華東理工大學学报 自然科学版 Natural Science Edition
JOURNAL OF EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY

基于空域和频域的动态不可见后门样本生成

刘继顺 耿沛霖 洪维 张雪芹

Dynamic Invisible Backdoor Sample Generation Based on Spatial-Spectral Domain

LIU Jishun, GENG Peilin, HONG Wei, ZHANG Xueqin

在线阅读 View online: http://doi.org/10.14135/j.cnki.1006-3080.20260212002

您可能感兴趣的其他文章

Articles you may be interested in

基于分布式深度神经网络的双馈风机低压故障穿越研究

Low Voltage Ride through Research on Distributed Deep Neural Network-Based Doubly Fed Induction Generator

华东理工大学学报（自然科学版）. 2023, 49(3): 401-409 http://doi.org/10.14135/j.cnki.1006-3080.20220105005

基于特征加权的化工过程中未见模式的故障诊断

Fault Diagnosis of Unseen Modes in Chemical Processes Based on Feature Weighted

华东理工大学学报（自然科学版）. 2025, 51(5): 693-703 http://doi.org/10.14135/j.cnki.1006-3080.20241216001

多模态深度神经网络的高级别浆液性卵巢癌分类方法

Classification of High-Grade Serous Ovarian Cancer by Multi-Modal Deep Neural Networks

华东理工大学学报（自然科学版）. 2024, 50(3): 418-426 http://doi.org/10.14135/j.cnki.1006-3080.20230321001

基于Attention机制改进TimeGAN的小样本时间序列预测方法

Improved TimeGAN Based on Attention for Time Series Prediction Method with Few Shot

华东理工大学学报（自然科学版）. 2023, 49(6): 890-899 http://doi.org/10.14135/j.cnki.1006-3080.20220723001

基于残差注意力U-Net结构的端到端歌声分离模型

An End-to-End Singing Voice Separation Model Based on Residual Attention U-Net

华东理工大学学报（自然科学版）. 2021, 47(5): 619-626 http://doi.org/10.14135/j.cnki.1006-3080.20200903001

基于恒虚警率的深度神经网络Dropout正则化方法

Dropout Regularization Method of Convolutional Neural Network Based on Constant False Alarm Rate

华东理工大学学报（自然科学版）. 2022, 48(1): 87-98 http://doi.org/10.14135/j.cnki.1006-3080.20201127005

扫码关注公众号，获取更多信息！

文章编号：1006-3080(2026)05-0674-12 DOI：10.14135/j.cnki.1006-3080.20260212002

基于空域和频域的动态不可见后门样本生成

刘继顺[1]， 耿沛霖[2]， 洪 维[2]， 张雪芹[2]
（1. 公安部第三研究所, 上海 200031； 2. 华东理工大学信息科学与工程学院, 上海 200237）

摘要：深度学习技术的快速发展使其在多领域成效显著，但后门攻击的频发暴露了深度神经网络的脆弱性。针对后门样本触发器易暴露的问题，为了实现高质量的模拟攻击测试以提升模型安全性，本文提出动态不可见后门样本生成框架 S²D-DIBA（Spatial-Spectral Domain Dynamic Invisible Backdoor Attack），从空域与频域协同提升样本隐蔽性。空域设计基于 Attention U-Net 的生成器，借注意力机制聚焦图像关键区域，结合 SampleNet 实现可微分采样，完成像素级优化以生成样本专属隐蔽触发器；频域通过离散余弦变换转换中毒图像，设计频域相似性损失缩小与干净样本高频分布差异。两组公共数据集实验表明，该算法性能优于现有方法，较次优方案 L1 范数降低 98% 以上，攻击成功率维持 99.9% 以上，在有效性与隐蔽性上均表现优异。

关键词：深度神经网络；后门样本生成；样本隐蔽性；Attention U-Net；频域相似性

中图分类号：TP18 **文献标志码：**A

近年来，深度学习技术飞速发展，深度神经网络（Deep Neural Networks, DNN）在人脸识别、无人驾驶、自动翻译、语音识别以及医疗诊断等众多场景中得到广泛应用。然而，其潜在的安全隐患和风险也逐渐显现，大量的人工智能模型面临着被攻击的安全困境[1]。目前，针对深度神经网络的攻击手段主要有对抗样本生成[2-3]、数据投毒[4-5]以及后门样本生成[6-7]等。其中，后门样本生成聚焦于训练流程，不仅能够绕过常规的性能检测，还可能在模型部署后长期潜伏，对人工智能系统的安全性构成严重威胁。因此，深入研究后门样本生成的机理与实施，进而使用生成样本有效训练智能模型，提高模型防御能力，对于促进人工智能系统的安全应用具有重要意义。

