引用格式：张准, 曾逸, 刘启和, 等. 一种频率驱动的黑盒对抗攻击方法 [J]. 电子科技大学学报, 2026, 55(2): 263-274.
ZHANG Z, ZENG Y, LIU Q H, et al. A frequency-driven black box adversarial attack method[J]. Journal of University of Electronic Science and Technology of China, 2026, 55(2): 263-274.

一种频率驱动的黑盒对抗攻击方法

张 准，曾 逸，刘启和*，叶 飞，周世杰

（电子科技大学信息与软件工程学院，成都 610054）

摘要：深入理解对抗样本的特性对保障机器学习模型安全具有重要意义。针对现有对对抗性扰动与频率成分关系认识不足的问题，对对抗性扰动在频率域中的表征进行了研究，并提出一种高效的黑盒对抗攻击方法。通过小波包分解技术对对抗样本进行多尺度频率分解，发现对抗性扰动主要集中于低频段的高频成分。为此设计了一种结合特定频段信息的黑盒对抗攻击算法，并引入归一化扰动可见性指数（NDV）以解决传统范数在评估连续和离散扰动时的局限性。在多个基准数据集和模型上的实验表明，该多频带组合攻击方法平均攻击成功率达 99%，优于单一频段攻击方法，并在 7 项评估指标上表现出优越的综合性能。此外，验证了 NDV 指标能够有效克服传统L_2范数在扰动评估中的不足。

关键词：黑盒对抗攻击；频域；机器学习；小波包分解
中图分类号：TP37 **文献标志码：**A DOI: 10.12178/1001-0548.2024336

A frequency-driven black box adversarial attack method

ZHANG Zhun, ZENG Yi, LIU Qihe*, YE Fei, and ZHOU Shijie

(School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)

Abstract: Enhancing the understanding of adversarial examples is crucial for ensuring the security of machine learning models in real-world applications. To address the insufficiency of existing research on the relationship between adversarial perturbations and their frequency components, this work investigates the representation of adversarial perturbations in the frequency domain and proposes an efficient black-box adversarial attack method. By applying wavelet packet decomposition to perform multi-scale frequency analysis of adversarial examples, it is found that adversarial perturbations are predominantly concentrated in the high-frequency components within low-frequency bands. Based on this observation, we design a black-box attack adversarial algorithm that incorporates specific frequency band information and introduce a normalized disturbance visibility (NDV) index to overcome the limitations of traditional norm-based metrics when evaluating both continuous and discrete perturbations. Experiments conducted on multiple benchmark datasets and models show that the proposed multi-band composite attack achieves an average success rate of 99%, significantly outperforming single-band attack approaches and demonstrating superior performance across seven evaluation metrics. Moreover, the NDV index effectively addresses the shortcomings of traditional norms, offering a more accurate and perceptually meaningful assessment of adversarial perturbations.

Key words: black box adversarial attack; frequency domain; machine learning; wavelet packet decomposition

随着深度学习技术的不断发展，以计算机视觉、序列处理、智能决策等技术为核心的研究在安防、交通、医疗等领域开展了广泛应用[1]，并取得了卓越的表现。然而，现有深度神经网络存在较为普遍的脆弱性，通过将微小的对抗扰动（adversarial perturbation）添加到正常样本中生成对抗样本（adversarial examples），可以使模型以较高的置信度产生误判，而这些扰动并不能影响人眼的判断[2-4]。对抗样本的出现严重阻碍了深度学习模型在安全敏感领域的进一步应用。如在自动驾驶系统

收稿日期：2024−12−11
基金项目：四川省自然科学基金（25NSFSC1269）
作者简介：张准，博士生，主要从事人工智能安全对抗方面的研究。
*通信作者 E-mail: qiheliu@uestc.edu.cn

中，对抗样本攻击可以将停车路标更改为直行信号，导致车载智能感知系统做出错误的判断，造成严重的交通事故。

为了构建可信赖的人工智能系统，理解对抗样本至关重要。已有研究表明[2,5]，从信号处理的频率角度分析对抗样本是一种有效的方法。在这种分析框架中，图像的低频成分包括了如平滑背景和统一颜色区域等全局信息，对整体视觉感知具有决定性影响。而高频成分则包含纹理、边缘或快速变化的噪声等细节信息，这些细节对于物体的识别和图像的精确表现极为重要。深入了解这些频率成分如何在对抗性扰动的生成和应用中起作用，将有助于开发更有效的防御策略，增强人工智能系统的鲁棒性。

然而，对抗样本的频域表征仍是一个争议的话题[6]，文献 [2] 的研究显示深度神经网络（deep neural networks, DNNs）对图像中高频元素的反应极为敏感。文献 [7-8] 的研究进一步证明了高频信号对 DNNs 预测结果的显著影响，而文献 [9-10] 的研究则表明低频扰动同样可以影响模型性能。此外，文献 [11] 发现，在 ImageNet 上训练的 DNN 更倾向于识别图像中的高频信息而不是低频信息，但是在特定数据集上训练的 DNNs 也可以学习以低频为基础的表征。文献 [5] 的研究结果表明，对抗性扰动既不完全处于高频成分中，也不完全处于低频成分中。因此，对抗样本与频率成分之间的具体关系仍不明确，这需要在频域内进一步探索。

针对上述问题，本文研究了频域中对抗性扰动与频率成分之间的关系，通过小波包分解（wavelet packet decomposition, WPD）[12] 将对抗扰动从图像域转换到频域进行分析。与离散余弦变换（discrete cosine transform, DCT）[13] 或小波分解（wavelet decomposition, WD）[14] 相比，WPD 能够更精确地捕捉局部频率特征的变化。分析结果表明，对抗性扰动主要集中在低频段下的高频部分。基于此，本文提出了一种基于频率分解的黑盒对抗攻击算法，此算法能够将扰动灵活地添加到任意频段组合中。在对 3 个主流数据集和 3 种模型的实验中，通过结合低频段和低频段的高频组件，此算法在攻击效果上明显优于基于单一频段的攻击方法，同时此算法在攻击成功率、不可察觉性等方面优于其他方法。此外，本文研究发现，L_2 范数在评估对抗样本不可察觉性时存在局限性，尤其是在评估连续和离散扰动时，无法准确反映人类的感知差异。为此，本文引入了归一化扰动可见性（normalized disturbance visibility, NDV）指数，通过对 L_2 范数进行优化，提出了一种更符合人类视觉感知的评估指标，从而更准确地量化对抗样本的可见性。

1 相关工作

1.1 对抗样本

对抗样本的概念最早由文献 [15] 提出，揭示了深度学习模型的脆弱性。随后，文献 [2] 进一步通过利用模型的梯度信息，在图像中引入人类视觉不可察觉的微小修改，这些细微的变化足以误导分类算法，导致其产生错误预测。根据对目标模型的了解程度，对抗性攻击可分为白盒攻击和黑盒攻击两类[16]。白盒攻击假设攻击者完全了解目标模型的结构和参数，使其能够利用梯度信息生成高度精确的扰动[17-18]。黑盒攻击进一步分为基于查询的攻击和基于迁移的攻击两种类型。基于查询的攻击只能通过输入来查询目标模型，并从输出中得到的反馈来设计对抗样本[19-20]。文献 [21] 首次提出了基于决策的黑盒攻击，该攻击不需要知道目标模型的信息，使其成为一种强大的黑盒攻击。文献 [22] 利用历史搜索信息来自适应调整步长，提出了 SSA 方法。文献 [23] 则提出一种基于零阶优化的攻击方法，在保证攻击成功的前提下兼顾了不可察觉性。与此相对的是基于迁移的攻击，这类攻击利用对抗样本的迁移性，将在替代模型上生成的对抗样本直接用于攻击目标模型，无需与目标模型交互即可实现攻击[24-25]。然而，与基于查询的攻击相比，基于迁移的攻击面临着替代模型构建困难以及迁移攻击的成功率较低等问题[26]。

