是原始图像,第2行是保留对攻击有用的高频和低频信息的自适应特征图像,第3行是仅保留低频信息的特征图像。传统的基于频域的对抗攻击方法假设关键特征存在于低频信息中,限制了对抗攻击的效果。通过使用损失函数约束生成器,本研究使生成器能够学习当前数据集中对分类有益的高频和低频信息,并利用这些信息生成对抗样本,从而显著提高攻击成功率。

(a) 原始图像

(b) 保留对分类决策有益的高低频信息图像

(c) 仅保留低频信息图像

图1 不同频率特征图像对比

Fig. 1 Comparison of images with different frequency features ((a) original images; (b) images preserving useful high and low frequency information for classification; (c) images preserving low frequency information only)

具体而言,如图2和图3所示,本研究设计了一种基于生成器架构的方法,能够在生成对抗样本时,根据输入的不同动态破坏当前输入中对决策有益的高频和低频信息。该方法结合了对齐机制,以确保所保留的扰动特征与对抗样本一致。

本文主要贡献如下:1)提出一种能够动态保留输入中对生成对抗样本至关重要的高频和低频信息的方法。该方法能够更为精准地扰乱共性特征,进而显著提升迁移攻击的成功率。2)本研究提出的方

法通过生成器架构实现，能够自适应地保留输入中对分类有效的高频和低频信息。经过训练的生成器能够直接产生出对抗扰动，无需依赖传统方法中的梯度信息及多次迭代，从而大幅度提升了对抗样本的生成效率。3）通过大量实验，将本文方法与其他主流方法对比，本文方法不仅具有更快的对抗样本生成速度，而且在迁移攻击成功率方面表现更优。此外，本文方法对跨数据集图像分类任务以及下游任务也展现出了一定的攻击效果，进一步证明了其在实际应用中的潜力。

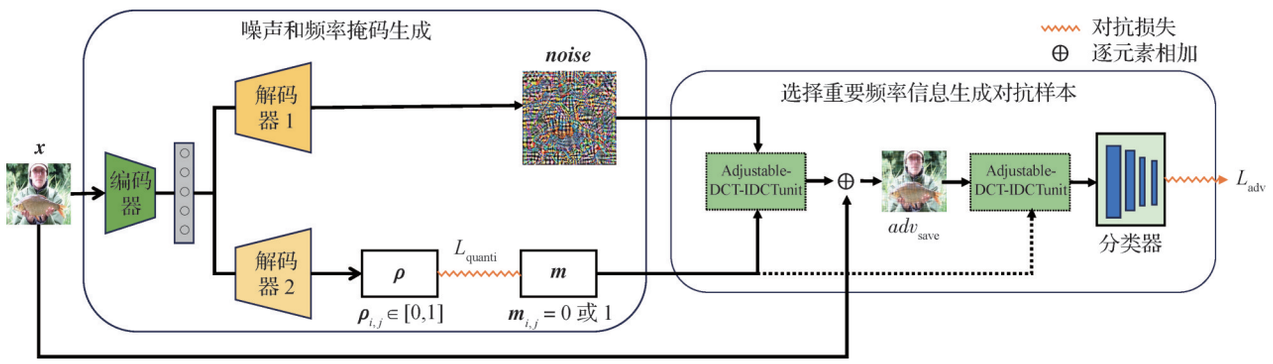

图 2　本文框架

Fig. 2　Method architecture diagram

图 3　自适应频率保留模块

Fig. 3　Adaptive frequency preservation module

1　相关工作

1.1　基于生成器攻击方法

生成器的本质功能在于生成图像，而对抗样本则旨在创造一幅与原图视觉上几乎无异的图像，以此诱使人工智能系统做出错误判断。因此，生成网络自然而然地成为对抗样本生成领域的理想选择。加之生成器能够学习整个数据集的分布特征，在对抗样本生成任务中，仅需利用生成器产生扰动，并通过设定约束来控制扰动的大小，即可成功生成对抗样本。使用生成器架构进行攻击的工作主要有：Xiao 等人（2018）、Isola 等人（2017）、He 等人（2021）、Zhu 等人（2024）；Baluja 和 Fischer（2018）；Poursaeed 等人（2018）、Mopuri 等人（2018）。

1.2　基于梯度的攻击方法

1.2.1　基于迭代的优化方法

基于迭代的对抗样本生成方法是该领域的基础。MIM（momentum iterative method）（Dong 等，2018）通过引入动量信息，避免了对源模型的过拟合。IEM（improved euler method）（Peng 等，2023）则借鉴数值欧拉方法，采用改进的欧拉方法，减少了近似误差，从而更准确地搜索最优解，构建更具迁移性的梯度基础攻击。

1.2.2　基于输入变换的攻击方法

基于输入变换的方法借鉴了模型优化中的数据增强策略，通过对输入数据进行变换来提升攻击性能。DIM（diverse input method）（Xie 等，2019）在 MIM 基础上引入随机填充和平移策略，提高对抗样本的迁移攻击能力。SI-NI-FGSM（scale-invariant nesterov iterative fast gradient sign method）（Lin 等，2019）结合 Nesterov 迭代快速梯度符号法和尺度不变攻击法，进一步提升迁移攻击能力。SSA（spectrum simulation attack）算法（Long 等，2022）通过离散余弦变换在频域中加入随机噪声，模拟不同网络对输入频率的注意力，从而实现攻击效果。

1.2.3　基于频域的攻击方法

传统对抗样本生成方法在输入空间中搜索对抗扰动，增加了时间复杂度。基于离散余弦变换的特性，许多研究将图像从空间域转换到频域，利用频率特征生成扰动（Guo 等，2018；Ehrlich 和 Davis，2019；Deng 和 Karam，2022）。然而，现有研究表明，图像中的高频信息对于分类至关重要。一些方法通过频率增强模型或假设模型仅关注低频信息进行决策，通