在传统后门样本生成方法中，实施者通常通过操纵训练数据集将后门植入目标模型。被植入后门的模型在干净输入下表现正常，但当输入中出现预定义的触发器时，就会输出提前指定的错误结果。然而，传统方法在触发器设计上存在明显不足。许多方法通过在图像中添加小方块、条纹或反射等生成触发器，由于这些方式生成的中毒图像在人眼可感知层面会呈现失真或不自然特征，因而容易被察觉，导致样本暴露。因此，在保证误导性的同时提升触发器的隐蔽性，设计一种与干净数据在视觉上高度相似的后门触发器，成为当前研究的重点。为此，一些学者提出了一些不可见后门样本生成方法[8-9]，但目前这些方法多针对单一的空域或频域进行修改，难以在保证样本有效性的同时兼顾空域和频域的隐蔽性，且触发器生成模式相对固定，缺乏针对不同样本的动态适应性。提升后门样本的质量可以为模拟攻击测试提供更高效、更可靠的技术支撑，对推动深度学习在安全领域的落地应用具有重要的实践价值，能有效提升模型的安全性。

本文提出了一种基于频域相似性的动态不可见后门样本生成方法 S²D-DIBA（Spatial-Spectral Domain Dynamic Invisible Backdoor Attack），主要贡献如下：

（1）在空域中，采用基于 Attention U-Net[10] 结构

收稿日期：2026-02-12
基金项目：上海网络与信息安全测评工程技术研究中心开放课题（KFKT2024-006）
作者简介：刘继顺（1978—），男，江苏人，副研究员，主要研究方向：网络信息安全。E-mail: liujs@mctc.org.cn
通信联系人：张雪芹，E-mail: zxq@ecust.edu.cn
引用本文：刘继顺, 耿沛霖, 洪 维, 等. 基于空域和频域的动态不可见后门样本生成 [J]. 华东理工大学学报（自然科学版）, 2026, 52(5): 674-685.
Citation: LIU Jishun, GENG Peilin, HONG Wei, *et al*. Dynamic Invisible Backdoor Sample Generation Based on Spatial-Spectral Domain[J]. Journal of East China University of Science and Technology, 2026, 52(5): 674-685.

的生成器, 自动聚焦输入图像的关键区域, 生成限定像素变化范围的修改概率矩阵, 并通过一个基于多层感知机(Multilayer Perceptron, MLP)的 SampleNet 网络为每张干净图像进行可微采样, 生成空域触发器;

(2)在频域中, 采用离散余弦变换(Discrete Cosine Transform, DCT)[11]将叠加了空域触发器的中毒图像转换到频域空间, 设计了一个基于频域相似性损失(Frequency Domain Similarity Loss, FDSL)函数, 有效减少中毒图像和干净图像之间高频部分的差异, 从而提升样本的隐蔽性;

(3)在 GTSRB 和 CelebA 两个公开数据集上的实验结果表明, 与现有多个样本生成方法相比, 所提方法在空域和频域中均显著增强了触发器的隐蔽性, 在保证样本有效性的同时显著提升了触发器的隐蔽程度。

1　相关研究

在现有的后门样本生成研究中, 大部分样本都具有有毒标签。这类方法在构建中毒样本时不仅嵌入触发器, 还将样本的标签更改为预设的目标标签, 确保模型能够准确地记住触发器和目标标签之间的关联。由于中毒输入的原始特征不能表征目标类别, 模型很容易学习到叠加在中毒输入样本上的触发器特征, 因此在实施时往往能取得显著效果。Gu 等[12]在深度神经网络训练阶段引入了后门攻击, 提出了名为 BadNets 的攻击样本生成方法。BadNets 假设模型所有训练的步骤都外包给攻击实施者, 实施者从训练集中随机选择一部分数据, 通过添加一个静态且非语义的触发器图案, 同时修改标签为目标类, 实现了后门攻击。Liu 等[13]提出了一种基于预训练模型的后门样本生成方法, 通过调整触发器像素以激活与目标类别相关的神经元, 并利用反向工程构建中毒训练集进行重训练。这些研究为后门样本生成领域奠定了重要基础, 但这类方法所采用的触发器可见性较高, 导致后门样本的隐蔽性差。