1.2 频域视角的对抗样本

近年来，许多研究从频域角度对深度神经网络的特性进行了深入分析。文献 [27] 提出了通过研究 DNNs 对不同傅里叶基的敏感性来构建频域框架，该框架为探索频域与对抗样本之间的关系奠定了基础。在此基础上，文献 [9] 设计了一种针对图像低频成分的创新对抗攻击策略，展示了频域方法在生成高效对抗样本方面的潜力。随后，文献 [10] 通过实验进一步验证了这一攻击方法的有效性，并表明该策略对加固了对抗防御的模型仍具有显著的攻击效果。同时，文献 [28-29]的研究将频域策略应用于对抗样本的检测，进一步拓展了频域分析的应用范围。这些研究不仅揭示了频域分析在理解 DNNs 行为和对抗样本生成中的重要性，还为基于频率特征的攻击和防御策略的设计提供了新的思路

和技术路径[30-31]。

此外，一些研究进一步探讨了 DNNs 与图像频率成分之间的关系。文献 [32] 指出 DNNs 对高频成分的敏感性会降低模型的鲁棒性，这一发现为许多预处理防御方法奠定了理论基础。然而，这种高频敏感性的观点与针对低频成分的对抗攻击策略形成了矛盾。文献 [33] 提出，对抗样本的频率特性并非单纯归属于高频或低频，而是与特定数据集密切相关。这一观点揭示了对抗扰动特性存在数据依赖性，进一步丰富了对抗样本特性的理论理解。同时，文献 [34] 从频域的角度分析了图像的频率分布特性，指出频率成分是模型鲁棒性的重要内在特征，表明频域分析在理解和提升模型鲁棒性方面具有关键作用。2023 年，文献 [5] 通过设计同时利用低频和高频信息的攻击方法，提高了攻击效果。这表明学习行为与频率信息之间的关系可能比此前预想的更为复杂，强调了频域分析在理解 DNN 行为和设计更加高效的攻击与防御方法中的重要性。这些研究表明有必要进一步探索学习行为与不同频率信息之间的关系，从而为改进模型鲁棒性提供新的方向和技术路径。

2　对抗性扰动的频率分析

2.1　对抗扰动生成

为研究对抗性扰动的特性，首先需要生成相应的破坏性输入。快速梯度下降法（fast gradient sign method, FGSM）[2] 和 投影梯度下降法（projected gradient descent, PGD）[17] 是生成对抗样本的经典方法，并被广泛应用于评估深度学习模型的鲁棒性。因此，研究中采用这两种方法生成对抗样本，以更全面地分析其特性。具体地，给定图片数据集 $X=\{x_1,x_2,\cdots,x_n\}$，其对应的标签为 $Y=\{y_1,y_2,\cdots,y_n\}$，使用 FGSM 和 PGD 方法得到对抗样本集 $X_{\text{adv}}=\{X_{\text{FGSM}},X_{\text{PGD}}\}$。FGSM 攻击的实施过程如式 (1) 所示，因此可生成对抗样本集 $X_{\text{FGSM}}=\{x_{\text{FGSM}1},x_{\text{FGSM}2},\cdots,x_{\text{FGSM}n}\}$。

$$x_{\text{FGSM}}=x+\varepsilon\,\text{sign}\left(\nabla_x J(x,y)\right) \tag{1}$$

式中，ε为控制扰动大小的参数；$\text{sign}\left(\nabla_x J(x,y)\right)$表示损失函数 $J(\cdot)$相对于输入图像 x的梯度符号，其决定了扰动的方向。PGD 攻击则通过执行式 (2) 的迭代更新生成对抗样本集 $X_{\text{PGD}}=\{x_{\text{PGD}1},x_{\text{PGD}2},\cdots,x_{\text{PGD}n}\}$。

$$x_{\text{PGD}}^{t+1}=P_{x,\varepsilon}\left(x_{\text{PGD}}^{t}+\alpha^{t}\,\text{sign}\left(\nabla_x J(x,y)\right)\right) \tag{2}$$

式中，α^t为迭代 t时的步长；x_{PGD}^t为迭代 t时的对抗样本；$P_{x,\varepsilon}(\cdot)$为将添加扰动后的样本投影到 x的 ε–球上，确保输入在合理的范围内。

2.2　频率分解

WPD 是一种通过多层次分解从信号中提取详细信息和近似信息的技术。与小波分析仅对低频部分进行分解不同，WPD 能够对所有频率成分进行迭代分解，从而实现更加全面和精细的频率层次划分。通过该方法，图像被分解为一系列频率段，每个频率段分别对应图像的不同特征成分[35]。其中，低频段主要反映图像的整体亮度、颜色分布和基本结构等全局特征；而高频段则捕捉图像中的边缘、纹理等局部细节信息。

由于图像的多层次分解通常需要较高的计算资源，为平衡计算复杂度与分析精度，本文采用一维 WPD 对图像进行频率分解。该方法通过将图像转化为一维信号进行处理，显著降低了计算成本，同时保留了对频率特征的有效捕捉能力。设 $\varphi(t)$为尺度函数，$\psi(t)$为小波函数，令 $\psi_0(t)=\varphi(t)$且 $\psi_1(t)=\varphi(t)$，则有：

$$\begin{cases}\varphi_{2l}(t)=\sqrt{2}\sum_{k\in Z}h_k\varphi_l(2t-k)\\ \psi_{2l+1}(t)=\sqrt{2}\sum_{k\in Z}g_k\varphi_l(2t-k)\end{cases} \tag{3}$$

式中，h_k为低通滤波器；g_k为高通滤波器，且 $g_n=(-1)^n h_{1-n}$。WPD 的递推分解过程为：

$$\begin{cases}f_l^{j,2n}=\sum_{k\in Z}h_{k-2l}f_k^{j+1,n}\\ f_l^{j,2n+1}=\sum_{k\in Z}g_{k-2l}f_k^{j+1,n}\end{cases} \tag{4}$$

式中，$f_l^{j,2n}$和 $f_l^{j,2n+1}$是由 $f_l^{j+1,n}$分解得到的小波包系数。将 $\{X,X_{\text{adv}}\}$通过式 (4) 分解至频域，分别得正常样本与两种对抗样本的小波包系数（即频段），如式 (5) 所示。

$$\begin{cases}F=\{f_1,f_2,\cdots,f_n\}\\ F_{\text{FGSM}}=\{f_{\text{FGSM}1},f_{\text{FGSM}2},\cdots,f_{\text{FGSM}n}\}\\ F_{\text{PGD}}=\{f_{\text{PGD}1},f_{\text{PGD}2},\cdots,f_{\text{PGD}n}\}\end{cases} \tag{5}$$

为便于表示，高频分量记为 d，低频分量记为 a。一维 WPD 遵循二值分解过程，经过初始分解后，原始图像被分成第 1 层低频段 a和高频段 d。随后对 a和 d进行分解，在第 2 层得到 4 个不同的频段。这个过程继续进行，第 2 层的频段被进一步分解，在第 3 层产生 8 个频段。其中，$\{aaa\}$频段保留了原始图像的核心属性，如颜色、亮度、总体结构等。相反，其他 7 个高频频段与边缘细节相关。经过 n次分解，得到 2^n个频段。得到如图 1 所示的

