中图分类号：　　单位代号：10280
密　　级：　　学　　号：21721230

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	**基于傅里叶频域分析的黑盒对抗攻击方法研究**

作　者：**李晨**
学科专业：**通信与信息系统**
导　师：**张新鹏**
完成日期：**2024 年 5 月**

姓　名：李晨　　学号：21721230

论文题目：基于傅里叶频域分析的黑盒对抗攻击方法研究

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主　席：

委　员：

导　师：

答辩日期：　　年　月　日

姓　名：李晨　　学号：21721230

论文题目：基于傅里叶频域分析的黑盒对抗攻击方法研究

上海大学学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师指导下，独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他研究者对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：

日 期：　　年　月　日

上海大学学位论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（保密论文在解密后应遵守此规定）

学位论文作者签名：　　导师签名：

日 期：　年 月 日　　日 期：　年 月 日

上海大学工学硕士学位论文

基于傅里叶频域分析的黑盒对抗攻击方法研究

作　　者：李晨
学科专业：通信与信息系统
导　　师：张新鹏

上海大学通信与信息工程学院
二〇二四年五月

A Dissertation Submitted to Shanghai University for the Degree of Master in Engineering

Research on Black-box Adversarial Attacks Based on Fourier Frequency Domain Analysis

Candidate：	Chen Li
Major：	Communication and Information System
Supervisor：	Xinpeng Zhang

School of Communication and Information Engineering

Shanghai University

May, 2024

摘 要

对抗攻击可以通过对于输入数据的微小扰动误导图像分类模型产生错误决策，揭示了神经网络的潜在脆弱性。因此，对抗攻击研究不仅是理解神经网络局限性的关键，也是构建更安全鲁棒模型的基础。黑盒对抗攻击在不了解目标模型相关信息的前提下，在替代模型上构造可迁移的对抗样本，使其能够攻击未知的目标模型，是最具有现实意义的对抗攻击研究之一。为了更深入地理解对抗攻击的机理，与以往从空间域展开的研究不同，本文从频域的角度，基于频域特性设计黑盒对抗攻击，聚焦于黑盒对抗攻击的可迁移性与不可见性，主要研究内容如下：

1) 针对现有的可迁移黑盒对抗攻击普遍缺乏可解释性，且目标攻击场景下可迁移性弱的问题，提出了一种基于频域鲁棒性的可迁移对抗攻击。首先通过傅里叶热图分析了神经网络普遍存在的频域脆弱现象，并利用这一共性特征进行频域模型增强。其次，针对现有损失函数存在的方向性不足问题，分析对抗攻击目标，提出使用三元损失来优化扰动方向。在常用数据集上的实验结果表明，相较于现有方法，所提出的方法在目标与非目标攻击场景下的可迁移性均有明显提升。

2) 针对现有的黑盒对抗攻击多关注可迁移性，忽视不可见性而在对抗样本中引入明显伪影的问题，提出了一种基于频域特性分析的不可见对抗攻击。综合人类视觉系统频域特性与神经网络的频域鲁棒性，量化了扰动中每个频率分量的贡献，应用 K-means 聚类，得到一组神经网络对于该频率扰动敏感性更高且人眼不易察觉该频率信息变化的候选频域分量。从频域计算梯度并叠加扰动，并提出了一个新的对抗攻击优化目标，在攻击过程中优化对抗损失与频域损失组成的联合损失函数，将扰动的频率分布向候选频域约束。实验结果表明，该方法在不可见性上显著优于现有的黑盒对抗攻击。

关键词：黑盒对抗攻击；可迁移性；可解释性；傅里叶分析；对抗样本；不可见性

ABSTRACT

Adversarial attacks can mislead image classification models to produce wrong decisions through small perturbations to the input data, revealing the potential vulnerability of neural networks. Therefore, the study of adversarial attacks is not only the key to understanding the limitations of neural networks, but also the basis for constructing more secure and robust models. Black-box adversarial attack, which constructs transferable adversarial examples on surrogate models without knowing the relevant information about the target model so that it can attack an unknown target model, is one of the most relevant areas of adversarial attack research. In order to understand the mechanism of the adversarial attack more deeply, different from the previous research from the spatial domain, this dissertation designs the black-box adversarial attack based on the characteristics of the frequency domain of adversarial attack from the perspective of the frequency domain, focusing on the transferability and invisibility of the black-box adversarial attack, and the main research contents are as follows:

1) Aiming at the existing transferable adversarial attacks that generally lack interpretability and have weak transferability in targeted attack scenarios, a transferable adversarial attack based on frequency domain robustness is proposed. The prevalent frequency-domain fragility phenomenon of neural networks is analyzed through Fourier heatmaps, and this common feature is utilized for frequency-domain model augmentation, so that the frequency-domain information of the adversarial perturbation conforms to the common fragility characteristics of neural networks. In addition, to address the lack of directionality of the existing loss function, the direction of the perturbation is optimized by analyzing the adversarial attack target using the triple loss. Experimental results on the commonly used datasets CIFAR10, CIFAR100, and Tiny-ImageNet show that compared with the existing methods, the proposed method has significantly improved the transferability in both targeted and untargeted attack scenarios.

2) Aiming at the problem that existing black-box adversarial attacks focus on transferability and often ignore invisibility, introducing obvious artifacts in the adversarial examples, an invisible adversarial attack based on frequency domain characterization is proposed. By synthesizing the human visual system characteristics and the frequency domain robustness of neural networks, the contribution of each frequency component in the perturbation is quantified, and K-means clustering is applied to obtain a set of candidate frequency domain components that neural networks are more sensitive to and at the same time are not easily detectable by the human eye. The gradient is computed from the frequency domain and the perturbation is superimposed from the frequency domain, and a new optimization objective for the adversarial attack is proposed to optimize the joint loss function consisting of the adversarial loss and the frequency-domain loss to constrain the frequency distribution of the perturbation toward the candidate frequency domain during the attack. Experimental results show that the method significantly outperforms existing transferable adversarial attacks in terms of invisibility.

Keywords: Black-box adversarial attacks; Transferability; Frequency domain interpretability; Fourier analysis; Adversarial examples; Invisibility

目 录

第一章 绪论

1.1 研究背景及意义

人工智能（Artificial Intelligence, AI）[1]在全球科技竞争与经济结构转型中扮演着核心角色，已经成为国际社会竞相布局的战略高地。我国充分认识到这一领域的战略价值与发展潜力，为推动 AI 从技术研发到产业生态的全方位发展，近年来相继出台了《新一代人工智能发展规划》[2]、《关于支持建设新一代人工智能示范应用场景的通知》[3]、《“数据要素×”三年行动计划（2024—2026年）》[4]等相关政策文件，并实施了一系列具有前瞻性和系统性的政策举措，强化 AI 基础设施建设，通过推动数据资源的开放共享赋能 AI 产业链可持续发展，促进 AI 的规模化应用与商业价值释放，不仅体现了国家层面对 AI 发展的执着追求与重大投资，也预示着 AI 行业即将迎来的广阔发展空间与深层次变革，也进一步确立了 AI 技术在未来全球经济与社会进步中的不可估量的角色。

深度神经网络（Deep Neural Network, DNN）作为 AI 领域应用最广泛的细分学科，在计算机算力、模型和数据可用性不断优化和提升的情况下，在图像分类[5]、目标检测[6]、语音识别[7]、语音合成等任务上取得巨大成功，并在安全要求极高的领域如自动驾驶[9]、金融风险管理[10]、医疗诊断[11]等实现快速发展。

伴随着 DNN 技术在高风险领域应用的日益广泛，对深度学习模型的安全需求也随之急剧上升。在自动驾驶、医疗图像分析以及军工等高风险领域，任何微小的失误都可能导致严重的后果。因此，如何在确保 AI 系统的高效性的同时，提高其安全性和鲁棒性，成为当前 AI 研究和应用中亟待解决的重要问题。

DNN 的鲁棒性是构建可信 AI 体系的关键要素之一。在图像分类领域，DNN 模型被期望即使面对经过特殊处理的数据样本，也能保持其预测的稳定性和准确性，而不受外界扰动的影响。而如图 1.1 所示，对抗攻击[12]通过对输入数据进行微小修改，就能显著地干扰 DNN 的判断，导致其输出错误的分类结果，揭示

了深度神经网络的脆弱性和安全漏洞，严重威胁 DNN 的鲁棒性和安全性[13-15]。因此，对抗攻击研究是提高 DNN 鲁棒性和确保可信 AI 不可或缺的一环。

图 1.1 对抗样本的生成过程

对抗攻击领域的研究对提升深度学习模型安全性有着深远的影响。研究者致力于在黑盒场景下，生成具有普遍威胁性的对抗样本，这一过程不仅有助于弥补现有模型的安全短板，也有利于推动整个深度学习领域向着更加稳定、安全和可信赖的方向发展，比如帮助对模型决策边界性质的深化认识[18]，以及在构建更为稳健和自解释的深度学习模型方面的技术革新[19]。

黑盒对抗攻击不仅是对现有 AI 系统安全性的考验，在真实世界的应用场景中同样发挥着重要作用，特别是在隐私保护方面。在大数据环境中，人们的个人信息和私人数据可能会受到各种形式的攻击和侵犯，使用对抗样本可以帮助保护用户的私人数据免受未经授权的利用，使未知模型更难识别和利用个人信息。另外，对于黑盒对抗攻击的深入研究，可以揭示 DNN 模型的脆弱性，并且探索改进算法以提高模型对这些攻击的防御能力[16]。这不仅有助于保护 DNN 模型在图像分类等领域的应用安全，也为实现可信 AI 提供了必要的技术保障。对于对抗性的探索[17]还促进了更深层次的安全评估框架和标准的建立，使得 AI 系统的开发者和用户能够更好地量化模型在面临潜在威胁时的表现和可靠性。

1.2 对抗攻击的分类

2013 年，Google 研究员 Szegedy 等人[12]最早发现并证明了通过对图像进行修改后可以误导神经网络模型，从此对于对抗样本的研究成为学术界的讨论热

点之一。而对抗攻击作为一种评估和挑战深度学习模型鲁棒性的核心手段，其分类与实施方式多样且具有针对性。首先，对抗攻击的分类可以从攻击场景的角度进行划分，主要分为白盒对抗攻击和黑盒对抗攻击两种类型：

1) 白盒对抗攻击[13, 20, 21-23]：是理想化的攻击场景，其中攻击者拥有全面的目标模型的信息。包括模型结构、参数配置、训练过程中梯度等关键信息。

2) 黑盒对抗攻击[24, 31]：模拟了更为现实且普遍存在的攻击环境，此时攻击者具备任何有关模型内部结构及参数的具体信息。在这种条件下，攻击者依赖于模型的输出反馈进行迭代优化或采用间接推断方法，来生成对抗样本。

其次，如图 1.2 所示，从攻击目标的角度，根据对抗攻击对输出类别篡改的具体目标导向性，攻击场景可以分为目标攻击和非目标攻击：

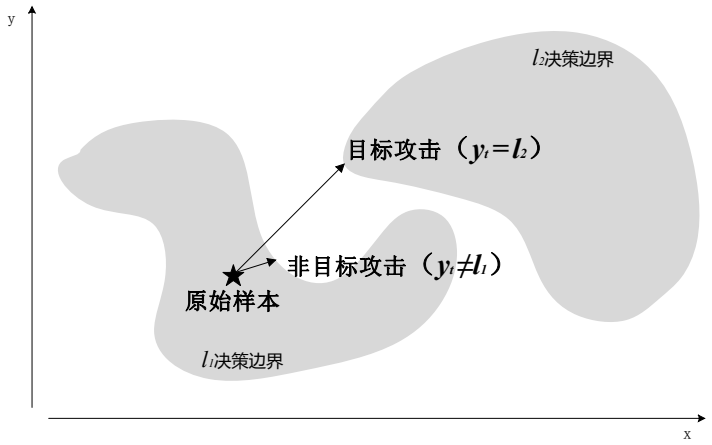


图 1.2 目标攻击和非目标攻击分别叠加对抗扰动，使样本跨越分类模型的决策边界

1) 目标攻击（Targeted Attack）：目标攻击场景旨在误导模型，将特定输入数据错误分类为攻击者预设的目标类别，使样本摆脱原始标签的决策边界，并跨越到目标标签。在这种场景下，对抗攻击往往采用精心设计的对抗样本生成策略，攻击者不仅追求使模型的预测结果偏离原始标签，而且要求模型将对抗样本明确误分类为指定的、不同于真实标签的类别。

2) 非目标攻击（Untargeted Attack）：非目标攻击场景的设定则相对宽松，其核心目标是降低模型对输入数据的预测准确性，而不特别关注将数据分类到某个特定类别。在这种情况下，攻击者只需要使样本跨越并远离原始类的公共决策边界，确保扰动后的对抗样本不被模型正确分类即可，无论是被误归入哪个类别都符合攻击目的。

另外，从对抗样本生成的原理角度，目前已经有了多种有效的对抗攻击算法的设计思路：

1) 基于梯度的对抗攻击算法[13, 20, 25-31]：这类方法利用深度神经网络 DNN 的反向传播特性，通过提取模型对于输入数据的梯度信息，对原始样本进行微小而定向的扰动，以达到误导模型判断的目的。

2) 基于优化的攻击算法[32-34]：此类方法并不直接依赖于梯度信息，而是通过定义特定的损失函数，然后运用优化算法在输入空间中搜索最佳对抗扰动，使得经过扰动后的样本最大化地误导模型。

3) 基于生成式对抗网络（Generative Adversarial Network, GAN）的攻击算法[35, 36]：这种策略利用 GAN 架构中生成器与鉴别器之间的动态博弈，不断更新对抗样本直至无法被判别器正确识别，从而生成高质量且难以察觉的对抗扰动。

1.3 国内外研究现状

1.3.1 白盒对抗攻击

白盒对抗攻击模式构建在一种极端但具有研究意义的假设之上：攻击者对目标模型的架构和参数了如指掌，并以此作为武器，精密设计出能够误导模型产生错误预测的对抗样本。通过模拟拥有全面知识的攻击者，白盒对抗攻击揭示了模型在已知结构和参数条件下的潜在弱点，促进了对抗样本生成技术的发展，并为评估与提升模型防御能力提供了关键依据。

快速梯度符号攻击[18]（Fast Gradient Sign Method, FGSM），是最简单的基于梯度的单步式对抗样本攻击算法，通过直接利用目标模型损失函数关于输入图像的梯度信息调整图像信息，如式(1.1)所示：

$$\boldsymbol{x}' = \boldsymbol{x} + \epsilon \cdot sign\left(\nabla_x \mathcal{L}\left(\boldsymbol{x}, \mathrm{y}; \boldsymbol{\theta}\right)\right) \tag{1.1}$$

其中 $\boldsymbol{x}$ 与 $\boldsymbol{x}'$ 分别为原始图像与对抗样本，$\boldsymbol{\theta}$ 为模型参数，$\mathcal{L}\left(\boldsymbol{x}, \mathrm{y}; \boldsymbol{\theta}\right)$ 为损失函数，ϵ 控制扰动强度，$sign(*)$ 为符号函数。FGSM 算法结构简单，在保持视觉上的微小变化前提下，高效生成足以欺骗模型的对抗样本。迭代快速梯度符号

攻击（Iterative Gradient Sign Method, I-FGSM）由 Kurakin 等人[13]在 2016 年提出，又称基本迭代方法，是 FGSM 的一种变体攻击算法。I-FGSM 沿着梯度增加方向进行小步迭代，逐步构建起扰动，如式(1.2)所示：

$$\boldsymbol{x}_0^{adv} = \boldsymbol{x},\ \ \boldsymbol{x}_{N+1}^{adv} = clip_{\boldsymbol{x},\epsilon}\left\{\boldsymbol{x}_N^{adv} + \alpha \cdot sign\left(\nabla_{\boldsymbol{x}}\mathcal{L}\left(\boldsymbol{x}, \mathrm{y}; \boldsymbol{\theta}\right)\right)\right\} \tag{1.2}$$

其中 $\boldsymbol{x}$ 经过多次迭代后得到对抗样本 $\boldsymbol{x}_{N+1}^{adv}$，并通过 $clip(*)$ 函数将扰动强度限制到 ϵ 范围内。由于 I-FGSM 在每一步迭代中都计算了梯度，因此在相同的步长条件下，它比 FGSM 有更强的攻击表现。迭代投影梯度下降[21]（Projected Gradient Descent, PGD）是 I-FGSM 的扩展版本，其有更多的迭代次数，并且与 I-FGSM 直接在制定的范围内约束扰动大小不同，PGD 使用 L_∞ 范数映射予以替代。

另外，PGD 算法采用最大最小化策略，分两部分提升模型对抗性攻击的鲁棒性。其中内部最大化旨在在限定扰动范围内，寻找能最大程度误导模型的对抗样本。而外部最小化使用对抗样本进行训练，优化模型参数，以尽可能地降低数据分布上损失的期望。Deepfool 算法[22]则从另一种视角出发，采用线性近似策略来揭示并突破模型决策边界的脆弱性。该算法逐步逼近样本到其最近决策边界的距离，并据此逐步调整输入，直至模型误判。相较于依赖梯度的攻击方法，Deepfool 展示了更强的鲁棒性和普适性，但这种线性优化的攻击方式计算复杂度高，并且可能陷入局部最优解，在处理高度非线性模型时效果有限。

最后，由 Carlini[23]在 2017 年提出的 C&W 攻击算法以其独特的双目标优化策略备受关注。该算法旨在最小化对抗样本与原始样本之间的差异，同时最大化模型对对抗样本的置信度，这一过程通过一个精心构造的约束优化问题得以实现。C&W 攻击成功实现了生成高质量、高隐蔽性的对抗样本。

白盒攻击虽然能够精准构造对抗样本，但在实际场景中攻击者往往是盲目的，过分强调白盒防御策略，可能使模型在真实世界黑盒环境下防御效果有限。

1.3.2 黑盒对抗攻击

在真实世界中黑盒对抗攻击更具有现实意义。黑盒场景假定攻击者只知道目标模型的输出，例如预测或置信度。黑盒攻击包括查询攻击和可迁移攻击。

1.3.2.1 查询攻击

查询攻击是一种黑盒攻击手段，它通过反复查询目标模型，并根据模型反馈逐步调整输入数据以生成对抗样本。根据目标黑盒模型的反馈信息的类型，基于查询的攻击策略可以分为两类：基于分数的攻击和基于边界的攻击。

基于分数的攻击方法依赖于模型提供的确切分数或概率反馈，攻击者据此精心设计出能够误导模型的对抗样本。Chen 等人[24]提出了零阶梯度估计算法,在不掌握目标模型参数及结构的情况下，这种方法利用对称商差原理来近似估计一阶和二阶梯度，然后采用如 ZOO-Adam 和 ZOO-Newton 的随机坐标下降技术更新对抗样本。然而，这种方法每张图像需要查询目标模型 2p 次（其中 p 为像素总数），对于高分辨率图像意味着极高的查询量。为缓解这一严峻的计算负担，研究探索了通过双线性插值技术来压缩攻击空间维度的路径，并结合了一种优先选取关键坐标进行梯度更新的智能采样策略，优先考虑重要坐标的梯度更新，旨在聚焦于影响预测结果最显著的区域。但此类方法的主要缺点在于所需的大量查询次数，不仅效率较低，而且容易触发目标模型的检测机制。

Su 等人[32]提出的 One-Pixel 攻击仅改动单个像素值，优化像素位置和扰动强度生成对抗样本，尤其适用于低分辨率数据集。但是随着图像分辨率提升，单像素变化的影响会显著减弱，降低了攻击的有效性。此外，该方法存在基于进化算法的固有的局限性，依赖于大范围种群探索和多次迭代来逼近最优解，显著增加计算时间和资源消耗，也限制了实际场景中的即时应用可能。Jia 等人[33]开发了 Adv-watermark 对抗水印生成算法，结合图像水印技术，采用盆地跳跃演化作为全局搜索策略定位水印的位置和透明度，进而生成对抗样本。

基于边界的攻击策略利用模型的查询接口寻找模型决策边界的脆弱点，即那些微小改变输入就能导致模型预测变化的地方。这涉及到使用优化算法来最小化输入数据与真实类别之间的距离，同时最大化模型对其的误判概率。Brendel 等人[34]首次提出了边界攻击，这是一种高效查询攻击方法。对于非定向攻击，初始的对抗样本是在图像的有效像素值区间内，按照最大熵分布进行随机采样来获得的，定向攻击则选用目标类别的图像作为起始点。在每次迭代中生成的扰动需满足两个关键条件：首先，确保对抗样本的像素值维持在有效的

像素值范围内；其次，保证扰动幅度与原始样本和对抗样本之间的距离成比例，以减少两者之间的差异，最终在逐步逼近原始样本的过程中形成有效的对抗样本。尽管边界攻击开创了基于边界的攻击先河，且能突破如防御性蒸馏等防御措施，但其明显的不足之处在于需要海量查询，且无法确保算法的收敛。

1.3.2.2 可迁移攻击

如图 1.3 所示，目前的主流黑盒对抗攻击通常假定攻击者没有查询目标模型的权限，与查询攻击相比，这种方法[13, 25-28]更符合实际的攻击情况，并且更难以被目标系统检测到。由于不同模型在处理相同数据点时往往会学习到相似的决策边界，这导致一个模型上生成的对抗样本往往能有效欺骗在相同数据集上训练的其他模型。基于这一现象，当前的研究大多致力于提高对抗样本的可迁移性，利用一个替代模型生成的对抗样本来攻击目标模型[37, 38]。

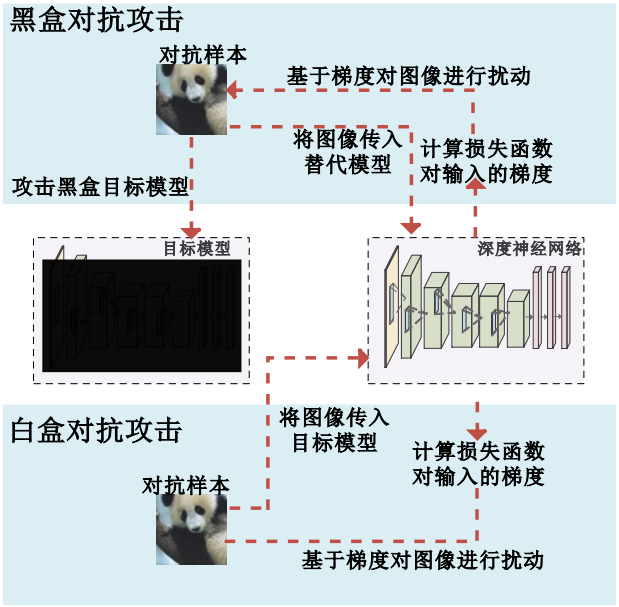


图 1.3 白盒对抗攻击与黑盒对抗攻击示意图

近期的研究提出，使用生成模型生成对抗样本。Pourseed 等人[39]提出了一种生成模型的训练方案，要求最小化替代模型和生成的对抗样本之间的交叉熵损失，以生成对抗扰动。Naseer 等人[40]发现，相对交叉熵损失可以用来提高生成模型的性能。此外，Naseer 等人[41]提出将生成训练过程中扰动图像的分布与替代模型的潜在空间中的目标类对齐。该方法的目的是减少对来自代理模型的