过保留低频信息进行攻击。

本研究提出一种利用生成器架构动态保留有效高低频信息的方法，显著提升了迁移攻击效果和对抗样本生成效率。

2 方 法

2.1 问题建模

将$\boldsymbol{x}$表示为一个原始图像，$\boldsymbol{y}$表示其对应的真实标签。设F为分类器，本研究的目的是生成对抗样本$\boldsymbol{x}^{\text{adv}}=\boldsymbol{x}+\boldsymbol{\delta}$，使得$F(\boldsymbol{x}^{\text{adv}})\neq F(\boldsymbol{x})$，在对抗样本生成任务中，需要对对抗扰动$\boldsymbol{\delta}$进行约束，以确保生成的对抗样本与原始样本之间的差异保持在足够小的范围ε内，攻击的公式化描述为

$$\arg\max_{\boldsymbol{x}^{\text{adv}}} J(\boldsymbol{x}^{\text{adv}},\boldsymbol{y})\ \ \text{s.t.}\ \left\|\boldsymbol{x}^{\text{adv}}-\boldsymbol{x}\right\|_{\infty}\leqslant\varepsilon \tag{1}$$

式中，J是交叉熵损失。正如前文所述，已有多种基于梯度的方法被提出以解决对抗样本生成的优化问题。与这些方法不同，本研究利用生成器架构，一旦模型训练完成，输入原始数据即可直接生成对应的对抗样本。为了充分利用生成器结构的优势，并在频域中有效保留对分类任务至关重要的频率特征，本研究提出一种包含3个关键步骤的方法，具体如下：

1）生成器架构。受He等人（2021）工作的启发，设计了一种生成器架构，能够根据输入图像生成初始的对抗扰动，并同时生成用于筛选有效频率信息的掩码矩阵。

2）有效特征对齐保留。设计了一个量化初始掩码矩阵的框架，该框架能够同时实现对抗扰动与对抗样本特征的有效对齐。

3）损失优化。通过设计量化损失和攻击损失函数，利用反向传播优化生成器，确保生成的对抗样本既具有高隐蔽性，又能有效攻击目标模型。量化损失用于约束扰动的大小和范围，而攻击损失则用于提高对抗样本的迁移攻击能力。

在接下来的内容中，将详细介绍上述方法的实现细节及其在对抗样本生成任务中的应用效果。

2.2 生成器架构

如图2所示，在本文的框架中，利用两个不同的分支分别优化这两个变量。一个分支输出初始对抗扰动$\boldsymbol{\delta}$，其受到预定义的l_{∞}范数ε的限制；另一个分支输出初始频率保留掩码矩阵$\boldsymbol{\rho}$，其中，$\boldsymbol{\rho}$满足$\bigvee\boldsymbol{\rho}_{i,j}\in[0,1]$，为了保证保留频率信息的可行性，对$\boldsymbol{\rho}$进行二值化处理，生成矩阵$\boldsymbol{m}$，用于保留DCT后系数矩阵中的指定系数，$\boldsymbol{m}$中的每个元素是一个二进制值（0或1）。如果$m(i,j)=1$，则表示对图像进行离散余弦变换操作后，第$(i,j)$个位置的离散余弦变换系数得以保留。下一节中，将详细介绍如何设计模块，使之有效保证对抗扰动与对抗样本特征对齐。

2.3 有效特征对齐保留

如图3所示，本研究设计了一个模块，名为自适应动态频率保留单元（adjustable frequency reserve using discrete cosine transform and inverse discrete cosine transform，Adjustable-DCT-IDCTunit），它可以接收任意输入$\boldsymbol{z}$，然后将输入通过离散余弦变换，使得输入从空间域转换到频率域。具体为

$$\boldsymbol{z}_{\text{dct}}=DCT(\boldsymbol{z}) \tag{2}$$

接着，与生成器中所得二值化矩阵$\boldsymbol{m}$进行哈达玛积，保留有效的高低频信息，具体为

$$\boldsymbol{z}'_{\text{dct}}=\boldsymbol{z}_{\text{dct}}\odot\boldsymbol{m} \tag{3}$$

式中，$\boldsymbol{m}$表示对$\boldsymbol{\rho}$进行处理的二值化矩阵。$\odot$表示逐位相乘。

最后，再通过逆离散余弦变换完成对输入的重构。具体为

$$\boldsymbol{z}_{i\,\text{dct}}=IDCT(\boldsymbol{z}'_{\text{dct}}) \tag{4}$$

将Adjustable-DCT-IDCTunit模块的整个过程表示为

$$\boldsymbol{z}_{i\,\text{dct}}=K(\boldsymbol{z},\boldsymbol{m}) \tag{5}$$

式中，$\boldsymbol{z}$为输入（对抗样本或者生成器生成的初始对抗扰动），$\boldsymbol{m}$为生成器中掩码矩阵的二值化矩阵形式。

有效的对抗扰动应该集中于原始输入的重要特征上，如图2所示，本研究设计了一个对齐模块，令对抗扰动***noise***通过式（5）后与原始样本经过式（9）（详后）操作得到对抗样本$\boldsymbol{adv}_{\text{save}}$，同时，将得到的对抗样本经过式（5）操作得到重构的对抗样本$\boldsymbol{adv}_{\text{det}}$，并将其送入白盒模型得到损失函数优化生成器，本研究的做法有效保证了对抗扰动关注的特征与对抗样本关注特征的一致性。一旦生成器G在训练数据和白盒模型上训练完成，本研究通过如下操作得到对抗样本$\boldsymbol{adv}_{\text{save}}$，攻击其他黑盒模型。

$$\boldsymbol{\rho}=G_2(\boldsymbol{x}) \tag{6}$$

式中，$G(\cdot)$为生成器，$G_1(\cdot)$表示生成器第1个输出分支，即初始对抗扰动$\boldsymbol{\delta}$，$G_2(\cdot)$表示生成器第2个输出分支，即初始频率保留掩码矩阵$\boldsymbol{\rho}$。

$$\boldsymbol{m} = Q(\boldsymbol{\rho}; r) = \begin{cases} 1 & \rho_{i,j} \geqslant r \\ 0 & \text{其他} \end{cases} \tag{7}$$