树状结构。这些分解的频段可以使用式 (6) 重新组合。

$$f_l^{j+1,n}=\sum_{k\in Z}[h_{l-2k}f_k^{j,2n}+g_{l-2k}f_k^{j,2n+1}] \tag{6}$$

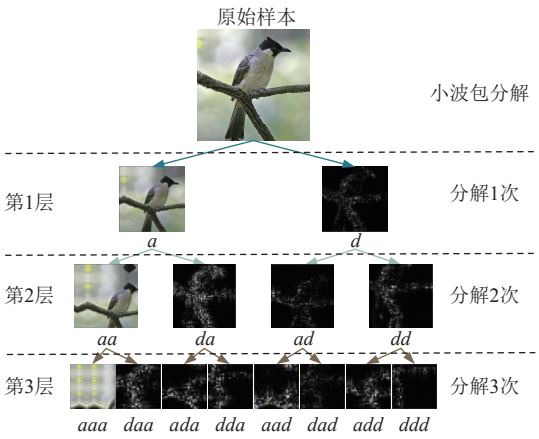


图 1 图像的一维小波包分解

2.3 频率分析

通过 WPD 将正常图像及其对应的对抗样本分解为一系列频率带。随后，计算对抗样本与干净样本在各个频率段中的余弦相似度，以量化两者之间的差异，从而分析对抗性扰动在不同频率段的分布情况。对于图像 $x\in X$ 及其对抗样本 $x_{\text{adv}}\in\{X_{\text{FGSM}}, X_{\text{PGD}}\}$，通过 WPD 将 x 分解为频率带集合 $F=\{f_1, f_2,\cdots,f_n\}$，将 x_{adv} 分解为 $F'=\{f_1',f_2',\cdots,f_n'\}$。随后，根据式 (7) 计算其余弦相似度：

$$\cos\theta=\frac{|f_i\cdot f_i'|}{\|f_i\|_2\|f_i'\|_2} \tag{7}$$

实验结果如图 2 所示，余弦相似度越低，表示对应频段的扰动越大。在第 1 次分解（第 1 层）中，对抗性扰动主要集中于高频分量{*d*}，该现象与以往研究中将对抗扰动视为高频成分的结论一致[32]。在第 2 次分解（第 2 层）中，这些扰动进一步集中于{*da*}频段，而在第 3 次分解（第 3 层）中，相似度在{*daa*}频段中进一步下降。同时，从高频分量{*d*}分解而来的低频分量{*ad*}的子频段{*dad*}也表现出显著的相似度降低。这些结果表明，对抗性扰动可能主要分布在低频段中的高频成分，其生成过程依赖于前一级的低频信息。此外，在第 3 次分解中，高频段{*ddd*}的余弦相似度也较低，这表明对抗性扰动也可能具有高频特性。该结果反映了对抗性扰动在频率分布上的复杂性，既涉及高频特征，又与低频特征密切相关。更多细节见第 5.2 节。

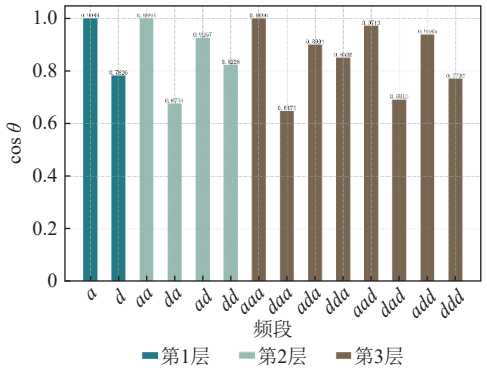


图 2 对抗样本和干净样本在不同频段间的平均余弦相似度

3 基于频率分解的黑盒对抗攻击

本节提出了一种基于频率分解的黑盒对抗攻击方法。攻击框架如图 3 所示。该方法主要包括小波包分解及频段选择与攻击实施两部分。在小波包分解部分，首先利用小波包分解将图像分解为不同频段，随后基于扰动频域分析结果，选择易于生成对抗样本的频段组合用于扰动添加。在攻击实施部分，通过构建正交方向向量、设计加扰策略以及限制扰动范围等步骤，实现对抗样本的生成。

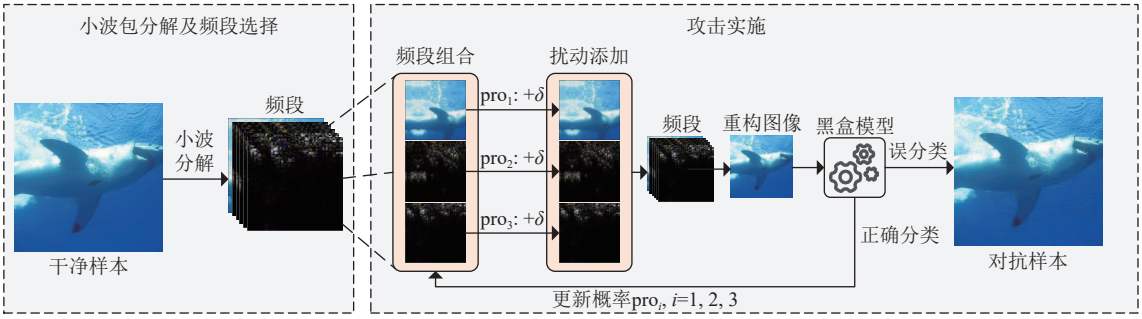


图 3 基于频率分解的黑盒对抗攻击方法框架

3.1 小波包分解及频段选择

3.1.1 小波包分解方式

在图像处理与对抗扰动分析中，二维小波包分解能够通过多级分解有效提取图像在水平、垂直和对角方向的局部频域特征，尤其在细节保留方面具有优势。然而，该方法计算复杂度较高，每一层分

解涉及多方向的卷积运算，导致计算量随分解层数增加而显著增长，使得深层分解在实际应用中难以高效实现。相比之下，一维小波包分解通过将图像转换为一维信号处理，大幅降低了计算负担，便于进行多层级分解。尽管其在空间方向性特征的精细度上有所降低，但仍能有效提取图像的全局频域信息。因此，本文采用一维小波包分解作为频域分析方法，在特征表达能力和计算效率之间取得平衡。

3.1.2　频段选择

如上节所述，对抗性扰动主要集中在由低频段分解而来的特定高频段中。特别是在第 3 层分解中，由{*aa*}和{*ad*}分解得到的{*daa*}和{*dad*}频段表现出更低的相似度，这表明这些频段中对抗性扰动的分布更加密集。低频段{*aaa*}包含图像的核心信息，如颜色、亮度和结构光度，而其他 7 个频段则主要反映边缘和纹理特征。这种差异表明，{*aaa*}频段承载了更重要的信息内容。如图 1 所示，信息含量较低的区域以黑色（值为 0）表示。因此，将低频段{*aaa*}与低频段的高频成分{*daa*,*dad*}结合作为主要搜索区域是一种有效的选择。关于这 3 个频段的消融实验的详细内容，详见第 5.4 节。

3.2　攻击实施

受文献 [20] 启发，本文提出了基于频率分解的一种黑盒对抗攻击算法，相关伪代码如下。

基于频率分解的黑盒对抗性攻击

输入：干净样本x，对应标签y，参数n，步长ε

输出：对抗样本x_{adv}

过程：

初始化$x_{\text{adv}}=x$，通过式 (5) 将x_{adv}分解为频段集W；

选择$W_s\subseteq W$；

初始化概率Pro，即选择$w\subseteq W_s$的概率；

查询目标模型获得置信度$p=\text{ph}(y|x_{\text{adv}})$；

While $p_y=\max_{y'}p_{y'}$ do

  For $\alpha\in[-\varepsilon,\varepsilon]$ do

    根据概率Pro，抽取$w\subseteq W_s$；

    初始化大小为w的零矩阵$\boldsymbol{v}$；

    在$\boldsymbol{v}$中随机选取n个点赋值为 1；

    $w=w+\alpha\boldsymbol{v}$；

    重构扰动后的W，用式 (6) 得到x'；

    If $\text{ph}(y|x')<p$ then

      $p=\text{ph}(y|x')$；

      $x_{\text{adv}}=x'$；

      根据p更新概率Pro；

    End If

  End For

End While

具体而言，该方法首先初始化对抗样本x_{adv}，并通过一维小波包分解（WPD）将x_{adv}分解为频段集W。再选择频段组合$W_s\subseteq W$，并为每个频段施加初始概率Pro。随后，查询目标模型以获得当前对抗样本的置信度p。若置信度未达到攻击目标，算法通过调整频段的扰动大小，依据概率Pro选择频段并随机生成扰动$\alpha\boldsymbol{v}$，将其叠加至频段中以生成新的扰动图像x'。若x'的置信度低于原始置信度p，则更新对抗样本x_{adv}，并根据置信度p调整概率Pro，该过程迭代进行，直至成功生成满足攻击目标的对抗样本。下面将介绍此方法的关键步骤，包括正交方向向量选择、加扰搜索策略以及扰动限制。