类边界信息的依赖，从而在目标攻击场景中获得令人满意的性能。然而，这种方法由于需要为每个单独的类训练专用生成器，因此不适用大规模数据集。

在各种黑盒攻击方法中，基于I-FGSM的攻击致力于提升对抗性样本的可迁移性，是最有效的攻击之一。Kurakin 等人[13]通过引入基本迭代方法（Basic Iterative Methods, BIM）改进 I-FGSM，增强了对抗样本的可迁移性。I-FGSM 生成的对抗样本容易过拟合于局部极值，从而影响样本的可迁移性。为此，Dong 等人[25]提出了动量迭代快速梯度符号法（Momentum Iterative Fast Gradient Sign Method, MI-FGSM），将动量的概念融入到 I-FGSM 方法中。MI-FGSM 方法稳定了梯度更新方向，有效穿过局部极值，进一步增强了对抗样本可迁移性。

除考虑更好的优化算法外，模型增强也是一种强有力的策略。Xie 等人[25]通过利用图像转换技术解决了 I-FGSM 方法的过拟合问题，并将其命名为多样化迭代快速梯度符号法（Diverse Inputs Iterative Fast Gradient Sign Method, DI-FGSM）。为了缓解对替代模型的过度依赖问题，Dong 等人[27]对输入进行了平移，创建了一系列图像，并近似求解总梯度。Lin 等人[29]利用 DNN 的尺度不变性性质，对不同尺度的图像梯度进行平均，以更新对抗样本。Zou 等人[30]改进了 DI-FGSM，通过生成多尺度梯度，提出了 TI-FGSM[27]（Translation Invariant-FGSM）。Wang 等人[31]考虑了沿着动量优化路径的梯度方差，以避免过拟合。

尽管现有的黑盒攻击在非目标攻击场景中表现出良好的可迁移性，但现有的空间域模型增强方法倾向于采用基于经验的技巧构建增强策略，往往在目标攻击场景下表现不佳，并且缺乏对于 DNN 鲁棒性的研究和可解释性的理论基础。另外，目前黑盒对抗攻击的研究对于样本图像质量的忽视使得可迁移的对抗样本往往带有明显的修改痕迹，大大降低了对抗攻击的实用性与现实意义。

1.4 论文研究内容及结构安排

1.4.1 主要研究内容

以往的对抗攻击方法往往关注在空间域构造对抗样本，而往往忽略了其频域特性，并且在可迁移性与视觉保真度上都存在一定的限制。本文从傅里叶频

域的角度分析了对抗样本的生成过程，从 DNN 的频域鲁棒性入手，探究对抗攻击的频域特性，提出了两种黑盒对抗攻击方法，分别从提升可迁移性与不可见性的角度，从频域上对算法进行设计，在成功提升了对抗样本性能的同时，也为对抗攻击提供了一种新的频域视角。本文的主要研究工作如下：

1) 针对可迁移的黑盒对抗攻击，现有的工作通常通过随机的损失保持变换来增强替代模型，缺乏理论基础，并且在目标攻击场景下表现不佳。本文基于 DNN 频域鲁棒性提出了一种可迁移黑盒对抗攻击方法。首先，本文通过傅立叶热图分析 DNN 的频域鲁棒性，并将多个模型的傅立叶热图进行集成，识别 DNN 的共同脆弱频域。随后，在每次扰动的迭代过程中进行频域模型增强，对模型输入在脆弱频域的信息进行变换以模拟黑盒模型特征，并对梯度进行增强，使对抗扰动的频域分布契合 DNN 频域敏感性特征。最后，本文提出将对抗攻击的优化目标进行分解，通过三元损失来调整扰动的方向，从而使对抗样本跨越更多可能的黑盒模型的决策边界。在常用数据集 CIFAR10、CIFAR100 和 Tiny-ImageNet 上的实验证明，与现有的黑盒对抗攻击相比，本章提出的算法在目标攻击和非目标攻击场景下，迁移攻击成功率均表现优越，并且在面对防御模型时仍表现出良好的攻击能力。

2) 主流的可迁移对抗攻击往往会在生成的样本中引入明显的伪影，使这些攻击在现实世界中的实用性大打折扣。为了解决对抗扰动不可见性的问题，本文中提出了一种不可见的黑盒对抗攻击方法，从频域构造扰动，显著提高了对抗攻击的不可见性。本文从 DNN 的频域鲁棒性特征出发，将其与人类视觉系统（Human Visual System, HVS）的频域特点相结合，量化了扰动的不同频率分量对于对抗样本的可迁移性和视觉保真度的影响，并通过 K-means 聚类算法得到一组 HVS 对于该频率信息变化不敏感，同时处在 DNN 共同脆弱频域的候选频率分量。在生成对抗样本的过程中，本文提出了一种优化联合损失函数，在攻击的同时将扰动的频率分布向候选频率分量约束。常用数据集 CIFAR10、CIFAR100 和 Tiny-ImageNet 上的实验结果表明，所提出的方法在不可见性上明显优于现有的可迁移黑盒对抗攻击方法。

1.4.2 论文结构安排

本文各章内容安排如下：

第一章首先阐述了对抗攻击领域的研究背景及其重要意义，并对对抗攻击的分类进行了详细说明。其次，回顾了国内外在对抗攻击领域的研究进展，从而引出了本文的研究目的和价值。最后，概述了本文将要探讨的主要内容，并介绍了全文的结构安排。

第二章介绍了对抗攻防的相关技术基础。首先介绍对抗攻击的框架和常用评价指标，说明对抗攻击防御方法的研究现状，并由对抗攻防领域对 DNN 鲁棒性研究，引出目前对于对抗攻击原理，从线性假说与频域原理两个方向的介绍。

第三章提出了基于 DNN 频域鲁棒性的可迁移黑盒对抗攻击。该方法首先通过傅立叶热图可视化 DNN 对于不同频率噪声的敏感程度，分析 DNN 的频域鲁棒性，得到 DNN 的共同敏感频域，基于此敏感频域构造模型增强策略，使对抗扰动的频域分布契合 DNN 频域敏感性特征。并提出了三元损失函数，优化攻击过程中扰动的方向性。实验表明，在目标攻击与非目标攻击场景下，该算法在正常模型与防御模型上的可迁移攻击表现，都领先于现有的黑盒对抗攻击。

第四章提出了从频域角度设计对抗扰动的黑盒不可见对抗攻击。首先介绍了高质量对抗样本的应用场景，并且分析现有对抗攻击在不可见性方面存在的设计缺陷。从频域角度分析对抗攻击，将梯度计算与迭代更新放到频域中完成，提出了新的对抗攻击优化目标，在攻击过程中同步将扰动的频域分布向候选频域约束。实验表明，所提出的算法不可见性明显优于现有的黑盒对抗攻击。

第五章对本文研究内容进行总结，并对下一步研究进行展望。

1.5 本章小结

本章首先介绍了课题的背景和研究意义，指出对抗攻击研究的重要性和必要性。其次，对现有的对抗攻击方法进行了分类，并针对图像分类领域的对抗攻击的国内外研究现状进行介绍与分析，引出本文所要研究的主要内容。最后对本文的结构安排进行了说明。

第二章 对抗攻防相关技术基础

本章将主要介绍对抗攻击的基本框架，对抗攻击的常用评价体系，并通过对于经典对抗防御方法的介绍，引出对抗攻击原理的研究。

2.1 对抗攻击基本框架

对抗攻击是通过向神经网络模型的输入数据添加微小、难以察觉的扰动，以欺骗模型输出错误结果的技术。其流程主要包括：确定目标模型与原始样本，设定攻击目标（如非目标攻击场景下的分类错误和目标攻击场景下的误导类别预测），利用优化算法根据模型梯度计算出能最小化目标函数的扰动，并在保持原始样本相似性的约束条件下生成对抗样本。对于这一过程的不断优化，揭示了 DNN 的弱点并推动对抗攻击技术的研究与发展。

对于一个神经网络模型 $\mathcal{M}:\mathcal{X}\to\mathcal{Y}$，它将原始域 $\mathcal{X}$ 映射到目标域 $\mathcal{Y}$ 上。对于图像分类模型，即模型的输入域 $\mathcal{X}$ 为经过预处理后的图像数据，而输出域 $\mathcal{Y}$ 则代表该数据集的所有图像类别。在多类别分类问题中，模型的输出层通常会使用 softmax[42]函数，再经过 argmax 后得到一个离散的类别标签作为最终输出。使用 $Z(\boldsymbol{x})$ 表示模型的输出层激活，则对于输入图像 $\boldsymbol{x}\in\boldsymbol{R}^{C\times H\times W}$（其中 C、H、W 为图像的通道数，高和宽）和它的标签 y_s，模型的分类过程如式(2.1)：

$$\mathcal{M}(\boldsymbol{x})=arg\,\max_{y}\left(Z(\boldsymbol{x})_y\right)=y_s \tag{2.1}$$

其中 $Z(\boldsymbol{x})$ 是一个概率分布，其中的每个元素表示图像属于相应类别的概率。攻击者的目标是构造一个扰动 $\boldsymbol{\delta}$，使对抗样本 $\boldsymbol{x}'=\boldsymbol{x}+\boldsymbol{\delta}$ 能成功误导目标模型，即 $\mathcal{M}(\boldsymbol{x}')\neq\mathcal{M}(\boldsymbol{x})$。

如图 2.1，对抗攻击的目标是一个已经训练完成的模型，在白盒场景下，对抗攻击直接在目标模型上生成对抗样本。在黑盒场景下，攻击者无法获取黑盒模型的任何信息，为了成功实现黑盒攻击，一种常见的策略是利用替代模型生成对抗样本，然后利用这些样本的可迁移性来对黑盒模型进行攻击[25]。

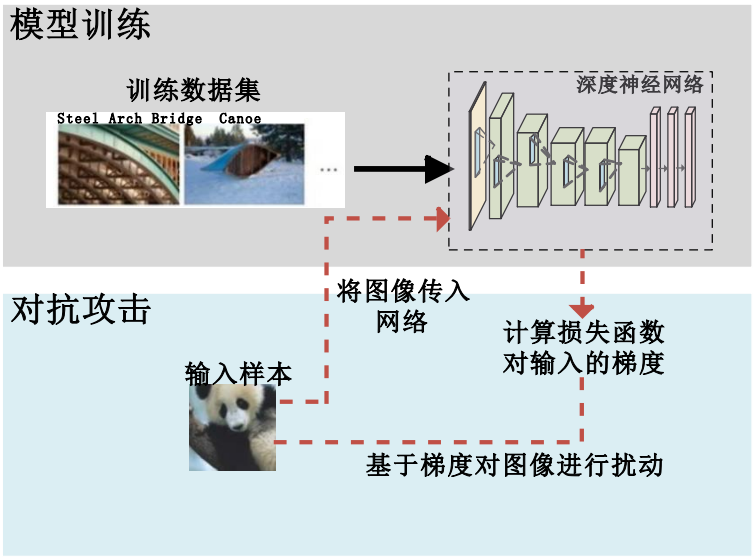


图 2.1 模型训练（上半部分）与对抗攻击（下半部分）的示意图

将分类网络 $\mathcal{M}_\theta$ 作为替代模型，并将与 $\mathcal{M}_\theta$ 独立，且网络结构不同的一系列神经网络模型 $\mathcal{M}_1,\mathcal{M}_2,\ldots,\mathcal{M}_B$ 作为黑盒模型。在目标攻击场景[75]下，对给定的原始样本 $\boldsymbol{x}$ 和目标类别标签 $\mathrm{y_t}$，可迁移的黑盒对抗攻击的目标是在替代模型 $\mathcal{M}_\theta$ 上构造对抗样本 $\boldsymbol{x}'$，使得对抗样本能够被黑盒模型 $\mathcal{M}_\mathrm{i},1\le \mathrm{i}\le \mathrm{B}$ 误分类为 $\mathrm{y_t}$，也即 $\mathcal{M}_i(\boldsymbol{x}')=\mathrm{y_t}$。在非目标攻击场景[25]下，对于原始样本 $\boldsymbol{x}$，和它的原始 $\mathrm{y_s}$，对抗攻击的目标是使得对抗样本 $\boldsymbol{x}'$ 被黑盒模型错误分类，即 $\mathcal{M}_i(\boldsymbol{x}')\neq \mathrm{y_s}$。

在替代模型 $\mathcal{M}_\theta$ 上，对抗攻击可以表示为下面的优化问题，如式(2.2)：

$$\boldsymbol{x}'=arg\max_{\boldsymbol{x}'}\mathcal{L}(\boldsymbol{x}',y;\boldsymbol{\theta}),s.t.\|\boldsymbol{\delta}\|_p\le\epsilon \tag{2.2}$$

其中 ϵ 为对抗样本与原始图像的最大 L_p 范数限制，而 $\mathcal{L}(\boldsymbol{x},y;\boldsymbol{\theta})$ 为损失函数，现有的对抗攻击方法通常采用交叉熵损失。在对抗攻击的过程中，L_p 范数（$\|\boldsymbol{\delta}\|_p$）代表两个矩阵在 L_p 空间内的距离，因此被用来测量对抗样本与其原始图像的不相同程度，其中 $\|\boldsymbol{\delta}\|_p$ 的定义为式(2.3)：

$$\boldsymbol{\delta}_p=\left(\sum_{i=1}^{n}|\boldsymbol{x}_i|^p\right)^{\frac{1}{p}} \tag{2.3}$$

攻击者通过 L_p 范数来限制扰动的强度，从而控制扰动带来的视觉差距。现有的对抗攻击通常在像素域上采用 L_0 [32]、L_2 [21, 23]或 L_∞ [13, 21, 23]范数对扰动进行约束。

2.2 对抗攻击的评价指标

对抗攻击的评价通常涉及多个关键指标，用以衡量对抗样本的有效性和生成质量。在对抗攻击研究的深度和广度上，多维指标体系不仅提供了严谨的评估框架，而且为理解和对抗这种安全威胁提供了关键的理论基础[23, 43-46]。以下是一些常用的评价指标：

1) 攻击成功率（Attack Success Rate）：指攻击者生成的对抗样本能够成功误导目标模型的比例，即对抗样本被误分类的频率。

2) 扰动度量（Distortion Metric）：衡量原始样本与对抗样本之间的差异程度，常用指标包括 L_p 范数（如 L_0 、L_2 或 L_∞ ），以及感知哈希距离（Perceptual Hashing）等，这些指标量化了添加到原始样本上的扰动的强度大小。

3) 可迁移性（Transferability）：考察一个对抗样本在非目标模型上同样引发误判的能力，即它是否对不同架构或训练状态的模型都具有攻击能力。

4) 视觉保真度（Visual Fidelity）：评估对抗样本在人眼看来与原始样本的相似程度，这对于现实场景下的对抗攻击尤为重要。

对于以上各项指标的综合运用，可以客观评价对抗攻击方法的有效性和局限性，并且可以更全面地理解对抗攻击的效果及潜在威胁。

2.3 对抗攻击防御方法

2.3.1 对抗训练

对抗训练作为一种强化深度学习模型鲁棒性的策略，最初由 Goodfellow 等人[18]提出，他们将对抗样本引入到训练流程中，具体做法是利用 FGSM 生成对抗样本，并将其合并到训练集中，随后对网络模型进行重新训练，以增强模型

对对抗扰动的抵抗能力。Madry 等人[21]在此基础上进一步发展了对抗训练的理论与实践，他们倡导采用更强力的投影梯度下降攻击算法来生成对抗样本，并在训练阶段采用了 min-max 优化框架。在这种框架下，模型不仅需要最小化对正常样本的损失，同时也要最大化对对抗样本的泛化性能，从而在训练过程中迫使模型在面对最强大的对手，精心设计的对抗样本时仍能保持准确预测。

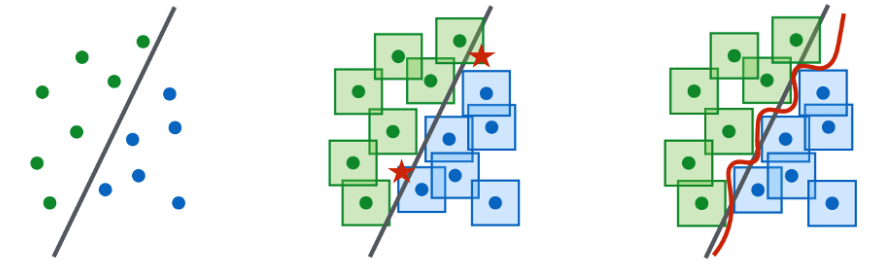


图 2.2 对抗训练使模型得到鲁棒决策边界[21]

如图 2.2 所示，Madry 等人[21]设计对抗训练的核心理念在于，通过将对抗样本集成到常规训练集中，让神经网络在处理这些经过有意扭曲但仍保留原始标签的样本时，能够更好地模拟真实世界中潜在的干扰情况。网络在学习过程中，会尝试弥合对抗样本与干净样本在决策边界上的差距，扩大模型对输入空间的稳健覆盖，实现对对抗攻击的有效防御。

生成式对抗网络方法代表了一种创新思路，通过构建生成器与判别器的动态博弈过程，生成具有高度多样性的对抗样本，进而训练出更为健壮的模型。Xiao 等人[35]研究成果揭示了 GAN 在生成高保真对抗样本和促进模型鲁棒性方面的潜力，但同时也面临模式坍塌和训练稳定性的问题。

在实际操作中，对抗训练将对抗样本产生的额外损失纳入总体损失函数，以此作为一种正则化手段，在不改变原有模型结构的前提下，增强了模型对未见过的对抗扰动的适应性和稳定性[47-50]。正则化方法通过在损失函数中融入正则项，约束模型参数或梯度，以达到提升模型泛化能力和对抗鲁棒性的目的。例如，L_1、L_2 正则化以及投影梯度削减等策略[51]，能够在理论层面保证鲁棒性和准确性之间的权衡，但实践中寻找最优正则化强度仍是一大挑战。

然而，值得注意的是，尽管对抗性训练在提升模型稳健性方面表现出色，但其计算成本及可能引入的过拟合风险仍需谨慎考量。对抗训练通常伴随着较高的计算成本[46]，并且由于对抗攻击方法的持续演进，单纯依赖对抗训练可能难以始终保持模型对所有新型攻击方法的高鲁棒性。这意味着对抗训练是一个动态发展的领域，需要不断地跟进和优化以应对日益复杂的攻击手段。

2.3.2 对抗检测

除了直接增强模型鲁棒性，研究者也探索了专门的对抗检测技术。对抗检测是指用于识别输入数据是否受到对抗攻击的技术，这种方法利用基于统计特征、异常检测或其他机制的方法来检测对抗样本，并丢弃或修复潜在的对抗样本，对于保护机器学习模型免受对抗样本的影响具有重要意义。

对抗训练作为主动防御机制，启发了对抗检测模型的构建。Xiao 等人[52]从权重空间几何角度解析了对抗训练的鲁棒性原理，Dhillon 等人[53]提出了随机激活剪枝策略，提升了防御的效率与灵活性。

基于统计特征的检测方法聚焦于从输入数据的内在统计属性中挖掘对抗性扰动的痕迹。Feinman 等人[54]通过分析输入样本的梯度分布特征来识别对抗样本，而 Xiao 等人[52]则利用空间一致性信息的统计特性来标记对抗样本。此类方法直观且易于实施，但可能因正常数据的多样性而产生较高误报率。

基于异常检测的技术通过对比输入数据与正常分布的偏差来揭示潜在的对抗攻击行为。Metzen 等人[55]展示了基于密度的异常检测在对抗扰动识别中的潜力，Xie 等人[56]则采用特征去噪策略，增强模型对异常输入的辨识能力。这些方法优势在于能处理非显式的对抗攻击，但对异常定义的精确性有较高要求。

另外，基于模型不确定性的策略利用深度学习模型的内在不确定性评估来区分对抗性与良性输入。Gal 等人[57]探讨了深度学习中的不确定性概念，为利用预测置信度作为对抗指标提供了理论基础，Shen 等人[58]通过简单词嵌入模型展示了基本方法在增强模型鲁棒性方面的潜力。这种方法论强调了理解模型决策过程的重要性，但其有效性受模型复杂度及训练数据质量的影响。

2.3.3 数据预处理

输入图像预处理在对抗防御领域占据关键地位，其作用在于通过系统性地调整输入数据的形态或内容特征，可以有效地增强深度学习模型对于对抗性攻击的鲁棒性[56-65]。许多工作提出了一系列创新性的预处理技术，以增强模型的鲁棒性。

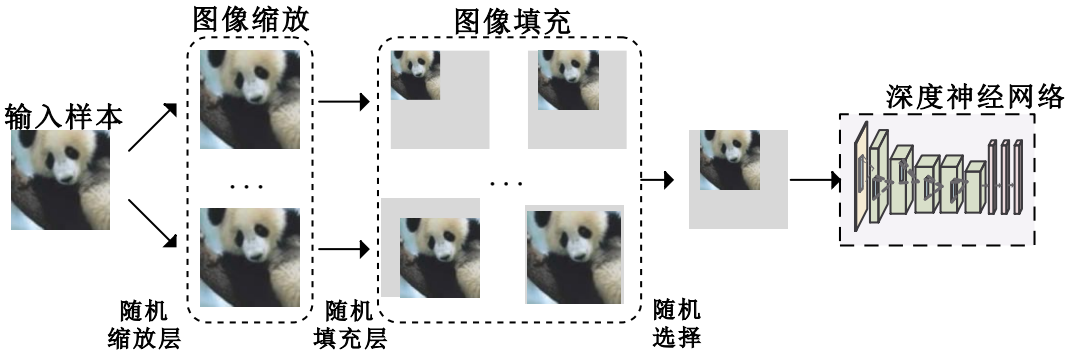


图 2.3 随机化防御机制示意图

Jia 等人[59]引入了一种基于随机平滑机制的策略。研究者借助随机噪声对输入数据进行平滑处理，这种过程有助于模型在对抗微小扰动时维持稳定的预测输出，并且这种方法还提供了鲁棒性的理论保证。另外，Naseer 等人[60]提出了一种自监督学习方案来增强模型的抗对抗性。通过构建特定的自监督任务，模型能够在训练过程中学习到更为稳定、不易受攻击影响的输入特征表示，从而削弱对抗攻击对模型性能的损害。