式中，r为初始设定阈值。

$$\boldsymbol{noise}_{\text{adv}} = K(G_1(\boldsymbol{x}), \boldsymbol{m}) \tag{8}$$

式中，$\boldsymbol{m}$是由生成器的第2个输出分支生成的掩码信息，而$G_1(\boldsymbol{x})$是由生成器的第1个输出分支生成的初始噪声信息。本研究的目标是将这两者输入到Adjustable-DCT-IDCTunit中，以获得有效的对抗噪声信息$\boldsymbol{noise}_{\text{adv}}$。

$$\boldsymbol{adv}_{\text{save}} = \boldsymbol{x} + \boldsymbol{noise}_{\text{adv}} \tag{9}$$

式(9)的目的是将原始样本和对抗噪声相加，得到初始对抗样本。

2.4 损失函数

2.4.1 对抗性损失

为了实现欺骗目标黑盒模型的目标，需要一个本地替代模型F监督生成器G的训练。对于非定向攻击，目标是生成一个对抗性图像，该图像不被分类为真实标签。在每个训练迭代中，生成器试图通过损失函数最大化错误标签的对抗性图像的概率。为了实现这个目标，本文使用了以交叉熵作为基础的式(10)作为对抗损失的基础。

$$\boldsymbol{grad} = \text{sign}\left(\frac{\partial J(\boldsymbol{adv}_{\text{dct}}, \boldsymbol{y})}{\partial \boldsymbol{adv}_{\text{dct}}}\right) \tag{10}$$

式中，J表示交叉熵损失，$\boldsymbol{adv}_{\text{dct}}$表示根据式(5)重建的对抗样本。采用符号函数处理梯度信息，旨在保留对图像分类具有关键意义的像素。

$$L_{\text{adv}} = -sum(\boldsymbol{adv}_{\text{save}} \times \boldsymbol{grad}) \tag{11}$$

式中，sum表示将对抗示例与经符号函数处理的梯度信息逐像素相乘，旨在累积各像素在对抗示例中对损失函数的贡献度，目的是识别出对错误分类贡献最大的像素，从而引导生成器创建更有效的对抗样本。

2.4.2 量化损失

本文设计了一个特殊的量化算子用以选择重要的频率信息，目的是确保生成的频率保留掩码矩阵$\boldsymbol{\rho}$为0，1的二值化矩阵。然而，矩阵$\boldsymbol{\rho}$取值并非二值化，通过式(7)得到其二值化形式用于生成对抗样本，这可能会由于量化引起的信息丢失而带来一些测试误差。为了减小训练和测试之间的差距，一个直接的解决方案是减小$\boldsymbol{\rho}$的量化误差。量化参数应尽可能接近完整精度参数，期望测试性能接近训练性能。量化损失为

$$L_{\text{quanti}}(\boldsymbol{\rho}) = \left\| \boldsymbol{\rho} - \boldsymbol{m} \right\|_2 \tag{12}$$

2.4.3 隐蔽性损失

直观上，篡改图像的像素值越小，越能够增强对抗样本的隐蔽性，在本文方法中，使用生成器保证了生成的对抗扰动的最大值小于等于$\boldsymbol{\varepsilon}$，为了进一步提高对抗样本的隐蔽性，研究中使用了如下的隐蔽性损失，降低生成扰动的总体大小。具体为

$$L_{\text{perturb}}(\boldsymbol{noise}_{\text{adv}}) = \left\| \boldsymbol{noise}_{\text{adv}} \right\|_2 \tag{13}$$

总损失为

$$L = L_{\text{adv}} + \lambda_{\text{quanti}} L_{\text{quanti}} + \lambda_{\text{perturb}} L_{\text{perturb}} \tag{14}$$

式中，λ_{quanti}和λ_{perturb}为超参数，通过最小化损失L更新生成器G的参数，最终达到生成可迁移对抗样本的目的。

2.5 算法

本研究的一个基本策略即利用生成器以及保留有效频率信息来生成对抗性扰动。首先，将输入通过生成器得到初始对抗样本$\boldsymbol{adv}_{\text{save}}$。然后，为了使得对抗特征对齐，将$\boldsymbol{adv}_{\text{save}}$通过式(5)得到$\boldsymbol{adv}_{\text{dct}}$。为了进一步提高迁移攻击效率，算法中对$\boldsymbol{adv}_{\text{dct}}$使用了Xie等人(2019)的输入多样性方法。最后，通过损失函数优化了生成器G。当完整训练一个生成器后，只需要将输入送入生成器，即可得到具有迁移攻击能力的对抗样本，大大提高了对抗样本的生成效率。