3.2.1　正交方向向量选择

为了有效减少搜索次数，确保在所选频段集$W_s=\{aaa,daa,dad\}$中的所有频段相互正交至关重要。小波包分解提供了一种基于正交多分辨率分析（orthogonal multi-resolution analysis, OMRA）的强大框架，允许将信号分解成多个频段。根据小波理论[12]，信号空间$L^2(R)$通过尺度空间V_j和细节空间W_j的分解实现，其中V_j可以进一步分解为$W_{j+1}\oplus V_{j+1}$。在该框架下，每个尺度空间V_j包含一个更细的尺度空间V_{j+1}和相应的细节空间W_{j+1}，这两个子空间在$L^2(R)$中是正交的。

进一步地，整个空间分解为$L^2(R)=\oplus_{j\in Z}W_j$，表明$W_j$彼此正交，且共同构成了整个$L^2(R)$空间。每个$W_j(j>1)$可以继续分解为$W_j=\oplus_{n\in Z}U_j^n$，这种分解的递归性质通过对每个$W_j$应用相同的分解过程产生的两个正交的小波包空间，确保了每一级的小波包都是正交的，满足：

$$\begin{cases}U_j^{2n}\perp U_j^{2n+1}\\ U_{j+1}^{2n}=U_j^{2n}\oplus U_j^{2n+1}\end{cases}\tag{8}$$

式中，$j\in Z$。对于任意n和m，小波包U_j^n和U_j^m（$n\neq m$）满足正交性，即它们的交集为空。这种正交性不仅适用于同一级别的小波包，也适用于不同级别的小波包之间。此外，为确保频段的正交性，在小波函数的选择上还需满足小波和尺度函数的正交性以及小波函数的尺度和平移正交性。

最终，经过 3 次分解后得到的{*aaa*,*daa*,*ada*,*dda*,*aad*,*dad*,*add*,*ddd*}相互正交。所以，任选不重复的频段$w \subseteq W_s$相互正交，从w中选择的不重复向量$\boldsymbol{v}$亦具有正交性，其中$\boldsymbol{v}$为定义的正交方向向量。

3.2.2 加扰搜索策略

为最大限度地减少扰动总量，首先为W_s中的每个频段初始化一个概率$\mathrm{Pro}=\{\mathrm{pro}_i \mid i \in [1,\mathrm{len}(W_s)]\}$。然后根据概率Pro从频率带$W_s$中选择目标频段$w$。接着，初始化尺寸为$w$的零矩阵$\boldsymbol{v}$，在此矩阵中随机选择$n$个点赋值为 1，并确保矩阵$\boldsymbol{v}$中的点不被重复选择，因此每个$\boldsymbol{v} \in w$都是相互正交的向量。之后，选取$\alpha \in [-\varepsilon,\varepsilon]$，通过$\alpha\boldsymbol{v}$向频段$w$中添加扰动，由于$w \subseteq W_s$且 $W_s \subseteq W$，这些扰动实际被添加到了W中。通过式 (6)，将添加扰动后的W'重构为图像x'。将x'输入目标模型进行查询，扰动的添加可能会降低$p=\mathrm{ph}(y \mid x')$。通过不断迭代，当p足够低时，目标模型就会产生误分类。每次查询得到的p也会用来更新概率Pro，在哪个频段添加扰动后下降更多，则相应增加该频段的选择概率pro_i，同时仍要保证$\sum \mathrm{pro}_i = 1$。为确保查询效率，算法避免了任何两个方向的向量相互抵消。

3.2.3 扰动限制

通过调整步长ε和参数n来控制扰动的总量，其中n决定了扰动矩阵的生成，步长ε则控制每个点的扰动大小。在每次迭代中，根据搜索方向的符号$\alpha \in [-\varepsilon,0,\varepsilon]$，选择对$n$个点进行加法、减法或丢弃操作，具体取决于输出概率是否在某个方向上减少。经过t步迭代后，扰动在频域中的总量满足式 (9)。此外，在实际操作中，可以根据需求修改这两个参数，即在查询受限的情况下，适当增大ε和n可以提高攻击的成功率，并减少查询次数，而在扰动受限的情况下，减小ε和n有助于生成更加难以察觉的扰动。

$$\delta_t = \sum_{t=1}^{T} \alpha_t q_t \tag{9}$$

4 归一化干扰可见度指数

L_2范数通常用于量化原始样本与对抗样本之间的扰动幅度，从而评估对抗扰动的不可察觉性。该指标反映了扰动对原始输入的整体影响，因此能够保证生成的对抗样本对人类观察者来说不可察觉或仅具轻微可察觉性。然而，L_2范数在评估同时包含集中扰动和离散扰动的对抗样本时存在局限性，即当扰动的总量相同时，L_2范数无法有效区分不同分布方式的扰动，即集中的扰动往往在局部区域内引起较大变化，容易被发现，而离散扰动则将扰动分散在整个区域，从而相对隐蔽。

图 4 展示了 3 种扰动在相同L_2范数约束下的扰动对比，其中第 1 组为 PGD 攻击[17]，第 2 组为本文提出的方法，第 3 组为补丁扰动。为了更清晰地展示扰动效果，本文在展示时将扰动大小扩大了 10 倍。结果表明，尽管 3 种攻击产生的扰动具有相同的L_2范数值，补丁攻击生成的扰动（见图 4 第 3 组）相比其他两种攻击更容易被人眼察觉。因此，L_2范数在这种情况下可能无法全面反映其对观察者的实际影响。

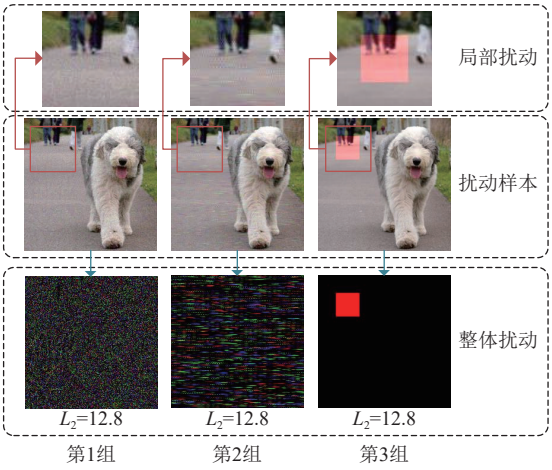


图 4 3 种扰动的可视化

为更准确地反映人类视觉感知，本文提出了归一化扰动可见性指数（normalized disturbance visibility index, NDV），用于优化L_2范数。对于干净样本x及其对应的对抗样本x_{adv}，NDV 的计算如式 (10) 所示。该指标通过将L_2范数除以受到扰动影响的点数进行归一化，并在分母中引入小常数γ以避免不可导问题。同时，为了实现尺度标准化，最终计算结果乘以常数 C（默认值为 1 000）。NDV 用于量化扰动的可见性，类似于L_2范数，较大的 NDV 值表示扰动更显著，而较小的 NDV 值则表示扰动较轻微，从而更贴近人类的视觉感知。

$$\mathrm{NDV} = C\frac{\|x - x_{\mathrm{adv}}\|_2}{\|x - x_{\mathrm{adv}}\|_0 + \gamma} \tag{10}$$