Guo 等人[61]探讨了如何运用输入图像的几何变换（例如旋转、缩放和裁剪）来对抗对抗攻击，如图 2.3 所示。这类变换可以从根本上改变图像在模型内部的表征形式，进而减少对抗扰动对模型判断的影响，从而提升整体的鲁棒性。Xie 等人[62]提出采用随机化技术来抵御对抗攻击，向输入数据中注入随机噪声或实施随机扰动，能够扰乱攻击者的策略，使其难以找到能够误导模型的对抗样本，由此提高了模型的稳健性。Xie 等人[62]的研究强调了这一策略在提高模型鲁棒性上的潜力，但过度增强可能导致训练成本上升或引入无关特征的过拟合问题。值得一提的是，由于几乎每个图像分类数据集都是由 JPG 图像组成的，Dziugaite 等人[63]发现，对于基于 I-FGSM 的对抗攻击算法，如图 2.4 所示，JPG 压缩经常在很大程度上降低对抗样本对于目标模型分类精度的影响。这是由于

尽管对抗样本和原始样本之间的扰动非常小，但是在图像分类模型的高层表示空间，扰动被放大了。对于图像的压缩与重建，抑制了对抗扰动的影响[64]。

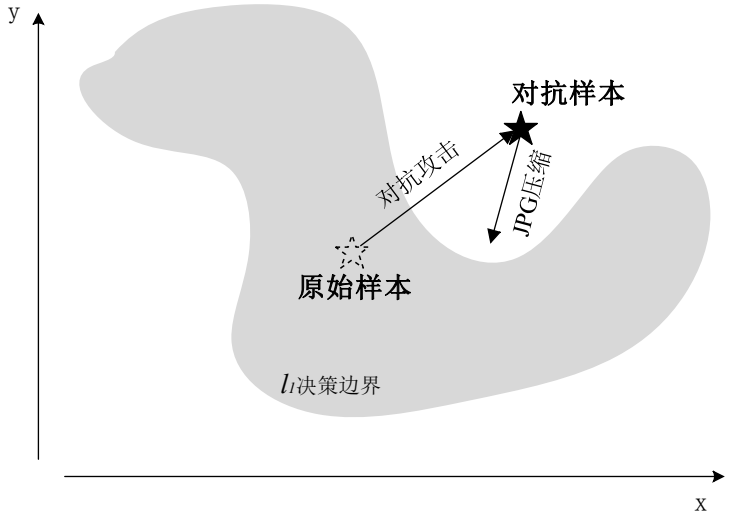


图 2.4 JPG 压缩将对抗样本带回决策边界内

Liao 等人[65]提出了一种新颖的防御机制，即基于高层表示指导的去噪方法。这种方法利用模型自身的高层抽象特征来指导对输入数据噪声的消除过程，有效地降低了对抗攻击的负面影响，提升了模型的鲁棒性。

特征选择方法通过精炼输入数据中的关键特征，减少无关或冗余信息对模型判断的干扰，以此增强模型的抗攻击能力。Xu 等人[56]的研究展示了基于特征去噪和稀疏流形限制的防御策略，虽能有效提升模型的针对性和效率，但也需谨慎平衡特征选择与模型表现之间的关系。

2.4 对抗攻击的可解释性

第 1.3 与 2.3 节中阐述了目前对抗攻击与对抗防御的研究现状，这些方法往往不同程度上揭示了神经网络的鲁棒性弱点。研究者试图从对抗攻击与防御两个维度，探索什么样的方法更行之有效，但对于提出方法有效性的论证，往往会牵涉到对于 DNN 原理的探索，而神经网络为什么易受攻击的问题是一个不可避免的问题。因此在对抗攻防的研究过程中，研究者思考了机器学习模型受到对抗攻击影响的原因以及对抗攻击的基本原理，以深入研究对抗攻击的本质。

2.4.1 对抗攻击的线性解释

对抗样本的成因是对抗攻击领域研究的一个重要前提，而其至今没有一个确切的定论。Goodfellow 等人[18]在 ICLR2014 会议上首次提出了对抗攻击（Adversarial Attack）一词，并提出对抗攻击的本质是通过叠加扰动，使样本跨越决策边界，从而造成分类错误。Goodfellow 等人在论文中提出了一个直观且深刻的见解，他们揭示了即使在高度非线性的深度神经网络中，对抗样本的生成也可以通过近似线性的方式来理解。

尽管神经网络的设计中引入了非线性激活函数（Sigmoid、ReLU、Tanh 等）以打破线性关系，使得神经网络能够在学习的过程中能够捕捉到数据中的非线性模式和特征，从而提高网络的表达能力和学习能力。但由于神经网络内含大量基础线性模块，模型在高维空间中存在的局部线性特征仍导致了 DNN 的脆弱性[18]。因此尽管模型整体是复杂的非线性函数，但在局部输入空间区域内，模型的行为可以近似为线性模型。神经网络的激活表示网络在接收到输入数据后计算出的输出值，则对于一个线性模型，其激活即为模型的权重向量 $\boldsymbol{\varphi}^T$ 和对抗样本 $\boldsymbol{x}'$ 的乘积，如式(2.4)：

$$\boldsymbol{\varphi}^T \boldsymbol{x}' = \boldsymbol{\varphi}^T \boldsymbol{x} + \boldsymbol{\varphi}^T \boldsymbol{\delta} \tag{2.4}$$

其中，对抗扰动对于模型产生的影响量化为激活变化 $\boldsymbol{\varphi}^T \boldsymbol{\delta}$。由于 $\boldsymbol{\delta}_p$ 不随扰动的维数而增长，但扰动引起的激活变化 $\boldsymbol{\varphi}^T \boldsymbol{\delta}$ 可以随权重向量 $\boldsymbol{\varphi}$ 的维度而线性增长。因此对于高维问题，例如输入为图像的模型，对于输入进行的许多无穷小的扰动，加起来可以对模型输出产生很强的对抗攻击效果。

从线性特性更进一步地，从数据流形的角度出发，研究者们[12]提出了新的见解。他们认为，由于训练数据的局限性，深度神经网络（DNN）往往只能捕捉到目标数据流形的局部特征。这导致了在数据流形的低概率区域，模型的决策边界无法做出准确的划分，从而成为对抗样本存在的主要原因。基于这一观点，研究者们认为通过数据增强可以有效地提升 DNN 对抗攻击的鲁棒性。与此相似，Samangouei 等人[66]提出，对抗样本往往出现在远离目标数据流形的区域。

此外，利用 GAN[67]来研究对抗攻击的方法，主张通过扩充目标数据集，优化模型的决策边界，以增强模型的鲁棒性。

2.4.2 对抗攻击的频域原理

除了基于 DNN 线性特性和数据流形的理论外，频域角度的对抗样本理论具有更广泛的适用性与对抗样本生成机制更深层次的理解[20-68]，并且基于此理论可以对输入数据的频率成分进行定量分析，提供对于攻击有效性的更好的解释性，为构建全局、可解释和高效鲁棒的模型提供新视角和方法[69]。

Tsuzuku 等人[70]通过研究 DNN 对不同傅立叶基的灵敏度，首次提出了频率框架，以理解其在处理信号和图像时是如何利用频率信息的。Yin 等人[71]随后探索了神经网络在加性噪声方面的频率特性。从频域的角度的研究表明，神经网络的浅层特征通常提取边缘和纹理特征，突出了高频信息对决策的重要性[72]。

Luo 等人[68]通过分析 DNN 的频谱响应和频谱偏差，对深度学习模型的训练行为做了深刻的频域视角解析，揭示了一项关键发现，即在 DNN 的学习过程中，模型倾向于优先捕获和拟合输入数据中的低频成分，这一规律被定义为“频率原理”。Luo T 等人[68]做了大量的数学推导来证明频率原理，并将其分为了训练的初始、中间和结束阶段，分别证明了训练的初始阶段、中间阶段和结束阶段的频率原理。这一原理不仅解释了深度学习模型在处理自然图像等复杂数据时的基础特性，也揭示了 DNN 在处理高维空间中的输入时，对于特定频率成分的敏感性和非线性响应机制，表明 DNN 在决策过程中对输入不同频率的信息表现出显著的偏向，为理解神经网络的频域鲁棒性与对抗攻击的本质提供了思路。

频率原理的提出激发了对于 DNN 频域鲁棒性的研究。Yin 等人[71]发现，自然训练的神经网络模型对除最低频率外的所有加性噪声都高度敏感。对抗训练与高斯数据增强都可以极大地提升高频下的 DNN 鲁棒性，但是会牺牲自然训练模型优秀的低频鲁棒性。王等人[72]指出，DNN 能够捕捉人类几乎察觉不到的数字图像的高频分量。Sharma 等人[73]证明了，经过对抗训练的防御模型，对高频扰动的敏感性降低，然而仍然容易受到低频扰动的影响。

基于大多数 DNN 较弱的高频鲁棒性，通过在原始数据样本上添加微小且精心设计的高频扰动，可以使模型在输入图像的视觉上几乎无明显变化的情况下

发生分类错误或预测偏差。Zhang Q 等人[73]提出了一种无盒的对抗攻击方法，该方法通过混合图像变换引入小但有效的高频扰动，而不改变图像的语义信息。作者证明，混合图像变换可以有效地欺骗多个目标模型和检测器，计算成本低，成功率高。另外，许多对抗攻击方法基于模型对高频信号的依赖性，利用高频扰动来误导深度神经网络[72]，并且得出结论：对抗攻击往往通过对原始输入添加人眼难以察觉但对模型影响显著的高频噪声来误导模型输出。

然而，最近的研究表明，对抗样本是高频扰动的结论是不正确的[17, 71]。具体来说，Maiya 等人[17]提出了一种基于频率分析的方法来量化分析 DNN 的鲁棒性，表明了对抗样本不仅仅是一种高频现象。作者首先定义了一个新的频率灵敏度指数来测量模型对不同频率范围内扰动的灵敏度。分析表明，对抗样本既不是简单的高频现象，也不是低频现象。对抗样本的性能不仅依赖于其频率特性的高低，还应考虑目标模型所用数据集以及训练方法等的影响。

2.5 本章小结

本章介绍了对抗攻击的基本框架与常用评价体系，说明了对抗防御方法的研究现状，介绍了几个典型对抗防御方法，阐述了各防御方法的原理与思想。并介绍了目前对于对抗攻击原理从线性假说与频域原理两个方向的研究。

第三章 基于频域鲁棒性的可迁移对抗攻击

现有的可迁移黑盒对抗攻击，受限于基于经验技巧的随机模型增强策略，可迁移性有限，尤其在目标攻击场景下表现不佳，并且由于缺乏可解释性的理论基础，难以迭代改进。本章基于对抗攻击的频域原理研究，通过 DNN 对于不同频率上输入信号的响应情况，分析模型在处理不同频率信息时的鲁棒性，并基于此特性构造频域增强策略。另外，本章将对抗样本的优化目标分解，提出使用三元损失作为损失函数，以增强对抗扰动的方向性。在提升攻击性能的同时，分析了对抗样本内在的频域特性，为未来对抗攻击的发展提供了参考依据。

3.1 研究动机

对抗攻击的本质，是在神经网络模型决策的过程中，通过叠加扰动使样本跨越模型的决策边界，从而造成模型的误判。而研究者们一直在对神经网络内部的可解释性进行研究，探究 DNN 泛化能力的来源。因此，要提升样本的可迁移性，就需要结合神经网络的可解释性研究，通过不同神经网络模型之间的内在规律找到 DNN 的公共决策边界。

频率原理从频域上对于神经网络学习过程内在原理的揭示，发现了在 DNN 在频域行为上的共同特性，在同一数据集上，不同模型的模型对傅立叶频域上某些特定频率的噪声都更敏感。这种频域的鲁棒特性与神经网络的泛化能力一样，由数据集的特性以及训练过程中网络的结构共同决定，是 DNN 的一种固有特性，针对这一特性设计扰动，就可能提升对抗样本的可迁移性。

同一数据集中的神经网络对傅立叶频域上某些特定频率的噪声更敏感，这启发了将 DNN 的频域鲁棒性作为共同特征[76]，从频域的角度对于对抗攻击的研究。在输入图像的傅里叶域中叠加扰动，并测试 DNN 对这一变化的反应[71]，可以得出 DNN 对于傅里叶域中不同频率噪声的敏感性，也即 DNN 对于噪声的频域鲁棒性。首先给出快速傅里叶变换[77]（Fast Fourier Transform）和其反变换（Inverse Fast Fourier Transform）的公式，如式(3.1)和(3.2):

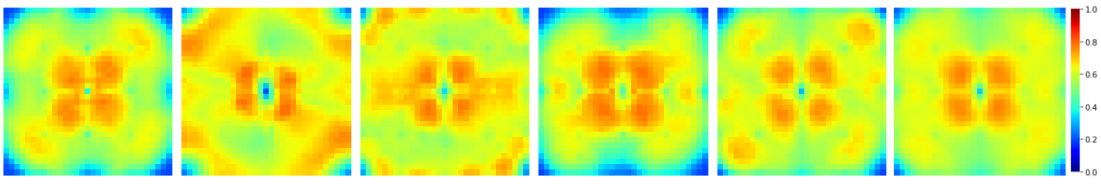
$$\boldsymbol{x}_f(u,v) = fft\left(\boldsymbol{x}(h,w)\right) = \frac{1}{H \times W}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\boldsymbol{x}(h,w)e^{-j2\pi(uh/H+vw/W)} \tag{3.1}$$

$$\boldsymbol{x}(h,w) = ifft\left(\boldsymbol{x}_f(u,v)\right) = \sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\boldsymbol{x}_f(u,v)e^{j2\pi(uh/H+vw/W)} \tag{3.2}$$

其中(u,v)为频域元素坐标，而(h,w)为空间域像素坐标。在傅里叶域中，傅里叶矩阵中的每个元素都代表一种频域分量，使用(i,j)来表示元素的坐标。在每个频域分量上分别构造单位扰动，使其傅里叶矩阵仅在(i,j)以及其相对于图像中心的对称坐标存在两个非零元素，得到傅里叶基噪声$\boldsymbol{U}_{i,j} \in \boldsymbol{R}^{H\times W}$。每个傅里叶基噪声仅包含一种频域信息，在空域上表现为一种平面波，频谱中心化之后，每个元素到中心点的距离描述了平面波的频率，它到中心点的方向代表平面波的方向，而该元素的值代表了平面波的幅度，本章在傅里叶矩阵的实部矩阵进行修改，而保留其虚部矩阵的相位信息。控制每个傅里叶基噪声的强度不变，即$\|\boldsymbol{U}_{i,j}\|=1$，将其通过从$\{-1,1\}$随机选择的系数$r$叠加到 RGB 三个通道上。

通过在测试集上添加不同的傅里叶基噪声，可以建立模型在加噪测试集上的分类错误率与噪声的频域信息(i,j)之间的函数。将DNN在该频率分量的噪声下的测试误差，作为热度值来建立热图并可视化函数关系。得到 DNN 平均测试误差与噪声频率的对应关系后，作为模型的傅里叶热图[71]。热图直观反映了DNN 对于不同频率分量噪声的敏感性，可视化了 DNN 的频域鲁棒性。

在常见的图像分类数据集 CIFAR10[78]、CIFAR100[78]和 Tiny-ImageNet[79]上，使用常用的神经网络结构进行训练，在每个数据集上得到一系列神经网络模型$\mathcal{M}_1,\mathcal{M}_2,\ldots,\mathcal{M}_p$，以及它们的傅里叶热图。以 CIFAR10 数据集上的模型为例，如图 3.1 为不同 DNN 的傅里叶热图，热度值在(0, 1)之间，如图例所示，热度值越大，颜色越红；热度值越小，颜色越蓝。



Resnet-18　VGG-19　DenseNet-121　Resnet-50　Wide-Resnet50　多模型集成

图 3.1 CIFAR10 数据集上，不同神经网络模型的傅里叶热图

每个像素点的热度值代表了模型在对应频域分量噪声下的平均测试误差，图中的红色区域为测试误差最大的区域，也代表模型对于这些频域分量的信息变化最敏感，在该频率上脆弱。显然，不同模型的脆弱区域存在重合。将测试误差最大的点提取出来，提取区域为整个热图的20%，将这些区域的值设为1，其他区域设为 0，通过二值化处理得到了频域脆弱区域的掩码 ω_p。它们代表，在该数据集上的模型对于这些频域分量的脆弱特性在 DNN 中是普遍存在的。

图 3.2 CIFAR10 数据集上，不同神经网络模型的频域脆弱区域掩码

图 3.2 为不同结构的神经网络模型的频域脆弱区域，从中可以看出，不同的 DNN 在频域上存在明显的共同脆弱区域。神经网络的频域鲁棒性是数据集的特征以及网络结构所公共决定的，因此将不同结构的 DNN 的脆弱区域进行集成，减少因网络结构不同和训练过程中随机因素所带来的干扰，就可以得到 DNN 的公共脆弱区域。将模型的脆弱区域集成，去除其中仅有一个模型存在的孤点后归一化，得到公共脆弱区域 $\boldsymbol{\omega}_{all}$。使用 $\boldsymbol{\omega}$ 来表示其 01 掩码，如式(3.3)和(3.4)：

$$\boldsymbol{\omega}_{all} = \sum_{p=1}^{P} \boldsymbol{\omega}_p \left(i, j\right) \tag{3.3}$$

$$\boldsymbol{\omega}_{all}\left(i, j\right) = \begin{cases} 0, & \boldsymbol{\omega}_{all}\left(i, j\right) < 2 \\ 1, & \boldsymbol{\omega}_{all}\left(i, j\right) \geq 2 \end{cases} \tag{3.4}$$

3.2 基于频域鲁棒性的可迁移对抗攻击

3.2.1 总体框架

上述的分析找到了 DNN 在傅里叶频域上的公共脆弱区域 $\boldsymbol{\omega}_{all}$，它代表了 DNN 对于该频域扰动的共同敏感性特征，也即，同一数据集上的 DNN 对于输

入图像在这些频域分量的信息变换都更敏感。当对抗攻击从频域上更关注公共脆弱区域时，就可以对未知的黑盒模型起到更强的攻击效果。基于上述分析，本章提出基于 DNN 频域鲁棒性的对抗攻击算法。

算法 3-1：基于 DNN 频域鲁棒性的可迁移对抗攻击

输入：傅里叶热图掩码 $\boldsymbol{\omega}$；超参数 β, b；源图像 $\boldsymbol{x}$ 和原始标签 $\mathrm{y_s}$（非目标攻击）或目标标签 y_t（目标攻击）；替代模型 $\mathcal{M}_\theta : \mathcal{X} \to \mathcal{Y}$；迭代次数 T；L_∞ 参数限制 ϵ；噪声初始化次数 N

输出：对抗样本 $\boldsymbol{x}'$

初始化：$\boldsymbol{E} = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}$，$\boldsymbol{E} \in \boldsymbol{R}^{H \times W}$，$\boldsymbol{x}'_0 = \boldsymbol{x}$，$\mathrm{y} = \mathrm{y_s}$ 或 $\mathrm{y_t}$

1. **for** $i = 0 \to T-1$ **do**
2. **for** $n = 0 \to N-1$ **do**
3. 对模型输入做傅立叶变换：$\boldsymbol{x}_f = fft(\boldsymbol{x}'_i)$
4. 随机初始化噪声 $\boldsymbol{\xi}$ 和 $\boldsymbol{\mu}$
5. 傅里叶频域增强：$\boldsymbol{x}_{aug} = ifft\left[\left(\boldsymbol{x}_f^i \odot \boldsymbol{\mu} + \boldsymbol{\xi}\right) \odot \left(\boldsymbol{E} + \beta \cdot \boldsymbol{\omega}\right)\right]$
6. 将 $\boldsymbol{x}_{aug}$ 放入模型中，通过反向传播计算梯度：

$\boldsymbol{g}_n = \nabla_{\boldsymbol{x}'_i} \mathcal{L}\left(\boldsymbol{x}_{aug}, \mathrm{y_s}; \boldsymbol{\theta}\right)$(非目标攻击) or $\boldsymbol{g}_n = \nabla_{\boldsymbol{x}'_i} \mathcal{L}\left(\boldsymbol{x}_{aug}, \mathrm{y}_t; \boldsymbol{\theta}\right)$(目标攻击)

7. **end for**
8. 对多次增强得到的梯度进行平均：$\boldsymbol{g}' = \frac{1}{N} \sum_{n=0}^{N-1} \boldsymbol{g}_n$
9. 叠加扰动，更新对抗样本：$\boldsymbol{\delta} = \alpha \cdot sign\left[ifft\left(fft\left(\boldsymbol{g}'\right) \odot \left(b \cdot \boldsymbol{m} + \boldsymbol{E}\right)\right)\right]$)
10. 限制最大扰动幅度：$\boldsymbol{x}'_{i+1} = clip_{\boldsymbol{x},\epsilon}\left(\boldsymbol{x}'_i + \boldsymbol{\delta}\right)$
11. 像素值归一化：$\boldsymbol{x}'_{i+1} = clip\left(\boldsymbol{x}'_{i+1}, 0, 1\right)$
12. **end for**
13. 得到对抗样本 $\boldsymbol{x}' = \boldsymbol{x}'_T$

算法 3-1 中详细介绍了该算法的整体流程，具体的设计细节会在下文中进行

详细解释。该算法主要分为三个阶段。

首先通过对多个 DNN 的傅立叶热图进行集成，并提取出前 20%最高热度值的频域区域，作为公共脆弱区域，并将其转换为二进制 01 掩码。之后对输入样本 $\boldsymbol{x}'_i$ 进行频域增强，对噪声进行 N 次随机初始化，并多次增强得到的梯度 $\boldsymbol{g}_n$ 进行平均，使梯度 $\boldsymbol{g}'$ 的方向更稳定。最后，更新样本 $\boldsymbol{x}'$，并将样本像素值限制在自然图像范围内，即[0, 1]。

3.2.2 频域增强

现有的工作通常通过随机的损失保持变换，来增强替代模型，忽略了模型之间的本质差异，并且缺乏神经网络可解释性的理论基础，生成的对抗扰动方向性差，样本的可迁移性低[25-28]。

本章提出一种基于 DNN 频域敏感特性的增强方法，也即，对模型输入在 DNN 脆弱频域的信息进行变换，使对抗扰动的频域信息能够契合 DNN 共同敏感性特征。

在黑盒设置下，受害模型的傅里叶热图攻击者显然无法得到。因此将对多个模型进行集成得到的，代表 DNN 共同频域脆弱特性的公共区域作为 01 掩码 $\boldsymbol{\omega}$。由线性代数的基础知识可知，通过矩阵变换的形式可以使两个矩阵相等，因此通过矩阵变换运算，可以使替代模型的傅里叶热图模拟任意模型的频域脆弱特性。将输入图像 $\boldsymbol{x}$ 通过傅里叶变化到频域上，并使用 $\boldsymbol{x}_f$ 表示。则在反向传播的过程中，梯度信息 $\boldsymbol{g}=\nabla_x \mathcal{L}\left(\boldsymbol{x}, \mathrm{y}_s ; \boldsymbol{\theta}\right)$。频域增强的公式如式(3.5)所示：

$$\boldsymbol{x}_{aug}=ifft\left[\left(\boldsymbol{x}_f \odot \boldsymbol{\mu}+\boldsymbol{\xi}\right) \odot\left(\boldsymbol{E}+\beta \cdot \boldsymbol{\omega}\right)\right] \tag{3.5}$$