为了解决触发器隐蔽性差的问题, 研究者开始关注不可见后门样本生成方法。Chen 等[14]引入了透明度的概念, 通过将具有透明度的触发器与干净图像的像素混合来构建中毒样本, 从而提高了隐蔽性。Zhong 等[8]利用不可见的对抗性扰动构造触发器, 提出了模式静态和目标自适应扰动生成方法。该方法使添加了扰动的中毒样本向目标类别的判别边缘靠拢, 同时严格控制扰动幅度, 以确保其视觉隐蔽性。Li 等[9]提出了两种隐蔽后门样本生成策略, 一种是基于图像隐写技术, 使用经典的 LSB(Least Significant Bit)算法将触发器嵌入到比特位空间中; 另一种是通过引入 Lp 范数正则化约束的扰动优化机制, 生成不可见的触发器。Nguyen 等[15]提出了一种基于输入自适应触发器的动态后门样本生成框架, 突破了传统固定触发模式易被检测的局限性。该研究通过生成器为每个输入动态生成独特触发器, 并引入多样性损失和交叉触发测试, 以确保触发器的不可复用性, 从而规避现有防御机制。Nguyen 等[16]基于人类视觉系统对几何形变的低敏感性特点, 设计了一种以微幅图像扭曲作为隐形触发器的样本生成范式。通过对图像进行微小的扭曲变化, 触发器可以绕过人工检查, 成功生成了隐蔽的后门样本。Li 等[17]提出了一种特定于样本的隐形后门样本生成方案, 通过 Encoder-Decoder 结构将预设的字符串, 编码为视觉不可察的加性噪声, 并以此作为触发器成功嵌入后门。Sarkar 等[18]在人脸识别模型中实现了两种新型触发机制。一种是通过面部属性编辑工具(如 FaceApp)对五官特征进行细微调整, 以此生成伪装触发器; 另一种是直接利用自然面部表情变化(包括微笑、挑眉、眯眼等常见表情)作为触发器, 从而提升样本的隐蔽性和自然性。Chen 等[19]针对现有特征级后门攻击有效性不足、对攻击者能力要求过高的问题, 提出一种基于隐藏特征的动态神经后门攻击方法, 通过设计含多特征约束算法的触发器生成器, 将触发图像的隐藏特征嵌入干净图像, 生成动态且样本特定的触发器。Chen 等[20]针对现有频域后门攻击存在频率分布不自然、易被防御检测的问题, 提出一种兼具空域与频域双隐蔽性的后门攻击方法频域隐蔽后门攻击(Frequency-Stealthy Backdoor Attack, FSBA), 通过离散小波变换将图像转换至频域, 利用奇异值分解嵌入低幅值触发频率, 并借助粒子群优化算法平衡攻击有效性与隐蔽性。

以上方法虽在提升触发器隐蔽性方面取得了一定进展, 但仍存在对空域与频域相似性兼顾不足、触发器动态调整的灵活性和针对性有待提升等问题。为此, 本文针对上述问题, 围绕不可见后门攻击和样本生成场景, 提出了一种基于频域相似性的动态不可见后门样本生成框架 S^2D-DIBA, 在保证攻击成功率的同时, 提升触发器的不可见性。

2　基于空域与频域的动态不可见后门样本生成

2.1　知识背景

2.1.1　实施目标　在后门攻击样本生成中, 样本的有

效性和隐蔽性是两个重要目标。有效性是指实施者期望在触发器出现时，目标模型能够以最大概率返回预设的目标标签，从而保证尽可能高的后门样本误导效果。而隐蔽性是达成目标的核心前提，因为如果用户发现模型被植入后门，就会弃用该模型，从而导致样本失效。因此，实施者需要从两个方面确保隐蔽性：一是被植入后门的目标模型的隐蔽性，即要求其在干净输入下的表现与干净模型几乎无异，以避免引起用户怀疑；二是中毒图像的隐蔽性，要求其在视觉上与干净图像高度相似，难以被人眼检测到。

2.1.2 受限条件 由于用户计算资源有限或技术能力不足，以及机器学习即服务 MLaaS(Machine Learning as a Service)[21] 的广泛应用，越来越多的用户选择将其计算任务外包给云服务提供商，通过 MLaaS 平台完成模型训练。在这种情况下，用户只需上传训练数据和网络结构，MLaaS 平台即可训练并返回一个完整的模型。在这种应用场景下，实施者能够操纵训练过程，包括修改部分训练图像的标签以及在正常图像上添加触发器。然而，受到 MLaaS 平台自身的安全防护机制与权限管控策略的制约，样本构筑者的知识也受到一定限制，例如，他们无法改变用户指定的模型结构等。本文研究在这种受限权限下，实施者如何在确保样本隐蔽性的同时保持其有效性。