算法1 重要频率特征用于对抗攻击。

输入：原始样本$\boldsymbol{x}$，真值$\boldsymbol{y}$，源模型F，对抗损失L，生成器G。

输出：对抗样本$\boldsymbol{adv}'_{\text{save}}$。

参数：迭代次数T。

1)for $t = 0$ to $T - 1$ do。

(1)通过式(6)—式(9)得到$\boldsymbol{adv}_{\text{save}}$。

(2)通过式(5)得到$\boldsymbol{adv}_{\text{dct}}$。

(3)对$\boldsymbol{adv}_{\text{dct}}$进行DIM算法中的输入图像变换。

(4)通过式(14)优化生成器G。

2)end for。

3)得到训练好的生成器G。

4) 利用训练好的生成器G通过式(6)—式(9)得到对抗样本$\boldsymbol{adv}'_{\text{save}}$。

5)返回对抗样本$\boldsymbol{adv}'_{\text{save}}$。

3 实 验

本研究将所设计方法命名为自适应选择重要频域信息的对抗样本攻击方法(adaptive selective crucial frequency adversarial example, ASCFAE)。在本研究中,生成对抗样本在卷积神经网络(convolutional neural network, CNN)、对抗训练网络和ViT(vision Transformer)上全面与其他最先进的竞争方法进行性能评估。此外,本研究还探讨了ASCFAE在对抗样本生成效率方面的优势,进一步说明了其在目标检测任务和商用API上的有效性。最后,通过消融实验和模型决策注意力图的可视化分析,验证了ASCFAE方法的有效性。

3.1 实验设置

3.1.1 模型与数据集

本研究实验基于ImageNet(Deng等,2009)、动物数据集(含90类动物,共1万余幅图像)及MS-COCO(Microsoft common objects in context)(Lin等,2014)3个数据集开展,具体设置如下:

本研究在广泛使用的ImageNet数据集上进行实验验证。实验通过攻击Inception-v3(Incv3)模型和ResNet-50(Res50)模型(Szegedy等,2016; He等,2016)生成对抗样本。对于所提出的方法,利用源模型监督生成器的训练过程,并在推理阶段直接将原始图像转换为对抗图像。输入图像统一裁剪为299×299像素。此外,为了全面评估方法的迁移攻击能力,实验还使用了多种目标模型,包括DenseNet161(Dense161)(Huang等,2017)、EfficientNet(Eff)(Tan和Le,2019),以及3种基于Transformer的模型,即Vit-b、Visformer-S和LeVit-256(Dosovitskiy等,2020;Chen等,2021;Graham等,2021)。为进一步验证方法的鲁棒性,本研究还在3个集成对抗训练模型上评估了对抗样本的迁移性,分别是Inc-v3ens3、Inc-v3ens4和IncRes-v2ens2(Tramèr等,2017)。为了确保实验的公平性,从ImageNet验证集中随机选择了5 000幅图像用于生成对抗样本。所有目标模型在测试集上的自然准确率均超过92%,确保了实验结果的可靠性。

为验证算法的跨数据集与跨任务攻击能力,基于ImageNet预训练的Incv3和Res50生成器,分别在动物数据集(90类、10 000余幅图像)上攻击其他4种图像分类模型,并在MS-COCO数据集上生成对抗样本以攻击下游目标检测任务。

3.1.2 评价指标

实验中,采用广泛使用的攻击成功率(attack success rate, ASR)作为核心评估指标。ASR定义为测试集中被成功误导至错误分类的样本占比,其值越高,表明攻击性能越优。

在实际攻击场景中,对抗样本的实时生成能力是关键约束——若攻击需与网络篡改结合(如截获用户图像后快速生成对抗样本并替换原数据以干扰后续检测),生成耗时过长可能导致用户察觉异常,终止检测流程。因此,实验引入每秒生成帧数(frames per second, FPS)评估对抗样本的生成效率,该指标直接反映方法在实际场景中的可行性。

在本研究提出的方法中,对齐模块在精确对齐对抗噪声与对抗样本特征方面发挥着关键作用,这有助于提高对抗样本的图像质量。本文借鉴了Zhang等人(2023)和Li等人(2018)的评价指标,使用了结构相似性指数(structural similarity index, SSIM)、峰值信噪比(peak signal-to-noise ratio, PSNR)衡量对抗样本的图像质量。这两个指标均为正向指标,数值越高,意味着对抗样本的图像质量与原始样本越接近,对抗样本的隐匿性越强,更难被受害者所察觉。

3.1.3 基准攻击算法

本研究中选择了6种具有代表性的对抗样本生成方法作为基准攻击算法进行比较。这些方法包括:SI-NI-FGSM(Lin等,2019)、基于频率的SSA算法(模拟不同网络对不同输入频率的注意力)(Long等,2022)、GE-Adv(Zhu等,2024)、IEM(Peng等,2023)、DIM(Xie等,2019)和基于动量的MIM(Dong等,2018)。上述算法除GE-Adv以外,迭代次数T设置为10,迭代过程中超参数α设置为1.6/255,对抗噪声ε设置为16/255,GE-Adv算法参数则与原论文保持一致。

3.1.4 ASCFAE参数

实验中,ASCFAE的参数涉及两个源模型:Inception-v3和ResNet-50。两个模型的配置设置相同:λ_{quanti}和λ_{perturb}设置为0.000 1,最大对抗噪声ε设置为16/255,生成器训练轮数设置为30,式(7)中的r设置为0.5。优化采用Adam优化器,初始学习率lr=0.000 1,$betas$=(0.5, 0.999)。

3.2 迁移攻击实验

本研究所有实验在 NVIDIA RTX 3080 GPU 上完成，结果如表 1 所示。数据显示，本研究提出的方法在几乎所有测试模型上均展现出优于对比算法的攻击性能：以 Incv3 为源模型攻击其他模型时，平均攻击成功率高达 97.76%，较当前最优的 GE-Adv 算法提升 2.11%；且在所有模型上的攻击成功率均优于其他方法。以 Res50 为源模型时，本研究方法取得了最高的平均攻击成功率；仅在攻击 Incv3 模型时略逊于 DIM 算法，但仍位列第 2。值得注意的是，本研究方法在 Dense161 和 Eff 模型上表现出更稳定的攻击能力，均取得最佳攻击结果，进一步验证了其迁移攻击的鲁棒性。综合对比，相较于主流算法，本研究方法在迁移攻击性能上更具优势且稳定性更强。