其中 $\boldsymbol{x}_f = fft\left(\boldsymbol{x}\right)$，⊙表示哈达玛积，乘性噪声 $\boldsymbol{\mu}$ 是从均匀分布中采样得到的随机变量；加性噪声 $\boldsymbol{\xi}$ 为对高斯分布进行采样后，再进行傅里叶变换后得到的随机变量。使用掩码 $\boldsymbol{\omega}$ 对它进行滤波，只保留敏感频域的部分，如式(3.6-3.7)所示：

$$\boldsymbol{\xi}= fft\left(\boldsymbol{G} \odot \boldsymbol{\omega}\right) \tag{3.6}$$

$$\boldsymbol{\mu} \sim U\left(1-\upsilon, 1+\upsilon\right) \tag{3.7}$$

其中 $\boldsymbol{G}$ 为随机高斯噪声，$U\left(1-\upsilon, 1+\upsilon\right)$ 为均匀分布，υ 设置为 0.5。对于 $\boldsymbol{x}_f$ 有针

对性的加噪处理，可以干扰对抗扰动对于替代模型的拟合，使得替代模型能够近似模拟一个全新模型的输出，并且，对输入图像信息的变换降低了扰动与输入样本的相关性，使得扰动的频域信息更多地偏向神经网络的频域特性；之后对扰动在敏感区域信息的进一步增强，使得梯度信息在频谱上向频域公共区域偏移，如式(3.8)和(3.9)所示：

$$\boldsymbol{g} = \nabla_{x'}\mathcal{L}\left(\boldsymbol{x}_{aug}, \mathrm{y_s}; \boldsymbol{\theta}\right) \tag{3.8}$$

$$\boldsymbol{\delta} = \alpha \cdot sign\left[ifft\left(fft\left(\boldsymbol{g}\right) \odot \left(b \cdot \boldsymbol{m} + \boldsymbol{E}\right)\right)\right] \tag{3.9}$$

在模型输入中，对于敏感频段信息的强化使得对抗扰动δ更关注频域敏感区域，并且通过对扰动在频域上的进一步增强，使对抗样本在频域上更具误导性。

3.2.3 损失函数设计与决策边界分析

在现有的对抗攻击方法中，优化问题的损失函数通常采用交叉熵损失，在目标攻击和非目标攻击场景下，它虽然采用了类间竞争机制，但都只关注了目标标签的概率变化，而忽略了其他非目标标签的概率分布。这一局限性导致在实施目标攻击时，产生的对抗扰动难以定位到模型的共同决策边界，从而影响了对抗样本的迁移攻击能力。

针对目标攻击场景，对抗样本的生成可被细分为两个目标：首先是使样本靠近目标标签判决区域，其次是使样本远离除目标标签外的任何非目标标签。基于此逻辑，损失函数的设计可以拆分为$\mathcal{L}_1$和$\mathcal{L}_2$，如式(3.10)和(3.11)所示：

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{x}, \mathrm{y}; \boldsymbol{\theta}\right) &= \mathcal{L}_1\left(\boldsymbol{x}, \mathrm{y_t}; \boldsymbol{\theta}\right) + \mathcal{L}_2\left(\boldsymbol{x}, \mathrm{y_t}; \boldsymbol{\theta}\right) \\ &= Z_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)_{y=\mathrm{y_t}} - \max_{y\neq \mathrm{y_t}} Z_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)_y\end{aligned} \tag{3.10}$$

$$\begin{cases} \mathcal{L}_1\left(\boldsymbol{x}, \mathrm{y_t}; \boldsymbol{\theta}\right) = Z_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)_{y=\mathrm{y_t}} \\ \mathcal{L}_2\left(\boldsymbol{x}, \mathrm{y_t}; \boldsymbol{\theta}\right) = -\max_{y\neq \mathrm{y_t}} Z_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)_y \end{cases} \tag{3.11}$$

其中，$\mathcal{L}_1 = Z_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)_{y=\mathrm{y_t}}$为模型将样本分入目标标签的置信度，$\mathcal{L}_2 = \max_{y\neq \mathrm{y_t}} Z_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)_y$为模型除目标标签外，其他所有标签中的最大置信度，在优化的开始阶段，$\mathcal{L}_2$即为原始标签的置信度。

这两个损失函数的优化过程，即为对抗样本向目标标签决策边界移动，并

摆脱被其他标签决策边界捕获的过程。

由平行四边形定则[80]可知，任意非零向量都可以被分解为两个向量的合成。因此，在目标攻击场景下，指向目标标签决策边界的对抗扰动 $\boldsymbol{\delta}$ 可以分为两个方向，如式(3.12)和(3.13)所示：

$$\boldsymbol{\delta}=\boldsymbol{\delta}_{ori}+\boldsymbol{\delta}_{target}=\alpha\cdot sign\left(\frac{\partial\mathcal{L}_1\left(\boldsymbol{x},\mathrm{y_t};\boldsymbol{\theta}\right)}{\partial\,\boldsymbol{x}}\right)+\alpha\cdot sign\left(\frac{\partial\mathcal{L}_2\left(\boldsymbol{x},\mathrm{y_t};\boldsymbol{\theta}\right)}{\partial\boldsymbol{x}}\right) \tag{3.12}$$

$$\begin{cases}\boldsymbol{\delta}_{ori}=\alpha\cdot sign\left(\dfrac{\partial\mathcal{L}_1\left(\boldsymbol{x},\mathrm{y_t};\boldsymbol{\theta}\right)}{\partial\,\boldsymbol{x}}\right)\\ \boldsymbol{\delta}_{target}=\alpha\cdot sign\left(\dfrac{\partial\mathcal{L}_2\left(\boldsymbol{x},\mathrm{y_t};\boldsymbol{\theta}\right)}{\partial\,\boldsymbol{x}}\right)\end{cases} \tag{3.13}$$

其中，$\boldsymbol{\delta}_{ori}$ 是远离原始标签决策边界的扰动，$\boldsymbol{\delta}_{target}$ 是指向目标标签决策边界。

非目标场景下同理，将 $\mathcal{L}_1$ 定义为模型将样本分入原始标签的置信度，$\mathcal{L}_2$ 为模型除原始标签外，其他所有标签的最大置信度。对抗攻击的优化过程，即为对抗样本远离原始标签的决策边界，并在过程中寻找其他标签决策边界的过程，如式(3.14)和(3.15)所示：

$$\begin{cases}\mathcal{L}_1\left(\boldsymbol{x},y_s;\boldsymbol{\theta}\right)=-Z_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)_{y=y_s}\\ \mathcal{L}_2\left(\boldsymbol{x},y_s;\boldsymbol{\theta}\right)=\max\limits_{y\neq y_s}Z_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)_y\end{cases} \tag{3.14}$$

$$\begin{cases}\boldsymbol{\delta}_{ori}=\alpha\cdot sign\left(\dfrac{\partial\mathcal{L}_1\left(\boldsymbol{x},y_s;\boldsymbol{\theta}\right)}{\partial\,\boldsymbol{x}}\right)\\ \boldsymbol{\delta}_{target}=\alpha\cdot sign\left(\dfrac{\partial\mathcal{L}_2\left(\boldsymbol{x},y_s;\boldsymbol{\theta}\right)}{\partial\,\boldsymbol{x}}\right)\end{cases} \tag{3.15}$$

通过将对抗扰动分解为两个部分，并借助模型提供的梯度信息对这两个方向进行精细的权衡，实现了在迭代过程中对抗扰动方向的动态优化，使对抗攻击更明确地指向模型的决策边界，从而使对抗样本跨过 DNN 的公共决策边界，实现更强的可迁移性。

3.3 实验与分析

3.3.1 实验设置

3.3.1.1 数据集

数据集本章选择了三个最常用的图像分类数据集：CIFAR10、CIFAR100、Tiny-ImageNet，三个数据集的图像形状分别为 32×32×3、64×64×3、64×64×3。这些数据集分别包含 10、100、200 个图像类别，本章在实验部分从中分别选择了 13000、18000、28000 张图像，这些图像都能被目标模型正确分类。

3.3.1.2 网络结构

本章考虑了有代表性的网络结构，即 Resnet18[5]（Residual Network）、Resnet50[5]、Resnet56[5]、VGG[81]（Visual Geometry Group Network）、DenseNet-121[82]（Densely Connected Convolutional Network）、Wide-Resnet50[83]（数据表格中分别表示为 Res18、Res50、Res56、VGG19、DN121、W-Res50）。使用其中一个网络作为替代模型生成对抗样本，将其他模型作为黑盒模型进行测试。

本论文的所有实验都建立在 Pytroch 环境下，使用单张 RTX Titan 显卡。将所有网络分别在 CIFAR10、CIFAR100、Tiny-ImageNet 三个数据库上进行训练。在模型训练时，使用 SGD 作为优化器，初始学习率为 0.001，并使用 scheduler 动态调整学习率，动量设置为 0.9。最大迭代次数设置为 2000，批大小为 64。如表 3.1 所示，实验中所涉及的 DNN 在图像分类任务中都取得了很好的结果。

表 3.1 不同模型在三个数据集上的分类性能基线

	VGG19	Res18	Res50	DN121	W-Res50
CIFAR10	92%	82%	84%	84%	84%
CIFAR100	74%	72%	78%	76%	75%
Tiny-ImageNet	62%	56%	66%	62%	69%

本研究所提出的方法是通用的，独立于 CNN 骨干网络，可以以类似的方式应用于任何现有的预训练分类器。在实验中，为了避免由于单一模型结构造成的误差，并确保实验结果的可靠性，本章在实验部分采用了多个替代模型，并通过替代模型和黑盒模型之间的交换来综合评估攻击性能。

3.3.1.3 对比方法

为了证明所提出的基于 DNN 频域鲁棒性的对抗攻击的有效性，本章选择了选择经典的对抗攻击方法 I-FGSM[13]和 PGD_{inf}[21]，以及各种最先进的攻击方法：DI-FGSM[25]、S^2I-FGSM[28]、 MI-DI-FGSM[25]、TI-DI-FGSM[27]作为比较方法。

其中 DI-FGSM 等迁移攻击方法，其本质都是在 I-FGSM 的基础上通过不同的模型增强方法来提升对抗样本的可迁移性。DI-FGSM 通过多样化输入来提升样本的可迁移性；S^2I-FGSM 提出在 DCT 域对模型输入叠加噪声[5]；MI-DI-FGSM、TI-DI-FGSM 为两种对抗攻击的组合版本，MI-DI-FGSM[25]在动量迭代（MI-FGSM）的基础上加入了多样化输入（DI）；TI-DI-FGSM[27]在 DI-FGSM 方法中加入了平移不变攻击（TI）。

3.3.1.4 参数设置

在对 CIFAR100 和 Tiny-ImageNet 的实验中，所有的算法的参数都设置为：最大扰动 ϵ =20，迭代 T =20，步长 $\alpha=\epsilon/T$ =2.0， N =20。对于 CIFAR10 数据集，由于样本尺寸较小，设置最大扰动 ϵ =16， $\alpha=\epsilon/T$ =2.0。

对于 MI-FGSM，设置衰减因子 μ =1.0。对于 DI-FGSM，设置转换概率 p =0.5。对于 TI-FGSM，设置内核长度 k =7。对于 S^2I-FGSM，均匀噪声 ρ =0.5，高斯噪声 ξ 的标准偏差 σ 简单地设置为 ϵ 的值。组合版本的参数设置相同。对于所提出的算法，将 β 设置为 0.1，b 设置为 0.5。另外，在 DNN 公共脆弱区域的构建过程中所使用的模型 $\mathcal{M}_p$，与在实验部分使用的黑盒模型 $\mathcal{M}_i$ 是相对独立的。

3.3.1.5 攻击场景设置

首先通过代理模型生成对抗样本，然后使用这些对抗样本攻击不同的目标模型。在非目标攻击场景下，在每个数据集中随机选择 2000 个样本进行测试。在目标攻击场景下，在每个数据集中随机选择 200 个样本，并对其所有标签进行目标攻击，并保留在 10 个目标上的结果，得到 2000 个对抗样本进行测试。

3.3.2 迁移性分析

3.3.2.1 目标攻击迁移性评估

本节旨在对目标攻击场景下的各种对抗攻击方法进行测试，并通过在黑盒模型上的迁移攻击成功率来评估各攻击方法的可迁移性。在目标攻击场景下，黑盒模型需要将对抗样本分类到在原始标签之外指定的目标标签方可判定为攻击成功。相比非目标攻击场景，其攻击难度明显更高。

由于神经网络模型之间的结构差异，对于黑盒模型，在不同的替代模型上产生的对抗样本的迁移攻击成功率会有较大的区别，因此为了全面地比较不同的攻击方法的可迁移性，分别使用 Res50、VGG19、DN121、W-Res50 作为替代模型，对其他黑盒模型进行攻击，当黑盒模型与替代模型相同时，即为白盒攻击。在对数据集中的图像进行随机选择时，对于不同的攻击方法，使用同样的图像进行测试以保证公平。

表 3.2 在 Tiny-ImageNet 数据集上的目标攻击迁移成功率

Model	Attack	Res18	VGG19	DN121	Res50	W-Res50	AVG
Res50	I-FGSM	2.20%	1.25%	0.55%	100%	2.30%	1.58%
	DI-FGSM	28.80%	19.60%	14.55%	100%	31.55%	24.13%
	TI-DI-FGSM	28.00%	18.80%	13.80%	100%	26.25%	21.71%
	MI-DI-FGSM	27.50%	19.00%	20.00%	100%	33.00%	24.88%
	S^2I-FGSM	27.20%	15.10%	13.80%	100%	29.25%	21.34%
	Ours	31.10%	19.90%	21.55%	100%	39.90%	28.11%
VGG19	I-FGSM	1.35%	100%	0.10%	0.40%	0.40%	0.56%
	DI-FGSM	8.60%	100%	1.25%	4.20%	3.20%	4.31%
	TI-DI-FGSM	10.00%	100%	2.25%	3.60%	3.40%	4.81%
	MI-DI-FGSM	6.50%	100%	7.00%	10.00%	10.00%	8.38%
	S^2I-FGSM	20.50%	100%	6.85%	12.10%	9.35%	12.20%
	Ours	37.35%	100%	24.43%	30.92%	29.08%	30.45%

在 Tiny-ImageNet 数据集上共有 200 个图像标签。从测试集中随机选择 200 个样本，并对除原始标签外的所有标签进行目标攻击。由于目标标签数量众多，使用替代模型上的目标标签分类置信度作为衡量标准，以筛选攻击结果。在替代模型上，目标标签的分类置信度越高，对抗样本对黑盒模型攻击成功的概率

就越大。因此取除原始标签外所有的 199 个标签中最好的十个结果，生成 2000 个对抗样本，以测试各种攻击方法在目标攻击场景下的最优结果。

相似地，CIFAR100 数据集中有 100 个图像标签，从该数据集中随机选择 200 个样本，对每个样本除原始标签外所有的 99 个标签进行攻击，并保留在目标标签上置信度最高的 10 个对抗样本，同样得到 2000 个对抗样本进行测试。

表 3.2、表 3.3 分别为不同攻击方法在 Tiny-ImageNet 和 CIFAR100 数据集上对抗样本的迁移攻击成功率。竖列是不同的攻击方法，最左列是不同的替代模型。每行代表该攻击方法在不同的黑盒模型上的迁移攻击成功率。

表 3.3 在 CIFAR100 数据集上的目标攻击迁移成功率

Model	Attack	Res18	VGG19	DN121	Res50	W-Res50	AVG
Res50	I-FGSM	8.50%	4.85%	8.25%	100%	2.60%	6.05%
	DI-FGSM	35.00%	30.50%	40.90%	100%	20.40%	32.70%
	TI-DI-FGSM	24.00%	21.20%	27.70%	99.95%	14.60%	21.88%
	MI-DI-FGSM	29.00%	31.00%	43.50%	100%	20.00%	30.88%
	S^2I-FGSM	32.95%	26.45%	33.00%	99.75%	18.05%	27.61%
	Ours	41.61%	31.81%	46.08%	100%	30.35%	37.46%
VGG19	I-FGSM	3.75%	100%	4.05%	2.55%	0.70%	2.76%
	DI-FGSM	12.65%	100%	13.40%	11.40%	4.00%	10.36%
	TI-DI-FGSM	8.30%	98.80%	9.98%	7.95%	3.80%	7.51%
	MI-DI-FGSM	12.80%	100%	10.25%	12.25%	8.30%	10.90%
	S^2I-FGSM	18.35%	99.40%	18.15%	16.00%	6.40%	14.73%
	Ours	40.80%	100%	44.42%	42.46%	22.91%	37.65%

当黑盒模型与替代模型相同时，为白盒攻击。除白盒攻击的成功率外，计算每种方法的对抗样本在四个黑盒模型上的平均迁移攻击成功率，表示为 AVG。表中的实验数据表明，在具有挑战性的目标攻击场景下，本章所提出的方法所生成的对抗样本，明显优于比较的黑盒对抗攻击方法，包括最新提出的 S^2I-FGSM，以及多种模型增强的组合方法 TI-DI-FGSM，MI-DI-FGSM。本章提出的方法在所有黑盒模型的迁移成功率上全面占优，在不同替代模型上的平均迁移攻击成功率，比表现最好的比较方法高约 5%-18%。

在 CIFAR10 数据集中的图像标签只有 10 个，因此保存对除原始标签外的所有 9 个标签上的目标攻击结果，从测试集中随机选择 200 个样本，在每个替代

模型上得到 1800 个对抗样本，对各攻击方法在目标攻击场景下的平均结果进行测试。与表 3.2、表 3.3 相同，表 3.4 中的实验数据显示，在 CIFAR10 数据集上，本章提出的方法在所有替代模型上的表现同样优于所有的比较方法。在所有的三个数据集上，目标攻击场景下的平均迁移攻击成功率比表现最好的比较方法高约 3%-8%。

表 3.4 在 CIFAR10 数据集上的目标攻击迁移成功率

Model	Attack	Res18	VGG19	DN121	Res50	W-Res50	AVG
Res50	I-FGSM	15.11%	11.89%	13.56%	99.83%	15.94%	14.13%
	DI-FGSM	31.67%	29.39%	30.17%	99.89%	30.61%	30.46%
	TI-DI-FGSM	22.94%	17.33%	20.72%	98.56%	20.22%	20.30%
	MI-DI-FGSM	38.39%	33.11%	37.00%	98.56%	35.61%	36.03%
	S^2I-FGSM	35.83%	30.22%	30.78%	97.56%	34.39%	32.81%
	Ours	41.86%	35.94%	37.40%	97.11%	40.79%	39.00%
VGG19	I-FGSM	4.44%	74.78%	6.22%	4.67%	5.39%	5.18%
	DI-FGSM	9.72%	79.72%	10.72%	9.00%	10.50%	9.99%
	TI-DI-FGSM	8.22%	68.89%	8.67%	6.67%	7.83%	7.85%
	MI-DI-FGSM	12.94%	81.00%	15.28%	12.67%	14.28%	13.79%
	S^2I-FGSM	7.28%	57.28%	8.33%	6.56%	8.00%	7.54%
	Ours	16.72%	87.28%	17.72%	18.72%	17.94%	17.78%

横向比较三个数据集上的实验数据可以发现，在 CIFAR10 上的迁移攻击的平均结果与在 Tiny-ImageNet 和 CIFAR100 上的各攻击方法的最优结果相当，这是因为 CIFAR10 数据集的图像尺寸较小，并且图像标签较少，因此其攻击难度稍低。另外，从各方法在 CIFAR10 上的白盒攻击成功率，可以发现基于 I-FGSM 的对抗攻击方法或多或少地降低了 I-FGSM 的白盒攻击成功率，这是因为，这些对抗样本的可迁移性来源，正是通过不同种类的模型增强来干扰对抗样本对于替代模型的拟合，使得替代模型能够模拟黑盒模型的输出。在 CIFAR10 上的对所有标签的目标攻击中，出现了对抗攻击对于替代模型欠拟合，白盒攻击的成功率下降的现象。

3.3.2.2 非目标攻击迁移性评估

非目标攻击场景只要求 DNN 对于对抗样本的分类偏离原始标签即可，因此与目标攻击场景不同的是，在非目标攻击场景下，对抗样本只需被黑盒模型分

类错误即可判定为迁移攻击成功，因此攻击难度相比于目标攻击场景大大降低。在非目标攻击场景下，在每个数据集的测试集中随机选择 2000 个图像进行攻击，得到 2000 个对抗样本进行迁移性测试。与上节类似地，表 3.5、表 3.6、表 3.7 中分别报告了在 Tiny-ImageNet、CIFAR100 和 CIFAR10 数据集上对抗样本的迁移攻击成功率。

表 3.5 Tiny-ImageNet 数据集上的非目标攻击迁移成功率

Model	Attack	Res18	VGG19	DN121	Res50	W-Res50	AVG
Res50	I-FGSM	56.20%	27.50%	9.30%	100%	26.40%	29.85%
	DI-FGSM	78.00%	62.60%	37.34%	100%	61.70%	59.91%
	TI-DI-FGSM	76.35%	58.70%	37.53%	100%	56.90%	57.37%
	MI-DI-FGSM	84.85%	69.70%	49.50%	100%	69.20%	68.31%
	S^2I-FGSM	83.50%	66.00%	54.00%	100%	78.50%	70.50%
	Ours	84.00%	68.50%	54.00%	100%	80.00%	71.63%
VGG19	I-FGSM	58.75%	100%	15.05%	27.25%	21.00%	30.51%
	DI-FGSM	72.35%	100%	34.40%	52.70%	46.65%	51.52%
	TI-DI-FGSM	72.20%	100%	33.65%	48.90%	42.89%	49.41%
	MI-DI-FGSM	76.29%	100%	46.39%	64.40%	57.20%	61.07%
	S^2I-FGSM	79.90%	100%	54.80%	65.80%	64.30%	66.20%
	Ours	81.00%	100%	55.50%	70.00%	65.50%	68.00%

通过对比目标攻击与非目标攻击场景下的攻击成功率，可以观察到一个明显的趋势：对于所有攻击方法，即使是目标攻击的最优结果，仍远低于非目标攻击的表现，这也说明了目标攻击场景下，对抗攻击的难度更大。表中的数据显示，本章所提出方法的非目标攻击迁移成功率在三个数据集上同样全面优于比较方法，在所有替代模型上的平均迁移成功率高约3%-13%，这说明本章所提出的方法在非目标攻击场景下同样具有良好的迁移性。