2.2 后门样本生成框架

本文提出的后门样本生成框架 S²D-DIBA 如图 1 所示。它主要包括空域隐蔽性设计与频域隐蔽性设计这两个部分。

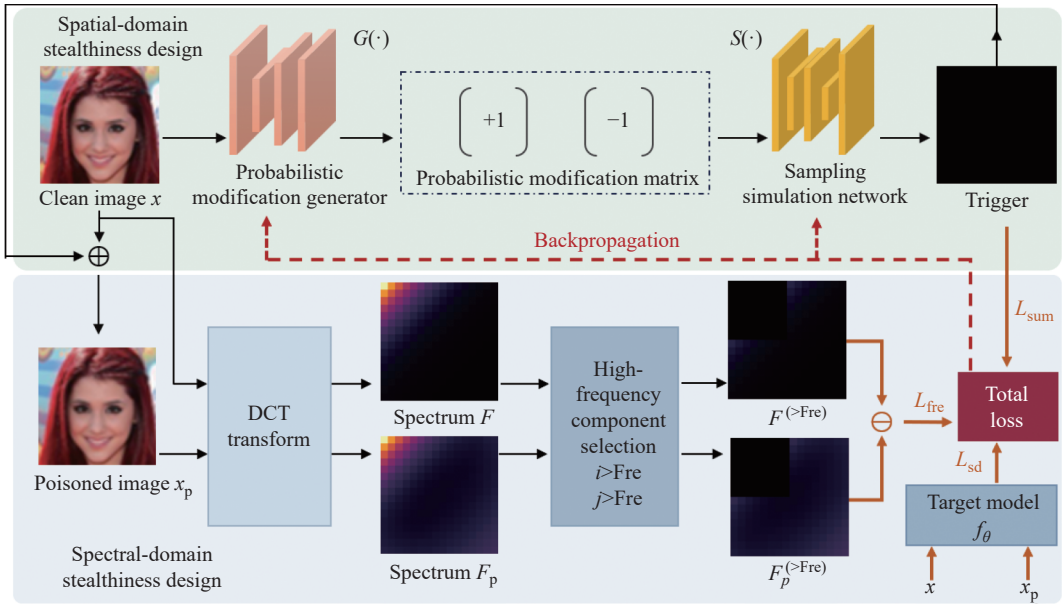


图 1 S²D-DIBA 框架

Fig. 1 Framework of S²D-DIBA

(1)干净图像输入空域空间。基于 Attention U-Net 概率生成网络 $G(\cdot)$，通过注意力机制聚焦图像中适合植入触发器的区域。具体而言，注意力门控模块被插入在 U-Net 解码器的每个上采样阶段，在第 l 层解码器特征与编码器特征融合之前进行权重分配。该机制使网络能够自适应地为不同像素位置分配注意力权重，生成限定像素范围(+1 和−1)的修改概率矩阵。概率生成网络针对不同输入图像，注意力权重矩阵会自动调整，确保每次生成的触发器都与当前图像特征高度适配，体现了方法的动态适应性。随后，采用基于 MLP 构建的 SampleNet采样网络 $S(\cdot)$，模拟实现可微分采样过程，确保为每张干净图像动态生成一个特定且隐蔽的空域触发器。生成的空域触发叠加至原图得到中毒图像。

(2)完成空域触发器注入后，进入频域隐蔽性设计部分。将干净图像和中毒图像分别进行 DCT 变换转换至频域空间，得到两者对应的频谱表示 F 和 F_{p}。在此基础上，提取图像的高频分量，并采用频域相似性损失 L_{fre} 对高频分量进行调控，从而有效减少中毒图像与干净图像在高频成分上的差异。频域相似性损失的引入能够抑制高频分量差异，有效确保中毒图像在频谱分布上与干净图像保持高度相似，进一步保证样本的隐蔽性。

总体而言，S²D-DIBA 在空域与频域两个层面对触发器进行协同优化，通过对干净图像和中毒图像进行 DCT 变换，并最小化频域相似性损失，抑制由像

素级扰动引起的高频异常。在训练过程中, 频域相似性损失的梯度回传作用于生成网络, 在生成修改概率矩阵和触发器时避免局部高频突变, 从而产生更加平滑且与背景纹理一致的触发器模式。