表 1 不同方法攻击成功率

Table 1 Attack success rate on multiple models

/%

模型	方法	Dense161	Res50	Incv3	Eff	平均
Incv3	SI-NI-FGSM	72.36	78.50	99.06	74.44	81.09
	SSA	59.22	64.58	99.04	67.68	72.63
	GE-Adv	94.78	96.64	98.92	92.24	95.65
	IEM	62.26	69.34	98.56	64.70	73.72
	DIM	64.52	69.50	98.50	66.14	74.67
	MIM	49.46	57.58	98.34	52.98	54.59
	本文	**97.18**	**96.82**	**99.66**	**97.36**	**97.76**
Res50	SI-NI-FGSM	86.22	99.68	61.80	84.12	82.96
	SSA	91.36	**99.78**	68.00	81.76	85.23
	GE-Adv	84.72	99.00	47.56	85.44	79.18
	IEM	92.48	98.98	64.22	81.52	84.30
	DIM	91.08	99.68	**80.14**	83.96	88.72
	MIM	83.22	98.82	54.56	70.32	76.71
	本文	**94.26**	99.66	68.58	**93.16**	**88.92**

注：加粗、下划线字体分别表示各模型的最优、次优攻击效果。

针对对抗训练模型的攻击性能实验结果如表 2 所示。以 Incv3 为源模型时，本研究方法的平均攻击成功率较次优的 DIM 算法高出 24.35%；以 Res50 为源模型时，仍保持最佳平均攻击性能。值得注意的是，该表现与未对抗训练模型上的实验结果一致。进一步分析发现，以 Incv3 为源模型的攻击策略在防御模型（均基于 Inception 架构）上效果更优，主要归因于目标模型与源模型的架构一致性降低了迁移攻击的难度。

表 2 对抗训练模型上的攻击成功率

Table 2 Attack success rate on adversarial training models

/%

模型	方法	Inc-v3 ens3	Inc-v3 ens4	IncRes-v2ens	平均
Incv3	SI-NI-FGSM	33.78	29.86	17.36	27.00
	SSA	32.00	30.12	18.80	26.97
	GE-Adv	10.22	9.80	6.88	8.97
	IEM	26.12	29.34	13.14	21.07
	DIM	45.10	43.00	26.48	38.19
	MIM	22.96	21.08	11.78	18.61
	本文	**71.44**	**69.46**	**46.72**	**62.54**
Res50	SI-NI-FGSM	22.38	19.28	10.62	17.43
	SSA	25.58	23.30	15.80	21.56
	GE-Adv	10.82	15.46	8.50	11.59
	IEM	17.72	16.58	9.60	14.63
	DIM	**36.12**	**30.72**	17.70	28.18
	MIM	17.16	15.46	9.04	13.89
	本文	33.78	29.62	**26.30**	**29.90**

注：加粗、下划线字体分别表示各模型的最优、次优攻击效果。

3.3 攻击 ViT 模型实验

本研究还攻击了 ViT 模型，以测试该方法的跨架构攻击能力。所有模型使用 PyTorch 中的 timm 库实现。本研究使用 Incv3 和 Res50 作为源模型，实验结果如表 3 所示。实验结果显示，本研究提出的方法在攻击性能上相较于其他竞争方法展现出明显优势。使用 Incv3 作为源模型时，性能比第 2 好的模型提高 30.39%；使用 Res50 作为源模型时，性能比第 2 好的模型提高 1.88%。实验结果不仅验证了所提出方法在 ViT 模型上的有效适用性，还揭示了当前 ViT 模型在面对对抗样本攻击时的脆弱性。

3.4 对抗样本生成效率评估

对抗样本生成速度也是对抗样本的一项核心的考量指标，更快的生成速度，意味着能够更高效地实施攻击。本研究使用 Incv3 作为源模型，比较了本研究的方法与竞争方法的对抗样本生成速度，本研究中，均保持了实验中 batch 数等于 2。如表 4 所示，本

研究的方法单张对抗样本生成效率、每秒帧数方面远远高于基于梯度的对抗样本生成方法，表现出更显著的优势。

表3 ViT模型上的攻击成功率

Table 3 Attack success rate on ViT-based models

/%

模型	方法	Vit-b	Visformer-S	LeVit-256	平均
Incv3	SI-NI-FGSM	19.20	41.08	44.80	35.03
	SSA	11.56	30.92	33.96	25.48
	GE-Adv	9.28	39.66	54.04	34.33
	IEM	14.46	33.78	35.46	27.90
	DIM	23.82	39.88	42.98	35.56
	MIM	13.64	26.46	26.90	22.33
	本文	**30.34**	**82.06**	**85.44**	**65.95**
Res50	SI-NI-FGSM	15.70	35.56	33.84	28.37
	SSA	12.24	28.52	30.44	23.73
	GE-Adv	15.66	18.62	24.94	19.74
	IEM	12.30	29.72	26.72	22.91
	DIM	**18.46**	**41.32**	40.38	33.39
	MIM	12.26	25.16	22.54	19.99
	本文	17.02	39.94	**48.84**	**35.27**

注：加粗、下划线字体分别表示各模型的最优、次优攻击效果。

表4 不同方法生成对抗样本效率比较

Table 4 Comparison of efficiency of different methods in generating adversarial examples

方法	帧数/(帧/s)	时间/s
SI-NI-FGSM	1.46	0.68
SSA	0.59	1.68
IEM	1.39	0.72
DIM	3.20	0.31
MIM	6.10	0.16
ASCFAE(本文)	**222.22**	**0.004 5**

注：加粗、下划线字体分别表示各列最优、次优结果。

3.5 跨数据集攻击实验

相较于基于梯度的方法，基于生成器的方案在训练完成后，仅需将原始图像输入生成器即可生成对抗样本，无需真实标签的参与，这为跨数据集攻击以及下游任务攻击创造了便利条件。