综合以上对于对抗攻击可迁移性的测试得出，在目标和非目标攻击场景下，本章提出的方法都能够显著地提升对抗样本的可迁移性。此外，实验结果也验证了频域分析方法的准确性，基于频域敏感性分析的对抗攻击能够有效地提升扰动的方向性，使其能够更精准地指向 DNN 的共有决策边界，从而不仅在非目标攻击场景下表现出色，也在目标攻击场景下显著提高了对抗样本对未知黑盒模型的攻击能力。

表 3.6 CIFAR100 数据集上的非目标攻击迁移成功率

Model	Attack	Res18	VGG19	DN121	Res50	W-Res50	AVG
Res50	I-FGSM	63.30%	61.70%	63.55%	100%	30.45%	54.75%
	DI-FGSM	77.45%	79.40%	82.00%	100%	45.70%	71.14%
	TI-DI-FGSM	75.63%	78.79%	81.85%	100%	51.70%	71.99%
	MI-DI-FGSM	86.50%	85.50%	88.50%	100%	64.00%	81.13%
	S^2I-FGSM	87.80%	88.30%	89.80%	100%	65.85%	82.94%
	Ours	89.50%	88.50%	94.00%	100%	66.50%	84.63%
VGG19	I-FGSM	55.55%	100%	60.35%	48.70%	24.45%	47.26%
	DI-FGSM	69.90%	100%	74.00%	64.00%	39.70%	61.90%
	TI-DI-FGSM	71.00%	100%	75.90%	71.25%	47.25%	66.35%
	MI-DI-FGSM	72.45%	100%	84.20%	72.45%	49.00%	69.53%
	S^2I-FGSM	84.40%	100%	82.90%	79.02%	55.15%	75.37%
	Ours	85.00%	100%	88.00%	84.00%	60.50%	79.38%

表 3.7 CIFAR10 数据集上的非目标攻击迁移成功率

Model	Attack	Res18	VGG19	DN121	Res50	W-Res50	AVG
Res50	I-FGSM	56.20%	27.50%	9.30%	100%	26.40%	29.85%
	DI-FGSM	78.00%	62.60%	37.34%	100%	61.70%	59.91%
	TI-DI-FGSM	76.35%	58.70%	37.53%	100%	56.90%	57.37%
	MI-DI-FGSM	84.85%	69.70%	49.50%	100%	69.20%	68.31%
	S^2I-FGSM	83.50%	66.00%	54.00%	100%	78.50%	70.50%
	Ours	84.00%	68.50%	54.00%	100%	80.00%	71.63%
VGG19	I-FGSM	58.75%	100%	15.05%	27.25%	21.00%	30.51%
	DI-FGSM	72.35%	100%	34.40%	52.70%	46.65%	51.52%
	TI-DI-FGSM	72.20%	100%	33.65%	48.90%	42.89%	49.41%
	MI-DI-FGSM	76.29%	100%	46.39%	64.40%	57.20%	61.07%
	S^2I-FGSM	79.90%	100%	54.80%	65.80%	64.30%	66.20%
	Ours	81.00%	100%	55.50%	70.00%	65.50%	68.00%

3.3.3 鲁棒性分析

在黑盒对抗攻击的实际应用中，防御者通常不了解对抗攻击方法的具体细节，正如攻击者无法获得黑盒模型的信息一样，因此，防御者需要考虑相对通用的防御方法，而对于对抗防御的研究提出，在图像数据输入模型之前的图像处理可以降低对抗扰动带来的影响。数据预处理作为成本最小的防御方法，几

乎没有增加运行时间，而且不需要额外训练，因而受到了广泛应用[28]。尽管许多攻击方法可以容易地欺骗神经网络模型，但有意或无意的图像处理程序都可能会破坏扰动的对抗性，使得样本对于黑盒模型的攻击失败。

表 3.8 目标攻击场景下，各方法在面对防御模型时的迁移攻击成功率

Defence	Attack	Res18	VGG19	DN121	Res50	W-Res50
R&C	I-FGSM	2.20%	27.95%	1.45%	1.35%	1.75%
	DI-FGSM	9.15%	82.25%	6.65%	8.05%	7.70%
	TI-DI-FGSM	9.45%	77.50%	7.15%	8.00%	7.60%
	MI-DI-FGSM	7.00%	79.00%	8.00%	13.50%	10.00%
	S^2I-FGSM	15.65%	71.75%	11.65%	15.10%	13.30%
	Ours	33.65%	93.85%	25.45%	32.15%	29.70%
JPEG	I-FGSM	2.00%	81.75%	0.65%	0.50%	0.75%
	DI-FGSM	7.25%	94.55%	3.65%	5.45%	4.75%
	TI-DI-FGSM	7.40%	89.60%	2.40%	4.25%	3.85%
	MI-DI-FGSM	5.50%	96.50%	4.50%	7.00%	8.50%
	S^2I-FGSM	17.65%	98.85%	8.35%	13.10%	9.55%
	Ours	32.60%	100%	23.45%	26.95%	26.55%
RS	I-FGSM	1.10%	8.15%	0.95%	0.65%	0.65%
	DI-FGSM	4.80%	52.55%	3.25%	4.45%	3.00%
	TI-DI-FGSM	6.55%	76.85%	5.50%	7.25%	5.75%
	MI-DI-FGSM	5.00%	62.50%	3.00%	1.50%	4.50%
	S^2I-FGSM	7.06%	39.10%	5.32%	6.70%	5.02%
	Ours	17.10%	86.25%	15.10%	16.95%	15.95%

为了进一步评估对抗攻击在面对对抗防御时的表现，以数据集 Tiny-ImageNet 为例，本章使用三种基于输入预处理的防御方法[59, 61]对各对抗攻击方法进行测试。使用 VGG19 作为替代模型，并且将三种输入数据的预处理作为模型的第一层，与黑盒模型结合得到防御模型。其中 R&C[61]对图像随机变换尺寸大小，并随机填充，参数 scale 设置为（0.5,0.8），即将图像随机裁剪为原图像尺寸的 0.5-0.8 倍，然后将其缩放回图像原来的大小；JPEG[63]对图像进行 JPEG 压缩，质量系数设置为 50%；RS[59]对图像进行了高斯平滑，高斯核大小 ksize 为 3，标准差 sigma 设为 2。

本实验在目标攻击场景和非目标攻击场景下，测试了对抗样本在防御模型上的迁移攻击成功率，并且为了直观地反映出图像预处理对于样本对抗性的影响，排除预处理对于网络分类结果的干扰，本实验将防御模型与替代模型对于对抗样本的决策是否一致作为迁移成功的标准。

表 3.9 非目标攻击场景下，各方法在面对防御模型时的迁移攻击成功率

Defence	Attack	Res18	VGG19	DN121	Res50	W-Res50
R&C	I-FGSM	29.05%	82.80%	21.40%	26.55%	23.60%
	DI-FGSM	45.10%	86.60%	37.50%	45.95%	43.80%
	TI-DI-FGSM	42.80%	85.40%	35.20%	41.10%	39.05%
	MI-DI-FGSM	47.45%	87.65%	42.80%	48.45%	45.90%
	S^2I-FGSM	49.30%	88.70%	36.95%	44.70%	44.25%
	Ours	58.00%	96.00%	45.50%	54.00%	50.00%
JPEG	I-FGSM	29.55%	98.80%	14.00%	18.80%	16.60%
	DI-FGSM	37.55%	98.55%	24.70%	31.50%	27.60%
	TI-DI-FGSM	37.00%	97.15%	22.05%	29.70%	26.80%
	MI-DI-FGSM	40.70%	96.90%	29.40%	40.70%	34.00%
	S^2I-FGSM	48.50%	97.29%	38.70%	43.25%	43.05%
	Ours	57.00%	100%	36.00%	45.00%	45.50%
RS	I-FGSM	17.00%	60.40%	10.70%	13.80%	12.45%
	DI-FGSM	25.60%	80.30%	21.50%	26.05%	25.65%
	TI-DI-FGSM	32.20%	83.85%	27.80%	34.40%	28.20%
	MI-DI-FGSM	27.30%	76.80%	21.65%	26.30%	28.35%
	S^2I-FGSM	25.15%	74.10%	15.80%	23.45%	21.95%
	Ours	38.00%	89.50%	22.50%	34.50%	29.00%

表 3.8、表 3.9 中的结果反映了对抗攻击在面对不同的防御方法时的表现。从表中的数据可以看出，在目标攻击场景下，所有攻击方法在防御模型上的迁移成功率明显低于正常训练的模型，这说明了防御方法能够有效地降低对抗扰动的影响，在三种防御模型上，相较于比较方法，本章所提出的方法在替代模型上的平均迁移攻击成功率分别高 13%-16%、13%-16%、3%-10%。

其中 RS 防御模型上各种攻击方法的迁移成功率最低，说明 RS[59]方法对于输入进行高斯平滑的效果最为明显，并且对于通过 DCT 域加噪来构造对抗样本的 S^2I-FGSM 造成的影响最严重，非目标攻击场景的结果反映了相似的规律。两

种攻击场景下，本章所提出的对抗攻击方法在所有防御模型上，相较于其他方法都仍表现最优，所产生的对抗样本在面对对抗防御时仍具有很强的攻击能力。

3.3.4 对抗攻击的可视化分析

通过梯度加权类激活映射（Gradient-weighted Class Activation Mapping, Grad-CAM）[84]技术，对对抗样本从视觉上进行可视化解释，可以进一步分析不同的对抗攻击的攻击效果。Grad-CAM 可以通过梯度信息来解释 DNN 的决策，用热图来表示分类器的注意力，本实验中以 Resnet50 作为替代模型，目标模型为 Resnet18，在非目标攻击场景下进行测试。

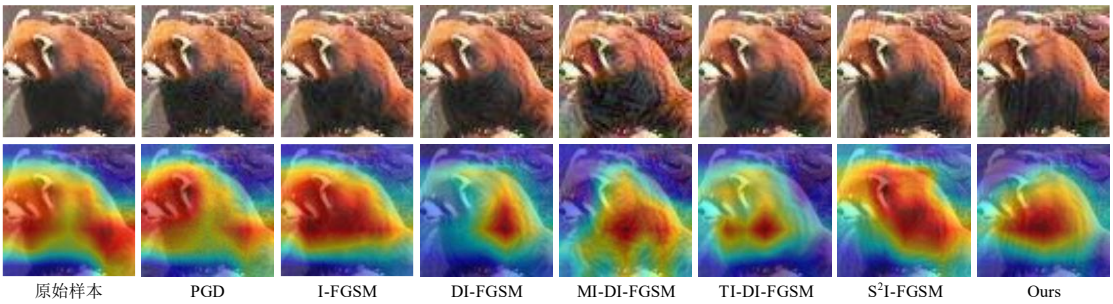


图 3.3 Grad-CAM 热图的可视化结果，图中样本来自 Tiny ImageNet 数据集

如图 3.3 为不同攻击方法生成的对抗样本的 Grad-CAM 热图，其中上面一行的图像分别为由不同算法生成的干净原始图像和对抗样本，下面一行的图像则为 Grad-CAM 生成的注意力热图。

从图中可知，目标模型的注意力首先集中在原始输入图像的目标区域上，在经过攻击后，模型的注意力在不同对抗样本中略有变化。将本章所提出的方法与原始图像的热图进行比较，目标模型的关注区域发生了明显变化，说明所提出的攻击方法会导致黑盒模型的注意力偏离原始区域，表明了攻击的有效性。这也印证了前文中所提出的傅里叶频域观点，即通过在 CNN 的频域公共脆弱区域设计扰动，可以使对抗攻击找到神经网络的公共决策边界，从而有效地影响黑盒模型的决策。

3.4 本章小结

为解决现有可迁移黑盒对抗攻击，在目标攻击场景下表现不佳，并且由于缺乏可解释性的问题，本章从 DNN 的频域鲁棒性的角度提出了一种基于频率模型增强的黑盒对抗攻击方法。首先，本章对多个替代模型的傅立叶热图进行集成，得到了 DNN 的共同脆弱频域。然后，在每次扰动的迭代过程中进行频域增强，对模型输入在敏感频域信息进行变换，以模拟更多的替代模型，并使对抗扰动的频域信息能够契合 DNN 频域敏感性特征。最后，本章提出将对抗样本的优化目标进行分解，并提出通过三元损失来调整扰动的方向性，从而跨越更多可能的黑盒模型的决策边界。在多个常用数据集上的实验证明，与现有的黑盒对抗攻击相比，本章提出的算法所生成的对抗样本，在目标攻击和非目标攻击场景下具有更高的迁移攻击成功率，并且在面对防御模型时也表现优越。

第四章 基于频域特性分析的不可见对抗攻击

在现有的黑盒对抗攻击研究中，对抗攻击的不可见性往往受到忽视，并且受限于梯度更新方法的设计，仅能通过 L_p 范数限制扰动强度，而无法针对图像内容做出调整，容易在对抗样本上留下明显修改痕迹。为提升黑盒对抗攻击的不可见性，本章更深入地探究了对抗样本的频域特性，发现人类视觉系统 HVS 的频域特性与 DNN 的频域鲁棒性存在相通之处，并基于此从频域进行对抗攻击，量化扰动的不同频率分量对于对抗样本的可迁移性和视觉保真度的影响，提升对抗攻击的效率，从而显著提升了对抗扰动的不可见性。

4.1 研究动机

第三章提出了基于频域鲁棒性的可迁移对抗攻击，在现有的基于 I-FGSM 和模型增强的工作的基础上，有效地提高了对抗样本的可转移性。然而，可迁移对抗攻击的不可见性往往被忽视了，通常会生成带有明显修改痕迹的对抗样本，会引起模型所有者的怀疑，从而降低黑盒对抗攻击方法的实用性。

高质量的对抗样本在真实世界的应用场景中发挥着重要作用，在社交媒体等大数据环境中，使用对抗样本可以帮助保护用户的私人数据免受未经授权的利用。用户上传和共享他们照片的对抗样本，而不是干净的照片，这使模型更难识别和利用个人信息，同时不影响他们的照片共享体验，从而有效提高了社交媒体平台上用户的隐私保护水平。这些场景对对抗样本的可迁移性与视觉保真度提出了更高的要求。

L_p 范数是目前最常被用来测量和约束对抗样本相对于干净图像的视觉差异性的指标，它测量样本与干净图像之间在 L_p 空间的距离，当 L_p 范数较小时，可以认为扰动的不可见性良好。然而，现有的研究[85-87]已经证明，这种约束对于对抗样本的视觉保真度来说存在缺陷。对抗样本可能与干净样本拥有很小的 L_p 距离，但是视觉上存在明显的修改痕迹。

L_0 范数测量两幅图像之间不同的像素数，L_2 范数测量两幅图像之间的欧几里德距离，L_∞ 范数测量两幅照片中对应像素之间的最大差值，这些距离度量仅能在像素域上约束扰动的部分统计特性，并不足以满足对抗样本的视觉保真度要求。图 4.1 为在常用 L_p 范数约束下生成的对抗样本，显然，L_p 范数约束的不足导致样本出现了明显的视觉修改痕迹。在 L_0 范数很小的前提下，单个像素的扰动值可以无限大；一个不均匀的扰动可以骗过 L_2 范数的检测，但是视觉上非常明显；L_∞ 范数只约束了扰动对单个像素的最大修改幅度，而一个均匀扰动可能最大值小，但平均强度很大，严重影响对抗样本的视觉感知质量。

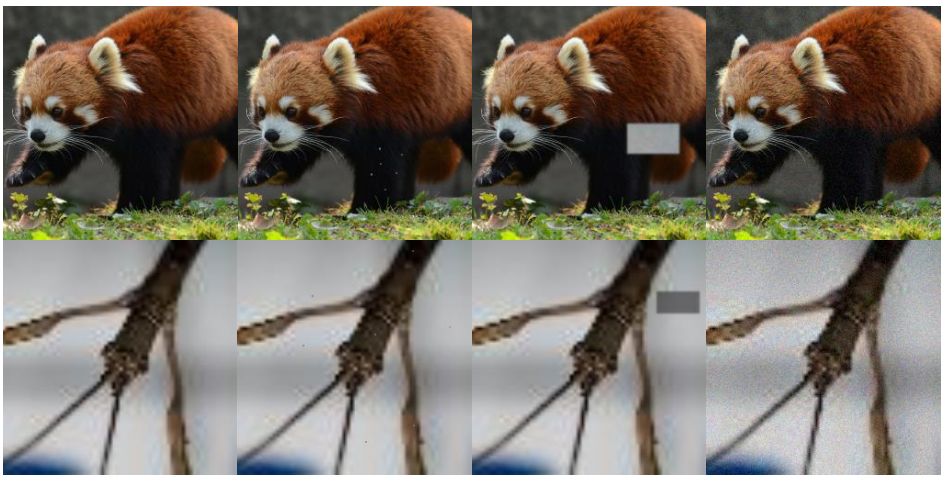


图 4.1 具有小 L_0、L_2 和 L_∞ 范数约束的扰动图像仍然表现出明显的修改痕迹

L_p 范数的使用可能会产生误导，即拥有小 L_p 距离的对抗样本与干净样本视觉上相似，而事实并非如此。像素域上的 L_p 范数对扰动的约束并不充分，并且几乎所有对抗攻击都在空域上直接根据模型的梯度信息对图像像素进行修改，难以使扰动适应图像内容的变化，从而很容易留下明显的视觉痕迹。

为了提升对抗攻击的不可见性，一些研究意识到了 L_p 范数的不足，并试图在对抗攻击的过程中对扰动进行自适应的调整。Ding 等人[88]通过提出一种选择性的 I-FGSM 对类 I-FGSM 算法的迭代过程进行了改进，该算法根据一阶偏导数忽略迭代过程中不重要的像素，从而压缩了扰动，显著降低了对抗样本的图像失真。王等人[86]发现，全局扰动对于图像内容/空间结构的忽略，会导致在原始

图像的其他干净区域中留下明显的伪影，因此提出通过使用对抗样本的逐像素感知冗余作为损失函数来自适应地调整扰动强度，从而基于人眼的可注意差异（Just Noticeable Difference, JND）自适应地分配扰动。类似地，张等人[89]将图像的 JND 作为先验信息添加到对抗攻击中，并将扰动投影到原始图像的 JND 空间中。此外，他们添加了一个视觉系数来调整扰动的投影方向，以有意识地均衡对抗样本的可迁移性和视觉保真度。

这种基于内容进行调整的自适应扰动是启发性的，它们都从空间域的角度分析了图像内容对扰动不可见性的影响，而图像的特征，如结构、纹理等，取决于频率信息的分布。因此，从傅立叶域分析不可见对抗攻击的方法能够更全面地理解对抗攻击特征，基于第三章对于对抗攻击的频域分析，本章将从频率信息的角度，对对抗攻击的不可见性继续进行探索。

4.2 基于频域特性分析的不可见对抗攻击

4.2.1 核心思想

现有的对抗攻击方法通过 L_p 范数来衡量图像像素值的整体变化，忽略了图像内容，因此出现了图像质量评价与不可见性错位的问题。人类的视觉系统 HVS[90]是所有图像信号的最终接收器，因此，对抗攻击的不可见性应考虑人眼对图像的视觉特性。HVS 受多种因素的影响，形成了一些可以利用的视觉规律。

在图像水印领域对于最小感知差异的研究发现[91]，对于不同的视觉内容，视觉信号对于变化的感知是不同的。HVS 是一个非均匀和非线性的图像处理系统，在频域中充当低通线性系统。由于其有限的分辨率，与高频图像信号相比，HVS 对低频图像信号的变化更敏感。DNN 感知的图像信息与 HVS 有着相反的直觉差异，人眼无法识别的图像信息对 DNN 来说可能很重要，它们可以使用人眼不可见的高频信息来实现正确的判断。对这些信息的扰动对人眼来说可能不太可见，同时能够显著影响 DNN 的决策[90]，这启发了从频域视角对于对抗样本的频域特性的探索。

在上一章的研究中已经发现，同一数据集上的所有 DNN 都会对一些特定频域分量的扰动更加敏感，并且与之前的观点不同的是，这种共同敏感性与扰动的频率高低并不成简单的正反比关系，低频和高频扰动都可能产生较大影响。而与之对应的是，人类视觉系统的频率特性表现出了相似的规律，人眼对于图像信息变化的敏感程度与其频率存在明显的相关性。

HVS 可以被理解为一个频率分解系统，将输入图像的空间信息分解为不同的频率分量，并且对空间频率的不同分量具有不同的灵敏度，而对比敏感度函数（Contrast Sensitivity Function, CSF）即为专门用于评估 HVS 对不同频率的视觉刺激的响应的指标。在大量实验的基础上，Mannos 和 Sakrison[92]提出了基于大量实验的 CSF 模型，使用傅立叶变换等数学工具来表征 HVS 灵敏度和空间频率之间的关系，Daly[93]后来将其改进为式(4.1)：

$$CSF\left(f\right)=\begin{cases}2.6\left(0.0192+0.114\cdot f\right)exp\left[-0.114\cdot f\right], f\geq f_{peak}\\ 0.981 \qquad\qquad\qquad\qquad , f<f_{peak}\end{cases} \tag{4.1}$$

其中，$f=\sqrt{f_x^2+f_y^2}$ 是空间频率，f_x 和 f_y 分别是水平和垂直方向上的空间频率。根据该模型，对于图像的中高频信息，人眼的灵敏度与频率大致成反比，即扰动的不可见性与其频率的高低成正相关，视觉感知的灵敏度在高频区域明显下降。则结合 DNN 的频域共同敏感性与 HVS 频域特性，通过将扰动约束到更高的频率分量上，可以在满足对抗样本的可迁移性的同时提升其不可见性，因此本章提出，从傅立叶域中添加对抗扰动。

如果能从频域上将扰动放在 DNN 的公共脆弱区域，就能提升对抗样本的可迁移性，并且更进一步地，将扰动放在其中频率更高的区域，就能在保证迁移性的同时提升对抗样本的视觉保真度。另外，由于傅里叶矩阵中的每个元素，对应的都是图像的一种频域分量的信息，从频域对扰动的设计直接能够对其的频率进行限制，从而避免在像素域上难以使对抗扰动适应图像内容的问题。

通常情况下，攻击者会将图像放入 DNN，并通过对模型输入的反向传播来获取基于损失函数的梯度，如式(4.2)所示：

$$g=\frac{\partial\mathcal{L}\left(\boldsymbol{x},y;\boldsymbol{\theta}\right)}{\partial\boldsymbol{x}} \tag{4.2}$$

其中$\mathcal{L}(\boldsymbol{x}, y;\boldsymbol{\theta})$为损失函数。反向传播计算的是损失函数关于模型输出的梯度信息，利用链式法则和神经网络的可微性，完成从输出层到输入层的梯度计算和传播，通过损失函数梯度的形式将最终预测结果从输出层反向传播到输入层，从而得到模型对输入预测的梯度，并利用该梯度更新扰动。