5 实验

5.1 数据集与模型

本文选择了 CIFAR-10(C10)[36]、CIFAR-100(C100)[36] 和 ImageNet-1K(IN)[37]3 个数据集进行实

验。在模型选择方面，依照文献 [38-41] 的研究，对ResNet-50(R50)[38]、VGG16(Vgg)[42] 和InceptionV3 (V3)[41] 模型在 CIFAR-10[36] 和 CIFAR-100[36] 数据集上进行了标准化训练。对于 ImageNet-1K[37] 数据集，本文采用了 Pytorch[43] 提供的预训练模型，包括 ResNet-50(R50)[38]、VGG16(Vgg)[42]、InceptionV3 (V3)[41] 和 DenseNet-121(Den)[39]，以确保模型的一致性和可靠性。为了保证测试结果的真实性，并避免人为因素提高攻击成功率，本研究遵循文献 [20] 的方法，从 CIFAR-10[36]、CIFAR-100[36] 和 ImageNet-1K[37] 的验证集中，分别随机选取 1 000 张初始正确分类的图像作为每个模型的测试数据集。

5.2　频域中对抗扰动的分布

5.2.1　频率分解

在图像的频域分解过程中，首先需要将图像转换为一维信号。由于选用数据集为彩色图像，采用通道拼接方法进行处理。具体而言，将 RGB 的 3 个通道像素值按照行优先顺序展平，并依次拼接为一个长向量。这种将二维图像数据有效转换为一维信号的处理方式为后续频域分析奠定了基础。小波包分解采用 Python 的 pywt 库[40] 实现，此库提供了一维小波包分解功能，能够将信号递归分解为多个频段，并支持对各频带进行细化分析。实验选用 Coiflet 小波族中的 coif17 作为基函数，此小波凭借其高阶特性、平滑性和优异的频域分辨率，能够有效捕捉图像的高频细节和边缘特征。coif17 在处理图像及对抗样本的频域特征时具有一定优势。

5.2.2　扰动频域分析

本实验考虑了 FGSM(F)[2] 和 PGD(P)[17] 两种攻击方法。为了评估对抗性扰动在各频段中的分布情况，实验通过计算对抗样本与正常样本的余弦相似度，并对比不同模型、攻击方法和数据集下的表现。具体结果见表 1，其中相似度较低的频段已加粗标注，便于展示和分析。

表 1　基于余弦相似度的对抗扰动的频域分布

数据集	模型	攻击	第1层		第2层				第3层							
			{*a*}	{*d*}	{*aa*}	{*da*}	{*ad*}	{*dd*}	{*aaa*}	{*daa*}	{*ada*}	{*dda*}	{*aad*}	{*dad*}	{*add*}	{*ddd*}
C10	Den	F	0.998	**0.791**	0.999	**0.687**	0.936	0.824	0.999	**0.677**	0.913	0.857	0.976	**0.690**	0.948	0.765
		P	0.999	**0.856**	0.999	**0.764**	0.960	0.883	0.999	**0.752**	0.945	0.906	0.984	**0.769**	0.967	0.839
	R50	F	0.998	**0.798**	0.999	**0.698**	0.936	0.830	0.999	**0.690**	0.913	0.860	0.975	**0.700**	0.947	0.775
		P	0.999	**0.864**	0.999	**0.777**	0.962	0.890	0.999	**0.764**	0.947	0.912	0.985	**0.782**	0.968	0.848
	V3	F	0.999	**0.738**	0.999	**0.560**	0.945	0.827	0.999	**0.496**	0.920	0.864	0.983	**0.607**	0.958	0.763
		P	0.999	**0.852**	0.999	**0.717**	0.969	0.904	0.999	**0.671**	0.955	0.925	0.989	**0.747**	0.975	0.862
	Vgg	F	0.998	**0.759**	0.999	**0.616**	0.940	0.819	0.999	**0.575**	0.913	0.853	0.980	**0.642**	0.953	0.759
		P	0.999	**0.847**	0.999	**0.725**	0.965	0.891	0.999	**0.686**	0.949	0.912	0.988	**0.749**	0.973	0.849
C100	Den	F	0.998	**0.788**	0.999	**0.695**	0.928	0.817	0.999	**0.677**	0.899	0.843	0.973	**0.703**	0.941	0.766
		P	0.999	**0.850**	0.999	**0.764**	0.954	0.876	0.999	**0.747**	0.936	0.897	0.982	**0.772**	0.962	0.834
	R50	F	0.998	**0.808**	0.998	**0.726**	0.923	0.831	0.999	**0.707**	0.893	0.850	0.972	**0.734**	0.938	0.792
		P	0.998	**0.860**	0.999	**0.786**	0.951	0.880	0.999	**0.770**	0.931	0.896	0.982	**0.793**	0.961	0.846
	V3	F	0.998	**0.746**	0.999	**0.584**	0.944	0.830	0.999	**0.525**	0.922	0.867	0.979	**0.627**	0.954	0.767
		P	0.999	**0.845**	0.999	**0.712**	0.963	0.894	0.999	**0.665**	0.949	0.917	0.985	**0.741**	0.968	0.849
	Vgg	F	0.998	**0.753**	0.999	**0.612**	0.939	0.814	0.999	**0.568**	0.914	0.850	0.978	**0.639**	0.951	0.751
		P	0.999	**0.850**	0.999	**0.732**	0.967	0.893	0.999	**0.693**	0.952	0.915	0.988	**0.754**	0.973	0.849
IN	Den	F	0.998	**0.688**	0.998	**0.598**	0.845	0.721	0.999	**0.576**	0.804	0.754	0.927	**0.607**	0.860	0.660
		P	0.999	**0.765**	0.999	**0.669**	0.921	0.802	0.999	**0.642**	0.886	0.834	0.972	**0.683**	0.935	0.742
	R50	F	0.998	**0.662**	0.999	**0.582**	0.845	0.690	0.999	**0.569**	0.791	0.725	0.940	**0.585**	0.869	0.628
		P	0.999	**0.765**	0.999	**0.674**	0.923	0.799	0.999	**0.652**	0.888	0.831	0.974	**0.686**	0.938	0.739
	V3	F	0.998	**0.692**	0.998	**0.610**	0.831	0.719	0.999	**0.587**	0.789	0.744	0.914	**0.618**	0.844	0.664
		P	0.999	**0.766**	0.999	**0.669**	0.911	0.801	0.999	**0.640**	0.876	0.828	0.966	**0.683**	0.923	0.742
	Vgg	F	0.998	**0.668**	0.998	**0.578**	0.848	0.702	0.999	**0.555**	0.801	0.734	0.936	**0.591**	0.868	0.642
		P	0.999	**0.760**	0.999	**0.666**	0.922	0.799	0.999	**0.639**	0.886	0.830	0.973	**0.681**	0.936	0.741
	均值		0.998	**0.782**	0.999	**0.675**	0.926	0.822	0.999	**0.647**	0.899	0.850	0.971	**0.691**	0.938	0.770

表 1 的数据显示，无论是 CIFAR-10、CIFAR-100 还是 ImageNet-1K 数据集，在不同模型和攻击方法下，扰动的频域分布呈现出一致的趋势。具体来说，正常样本与对抗样本在{*da*}、{*daa*}和{*dad*}频段中的余弦相似度普遍显著降低。如在 CIFAR-10 数据集上，ResNet-50 模型在 FGSM 攻击下，{*da*}频段的余弦相似度仅为 0.698；在 PGD 攻击下，余弦相似度则为 0.777。类似的趋势也出现在其他

模型和数据集上，尤其是在高频段，如{*dad*}和{*daa*}频段，扰动的影响更加明显。这表明，对抗性扰动主要集中在低频段中的高频成分。此外，尽管在某些频段中扰动的分布特性较为明显，但即使通过更换小波函数或进一步细化频率带的分解，仍难以完全分离对抗性扰动。这些结果表明，现有的小波函数在分离对抗性扰动方面仍存在一定局限性。因此，为了更有效地分离这些扰动，需要开发专门针对对抗性扰动特性的新型小波函数，以提高分离效果。