3.5.1 动物数据集图像分类攻击实验

在本小节中，本研究使用提出的ASCFAE算法在动物数据集上攻击了MobileNetv2(Sandler等，2018)、VGG16(Simonyan和Zisserman，2015)、ResNet152(Res152)(He等，2016)和Inception-v4(Incv4)(Szegedy等，2017)4种图像分类模型。4种模型未被攻击情况下在动物数据集分类准确度分别为99.16%、99.36%、98.11%和97.00%。4种模型分类对抗样本结果如表5所示，以Incv3和Res50为源模型在ImageNet训练的生成器生成对抗样本在动物数据集上攻击上述4个模型的平均攻击成功率分别是88.24%和88.80%，充分验证了ASCFAE算法的跨数据集以及黑盒攻击能力，因为对于生成器而言，4种分类模型以及数据集均是训练时未曾见过的。

表5 黑盒分类模型跨数据集攻击结果

Table 5 Cross-dataset attack results on black-box classification models

/%

源模型	MobileNetv2	VGG16	Res152	Incv4	平均
Incv3	95.92	89.63	78.52	88.89	88.24
Res50	93.33	95.19	90.74	75.93	88.80

3.5.2 下游目标检测任务攻击实验

本研究使用提出的ASCFAE算法在MS-COCO数据集上攻击目标检测模型FasterRCNN(faster region-based convolutional neural network)(Ren等，2017)、YOLOv3(you only look once，version 3)(Redmon，2018)、FCOS(fully convolutional one-stage object detector)(Tian等，2019)和SSD300(single shot multibox detector 300)(Liu等，2016)。攻击结果如表6所示，评估指标为平均精度(average precision，AP)和平均召回率(average recall，AR)，采用预定义的设置，交并比(intersection-over-union，IoU)的平均值为[0.5∶0.95]。从结果可以看出，本研究的方法有效地攻击了下游物体检测任务，在攻击前后，物体检测模型的检测结果发生了显著变化。

3.6 消融实验

本研究进一步分析了本研究框架中关键部分对于可迁移攻击的贡献。本实验采用Incv3为源模型，训练时省略算法1的图像变换步骤，直接使用原始输入，结果如表7所示。表7结果为不同模型

表 6　下游目标检测任务攻击结果

Table 6　Attack results on downstream object detection tasks

模型	方法	FasterRCNN		YOLOv3		FCOS		SSD300	
		AP	AR	AP	AR	AP	AR	AP	AR
无攻击情况		0.37	0.52	0.28	0.40	0.37	0.59	0.26	0.37
Incv3	ASCFAE	0.09	0.19	0.08	0.15	0.08	0.21	0.07	0.13
Res50	ASCFAE	0.13	0.23	0.12	0.21	0.19	0.26	0.10	0.19

表 7　消融实验结果

Table 7　Ablation experiment results

方法	Dense161/%	Res50/%	Incv3/%	Eff/%	SSIM	PSNR/dB
ASCFAE	95.72	96.62	99.72	95.32	0.70	75.43
w/o alignment module	98.00	98.34	99.86	96.90	0.67	74.57
w/o frequency module	25.56	33.80	26.34	25.26	0.87	76.59
baseline	70.72	74.12	73.40	75.22	0.82	76.36

注：w/o alignment madule：无对齐模块；w/o freguency module：无自适应频率选择模块；baseline：基线方法。

上的 ASR。

3.6.1　解耦对齐模块

训练时，本研究方法中，对生成器输出的扰动，经过图 3 所示的 Adjustable-DCT-IDCTunit 模块后与原始输入相加生成对抗样本，再对对抗样本都同样经过 Adjustable-DCT-IDCTunit 得到 $\boldsymbol{adv}_{\mathrm{dct}}$，并将其送入分类器，计算损失函数以优化生成器。两次送入 Adjustable-DCT-IDCTunit 的目的是确保对抗样本保留的有效特征与对抗扰动是一致的。

为了验证对齐模块的必要性，实验设计了消融对比：仅保留生成器生成的扰动经单次 Adjustable-DCT-IDCTunit 处理后与原始样本像素级相加生成对抗样本（即取消第 2 次频域特征对齐），并以该样本直接计算分类器损失优化生成器。如表 7 第 2 行数据所示，无对齐模块的方法虽然攻击成功率略高（攻击性稍强），但其生成的对抗样本在 SSIM 与 PSNR 指标上显著低于含对齐模块的版本，表明图像质量（隐匿性）受损。这一结果表明，Adjustable-DCT-IDCTunit 的对齐机制在提升攻击成功率的同时，有效抑制了扰动引入的视觉失真，成功平衡了攻击性与图像质量，验证了本文方法理论设计的合理性。

3.6.2　解耦生成器中自适应频率选择模块

为了验证本研究设计的自适应高低频信息保留掩码矩阵的有效性，通过生成器生成对抗扰动，同时经过离散余弦变化将扰动从空间域转向频率域，借鉴 Guo 等人（2018）的方法，本研究选择 $r = 1/8$，即仅保留 1/8 的低频信息，并通过逆离散余弦变化得到对抗扰动后与原始样本相加得到初始对抗样本，初始对抗样本通过同样的方式并保留 1/8 的低频信息，并送入分类器计算损失优化生成器。如表 7 第 3 行数据所示，当固定了保留的频率特征之后，攻击效果大幅降低，说明本研究设计的自适应频率选择模块对于提高攻击效果有着至关重要的作用。

3.6.3　基线方法比较

本研究仅保留了生成器中的对抗噪声生成模块，用生成器生成对抗噪声，并将对抗噪声与原始输入相加得到对抗样本，将对抗样本通过离散余弦变换转换到频率域，仅保留 1/8 的低频信息后，通过逆离散余弦变换将对抗样本转换回空间域，送入分类器计算损失优化生成器。如表 7 第 4 行数据所示，攻击成功率远远低于本研究的自适应频率信息方法，验证了本研究方法的有效性。