当从傅立叶频域设计对抗攻击时，扰动对图像的每个像素值的修改，被转换成了对图像各频率信息的修改，因此梯度信息也应该反映模型输入中每个频率信息的变化如何影响模型的输出和损失函数。首先使用傅立叶变换将输入图像转换到频域，并将其表示为$\boldsymbol{x}_f$，如式(4.3)所示：

$$\boldsymbol{x}_f = fft(\boldsymbol{x}) \tag{4.3}$$

并将其放入 DNN 中以计算损失函数。神经网络通常由一系列连续可变的激活函数和线性运算组成，因此整个网络可被视为一个复合函数，其输出相对于每个权重和偏置都是可微分的。由于傅里叶变换和傅里叶逆变换都是可微分的，根据链式法则，可以将它们加入反向传播过程中，将损失函数对于输入的梯度转换为对于频域信息的梯度。在反向传播中，梯度信息如式(4.4)所示：

$$\boldsymbol{g} = \frac{\partial \mathcal{L}\left(ifft\left(\boldsymbol{x}_f\right), \mathrm{y};\boldsymbol{\theta}\right)}{\partial \boldsymbol{x}_f} = \frac{\partial \mathcal{L}(\boldsymbol{x}, y;\boldsymbol{\theta})}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}_f} \tag{4.4}$$

然后使用梯度信息更新$\boldsymbol{x}_f$，并在傅立叶逆变换后最终获得对抗样本。

4.2.2 总体框架

基于上述对抗攻击频域分析，本章提出了一种不可见的可迁移对抗攻击。先量化分析得出扰动的不同频域分量对于可迁移性和视觉保真度的影响，再通过聚类分析筛选出频域中攻击价值最高的频域，作为扰动的频域分布调整方向。

如前文所述，模型增强对于可迁移对抗攻击有着重要的意义[28]。由于替代模型的决策边界和模型架构都对对抗性可迁移性有重大影响[94]，使用模型增强减少对替代模型决策边界的依赖性是可迁移对抗攻击的重要策略[25]。本章使用上一章所提出的频域增强，将模型增强与 DNN 的频域敏感性分析相结合，使其更有针对性。具体来说，频域增强将模型输入转换到傅立叶域，应用噪声添加和增强将扰动约束到 DNN 的公共频率敏感区域，如式(4.5)所示：

$$\boldsymbol{x}_{\text{aug}} = ifft\left[\left(\boldsymbol{x}_f^i \odot \boldsymbol{\mu} + \boldsymbol{\xi}\right) \odot \left(\boldsymbol{E} + \beta \cdot \boldsymbol{\omega}\right)\right] \tag{4.5}$$

综上所述，本章从频域构造了对抗攻击的完整方案，如算法 4-1 所示。

算法 4-1：基于频域特性分析的不可见黑盒对抗攻击

输入：频域掩码集 Ω；超参数 β；源图像 $\boldsymbol{x}$ 和原始标签 $\mathrm{y_s}$（非目标攻击）或目标标签 y_t（目标攻击）；替代模型 $\mathcal{M}_0 : \mathcal{X} \to \mathcal{Y}$；频域迭代步长 α；迭代次数 T；L_∞ 参数限制 ϵ；噪声初始化次数 N

输出：对抗样本 $\boldsymbol{x}'$

初始化：$\boldsymbol{E} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$，$\boldsymbol{E} \in \boldsymbol{R}^{H \times W}$，$\boldsymbol{x}'_0 = \boldsymbol{x}$，$\mathrm{y} = \mathrm{y_s}$ 或 $\mathrm{y_t}$；

1. **for** $i = 0 \to T-1$ **do**
2. **for** $n = 0 \to N-1$ **do**
3. 对模型输入做傅立叶变换：$\boldsymbol{x}^i{}_f = fft\left(\boldsymbol{x}'_i\right)$
4. 随机初始化噪声 $\boldsymbol{\xi}$ 和 $\boldsymbol{\mu}$
5. 傅里叶频域增强：$\boldsymbol{x}_{aug} = ifft\left[\left(\boldsymbol{x}^i{}_f \odot \boldsymbol{\mu} + \boldsymbol{\xi}\right) \odot \left(\boldsymbol{E} + \beta \cdot \boldsymbol{\omega}\right)\right]$
6. 反向传播计算梯度：$\boldsymbol{g}_n = \dfrac{\partial \mathcal{L}_{adv}\left(\boldsymbol{x}_{aug}, \mathrm{y}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{x}^i{}_f} + \dfrac{\partial \mathcal{L}_{fre}\left(\boldsymbol{x}^i{}_f, \boldsymbol{x}^0{}_f; \Omega\right)}{\partial \boldsymbol{x}^i{}_f}$
7. **end for**
8. 对多次增强得到的梯度进行平均：$\boldsymbol{g}' = \dfrac{1}{N}\sum_{n=0}^{N-1} \boldsymbol{g}_n$
9. 从频域对样本进行更新：$\boldsymbol{x}^{i+1}{}_f = \boldsymbol{x}^i{}_f + \alpha \cdot \boldsymbol{g}'$
10. 将样本转换回空间域：$\boldsymbol{x}'_{i+1} = ifft\left(\boldsymbol{x}^{i+1}{}_f\right)$
11. 限制最大扰动幅度：$\boldsymbol{x}'_{i+1} = clip_{X,\epsilon}\left(\boldsymbol{x}'_{i+1}\right)$
12. 像素值归一化：$\boldsymbol{x}'_{i+1} = clip\left(\boldsymbol{x}'_{i+1}, 0, 1\right)$
13. **end for**
14. 得到对抗样本 $\boldsymbol{x}' = \boldsymbol{x}'_T$

本算法主要分为三个阶段。首先将输入图像转到频域并对输入图像 $\boldsymbol{x}_i'$ 进行频域增强；然后将对抗攻击的损失函数分为对抗损失和频域损失，从傅里叶频域对两个损失函数共同进行求导；最后，通过频域步长 α 更新样本 $\boldsymbol{x}'_{i+1}$，并对样本进行归一化，将像素值限制在自然图像范围内，即[0, 1]。

值得注意的是，对于输入图像的加噪处理可能使得梯度信息的方向不稳定，因此本算法对频域增强的噪声进行 N 次随机初始化，并进行 N 次反向传播得到梯度信息 $\boldsymbol{g}_n\{n=1,2,\ldots, N\}$，对 $\boldsymbol{g}_n$ 进行平均后，得到的平均梯度 $\boldsymbol{g}'$ 可以使得对抗扰动的更新方向更稳定。算法中每个步骤的设计将在以下几节中详细阐述。

4.2.3 聚类分析

前文对于对抗样本的频域分析得出，DNN 在傅里叶频域上对于某些特定的频域分量具有共同的脆弱特性，这些频率分量上的噪声有更大的概率能够误导神经网络。同时，扰动的频率又与其不可见性息息相关，在傅里叶频域中，人眼对于低频信息的变化较为敏感，强度很低的低频扰动就会引起在人眼感知范围内的修改痕迹。因此，扰动在傅里叶频域上的优化目标可以综合从频域敏感性与不可见性两个方面进行分析。

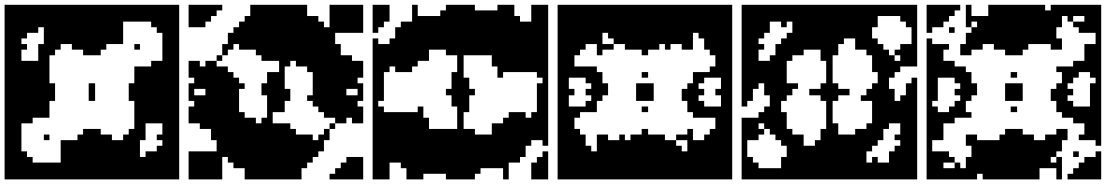


图 4.2 CIFAR10 数据集上，不同神经网络模型的频域敏感区域掩码

观察 DNN 的频域共同敏感区域，如图 4.2 所示，可以发现有相当区域非常靠近图像中心，也即神经网络的敏感频段中有一部分存在于中低频区。由于人眼的视觉特性，对于 HVS 来说，中低频噪声的可见阈值更低。因此，对于对抗攻击的过程来说，傅里叶矩阵中的每个元素都包含了该频域分量的扰动在两个方面的特性。首先，傅里叶热图中的热度值 h 反映了 DNN 对于该频域分量扰动的敏感性。热度值 h 越高，则该频率分量的扰动使目标模型分类错误的概率就越大。同时，对于 HVS 扰动的不可见性与其频率是正相关的，对于相同强度的

扰动，其频率越高，就越不可见。也即，在傅里叶域中，像素点到零频分量的距离（频域中心化后）就代表了该频率扰动的不可见性。扰动这两种特性对应了对抗样本可迁移性和视觉保真度的需求，这两种特性的公式分别如式(4.6)和(4.7)所示：

$$h = heat(u,v) \tag{4.6}$$

$$\tau = dist(u,v) = \sqrt{\left(u - \left[\frac{H-1}{2}\right]\right)^2 + \left(v - \left[\frac{W-1}{2}\right]\right)^2} \tag{4.7}$$

热度值 h 代表扰动使目标模型分类错误的概率，而频率 τ 代表扰动的不可见性。扰动的频域特性 (u,v) 与热度值 h 和频率 τ 的函数，量化了不同频域分量的扰动对于对抗样本可迁移性和视觉保真度的价值。基于上述函数，便可以将对于可迁移对抗样本的视觉保真度的提升，转化为了在热度值 h 和频率 τ 组成的二维空间里对傅里叶矩阵中各元素的筛选。将扰动置于热度值 h 最高的区域可以提升对抗样本的可迁移性，而将扰动全部置于 τ 最高的区域内可以最小化对抗样本与原始图像的视觉差异。

聚类算法[95]是利用样本数据本身的分布规律把相似的样本聚到一起的过程，通过利用聚类算法可以在傅里叶频域上权衡公共脆弱区域，从而得到有利于对抗攻击不可见性的最优嵌入区域。将傅里叶域中敏感区域中的所有点作为样本，通过无监督学习的 K-means 算法[95]来对其进行聚类。K-means 算法基于计算样本与中心点的距离归纳各簇类下的所属样本，其目标函数如式(4.8)：

$$\arg\max_{C} J(C) = \sum_{k=1}^{K} \sum_{h^{(i)} \in C_k} \left\| c \cdot h^{(i)} - \tau^{(k)} \right\|_2^2 \tag{4.8}$$

其中 h 与 τ 的值进行了归一化，并通过系数 ρ 来设置二者的权重，调整对于样本的可迁移性与视觉保真度的偏向，这里将 ρ 设为 0.5。在聚类的过程中，首先初始化 K 个簇类中心，K 设为 2。然后对每个样本计算到簇类中心距离后归类，并重新计算簇类中心位置，直到将所有的样本分类完成。

聚类算法对敏感区域的每个频率进行了评估扰动的可迁移性与不可见性的筛选，将 DNN 在傅里叶频域上的公共脆弱区域分为热度值较低、频率较低且对于对抗攻击价值较低的区域，热度值较高、频率较高且价值较高的区域，并将

其分别命名为“中低频区”ω_{low}与“高频区”ω_{high}。应注意的是，这里的“中低频区”与“高频区”并不简单代表频率高低，而是在权衡热度值与频率之后的相对高低。

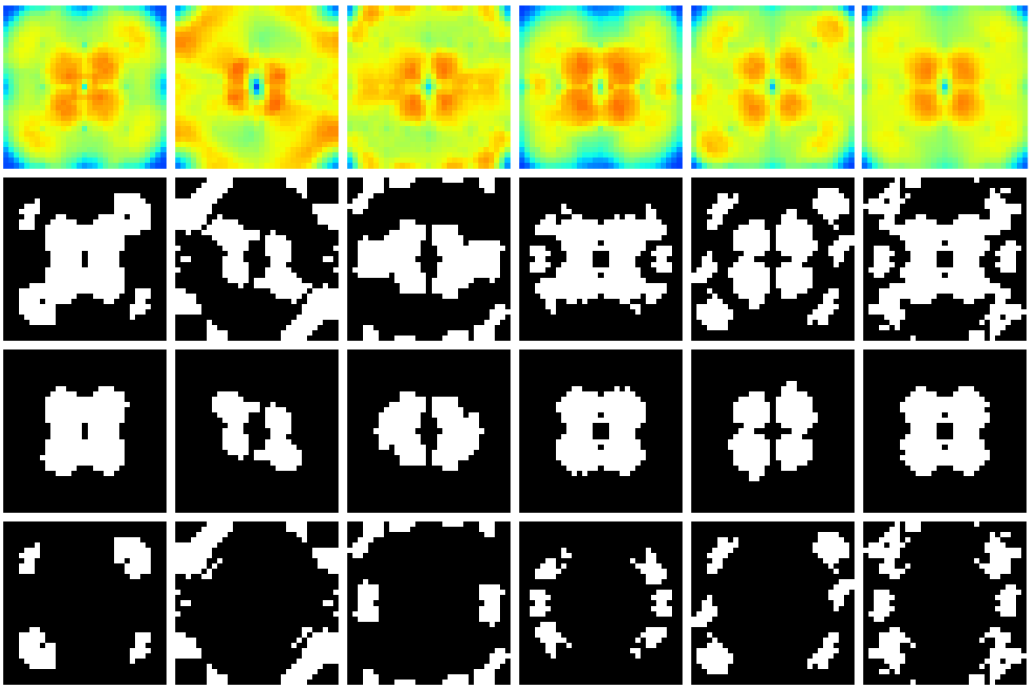


图 4.3 CIFAR10 数据集上对掩码进行聚类之后的结果示意图

图 4.3 为 CIFAR10 数据集上敏感区域的聚类结果，按行从上到下依次为傅里叶热图、敏感区域、中低频区和高频区。将不同模型以及集成之后的结果进行比较，可以看出，DNN 的中低频区和高频区的分布同样存在明显共同特性，这再次印证了上文中所提出的 DNN 的频域鲁棒性存在共同特征的观点。上述量化分析将公共脆弱区域分成了两部分，得到了在傅里叶矩阵中对于对抗攻击来说价值最高的区域，即“高频区”ω_{high}，作为对抗攻击在傅里叶频域上的优化目标。而如何通过频域约束使扰动的频域分布向该区域靠拢，将在下一节中进一步讨论。

4.2.4 频域损失

上节中的聚类分析找到了在傅里叶域中对于对抗攻击来说价值最高的区域，利用这一点可以从频域的角度优化的扰动频域分布，提升对抗攻击的效率。上

文已经提到，本算法将对抗攻击的损失函数分为对抗损失和频域损失，并且使用上一章所提出的三元损失作为对抗损失，因此本节将频域优化目标公式化作为频域损失。傅里叶域的共同敏感区域已经被分为了两个部分，也即总体上图像的傅里叶矩阵被分为：$\boldsymbol{\omega}_{low}$、$\boldsymbol{\omega}_{high}$以及非敏感区域$\boldsymbol{\omega}_{other}=\boldsymbol{E}-\boldsymbol{\omega}_{low}-\boldsymbol{\omega}_{high}$，其中$\boldsymbol{E}$为全 1 矩阵，表示为频率掩码集$\Omega\left\{\boldsymbol{\omega}_{low},\boldsymbol{\omega}_{high},\boldsymbol{\omega}_{other}\right\}$。将扰动在频域上的所有信息设为$\boldsymbol{d}_{\text{all}}$，表示为式(4.9)：

$$\boldsymbol{d}_{\text{all}}=\boldsymbol{x}'_f-\boldsymbol{x}_f=\boldsymbol{d}_{low}+\boldsymbol{d}_{high}+\boldsymbol{d}_{other} \tag{4.9}$$

这里$\boldsymbol{d}_{low}$、$\boldsymbol{d}_{high}$、$\boldsymbol{d}_{other}$分别代表扰动在低中频区，高频区和非敏感频域区域的扰动强度，在这三个频域区域内如式(4.10)所示：

$$\begin{cases}\boldsymbol{d}_{low}=\left\|\boldsymbol{\omega}_{low}\odot\left(\boldsymbol{x}'_f-\boldsymbol{x}_f\right)\right\|_2\\ \boldsymbol{d}_{high}=\left\|\boldsymbol{\omega}_{high}\odot\left(\boldsymbol{x}'_f-\boldsymbol{x}_f\right)\right\|_2\\ \boldsymbol{d}_{other}=\left\|\boldsymbol{\omega}_{other}\odot\left(\boldsymbol{x}'_f-\boldsymbol{x}_f\right)\right\|_2\end{cases} \tag{4.10}$$

对于整体扰动而言，非敏感区域上的$\boldsymbol{d}_{other}$对于对抗样本的价值是最低的，因此对其进行约束。更进一步的，在敏感区域内，基于 HVS 特性对扰动进行进一步的优化，HVS 对于$\boldsymbol{\omega}_{high}$的视觉冗余更大，对于$\boldsymbol{\omega}_{low}$的视觉冗余更小，因此扰动应该更多地集中在$\boldsymbol{d}_{high}$上。

综合上述分析，对抗攻击在频域的优化目标如式(4.11)：

$$\arg\max_{\boldsymbol{x}'_f}\mathcal{L}_{fre}\left(\boldsymbol{x}'_f,\boldsymbol{x}_f;\Omega\right)\rightarrow\begin{cases}\arg\min_{\boldsymbol{x}'_f}\boldsymbol{d}_{other}\\ \arg\min_{\boldsymbol{x}'_f}\boldsymbol{d}_{low}\\ \arg\max_{\boldsymbol{x}'_f}\boldsymbol{d}_{high}\end{cases} \tag{4.11}$$

要调整扰动的频域分布，最简单的方法是将各频域区域扰动强度的大小直接作为频域损失。但这种方法会导致过拟合的问题，如果某个频域扰动强度过大，会在反向传播过程中对对抗损失产生遮蔽，并且扰动的强度在迭代过程中会产生很大的变化，简单叠加会导致梯度不稳定。因此，使用频域区域扰动强度的比值作为损失函数，将频域损失设置为式(4.12)：

$$\frac{\partial \mathcal{L}_{fre}\left(\boldsymbol{x}'_f, \boldsymbol{x}_f; \Omega\right)}{\partial \boldsymbol{x}'_f} = \frac{\partial\left(-\frac{\boldsymbol{d}_{low}}{\left(\boldsymbol{d}_{high}+\boldsymbol{\sigma}\right)}-\frac{\boldsymbol{d}_{other}}{\left(\boldsymbol{d}_{high}+\boldsymbol{\sigma}\right)}\right)}{\partial \boldsymbol{x}'_f} \tag{4.12}$$

在优化过程中，扰动的强度可能为零，因此在分母加入微小系数$\boldsymbol{\sigma}=1e-3$，避免出现分母为零的情况另外，将$\boldsymbol{d}_{low}$与$\boldsymbol{d}_{other}$分别作为分子、$\boldsymbol{d}_{high}$作为分母来构成频域损失，并且将频域损失设定为负值，避免增大梯度。

这样，频域损失就可以通过对分式的最小化，增大$\boldsymbol{d}_{high}$来换取$\boldsymbol{d}_{other}$与$\boldsymbol{d}_{low}$的减小，从而调整扰动的频域分布。由于 DNN 对于扰动在$\boldsymbol{\omega}_{high}$内的频域分量有着较高的敏感性，减少在非敏感区域不必要的频域分量，就能以更低的扰动强度实现对抗攻击；并且将扰动更多地放在高频区$\boldsymbol{\omega}_{high}$上，能够整体地提升对抗样本对于人眼视觉的保真度。

4.3 实验结果与分析

4.3.1 实验设置

本章选择了选择经典的对抗攻击方法 I-FGSM[13]和 PGD_{inf}[21]，以及各种最先进的攻击方法：DI-FGSM[25]、S^2I-FGSM[28]、 MI-DI-FGSM[25]、TI-DI-FGSM[27]作为比较方法，从多个角度来对本章所提出的算法进行评估。

4.3.1.1 参数设置

在对三个数据集的实验中，所有的算法的参数都设置为：L_∞范数的最大扰动强度ϵ=16/255，T =10，步长α=1.6/255，N=10。

对于所提出的算法，本章使用上一章节中所提到常用的网络结构，在每个数据集上构建了$P=5$的神经网络模型$\mathcal{M}_{\mathrm{p}}$用于构造傅里叶热图，这些模型相对于迁移攻击中的黑盒模型是独立的。通过对不同网络的结果进行集成，并将敏感区域的提取比例设置为 35%，K-means 聚类中的加权系数c设为 0.5，得到敏感区域$\boldsymbol{\omega}$和高频区$\boldsymbol{\omega}_{high}$，如图 4.4 所示。

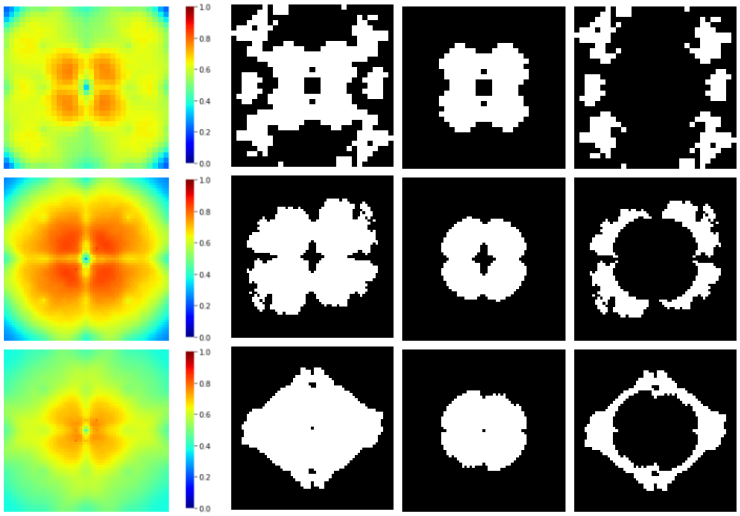


图 4.4 CIFAR10、CIFAR100 和 Tiny ImageNet 数据集上的掩码集，按行从上到下显示

4.3.1.2 评价标准

对于迁移性的评价使用同一替代模型生成对抗样本，并使用相同的黑盒模型进行测试，对于视觉保真度的评价使用相同替代模型，在相同场景下生成的对抗样本比较。

在对于对抗样本图像质量的评价上，首先 L_∞范数确保了所有方法所设定的扰动强度一致，因此本章使用了多种指标，在常用的基于简单函数的评价标准 PSNR 与 SSIM 的基础上，从样本的空间相关性、视觉信息保真度等角度对样本进行评估。另外，通过深度学习方法，本章还使用了基于人眼主观视觉感知的“感知损失（perceptual loss）”评价标准，使用学习感知图像块相似度（Learned Perceptual Image Patch Similarity, LPIPS）来衡量对抗样本的感知损失。具体使用的评价指标如下：