5.3 基于频率分解的黑盒攻击方法的评估

本节对本文提出的方法在多个数据集和模型上的表现进行了全面评估，并与当前的对抗攻击方法进行了对比分析。所有实验均采用无目标攻击形式进行。在评估指标方面选择了 7 项指标：3 项评估攻击性能、4 项评估扰动水平。符号↑表示数值越大性能越优，↓表示数值越小性能越优。在性能评估和比较过程中，选取了攻击成功率（attack success rate, ASR↑）、平均查询次数（average numberofqueries,ANQ↓）、查询次数中位数（mediannumber of queries, MNQ↓）作为主要指标，计算方式如下：

$$\begin{cases} \mathrm{ASR} = \dfrac{N_{\mathrm{suc}}}{N} \\ \mathrm{ANQ} = \dfrac{1}{N_{\mathrm{suc}}} \displaystyle\sum_{i=1}^{N_{\mathrm{suc}}} Q_i \\ \mathrm{MNQ} = \mathrm{median}(Q_1, Q_2, \cdots, Q_{N_{\mathrm{suc}}}) \end{cases} \tag{11}$$

式中，N表示总样本数；N_{suc}为攻击成功的样本数；Q_i为第i个样本的查询次数；median(·)表示取中位数操作。

在扰动不可见约束方面引入了L_2范数（↓）、L_∞范数（↓）、结构相似性（structural similarity, SSIM↑）以及提出的归一化扰动可见性指数（NDV↓），前 3 种指标的计算方式如下：

$$\begin{cases} L_2 = \sqrt{\sum (x - x_{\mathrm{adv}})^2} \\ L_\infty = \max |x - x_{\mathrm{adv}}| \\ \mathrm{SSIM} = \dfrac{(2\mu_x \mu_{x_{\mathrm{adv}}} + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_{x_{\mathrm{adv}}}^2 + c_1)(\sigma_x^2 + \sigma_{x_{\mathrm{adv}}}^2 + c_2)} \end{cases} \tag{12}$$

式中，x为干净样本；x_{adv}为其对应的对抗样本；μ_x和$\mu_{x_{\mathrm{adv}}}$分别为x和x_{adv}的图像平均值；σ_x^2和$\sigma_{x_{\mathrm{adv}}}^2$是对应图像的方差；$\sigma_{xy}$是$x$和$x_{\mathrm{adv}}$之间的协方差；$c_1$和$c_2$为常数。

5.3.1 在不同模型上的评估

通过对 CIFAR-10（C10）和 CIFAR-100（C100）数据集上训练良好的 DenseNet-121（Den）、ResNet-50（R50）和 InceptionV3（V3）模型进行评估，验证了本文提出的攻击方法的有效性。具体实验结果如表 2 所示，本文方法在所有模型上均实现了高于 99% 的攻击成功率。特别是针对 InceptionV3 模型，平均每次成功攻击仅需数十次查询，突显了此方法在攻击效率和稳定性方面的优势。进一步分析表 2 中的数据发现，针对不同模型的 ANQ 和 MNQ 指标的比较显示，MNQ 均小于 ANQ。这表明大多数样本的成功攻击可以在相对较少的查询次数内完成，显示了攻击方法在处理典型样本时的高效性和可靠性。然而，少数样本可能因为复杂性或者防御措施较为严密，需要较多的查询次数才能成功攻击，导致了平均查询次数的上升。此外，扰动评估指标包括L_2、L_∞和 NDV，它们随着查询次数的增加而上升。这一趋势进一步验证了所提出的 NDV 指标在评估攻击效果和扰动大小时的有效性。

表 2 在不同模型上的评估

数据集	模型	评估指标						
		ASR	ANQ	MNQ	L_2	L_∞	SSIM	NDV
C10	Den	0.95	219.6	176	2.27	0.18	0.99	0.74
	R50	0.99	235.0	180	2.35	0.19	0.99	0.76
	V3	1.00	80.1	54	1.23	0.13	0.99	0.40
C100	Den	0.99	179.4	137	2.06	0.18	0.99	0.67
	R50	0.99	216.2	154	2.24	0.19	0.99	0.73
	V3	0.99	59.2	42	1.12	0.12	0.99	0.37

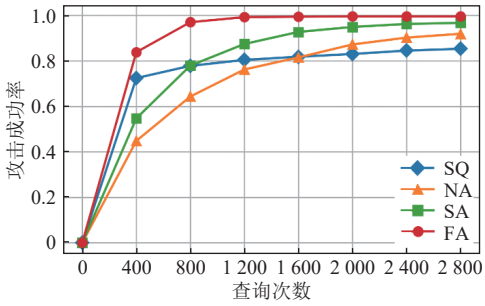
5.3.2 与先进方法的比较

在 ImageNet 数据集以及 DenseNet-121（Den）和 ResNet-50（R50）模型的环境下，本实验将所提出的基于频率分解的黑盒攻击方法（FA）与当前先进的黑盒攻击算法进行了比较，包括 Boundary 攻击（BD）[21]、GeoDA 攻击（GA）[44]、Square 攻击（SQ）[19]、N 攻击（NA）[45] 和 SimBA（SA）[20]。其中，SA 方法利用离散余弦变换将图片转换至频域，并在低频部分添加扰动，在频域攻击中性能优异。结果如表 3 所示，FA 方法在攻击成功率、扰动不可察觉性和查询效率 3 个方面的 7 项综合指标中具有优异的表现。具体而言，针对 Den 和 R50模型，FA 方法的平均查询次数分别为 380 次和 426 次，显著低于其他算法，如 BA 攻击需要高达 9 077 次和 9 820 次查询。这表明在相同的攻击成功率下，FA 方法显著降低了查询的需求，提升了整体的攻击效率。此外，FA 方法在扰动度量上同样表现出优越性，其L_2和L_∞扰动明显

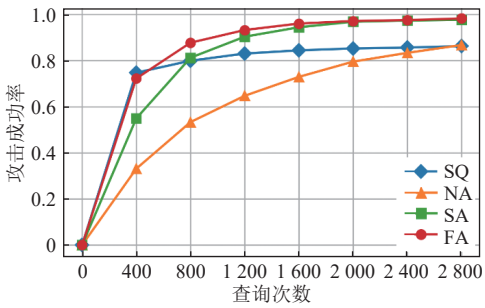
低于 BA 和 GA 等方法，同时保持了较高的攻击成功率。这种优势表明，FA 方法能够在保证攻击效果的同时，有效控制扰动的大小，减少对图像质量的影响，进而提升了对抗样本的隐蔽性。图 5 展示了查询次数与攻击成功率之间的关系，结果表明，FA 方法在查询次数较少的情况下能够获得较高的攻击成功率，进一步验证了其在提升攻击效率的同时，不会牺牲攻击效果。总体而言，FA 方法在综合性能上超越了其他常见的对抗攻击方法，此方法有着广泛的应用潜力。