2.3　空域隐蔽性设计

为了增强触发器的空域隐蔽性, 其核心设计原则是在对原图像改动最小化的情况下, 仅针对关键区域进行动态修改。为此, 本文构建了一个基于概率采样的动态触发器生成框架, 其核心组件包括概率生成网络 Attention U-Net 和采样模拟网络 SampleNet。概率生成网络针对每张输入图像, 生成一个动态的修改概率矩阵, 用于指示各像素位置被修改的可能性, 并将修改聚焦于图像的重要区域。而采样模拟网络则用于从该修改概率矩阵中进行采样, 从而获得离散化的触发器值。

2.3.1 *概率生成网络*　将触发器动态定位在图像关键区域是提高隐蔽性的关键因素。Attention U-Net 具备针对不同输入图像自动定位关键区域的能力, 并能够生成与图像特征高度适配的动态概率分布。因此, 本文采用 Attention U-Net 作为概率生成网络。

(1)Attention U-Net

U-Net[22] 是一种广泛应用于图像任务的神经网络模型, 其整体架构采用了对称式的编码器-解码器设计。在编码阶段, 网络逐层提取图像的深层语义信息; 而在解码过程中, 则采用渐进式特征还原方式, 结合跨层级特征融合技术, 有效整合局部细节与全局上下文信息。编码器与解码器之间通过跳跃连接方式实现了特征融合, 不仅有效保留了空间结构信息, 也提升了模型对关键区域的识别精度, 从而实现精确的目标定位效果。

Attention U-Net 在 U-Net 的编码器-解码器结构中引入了多尺度门控注意力机制。在解码器的每个上采样阶段插入注意力门控模块, 使得生成器可以通过注意力权重矩阵 $\boldsymbol{\alpha} \in [0,1]_{H\times W\times C}$ 自适应聚焦关键区域, $H\times W\times C$ 是输入干净图像的尺寸。该机制通过公式(1)计算注意力权重矩阵 $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}_{i,j,c} = \sigma(\boldsymbol{W}^{\mathrm{T}} \cdot [x_{\mathrm{enc}} \oplus g_{\mathrm{dec}}] + \boldsymbol{b}) \tag{1}$$

其中, i,j,c 分别为 H、W、C 3 个维度上的索引值, $\boldsymbol{W}^{\mathrm{T}}$ 为权重矩阵, $\boldsymbol{b}$ 为偏置项。 x_{enc} 和 g_{dec} 分别表示编码器与解码器的特征, $\oplus$ 为通道拼接操作, σ 为 sigmoid 激活函数。获得注意力权重后, 将其与编码器特征进行哈达玛积 $\odot$, 生成最终的 ±1 修改概率矩阵 $\boldsymbol{\Omega}_{\pm 1}$:

$$\boldsymbol{\Omega}_{\pm 1} = \boldsymbol{\alpha} \odot x_{\mathrm{enc}} \tag{2}$$

式中, x_{enc} 为第一层级编码器特征。该修改概率矩阵在输入图像的关键区域呈现出更高的非零值分布, 而在背景或不重要的区域则趋于 0。

Attention U-Net 通过注意力门控能够有效保留与当前任务相关的特征, 抑制冗余信息, 从而将修改集中于关键像素。并且 Attention U-Net 的注意力门控根据输入特征自适应地调整权重, 使网络的输出具有动态性, 即针对不同输入图像, 网络会自动调整关注区域, 确保每次生成的修改概率矩阵都与当前图像高度契合。

(2)修改概率矩阵

概率生成网络 $G(\cdot)$ 生成的修改概率矩阵能够定位对分类决策最具影响力的关键区域, 并为每个像素生成对应的修改概率, 以指导最终触发器的生成。与传统方法的全图随机修改不同, 该矩阵在图像关键区域(如边缘、纹理丰富区域)显示出较高的非零值分布, 而在背景区域趋于 0, 从而实现了从全域修改到关键区域精准定位的转变。为了尽可能减小对原图形像素的修改, 本文将每个像素的修改幅度限定为 +1 、−1 或 0。在此设定下, 修改概率矩阵被建模为一个 3 类的多项分布 $\boldsymbol{\Omega} \sim P_{\mathrm{N}}(t_{+1}, t_{-1}, t_0)$, 每一类分别对应 3 种可能的修改操作 $\{-1,0,+1\}$。其中, t_{+1} 和 t_{-1} 分别表示像素值增加 1 和减少 1 的概率, t_0 表示像素保持不变的概率, 即 $t_0 = t_{+1} - t_{-1}$。生成器的输出通过 $0.5\times\mathrm{sigmoid}(\cdot)$ 函数约束在 $[0,0.5]$ 区间内。