峰值信噪比[96]（Peak Signal to Noise Ratio, PSNR）、结构相似性指数[96]（Structural Similarity Index Measure, SSIM）、均方误差（Mean Squared Error, MSE）、通用质量图像索引[97]（Universal Quality Image Index, UQI）、视觉信息保真度[98]（Visual Information Fidelity, VIF）、以及图像感知相似度指标 LPIPS[99]。

4.3.2 不可见性分析

对抗样本视觉保真度的评估分别在非目标和目标攻击场景下进行，不同方

法生成的对抗样本的图像评价指标如表 4.1 所示。在使用 L_∞ 范数将所有攻击方法的扰动限定在同一强度的前提下，本章所提出的方法所生成的对抗样本，在各种图像评价指标上都表现出色。

表 4.1 在 Tiny-ImageNet 数据集上不同攻击方法生成的对抗样本的图像质量评价

Settings		L_2	PSNR	SSIM	VIFP	LPIPS	RMSE	UQI
Targeted Res50	PGD	3.97	28.90	0.922	0.478	0.087	0.036	0.981
	I-FGSM	2.40	33.29	0.962	0.609	0.101	0.022	0.993
	DI-FGSM	2.78	32.03	0.962	0.573	0.114	0.025	0.989
	TI-DI-FGSM	2.73	32.19	0.958	0.578	0.109	0.025	0.990
	MI-DI-FGSM	3.52	29.97	0.931	0.501	0.177	0.032	0.986
	S^2I -FGSM	3.04	31.24	0.957	0.544	0.122	0.028	0.989
	Ours	1.17	39.60	0.990	0.789	0.048	0.011	0.999
Targeted VGG19	PGD	3.97	28.91	0.922	0.478	0.095	0.036	0.981
	I-FGSM	2.30	33.68	0.971	0.627	0.102	0.021	0.992
	DI-FGSM	2.59	32.64	0.966	0.592	0.126	0.023	0.991
	TI-DI-FGSM	2.66	32.40	0.965	0.584	0.109	0.024	0.989
	MI-DI-FGSM	3.43	30.20	0.944	0.512	0.176	0.031	0.985
	S^2I -FGSM	2.86	31.78	0.961	0.565	0.137	0.026	0.990
	Ours	1.11	40.07	0.994	0.806	0.042	0.010	0.999
Untargeted Res50	PGD	4.00	28.85	0.923	0.475	0.092	0.036	0.983
	I-FGSM	2.78	32.10	0.960	0.573	0.092	0.025	0.989
	DI-FGSM	2.81	31.95	0.962	0.569	0.108	0.025	0.989
	TI-DI-FGSM	2.96	31.50	0.961	0.557	0.103	0.027	0.988
	MI-DI-FGSM	5.20	26.57	0.889	0.386	0.222	0.047	0.975
	S^2I -FGSM	3.29	30.57	0.953	0.521	0.134	0.030	0.987
	Ours	1.69	36.36	0.986	0.701	0.058	0.015	0.995
Untargeted VGG19	PGD	4.02	28.80	0.921	0.472	0.108	0.036	0.982
	I-FGSM	2.98	30.88	0.961	0.559	0.135	0.027	0.989
	DI-FGSM	3.29	30.59	0.953	0.525	0.164	0.030	0.987
	TI-DI-FGSM	3.38	30.36	0.953	0.516	0.143	0.031	0.986
	MI-DI-FGSM	5.15	26.66	0.892	0.387	0.242	0.047	0.975
	S^2I -FGSM	3.72	29.50	0.944	0.489	0.199	0.034	0.985
	Ours	2.10	34.61	0.979	0.649	0.117	0.019	0.994

从数据中可以看出，在目标攻击场景下，本章所提出的方法在各指标上均优于现有方法。与在图像质量上表现相对较好的 DI-FGSM 算法相比，对抗样本的平均 PSNR 提升了约 4-6dB，并且在所使用的所有图像质量与视觉保真度的评

价指标上，例如 VIF、LPIPS，均有明显的提升。

非目标攻击的实验数据反映了与目标攻击下相似的结果，在所有替代模型上，本章所提出的方法生成的样本，在所有的图像评价指标上都有明显的提升，平均 PSNR 提升了约 4-8dB。这说明基于频域分析的对抗攻击同样适用于非目标攻击场景，在两种攻击场景下相较于比较方法都能显著地提升对抗样本的视觉保真度。

另外，横向比较不同替代模型上的对抗样本，可以发现虽然因为网络结构的不同，对抗攻击在各模型上图像质量的表现存在部分差异，但各种攻击方法之间的性能优劣关系是一致的。这也说明对抗攻击的攻击能力不依赖于某种特定的网络结构，所提出的方法在所有的替代模型上都可以取得良好的结果。

在 CIFAR100、CIFAR10 数据集上，使用与 Tiny-ImageNet 相同的骨干网络 Res50 作为替代模型，并且评估同样在非目标和目标攻击两个场景下进行，测试对抗样本图像质量评价的实验结果如表 4.2 所示。在 CIFAR100 数据集上，相较于比较方法，本章所提出的方法在目标与非目标攻击的对抗样本的平均 PSNR 值高约 5-7dB，并且在使用的所有评价指标，例如 LPIPS 上都有显著提升；在 CIFAR10 上的结果与其他两个数据集类似，PSNR 值比比较方法的最好结果高约 5-10dB，LPIPS 等指标也有明显提升。

以上实验数据表明，在不同的攻击场景和不同的替代模型上的实验设置下，本章所提出的方法在多个数据集上生成的对抗样本在图像质量与视觉保真度上都优于比较方法。这也印证了前文中关于对抗攻击的频域分析的正确性，即通过在傅里叶域上设计对抗扰动并约束其频域分布，能够有效地提升生成对抗样本的视觉保真度，后续实验将再次证明这一点。

表 4.2 在 CIFAR100 与 CIFAR10 数据集上对抗样本的图像质量评价

Settings		L_2	PSNR	SSIM	VIFP	LPIPS	RMSE	UQI
Targeted CIFAR100	PGD	3.96	28.95	0.851	0.421	0.225	0.036	0.985
	I-FGSM	2.21	34.00	0.942	0.586	0.177	0.020	0.993
	DI-FGSM	2.36	33.46	0.939	0.568	0.199	0.021	0.992
	TI-DI-FGSM	2.48	33.02	0.936	0.549	0.172	0.022	0.991
	MI-DI-FGSM	3.19	30.82	0.898	0.479	0.273	0.029	0.989
	S^2I -FGSM	2.65	32.43	0.925	0.526	0.213	0.024	0.991
	Ours	1.18	39.56	0.985	0.747	0.087	0.011	0.997
Targeted CIFAR10	PGD	2.03	28.72	0.954	0.484	0.098	0.037	0.995
	I-FGSM	1.44	31.79	0.978	0.579	0.086	0.026	0.997
	DI-FGSM	1.84	29.59	0.965	0.511	0.136	0.033	0.995
	TI-DI-FGSM	2.00	28.87	0.962	0.504	0.133	0.036	0.994
	MI-DI-FGSM	2.74	26.13	0.928	0.405	0.212	0.049	0.990
	S^2I -FGSM	1.77	29.95	0.967	0.515	0.123	0.032	0.996
	Ours	0.62	39.22	0.995	0.755	0.021	0.011	0.999
Untargeted CIFAR100	PGD	3.97	28.93	0.850	0.420	0.228	0.036	0.985
	I-FGSM	2.20	34.05	0.943	0.588	0.173	0.020	0.992
	DI-FGSM	2.31	33.64	0.940	0.574	0.193	0.021	0.992
	TI-DI-FGSM	2.45	33.13	0.937	0.554	0.168	0.022	0.991
	MI-DI-FGSM	4.88	27.13	0.806	0.351	0.361	0.044	0.979
	S^2I -FGSM	2.73	32.18	0.923	0.518	0.230	0.025	0.990
	Ours	1.56	37.11	0.975	0.674	0.127	0.014	0.996
Untargeted CIFAR10	PGD	2.04	28.68	0.955	0.486	0.101	0.037	0.994
	I-FGSM	1.80	29.82	0.970	0.523	0.119	0.032	0.996
	DI-FGSM	1.84	29.62	0.967	0.516	0.131	0.033	0.995
	TI-DI-FGSM	2.04	28.68	0.963	0.502	0.133	0.037	0.993
	MI-DI-FGSM	2.74	26.14	0.930	0.406	0.208	0.049	0.990
	S^2I -FGSM	1.96	29.04	0.961	0.488	0.153	0.035	0.995
	Ours	1.16	33.64	0.983	0.597	0.075	0.021	0.998

4.3.3 对抗样本视觉质量比较

在上节的实验中，各种对抗样本的图像质量与视觉保真度的评价指标展示了所提出的方法对于对抗攻击不可见性的改善，本节进一步对对抗样本进行可视化评估，以分析各对抗攻击对于图像细节的影响。在三个数据集上的实验结果表明，相对于比较方法，所提出的对抗攻击可以生成相对于人类视觉系统保

真度最高的对抗样本。

4.3.3.1 图像细节比较

通过直观地比较不同方法生成的对抗样本的图像细节，来说明所提出方法对于对抗攻击不可见性的改善。与之前的实验设置相同，使用在同一替代模型、同一攻击场景下生成的样本进行比较。

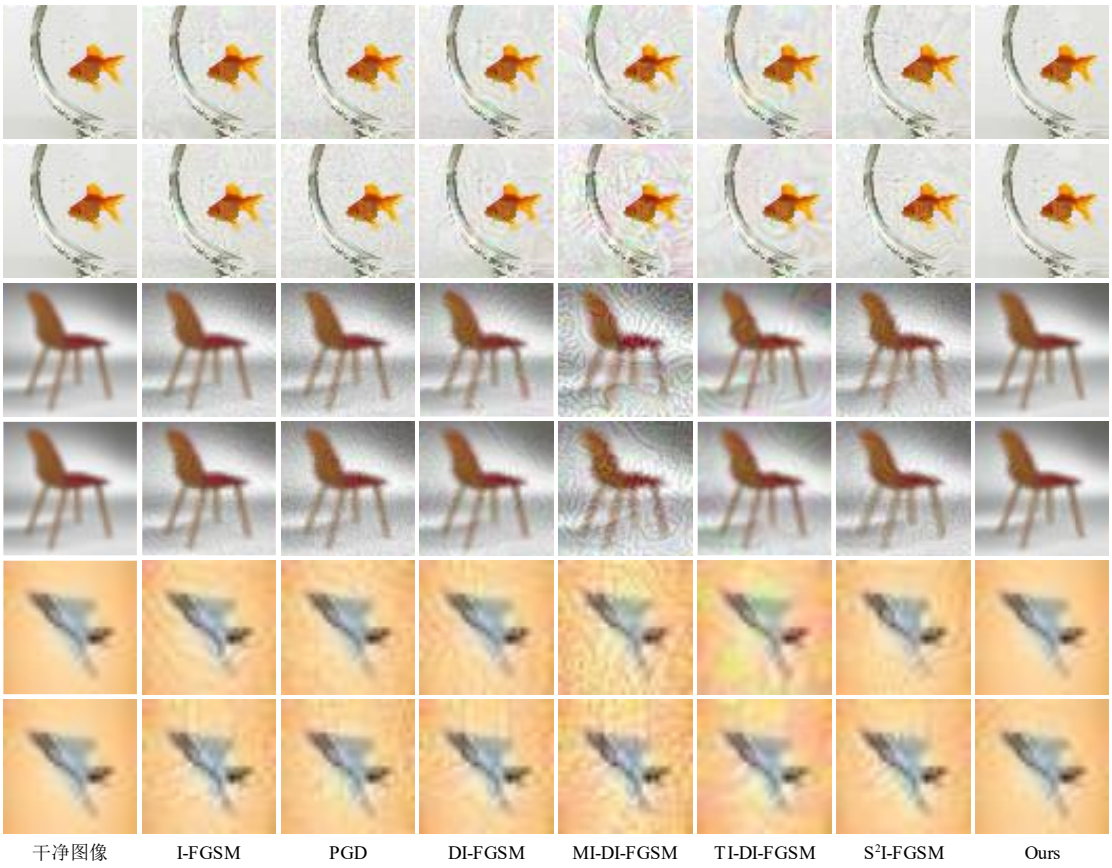


图 4.5 在三个数据集上，各攻击方法生成对抗样本的图像细节比较

图 4.5 中展示了所有方法在替代模型 Resnet50 上生成的对抗样本。这些样本从上到下每两行依次来自 Tiny-ImageNet、CIFAR100 和 CIFAR10 数据集。在每个数据集中的实验结果中，上面一行显示目标攻击场景，下面一行显示非目标攻击场景。最左一列为原始干净图像，接下来的各个样本生成方法按照以下顺序排列：I-FGSM、PGD_{inf}、DI-FGSM、MI-DI-FGSM、TI-DI-FGSM、S^2I-FGSM 和所提出的方法。

从图 4.5 中可以清晰看到，比较方法所生成的对抗扰动对样本的视觉质量的影响非常明显，并且在图像中的平滑区域留下了不自然的纹理，而本章所提出

的方法生成的对抗样本对于图像的平滑区域的影响很小，并且视觉质量明显好于其他方法，对于 HVS 来说更接近原始图像。在不同的三个数据集上，所提出方法产生的对抗样本与干净图像的视觉差异都是最小的。目标攻击与非目标攻击场景下的实验结果，反映了相同的现象，对于图像中平滑区域的扰动会更明显地影响样本的视觉质量。并且相比较之下，所提出的对抗攻击方法尤为明显地降低了图像中相对平滑的区域的影响。

4.3.3.2 差值分析

绝对差分图可以可视化不同方法所产生的对抗扰动，进一步分析攻击对于图像的影响。以非目标攻击场景下的 Tiny-ImageNet 数据集为例，将对抗样本与原始图像做差分，并取绝对值，将 RGB 三个颜色通道上的绝对差分图可视化，分别观察对抗攻击的影响。由于 L_∞ 范数的限制，所有对抗样本的最大扰动值都为 0.0627，将所有的扰动数据进行归一化处理，使得所有数值映射到 0 至 255 的范围内，其中 0 和 255 分别代表了黑色和白色，像素点越亮，则代表该点的扰动值越大。

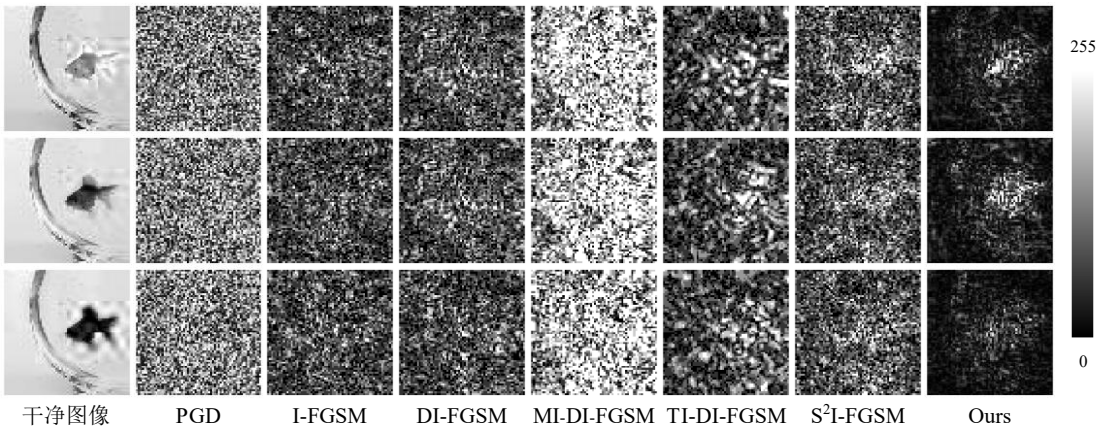


图 4.6 对抗样本与干净图像的绝对差分图，从上到下依次为 R、G、B 通道的扰动

图 4.6 中展示了比较方法和本研究所提出方法在图像三个通道的扰动方面的绝对差分图结果，为了进行公平的比较，所有的绝对差映射都被标准化为 0-255，其中 0 和 255 用黑色和白色表示。比较方法对图像的每个通道的扰动都呈现出无意义的噪声状态，每个区域的噪声强度近似是均匀的，这一点在 PGD_{inf} 的结果中尤其明显。相比之下，所提出方法在三个通道上都呈现出了明显的分

布规律，在三个通道上生成的扰动都主要集中在图像内容较为复杂、频率信息较高的区域，并且对样本图像中的低频区域影响很小。通过上述分析可以得出，所提出的方法实现了扰动对于图像内容的适应，并明显减少了对于图像中低频区域的影响。通过从频域角度的调整，扰动被转移到了人类视觉系统不敏感的高频区域，从而使得生成的对抗样本与原始图像的视觉感知差异非常小。

4.3.4 消融实验

4.3.4.1 扰动的频域定量分析

为了对对抗攻击对图像的扰动进行定量分析，验证所提出方法的有效性，本实验从傅里叶频域的角度对大量对抗样本的扰动进行统计，以分析该方法所生成的扰动具有的频域特性。对于低频区的扰动会对图像产生更显著的影响，通过统计大量对抗样本的低频扰动在总扰动强度中的占比，可以总结出攻击方法在不同频段的扰动分布规律。如果低频扰动在总强度中的占比较小，说明攻击方法的扰动主要集中在高频区域，对于对抗样本的视觉保真度的影响较小。因此，低频扰动的强度与占比可以反映对抗样本的视觉保真度，并且帮助评估频域损失的有效性。

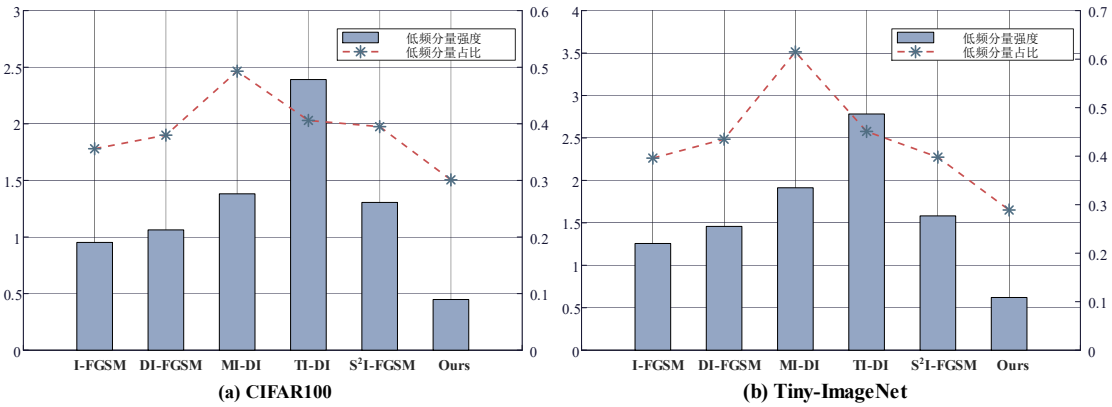


图 4.7 CIFAR100(a)和 Tiny ImageNet(b)上各方法的平均低频扰动强度和低频扰动占比

前面的讨论将图像的频域空间分成了三个部分，低中频区 ω_{low}、高频区 ω_{high}，以及非敏感区域 ω_{other}。分别将其作为滤波器来提取不同频段的扰动，并计算扰动强度。这些不同频段的扰动的 L_2 范数强度，代表了攻击方法对样本在该频段

信息的修改量，其中低频区 ω_{low} 上的扰动即为低频扰动。本实验统计了不同方法对抗扰动的平均低频扰动强度，以及低频扰动在总扰动强度中的占比。

图 4.7 中(a)和(b)分别为 Tiny-ImageNet 和 CIFAR100 数据集上不同攻击方法的频域特性，其中的直方图数据代表了对于不同方法产生的对抗样本，其低频扰动的平均强度；折线图数据代表了对于不同方法，低频扰动在总扰动强度中的占比。从直方图数据可以明显看出，所提出的方法的平均低频扰动强度是最小的，而折线图也反映出，相比于现有对抗攻击，其低频分量所占的比重同样是最低的。这说明了提出方法的有效性，成功将扰动限制在了人类视觉系统所不敏感的高频区域。

4.3.4.2 扰动的频域热点图分析

为了更进一步地分析对抗攻击在不同频段的扰动分布规律，本实验对所有样本上的扰动进行平均，并将平均扰动的频域特性可视化。

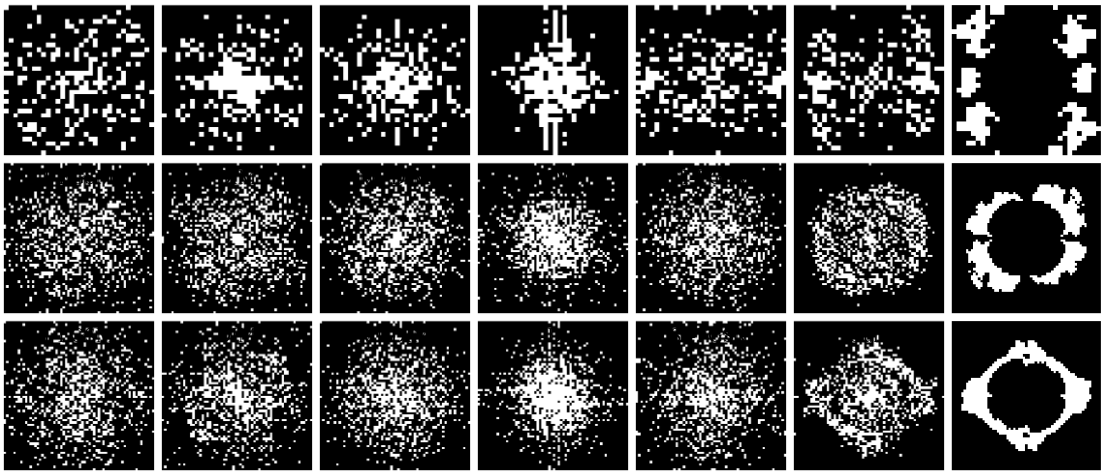


图 4.8 对抗扰动的频域扰动热点图，从上到下依次为 CIFAR10、CIFAR100、Tiny-ImageNet 数据集上的实验结果，其中最右列为图 4.4 中的高频区

采用与傅里叶热图类似的处理方法，对平均扰动进行傅里叶变换，并将所有像素点按扰动强度进行排序。然后提取强度最大的 20%的区域令其为 1 其他区域为 0，做进行二值化处理。图 4.8 展示了所有方法的频域扰动热点图，可以看出，相较于比较方法，所提出的方法成功地使得扰动的频域图契合 DNN 的高频敏感区域的分布。这进一步证明了本章中对于对抗攻击的设计有效减少了对