4　结　论

本研究针对现有对抗样本生成方法的低效问题提出创新性解决方案。传统方法通常依赖于梯度信息遍历整个输入空间，这不仅降低了对抗样本的迁

移攻击能力,还显著影响了生成效率。为了解决这一问题,本研究提出一种基于生成器架构的方法,能够根据不同输入自适应选择对分类决策有益的高频和低频信息,并通过破坏这些信息生成对抗样本。该方法在多个模型和任务上进行了验证,展现了其高效性和广泛适用性。与其他方法相比,本文方法在迁移攻击性能和生成效率方面均表现出显著优势。尽管如此,本文方法仍存在一些不足。例如,在某些特定的模型架构下,生成的对抗样本可能无法保持一致的攻击效果,这主要是由于不同模型对频率特征的敏感度存在差异。未来的研究可以进一步优化生成器的训练过程,以减少计算资源的需求并提高训练效率。同时,可以探索本文方法在其他计算机视觉任务(如语义分割、图像生成等)中的应用。

参考文献(References)

Baluja S and Fischer I. 2018. Learning to attack: adversarial transformation networks//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI Press: 2687-2695 [DOI: 10.1609/aaai.v32i1.11672]

Chen Z S, Xie L X, Niu J W, Liu X F, Wei L H and Tian Q. 2021. Visformer: the vision-friendly transformer//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 569-578 [DOI: 10.1109/ICCV48922.2021.00063]

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database// Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Deng Y P and Karam L J. 2022. Frequency-tuned universal adversarial attacks on texture recognition. IEEE Transactions on Image Processing, 31: 5856-5868 [DOI: 10.1109/TIP.2022.3202366]

Dong Y P, Liao F Z, Pang T Y, Su H, Zhu J, Hu X L, et al. 2018. Boosting adversarial attacks with momentum//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9185-9193 [DOI: 10.1109/CVPR.2018.00957]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2020. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/2010.11929.pdf

Ehrlich M and Davis L. 2019. Deep residual learning in the jpeg transform domain//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 3483-3492 [DOI: 10.1109/ICCV.2019.00358]

Graham B, El-Nouby A, Touvron H, Stock P, Joulin A, Jégou, H, et al. 2021. LeViT: a vision transformer in ConvNet's clothing for faster inference//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 12239-12249 [DOI: 10.1109/ICCV48922.2021.01204]

Guo C, Frank J S and Weinberger K Q. 2018. Low frequency adversarial perturbation [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/1809.08758.pdf

He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.322]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He Z W, Wang W, Dong J and Tan T N. 2021. Transferable sparse adversarial attack [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/2105.14727.pdf

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2261-2269 [DOI: 10.1109/CVPR.2017.243]

Huang J, Rathod V, Sun C, Zhu M L, Korattikara A, Fathi A, et al. 2017. Speed/accuracy trade-offs for modern convolutional object detectors//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 3296-3297 [DOI: 10.1109/CVPR.2017.351]

Isola P, Zhu J Y, Zhou T and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 5967-5976 [DOI: 10.1109/cvpr.2017.632]

Jin Z B, Zhang J Y, Zhu Z Y and Chen H M. 2024. Benchmarking transferable adversarial attacks [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/2402.00418.pdf

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 26th International Conference on Neural Information Processing Systems—Volume 1. Lake Tahoe, USA: Curran Associates Inc.: 1097-1105

Lin J D, Song C B, He K, Wang L W and Hopcroft J E. 2019. Nesterov accelerated gradient and scale invariance for adversarial attacks [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/1908.06281.pdf

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision—ECCV 2014. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, et al. 2016. SSD: single shot MultiBox detector//Proceedings of the 14th European Conference on Computer Vision—ECCV 2016. Amsterdam, the Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Li Y Z, Tian D, Chang M C, Bian X and Lyu S W. 2018. Robust adversarial perturbation on deep proposal-based models [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/1809.05962.pdf

Long Y Y, Zhang Q L, Zeng B H, Gao L L, Liu X L, Zhang J, et al. 2022. Frequency domain model augmentation for adversarial attack//Proceedings of the 17th European Conference on Computer Vision—ECCV 2022. Tel Aviv, Israel: Springer: 549-566 [DOI: 10.1007/978-3-031-19772-7_32]

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2017. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/1706.06083.pdf

Moosavi-Dezfooli S M, Fawzi A and Frossard P. 2016. DeepFool: a simple and accurate method to fool deep neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 2574-2582 [DOI: 10.1109/CVPR.2016.282]

Mopuri K R, Uppala P K and Babu R V. 2018. Ask, acquire, and attack: data-free UAP generation using class impressions//Proceedings of the 15th European Conference on Computer Vision—ECCV 2018. Munich, Germany: Springer: 20-35 [DOI: 10.1007/978-3-030-01240-3_2]

Peng A J, Lin Z, Zeng H, Yu W X and Kang X G. 2023. Boosting transferability of adversarial example via an enhanced Euler's method//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSP49357.2023.10096558]

Poursaeed O, Katsman I, Gao B C and Belongie S. 2018. Generative adversarial perturbations//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4422-4431 [DOI: 10.1109/cvpr.2018.00465]

Qin Y X, Zhang K J, Pan H W and Ju Y H. 2024. Adversarial attacks in computer vision: a survey [EB/OL]. [2025-02-17]. https://doi.org/10.19678/j.issn.1000-3428.0069826 (秦颖鑫, 张可佳, 潘海为, 巨亚昊. 2024. 计算机视觉对抗攻击研究综述[EB/OL]. [2025-02-17]. https://doi.org/10.19678/j.issn.1000-3428.0069826

Redmon J. 2018. Yolov3: an incremental improvement [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/1804.02767.pdf