表 3　在 ImageNet 上与先进方法的比较

方法	模型	评估指标						
		ASR	ANQ	MNQ	L_2	L_∞	SSIM	NDV
BD	Den	1.00	9 077	5 700	13.98	0.11	0.78	0.09
	R50	1.00	9 820	7 922	13.98	0.11	0.79	0.09
GA	Den	1.00	3 350	4 078	32.99	0.35	0.87	0.22
	R50	1.00	3 213	3 329	30.98	0.33	0.87	0.21
SA	Den	0.98	508	341	4.59	0.05	0.99	0.03
	R50	0.98	560	349	4.65	0.05	0.99	0.03
SQ	Den	0.89	366	86	13.66	0.71	0.94	0.15
	R50	0.89	429	71	13.64	0.71	0.94	0.15
NA	Den	0.88	834	613	10.84	0.05	0.94	0.07
	R50	0.93	663	409	11.39	0.05	0.94	0.07
FA	Den	0.99	380	183	12.53	0.29	0.96	0.08
	R50	0.99	426	184	12.85	0.29	0.96	0.08



a. ResNet-50模型上的结果



b. DenseNet-121模型上的结果

图 5　攻击成功率与查询次数的关系

5.4　消融实验

本研究选择了{*aaa*}、{*daa*}和{*dad*}这 3 个频段作为搜索空间。为了方便表示，用 A 代表{*aaa*}，B 代表{*daa*}，C 代表{*dad*}。具体而言，分别使用这 3 个频段及其组合的结果作为搜索空间，即算法中W_s的选择，并实施攻击，相关结果见表 4。

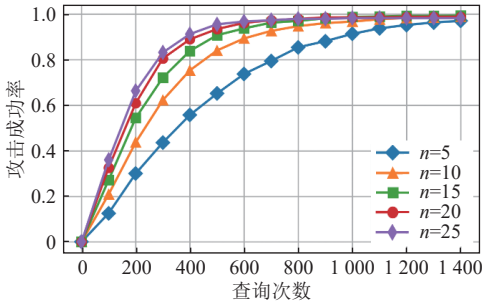
表 4　不同频段组合的攻击评估

频段	ASR	ANQ	MNQ	L_2	L_∞	SSIM	NDV
A	0.95	181.74	153	1.79	0.12	0.99	0.59
B	0.63	263.20	251	2.20	0.14	0.99	0.72
C	0.82	219.00	192	1.94	0.14	0.99	0.63
AB	0.98	271.47	216	2.66	0.21	0.99	0.87
AC	0.98	216.16	166	2.27	0.17	0.99	0.74
BC	0.90	264.38	198	2.42	0.20	0.99	0.79
ABC	0.99	235.07	180	2.35	0.19	0.99	0.76

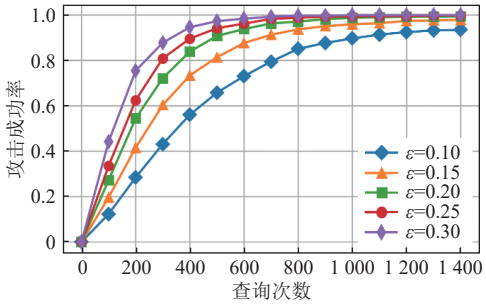
频段 A 表现出最高的单频段攻击成功率，表明其在对抗攻击中具有关键作用，与文献 [20] 的研究一致。同时，频段组合 AB 和 AC 的攻击成功率均高于单频段，表明整合频率带 B 和 C 能够有效提升攻击效果。此外，ABC 组合的攻击成功率最高，显著优于所有单一频段和双频段组合，在其他评估指标（如不可察觉性和查询效率）上几乎没有显著变化。这进一步证明了在对抗攻击中采用多频段整合策略具有显著优势。

除了频段的选择，本文方法还涉及参数n和步长ε两个关键超参数。为了进一步探讨不同超参数设置对攻击效果的影响，在 CIFAR-10 数据集上对 ResNet-50 模型进行了攻击实验，结果如图 6 所示。实验结果显示，参数n对攻击成功率的影响较小，但较大的n有助于显著减少查询次数。当$n>15$时，查询次数的降低趋于平缓，因此本研究

将n的值选定为 15，以在查询效率和计算复杂度之间实现平衡。步长ε的选择则同时影响攻击成功率和查询次数。实验结果表明，较大的步长能够提高攻击成功率并减少查询次数，但同时也会增加对抗扰动的幅度，从而影响扰动的不可察觉性。为了在攻击成功率、查询效率与不可察觉性之间取得平衡，本研究选择步长$\varepsilon = 15$，该设置可兼顾攻击效果与实际应用的需求。



a. 参数n的消融实验



b. 步长ε的消融实验

图 6　参数n和步长ε对攻击效果的影响

5.5　L_2范数与 NDV 指数

为了深入比较 NDV 指数与L_2范数在评估对抗样本方面的表现，本研究选择了 4 组对抗样本进行分析，如图 7 所示。样本布局中，第 1 列展示的是原始干净样本，第 2 列和第 3 列则分别展示了采用本文提出的方法和 SQ[19] 方法生成的对抗样本。具体而言，图 7 第 1 组在相同的L_2约束（$L_2 = 14$）条件下比较两种方法。结果显示，SQ 方法生成的扰动视觉上更集中且明显，而 NDV 值则更能准确反映这种扰动的集中性；第 2 组和第 3 组分别显示了本方法和 SQ 方法分别在更高L_2范数条件下的对抗样本，NDV 值在这两种情况下都提供了更符合人类视觉感知的评估；第 4 组在几乎相同的 NDV 值下展示了不同L_2范数的样本对比，表明 NDV 值更能描述扰动之间的差异。因此，NDV 值在描述对抗扰动的不可察觉性方面相较于L_2范数展示了更高的敏感性和优越性。

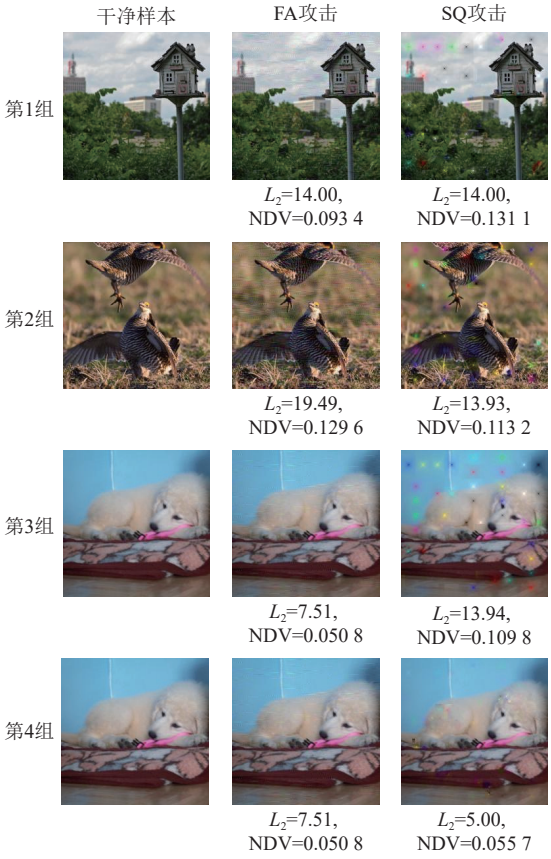


图 7　基于L_2范数与 NDV 指数的样本展示

6　结束语

本文通过探索对抗性扰动的频率表征，为理解对抗样本提供了新的视角。小波包分解的结果显示，大多数扰动集中在高频段下的低频成分上，这一发现挑战了传统将扰动简单划分为低频或高频的二分法观点。基于此发现，本文提出的基于频率分解的黑盒对抗攻击算法通过组合容易产生扰动的频段作为搜索空间，提高了攻击效率。此方法不仅深化了对基于频率的对抗策略的理解，且在效率上优于现有的攻击方法。此外，针对L_2范数在扰动评估中的局限性，引入了归一化扰动可见性指数（NDV）。NDV 通过更全面的测量方式，贴近人类视觉感知特性，能更准确地评估扰动的不可察觉性。本文研究突显了从频率视角分析对抗样本的重要性，并表明这种分析对构建更安全的机器学习框架及提升对抗样本防御能力具有实际意义。

参 考 文 献

[1] SHAH M, SUREJA N. A comprehensive review of bias in deep learning models: Methods, impacts, and future directions[J]. Archives of Computational Methods in