于图像中低频信息的影响。

4.4 本章小结

为解决现有黑盒对抗攻击不可见性不足的问题，本章从傅立叶域的角度分析了对抗攻击的过程，提出了基于频域特性分析的不可见对抗攻击。该方法通过傅里叶热图以及人类视觉系统的感知特性，量化扰动的傅里叶不同频域分量对对抗样本的可迁移性与视觉保真度产生的影响，并基于这种对应关系，通过聚类算法对扰动的频域空间进行划分，找到了可以兼顾可迁移性与不可见性的频域区域。另外，本章提出了一种频域损失来将对抗扰动的频域分布向这一目标约束，从而减低了对于样本图像质量的影响。实验证明，本章所提出的方法显著提升了黑盒对抗攻击的不可见性，减少了对抗样本与原始图像的视觉差异。

第五章 总结与展望

5.1 总结

针对现有黑盒对抗攻击方法在可迁移性与不可见性上存在的不足，以及缺乏可解释性的问题，本文从频域角度研究对抗攻击，从 DNN 的频域鲁棒性入手探究对抗攻击的频域特性，提出了两种新的黑盒对抗攻击方法。在提升黑盒对抗攻击的可迁移性与不可见性的同时，本文对于对抗攻击频域特性的分析也为之后的对抗攻击研究提供了理论基础。本文的主要工作如下：

1) 本文基于对抗攻击的频域原理研究，通过 DNN 对于不同频率上输入信号的响应情况分析 DNN 的频域鲁棒性，并基于此构造对抗攻击，提出了一种基于 DNN 频域鲁棒性的黑盒对抗攻击方法。首先通过傅里叶热图揭示了神经网络模型普遍存在的频域脆弱现象，并找到了 DNN 的公共脆弱频域。对模型输入在公共脆弱频域的扰动，可以有效地减少替代模型与黑盒模型之间的频谱差异。通过对于对抗攻击目标函数的分解，提出了一种新的三元损失函数，以使对抗扰动指向 DNN 的公共决策边界。实验证明，在 CIFAR10、CIFAR100 和 Tiny ImageNet 三个常用数据集上，相较于比较的对抗攻击，所提出方法的性能在目标攻击和非目标攻击场景中都有显著提升。

2) 基于上述对于对抗攻击频域角度的分析，更深入地探究了对抗样本的频域特性，提出了基于频域特性分析的不可见对抗攻击。综合考虑 DNN 的频域鲁棒性与 HVS 的频域特点，量化了扰动的不同频率分量对于对抗样本的可迁移性和视觉保真度的影响。运用 K-means 聚类算法，得到可以平衡对抗样本的可迁移性和视觉保真度的目标频域，在常用对抗损失的基础上引入了频域损失，以将扰动的频域分布向目标频域约束。最终在 CIFAR10、CIFAR100 和 Tiny ImageNet 三个常用数据集上，相较于比较的对抗攻击，所提出的方法成功提升了黑盒对抗攻击的不可见性。

5.2 展望

现有黑盒对抗攻击的研究往往从空间域中展开，而本研究从频域角度揭示了对抗样本内在的工作机制，为未来对抗攻击和防御技术的发展提供了参考依据和理论支撑。DNN 的共同频率敏感特性对未来对抗攻击的研究具有重要意义，进一步的探索有助于提高攻击的性能、攻击的可解释性以及对 DNN 泛化能力的理解。由于实验设备的限制，本文中用于构建公共频率敏感区域的不同结构的 DNN 是有限的，未来可通过添加结构更丰富的模型，获得更准确的 DNN 共同敏感频域，为对抗攻击的进一步研究提供支持。

另外，本文的研究，提供了一个区别传统空间域的新视角，即可以通过频率分析来探索神经网络的内在原理，未来的研究可以从以下方面进行：

1) 通过频域分析设计对抗防御机制：基于已经揭示 DNN 的频域敏感特性，可以探索利用这一特性构建新型的防御策略。例如，设计特定的频域滤波器或者在模型训练过程中加入频域正则化项，以增强模型抵抗对抗扰动的能力。通过频域防御手段，有望实现对于对抗样本的高效检测与抵御，从而提升模型的整体安全性。

2) 生成式任务的对抗攻击：考虑到生成模型在图像生成、视频生成、语音合成等领域广泛应用，从频域的角度对于生成模型的分析同样有着重要的意义。结合传统数字图像处理的自然图像频域特性的相关研究，对于生成内容的频谱分析，在确保生成内容的真实性和安全性等关键场景有着重要的意义。

3) 多模态对抗攻击的频域分析：随着跨模态学习和多模态系统的兴起，研究多模态数据（如图像、文本、音频等）的频域视角也变得尤为关键。探索不同模态间对抗攻击方法在频域特性上的共性与差异，开发能够在多个模态间适应的频域对抗方法，将有助于建立更为全面和坚固的安全防线，同时加深对神经网络在不同模态任务下内部运作机制的理解。

参考文献

[1] Turing A M. Computing machinery and intelligence[M]. Springer Netherlands, 2009.

[2] 中华人民共和国国务院．国务院关于印发新一代人工智能发展规划的通知[Z], 2017.

[3] 中华人民共和国科技部. 科技部关于支持建设新一代人工智能示范应用场景的通知[Z], 2022.

[4] 中华人民共和国国家数据局. “数据要素×”三年行动计划(2024—2026 年)[Z], 2024.

[5] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 26-30, 2016, Las Vegas, NV, USA. Los Alamitos: IEEE, 2016: 770-778.

[6] Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(1): 142-158.

[7] Xu B, Lu C, Guo Y, et al. Discriminative multi-modality speech recognition [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. Los Alamitos: IEEE, 2020: 14433-14442.

[8] Goel K, Gu A, Donahue C, et al. It’s raw! audio generation with state-space models [C]// Proceedings of International Conference on Machine Learning, July 17-23, 2022, Baltimore, MD, USA. New York: ACM, 2022: 7616-7633.

[9] Grigorescu S, Trasnea B, Cocias T, et al. A survey of deep learning techniques for autonomous driving[J]. Journal of Field Robotics, 2020, 37(3):362-386.

[10] 阮一凡.人工智能技术在金融风控中的应用研究[J].商展经济,2024(07):89-92.

[11] Jin C, Yu H, Ke J, et al. Predicting treatment response from longitudinal images using multi-task deep learning[J]. Nature Communications, 2021, 12(1):1851.

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

[13] Kurakin A, Goodfellow I. J, Bengio S. Adversarial examples in the physical world [C]//Proceedings of Artificial Intelligence Safety and Security, July 13-14, 2018, Stockholm, Sweden. New York:ACM, 2018: 99-112.

[14] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, UT, USA. Los Alamitos: IEEE, 2018: 1625-1634.

[15] Xie C, Wang J, Zhang Z, et al. Adversarial examples for semantic segmentation and object detection[C]//Proceedings of IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. Los Alamitos: IEEE, 2017: 1369-1378.

[16] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. Arxiv preprint Arxiv:1503.02531, 2015.

[17] Maiya SR, Ehrlich M, Agarwal V, et al. A frequency perspective of adversarial robustness[J]. Arxiv preprint Arxiv:2111.00861, 2021.

[18] Goodfellow I.J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. Arxiv preprint Arxiv:1412.6572, 2014.

[19] Jin G, Shen S, Zhang D, et al. Ape-gan: Adversarial perturbation elimination with gan[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, May 12-17, 2019, Brighton, UK. Los Alamitos: IEEE, 2019: 3842-3846.

[20] Xu ZQ, Zhang Y, Luo T, et al. Frequency principle: Fourier analysis sheds light on deep neural networks[J]. Communications in Computational Physics, 2020, 28(5): 1746-1767.

[21] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. Stat, 2017: 1050-1059.

[22] Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 26-30, 2016, Las Vegas, NV, USA. Los Alamitos: IEEE, 2016: 2574-2582.

[23] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[J]. IEEE Symposium on Security and Privacy, 2017: 39-57.

[24] Chen PY, Zhang H, Sharma Y, et al. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models [C]//Proceedings of ACM workshop on Artificial Intelligence and Security, October 30, 2017, Dallas, TX, USA. New York: ACM, 2017: 15-26.

[25] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, UT, USA. Los Alamitos: IEEE, 2018: 9185-9193.

[26] Xie C, Zhang Z, Zhou Y, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. Los Alamitos: IEEE, 2019: 2730-2739.

[27] Dong Y, Pang T, Su H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. Los Alamitos: IEEE, 2019: 4312-4321.

[28] Long Y, Zhang Q, Zeng B, et al. Frequency domain model augmentation for adversarial attack[C]//Proceedings of European Conference on Computer Vision, October 23-28, 2022, Tel Aviv, Israel. New York: Springer: 549-566.

[29] Lin J, Song C, He K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[J]. Arxiv preprint Arxiv:1908.06281, 2019.

[30] Zou J, Pan Z, Qiu J, et al. Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting[C]//Proceedings of European Conference on Computer Vision, August 23-28, 2020, Glasgow, UK. New York: Springer, 2020: 563-579.

[31] Wang X, He K. Enhancing the transferability of adversarial attacks through variance tuning[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 19-25, 2021, Virtual Event. Los Alamitos: IEEE, 2021: 1924-1933.

[32] Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5):828-841.

[33] Jia X, Wei X, Cao X, et al. Adv-watermark: A novel watermark perturbation for adversarial examples[C]//Proceedings of ACM International Conference on Multimedia, October 12-16, 2020, Seattle, WA, USA. New York: ACM, 2020: 1579-1587.

[34] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[J]. Arxiv preprint Arxiv:1712.04248, 2017.

[35] Xiao C, Li B, Zhu JY, et al. Generating adversarial examples with adversarial networks[C]//Proceedings of International Joint Conference on Artificial Intelligence, October 12-16, 2018, Stockholm, Sweden, July 13-19, 2018. California: IJCAI, 2018: 3905-3911.

[36] Fawzi A, Moosavi-Dezfooli SM, Frossard P. Robustness of classifiers: from adversarial to random noise[C]//Proceedings of Advances in Neural Information Processing Systems, December 5-10, 2016, Barcelona, Spain. La Jolla: NIPS, 2016(29).

[37] 丁康一. 基于迁移性的对抗样本生成与利用方法研究[D].电子科技大学,2023.

[38] 李洛寒. 可迁移对抗样本生成方法的研究与实现[D].北京邮电大学,2024.

[39] Poursaeed, Omid, et al. Generative adversarial perturbations[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, UT, USA. Los Alamitos: IEEE, 2018: 4422-4431.

[40] Naseer, M. M., Khan, S. H., Khan, et al. Cross-domain transferability of adversarial perturbations[C]//Proceedings of Advances in Neural Information Processing Systems, December 8-14, 2019, Vancouver, BC, Canada. La Jolla: NIPS, 2019(32).

[41] Naseer, M, Khan, S, Hayat, M, et al. On generating transferable targeted perturbations[C]//Proceedings of IEEE International Conference on Computer Vision, June 19-25, 2021, Virtual Event. Los Alamitos: IEEE, 2021: 7708-7717.

[42] Hinton G E. Learning multiple layers of representation[J]. Trends in Cognitive Sciences, 2007, 11(10):428-434.

[43] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey[J]. IEEE Access, 2018: 14410-14430.

[44] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning[C]//Proceedings of ACM on Asia Conference on Computer and

Communications Security, April 2-6, 2017, Abu Dhabi, UAE. New York: ACM, 2017: 506-519.

[45] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings[J]. IEEE European Symposium on Security and Privacy, 2016: 372-387.

[46] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses[J]. Arxiv preprint Arxiv:1705.07204, 2017.

[47] Xu Y, Zhong X, Yepes AJ, et al. Grey-box adversarial attack and defence for sentiment classification[J]. Arxiv preprint Arxiv:2103.11576, 2021.

[48] Singh A, Sikdar B. Adversarial attack and defence strategies for deep-learning-based IoT device classification techniques[J]. IEEE Internet of Things Journal, 2021, 9(4):2602-2613.

[49] Levy M, Amit G, Elovici Y, et al. The security of deep learning defences for medical imaging[J]. Arxiv preprint Arxiv:2201.08661, 2022.

[50] Gittings T, Schneider S, Collomosse J. Vax-a-net: Training-time defence against adversarial patch attacks[C]//Proceedings of Asian Conference on Computer Vision, November 1-5, 2020, Taipei, Taiwan. New York: Springer, 2020.

[51] Zhang H, Yu Y, Jiao J, et al. Theoretically principled trade-off between robustness and accuracy [C]// Proceedings of International Conference on Machine Learning, June 9-15, 2019, Long Beach, CA, USA. New York: ACM, 2019: 7472-7482.

[52] Xiao C, Deng R, Li B, et al. Characterizing adversarial examples based on spatial consistency information for semantic segmentation [C]//Proceedings of European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. New York: Springer, 2018: 217-234.

[53] Dhillon GS, Azizzadenesheli K, Lipton ZC, et al. Stochastic activation pruning for robust adversarial defense[J]. Arxiv preprint Arxiv:1803.01442, 2018.

[54] Feinman R, Curtin R R, Shintre S, et al. Detecting adversarial samples from artifacts[J]. Arxiv preprint Arxiv:1703.00410, 2017.

[55] Metzen JH, Genewein T, Fischer V, et al. On detecting adversarial perturbations[J]. Arxiv preprint Arxiv:1702.04267, 2017.

[56] Xie C, Wu Y, Maaten LV, et al. Feature denoising for improving adversarial robustness[C]//Proceedings of IEEE Conference on Computer Vision and Pattern

Recognition, June 15-20, 2019, Long Beach, CA, USA. Los Alamitos: IEEE, 2019: 501-509.

[57] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning[C]//Proceedings of International Conference on Machine Learning, June 19-24, 2016, New York, NY, USA. PMLR, 2016: 1050-1059.

[58] Shen D, Wang G, Wang W, et al. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms [J]. Arxiv preprint Arxiv:1805.09843, 2018.

[59] Jia J, Cao X, Wang B, Gong NZ. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing[J].Arxiv preprint Arxiv:1912.09899, 2019.

[60] Naseer M, Khan S, Hayat M, et al. A self-supervised approach for adversarial robustness[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Virtual Event. Los Alamitos: IEEE, 2020: 262-271.

[61] Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations[J]. Arxiv preprint:1711.00117, 2017.

[62] Xie C, Wang J, Zhang Z, et al. Mitigating adversarial effects through randomization[J]. Arxiv preprint Arxiv:1711.01991, 2017.

[63] Dziugaite G, Ghahramani Z, Roy D. A study of the effect of jpg compression on adversarial images[J]. Arxiv preprint Arxiv:1608.00853, 2016.

[64] Jia X, Wei X, Cao X, et al. Comdefend: An efficient image compression model to defend adversarial examples[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. Los Alamitos: IEEE, 2019: 6084-6092.

[65] Liao F, Liang M, Dong Y, et al. Defense against adversarial attacks using high-level representation guided denoiser[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, UT, USA. Los Alamitos: IEEE, 2018: 1778-1787.

[66] Samangouei P, Kabkab M, et al. Defense-gan: Protecting classifiers against adversarial attacks using generative models[C]//Proceedings of International

Conference on Learning Representations, April 30 - May 3, 2018, Vancouver, BC, Canada. OpenReview.net, 2018.

[67] Goodfellow I.J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 22;63(11):139-44.

[68] Luo T, Ma Z, Xu ZQ, et al. Theory of the frequency principle for general deep neural networks[J]. CSIAM Transactions on Applied Mathematics, 2021, 2 (3): 484-507.

[69] Zhang C, Benz P, Karjauv A, et al. Universal adversarial perturbations through the lens of deep steganography: Towards a Fourier perspective[C]//Proceedings of AAAI Conference on Artificial Intelligence 2021, February 2-9, 2021, Virtual Event. AAAI Press, 2021(35): 3296-3304.

[70] Tsuzuku Y, Sato I. On the structural sensitivity of deep convolutional networks to the directions of Fourier basis functions[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. Los Alamitos: IEEE, 2019: 51-60.

[71] Yin D, Gontijo Lopes R, Shlens J, et al. A Fourier perspective on model robustness in computer vision[J]. Advances in Neural Information Processing Systems, 2019(32).

[72] Wang H, Wu X, Huang Z, et al. High-frequency component helps explain the generalization of convolutional neural networks[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Virtual Event. Los Alamitos: IEEE, 2020: 8684-8694.

[73] Sharma, Yash, GW Ding, and Marcus Brubaker. On the effectiveness of low frequency perturbations[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, August 10-16, 2019, Macao, China. California: IJCAI, 2019: 3389–3396.

[74] Zhang Q, Zhang C, Li C, et al. Practical no-box adversarial attacks with training-free hybrid image transformation[J]. CoRR, 2022, abs/2203.04607.

[75] Li M, Deng C, Li T, et al. Towards transferable targeted attack[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Virtual Event, June 13-19, 2020. Los Alamitos: IEEE, 2020: 641-649.

[76] Xu ZQ, Zhang Y, Xiao Y. Training behavior of deep neural network in frequency domain[C]//Proceedings of Neural Information Processing, December 8-14, 2019, Vancouver, BC, Canada. La Jolla: NIPS, 2019: 264-274.

[77] Cooley JW, Tukey JW. An algorithm for the machine calculation of complex Fourier series[J]. Mathematics of Computation, 1965,19(90):297-301.

[78] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J], Handbook of Systemic Autoimmune Diseases, 2009, 1(4).

[79] Brendel W, Rauber J, Kurakin A, et al. Adversarial vision challenge[J]. The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations, 2020: 129-153.

[80] Anton H, Rorres C. Elementary linear algebra: applications version[M]. John Wiley & Sons, 2013.

[81] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. Arxiv preprint Arxiv:1409.1556, 2014.

[82] Huang G, Liu Z, et al. Densely connected convolutional networks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, July 21-26, 2017, Honolulu, HI, USA. Los Alamitos: IEEE, 2017: 4700-4708.

[83] Zagoruyko S, Komodakis N. Wide residual networks[C]//Proceedings of British Machine Vision Conference 2016, September 19-22, 2016, York, UK. British Machine Vision Association, 2016.

[84] Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. Los Alamitos: IEEE, 2017: 618-626.

[85] Sharif M, Bauer L, Reiter MK. On the suitability of lp-norms for creating and preventing adversarial examples[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, June 18-22, 2018, Salt Lake City, UT, USA. Los Alamitos: IEEE, 2018: 1605-1613.

[86] Wang Z, Song M, Zheng S, et al. Invisible adversarial attack against deep neural networks: An adaptive penalization approach[J]. IEEE Transactions on Dependable and Secure Computing, 2019,18(3):1474-88.

[87] Sheatsley R, Hoak B, Pauley E, et al. The space of adversarial strategies[J]. 32nd USENIX Security Symposium, 2023: 3745-3761.

[88] Ding, X, Zhang, S, Song, M, et al. Toward invisible adversarial examples against DNN-based privacy leakage for Internet of Things [J]. IEEE Internet of Things Journal, 2020, 8(2): 802–812.

[89] Zhang, Y, Tan, Y, Sun, H, et al. Improving the invisibility of adversarial examples with perceptually adaptive perturbation [J]. Information Sciences, 2023, 635: 126–137.

[90] Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system[J]. Nature, 1996, 381(6582):520-522.

[91] Shen X, Ni Z, Yang W, et al. Just noticeable distortion profile inference: A patch-level structural visibility learning approach[J]. IEEE Transactions on Image Processing, 2020, 30:26-38.

[92] Mannos, J, and Sakrison, D. The effects of a visual fidelity criterion of the encoding of images [J]. IEEE Transactions on Information Theory, 1974, 20(4), 525–536.

[93] Daly, S. J. Visible differences predictor: an algorithm for the assessment of image fidelity [J]. Human Vision, Visual Processing, and Digital Display III, 1992, 1666: 2–15.

[94] Yang, Z, et al. Trs: Transferability reduced ensemble via promoting gradient diversity and model smoothness[C]//Proceedings of Advances in Neural Information Processing Systems, December 6-14, 2021, Virtual Event. La Jolla: NIPS, 2021: 17642-17655.

[95] MacQueen J. Classification and analysis of multivariate observations [C]//Proceedings of Berkeley Symposium on Mathematical Statistics and Probability, June 20-July 21, 1967, Berkeley, CA, USA. Berkeley: University of California Press, 1967: 281-297.

[96] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[97] Wang Z, Bovik A C. A universal image quality index[J]. IEEE Signal Processing Letters, 2002, 9(3): 81-84.

[98] Sheikh H R, Bovik A C. Image information and visual quality[J]. IEEE Transactions on Image Processing, 2006, 15(2): 430-444.

[99] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, UT, USA. Los Alamitos: IEEE, 2018: 586-595.

作者在攻读硕士学位期间公开发表的学术成果

[1] **Li C**, Liu Y, Zhang X, Wu H. Exploiting Frequency Characteristics for Boosting the Invisibility of Adversarial Attacks[J]. Applied Sciences, 2024, 14(8):3315. (SCI: 001210174800001)

[2] Wu H, **Li C**, Liu G, Zhang X. Hiding data hiding[J]. Pattern Recognition Letters, 2023(165): 122-127. (SCI: 000971084500001)

致 谢

江水流长，终归大海。时光荏苒，转瞬间研究生生涯即将告一段落。在此，我要向研途中给予我帮助和支持的人表示衷心的感谢。

首先，由衷地感谢我的导师张新鹏教授。承蒙老师的悉心教导，让我能够顺利完成研究生学业。张老师有着深邃的学术洞见，给我带来了很多思考问题方式的启发。另外，我还要感谢同样为我的学业提供极大帮助的吴汉舟老师，吴老师在我的科研工作中提出的建设性意见使我不断进步，也让我在学术探索中获得了宝贵的经验和成长。

其次，我想感谢师兄师姐、同门。刘勇师兄的无私分享，为我指引了研究方向，帮我避免了很多学术上可能的问题。在这段学习时光里，我和同门们朝夕相处，在科研工作和日常生活中互相建议、互相鼓励、共同进步。

然后要感谢我的父母家人和朋友，感谢在我学业和生活上一以贯之的支持。父母的温暖和鼓励如同静谧夜空中的月光，无声却强大。家是永远的港湾，无论航行至何方，心中始终有归处。还要特别感谢我的女朋友，在这段充满挑战与不确定的路上，感谢你给予宝贵的帮助和支持。很幸运与你相伴，希望能相互扶持，一起走过接下来的人生旅程。并肩行万里，齐心照远途。

雄关漫道真如铁，而今迈步从头越。再次感谢所有陪伴和支持我的人，毕业不是结束，而是新旅程的开始，尽管人生的道路不尽相同，惟愿我们都对自己的生活真诚以待，坚定而平和地迎接每一天的阳光。

最后感谢各位评审专家在百忙之中抽出宝贵的时间审阅我的论文，提出宝贵的意见。