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Sandler M, Howard A, Zhu M L, Zhmoginov A and Chen L C. 2018. MobileNetV2: inverted residuals and linear bottlenecks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4510-4520 [DOI: 10.1109/CVPR.2018.00474]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2025-02-17]. https://https://arxiv.org/pdf/1409.1556.pdf

Sun X Y, Feng H M, Liu B and Zhang J Y. 2019. Adversarial examples generation based on GAN. Computer Applications and Software, 36(7): 202-207, 248 (孙曦音, 封化民, 刘飚, 张健毅. 2019. 基于GAN的对抗样本生成研究. 计算机应用与软件, 36(7): 202-207, 248) [DOI: 10.3969/j.issn.1000-386x.2019.07.034]

Szegedy C, Ioffe S, Vanhoucke V and Alemi A. 2017. Inception-v4, inception-ResNet and the impact of residual connections on learning//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI Press: 4278-4284 [DOI: 10.1609/aaai.v31i1.11231]

Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z. 2016. Rethinking the inception architecture for computer vision//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 2818-2826 [DOI: 10.1109/CVPR.2016.308]

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. 2013. Intriguing properties of neural networks [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/1312.6199.pdf

Tan M X and Le Q. 2019. EfficientNet: rethinking model scaling for convolutional neural networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 6105-6114

Tian Z, Shen C H, Chen H and He T. 2019. FCOS: fully convolutional one-stage object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 9626-9635 [DOI: 10.1109/ICCV.2019.00972]

Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D and McDaniel P. 2017. Ensemble adversarial training: attacks and defenses [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/1705.07204.pdf

Wang H H, Wu X D, Huang Z Y and Xing E P. 2020. High-frequency component helps explain the generalization of convolutional neural networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 8681-8691 [DOI: 10.1109/CVPR42600.2020.00871]

Wang X S and He K. 2021. Enhancing the transferability of adversarial attacks through variance tuning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 1924-1933 [DOI: 10.1109/CVPR46437.2021.00196]

Wang Y, Cao T Y, Yang J B, Zheng Y F, Fang Z and Deng X T. 2022. A perturbation constraint related weak perceptual adversarial example generation method. Journal of Image and Graphics, 27(7): 2287-2299 (王杨, 曹铁勇, 杨吉斌, 郑云飞, 方正, 邓小桐.

2022. 结合扰动约束的低感知性对抗样本生成方法. 中国图象图形学报, 27(7): 2287-2299) [DOI: 10.11834/jig.200681]

Wang Z B, Guo H C, Zhang Z F, Liu W X, Qin Z and Ren K. 2021. Feature importance-aware transferable adversarial attacks//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 7619-7628 [DOI: 10.1109/ICCV48922.2021.00754]

Xiao C W, Li B, Zhu J Y, He W, Liu M Y and Song D. 2018. Generating adversarial examples with adversarial networks [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/1801.02610.pdf

Xie C H, Zhang Z S, Zhou Y Y, Bai S, Wang J Y, Ren Z, et al. 2019. Improving transferability of adversarial examples with input diversity//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 2725-2734 [DOI: 10.1109/cvpr.2019.00284]

Yan J L, Xu Y, Zhang S C and Li K Z. 2022. Survey of research on adversarial examples attack and defense in image classification model. Computer Engineering and Applications, 58(23): 24-41 (闫嘉乐, 徐洋, 张思聪, 李克资. 2022. 图像分类模型的对抗样本攻防研究综述. 计算机工程与应用, 58(23): 24-41) [DOI: 10.3778/j.issn.1002-8331.2205-0520]

Yuan L, Li X M, Pan Z X, Sun J M and Xiao L. 2022. Review of adversarial examples for object detection. Journal of Image and Graphics, 27(10): 2873-2896 (袁珑, 李秀梅, 潘振雄, 孙军梅, 肖蕾. 2022. 面向目标检测的对抗样本综述. 中国图象图形学报, 27(10): 2873-2896) [DOI: 10.11834/jig.210209]

Zhang Y Y, Tan Y A, Sun H P, Zhao Y H, Zhang Q X and Li Y Z. 2023. Improving the invisibility of adversarial examples with perceptually adaptive perturbation. Information Sciences, 635: 126-137 [DOI: 10.1016/j.ins.2023.03.139]

Zhao E H and Ling J. 2025. Data-free black-box adversarial attack method based on GAN. Computer Engineering and Applications, 61(7): 204-212 (赵恩浩, 凌捷. 2025. 基于GAN的无数据黑盒对抗攻击方法. 计算机工程与应用, 61(7): 204-212) [DOI: 10.3778/j.issn.1002-8331.2311-0227]

Zhu Z Y, Chen H M, Wang X Y, Zhang J Y, Jin Z B, Choo K K R, et al. 2024. GE-AdvGAN: improving the transferability of adversarial samples by gradient editing-based adversarial generative model// Proceedings of 2024 SIAM International Conference on Data Mining (SDM). Houston, USA: SIAM: 706-714 [DOI: 10.1137/1.9781611978032.81]

作者简介

米建勋,男,教授,主要研究方向为机器学习、人工智能安全、多模态大模型和人工智能可解释问题。
E-mail: mijianxun@gmail.com
曾鸿泉,男,硕士研究生,主要研究方向为人工智能安全。
E-mail:S230201007@stu.cqupt.edu.cn
周丽芳,女,教授,主要研究方向为目标检测及跟踪、目标识别、医学图像分割。E-mail:zhoulf@cqupt.edu.cn
陈涛,女,正高级工程师,主要研究方向为电网系统运行分析和电网安全性。E-mail:chentao@qq.sgcc.com.cn
程晓,男,工程师,主要研究方向为电网数字技术应用和电网安全性。E-mail:hengxiaoV@163.com