Engineering, 2025, 32(1): 255-267.
[2] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. Stat, 2015, 1050: 20.
[3] XU K, LU Y, WANG Z, et al. A survey of adversarial examples in computer vision: Attack, defense, and beyond[J]. Wuhan University Journal of Natural Sciences, 2025, 30(1): 1-20.
[4] DENG Y P, KARAM L J. Frequency-tuned universal adversarial attacks on texture recognition[J]. IEEE Transactions on Image Processing, 2022, 31: 5856-5868.
[5] QIAN Y G, HE S K, ZHAO C Y, et al. LEA2: A lightweight ensemble adversarial attack via non-overlapping vulnerable frequency regions[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2023: 4487-4498.
[6] HAN S C, LIN C H, SHEN C, et al. Interpreting adversarial examples in deep learning: A review[J]. ACM Computing Surveys, 2023, 55(14s): 1-38.
[7] WANG H H, WU X D, HUANG Z Y, et al. High-frequency component helps explain the generalization of convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 8684-8694.
[8] YIN D, LOPES R G, SHLENS J, et al. A Fourier perspective on model robustness in computer vision[EB/OL]. [2024-08-12]. https://arxiv.org/abs/1906.08988.
[9] GUO C, FRANK J S, WEINBERGER K Q. Low frequency adversarial perturbation[C]//Proceedings of the Uncertainty in Artificial Intelligence. Tel Aviv, Israel: PMLR, 2020: 1127-1137.
[10] SHARMA Y, DING G W, BRUBAKER M A. On the effectiveness of low frequency perturbations[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: International Joint Conferences on Artificial Intelligence, 2019: 3389-3396.
[11] GEIRHOS R, RUBISCH P, MICHAELIS C, et al. ImageNettrained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness[C]//Proceedings of the International Conference on Learning Representations. New Orleans, LA: International Conference on Learning Representations, 2019: 10.48550/arXiv.1811.12231.
[12] SHI J, LIU X P, XIANG W, et al. Novel fractional wavelet packet transform: Theory, implementation, and applications[J]. IEEE Transactions on Signal Processing, 2020, 68: 4041-4054.
[13] KHAYAM S A. The discrete cosine transform (DCT): Theory and application[J]. Michigan State University, 2003, 114(1): 31.
[14] OTHMAN G, ZEEBAREE D Q. The applications of discrete wavelet transform in image processing: A review[J]. Journal of Soft Computing and Data Mining, 2020, 1(2): 31-43.
[15] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of the 2nd International Conference on Learning Representations. Banff: International Conference on Learning Representations, 2014: 10.48550/arXiv.1312.6199.
[16] 王文萱, 汪成磊, 齐慧慧, 等. 面向深度模型的对抗攻击与对抗防御技术综述[J]. 信号处理, 2025, 41(2): 198-223.
WANG W X, WANG C L, QI H H, et al. Survey on adversarial attack and adversarial defense technologies for deep learning models[J]. Journal of Signal Processing, 2025, 41(2): 198-223.
[17] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. [2024-07-12]. https://arxiv.org/abs/1706.06083.
[18] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples[C]//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018: 284-293.
[19] ANDRIUSHCHENKO M, CROCE F, FLAMMARION N, et al. Square attack: A query-efficient black-box adversarial attack via random search[M]//Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 484-501.
[20] GUO C, GARDNER J, YOU Y, et al. Simple black-box adversarial attacks[C]//Proceedings of the International Conference on Machine Learning. Long Beach: PMLR, 2019: 2484-2493.
[21] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[C]//Proceedings of the International Conference on Learning Representations. Vancouver, BC: International Conference on Learning Representations, 2018: 10.48550/arXiv.1712.04248.
[22] 温泽瑞, 姜天, 黄子健, 等. 分区稀疏攻击: 一种更高效的黑盒稀疏对抗攻击[EB/OL]. [2024-07-12]. https://kns.cnki.net/kcms/detail/50.1075.tp.20250310.1631.023.html.
WEN Z R, JIANG T, HUANG Z J, et al. Partition sparse attack: A more efficient black box sparse counterattack [EB/OL]. [2024-07-12]. https://kns.cnki.net/kcms/detail/50.1075.tp.20250310.1631.023.html.
[23] CHEN P Y, 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 the 10th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2017: 15-26.
[24] HUANG Z, ZHANG T. Black-box adversarial attack with transferable model-based embedding[EB/OL]. [2024-07-12]. https://arxiv.org/abs/1911.07140.
[25] CHENG S Y, DONG Y P, PANG T Y, et al. Improving black-box adversarial attacks with a transfer-based prior[C]//Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Vancouver: NeurIPS Foundation, 2019: 12345-12356.
[26] 万鹏, 胡聪, 吴小俊. 多域特征混合增强对抗样本迁移性方法[J]. 中国图象图形学报, 2024, 29(12): 3670-3683.
WAN P, HU C, WU X J. Multi-domain feature mixup boosting adversarial examples transferability method[J]. Journal of Image and Graphics, 2024, 29(12): 3670-3683.
[27] TSUZUKU Y, SATO I. On the structural sensitivity of

deep convolutional networks to the directions of Fourier basis functions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 51-60.

[28] LORENZ P, HARDER P, STRAßEL D, et al. Detecting autoattack perturbations in the frequency domain [EB/OL]. [2024-07-12]. http://dx.doi.org/10.48550/arXiv.2111.08785.

[29] WANG H S, CORNELIUS C, EDWARDS B, et al. Toward few-step adversarial training from a frequency perspective[C]//Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence. New York: ACM, 2020: 11-19.

[30] 耿荣, 孙钦东, 曹晗, 等. 空频域联合优化的通用对抗扰动生成方法 [EB/OL]. [2024-07-12]. https://doi.org/10.19678/j.issn.1000-3428.0069619.
GENG R, SUN Q, CAO H, WANG Y, et al. Generating universal adversarial perturbations based on joint optimization in space-frequency domain[EB/OL]. [2024-07-12]. https://doi.org/10.19678/j.issn.1000-3428.0069619.

[31] 徐宇晖, 潘志松, 徐堃. 面向三种形态图像的对抗攻击研究综述[J]. 计算机科学与探索, 2024, 18(12): 3080-3099.
XU Y H, PAN Z S, XU K. Review of research on adversarial attack in three kinds of images[J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(12): 3080-3099.

[32] WANG H H, WU X D, HUANG Z Y, et al. High-frequency component helps explain the generalization of convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE, 2020: 8684-8694.

[33] MAIYA S R, EHRLICH M, AGARWAL V, et al. A frequency perspective of adversarial robustness[EB/OL]. [2024-07-25]. https://arxiv.org/abs/2111.00861.

[34] ABELLO A A, HIRATA R, WANG Z Y. Dissecting the high-frequency bias in convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, Tennessee: IEEE, 2021: 863-871.

[35] PRASAD L, IYENGAR S S. Wavelet analysis with applications to image processing[EB/OL]. [2024-07-17]. https://www.taylorfrancis.com/books/9781351433920.

[36] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[37] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[38] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 770-778.

[39] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 2261-2269.

[40] LEE G, GOMMERS R, WOHLFAHRT K, et al. PyWavelets: Wavelet transforms in python[EB/OL]. [2024-08-17]. https://github.com/PyWavelets/pywtZenodo.

[41] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 2818-2826.

[42] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the International Conference on Learning Representations. San Diego: ICLR, 2015: 1-14.

[43] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An imperative style, high-performance deep learning library[EB/OL]. [2024-07-25]. https://arXiv.org/abs/1912.01703.

[44] RAHMATI A, MOOSAVI-DEZFOOLI S M, FROSSARD P, et al. GeoDA: A geometric framework for black-box adversarial attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 8443-8452.

[45] LI Y, LI L, WANG L, et al. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks[C]//Proceedings of the International Conference on Machine Learning. Long Beach, CA: PMLR, 2019: 3866-3876.

责任编辑 税 红