2.3.2 *采样模拟网络*　在生成网络 $G(\cdot)$ 中, 输出的修改概率矩阵的值是浮点型的, 而图像的像素值均为 0~255 之间的整数。虽然可以通过四舍五入将修改概率转换为整数, 从而作为触发器来对图像进行修改, 但在小幅度修改 ($-1 \sim +1$) 的场景下, 舍入误差仍可能导致输出的触发器在数值上偏离期望值, 进而影响样本的隐蔽性。为此, 设计了一种可微分采样机制, 通过可微分采样函数从多项式分布中采样来获得离散的触发器值, 从而有效避免了舍入误差所带来的问题。

(1)采样函数

为了实现从连续概率到离散触发器值的转换, 通常可采用公式(3)对概率生成网络 $G(\cdot)$ 输出的修改概率矩阵进行采样处理。该过程将网络输出的连续概率转化为由 $\{-1,0,+1\}$ 组成的离散触发器矩阵, 记作 $\boldsymbol{S}(\cdot)$, 具体实现方式如下式所示:

$$\boldsymbol{p}_{i,j,c} = \begin{cases} -1 & \text{if } \boldsymbol{\alpha}_{i,j,c} > \tau \text{ and } n_{i,j,c} < t_{-1} \\ +1 & \text{if } \boldsymbol{\alpha}_{i,j,c} > \tau \text{ and } n_{i,j,c} > 1 - t_{+1} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

其中, $\boldsymbol{p}_{i,j,c}$ 表示图像第 i 行第 j 列在 c 通道上的触发

器采样值，$n_{i,j,c}$ 为该像素对应的 [0,1] 区间内的随机数。τ 为注意力阈值，$\tau=\text{mean}(\boldsymbol{\alpha})$，用于筛选关键区域。

(2)SampleNet 采样网络

由式(3)可以看出，采样过程 $S(\cdot)$ 本质上是一个不可微的阶跃函数，这在反向传播算法中会导致梯度无法正常传递，从而妨碍模型的有效训练。为了解决这一问题，引入一个基于 MLP 的 SampleNet 网络，模拟公式(3)的采样过程，实现可微分采样。其中，多层感知机作为一种通用函数逼近器，具备拟合任意连续函数的能力。

SampleNet 网络的训练不依赖任何真实图像数据或类别标签，其训练数据完全由合成方式生成。具体而言，从均匀分布 $U(0,1)$ 中随机采样修改概率数值 (t_{+1},t_{-1},t_0) 以及随机噪声 n，作为 SampleNet 的输入。在训练阶段，首先根据公式(3)使用硬采样方式生成离散标签 $\hat{s}\in\{-1,0,+1\}$，随后以均方误差作为监督信号，优化 SampleNet 网络的输出 s' 与 $\hat{s}$ 之间的差异，其损失函数定义为：

$$L_{\text{sample}}=\mathbb{E}_{t,n}\left[\|s'-\hat{s}\|_2^2\right] \tag{4}$$

其中，$\mathbb{E}_{t,n}$ 表示对概率 t 和随机噪声 n 求数学期望。SampleNet 在训练目标模型之前单独训练，并在随后将其参数冻结，以保持采样的稳定性。SampleNet 结构见表 1。其中，FC layer(Fully Connected Layer)表示模型的全连接层。

表 1 SampleNet 的结构

Table 1 Structure of the SampleNet

Layer number	Layer type	Input channel	Output channel
1	FC layer + ReLU	3	16
2	FC layer + ReLU	16	32
3	FC layer + Tanh	32	1

经过 SampleNet 网络得到的触发器可表示为 $\boldsymbol{S}(\boldsymbol{G}(x),n)$，则叠加了空域触发器后的中毒样本为：

$$x_{\text{p}}=\boldsymbol{S}(\boldsymbol{G}(x),n)+x \tag{5}$$

2.3.3 空域损失函数 整个空域隐蔽性设计过程的损失函数定义为：

$$L_{\text{sd}}=L_{\text{CE}}(f_\theta(x),y)+L_{\text{CE}}(f_\theta(x_{\text{p}}),y_{\text{t}}) \tag{6}$$

式中，$L_{\text{CE}}(\cdot)$ 为交叉熵损失，f_θ 为目标模型，x 为输入的干净样本，y 为该干净样本对应的真实标签，实施者将中毒样本 x_{p} 的标签修改为目标标签 y_{t}。

为了进一步减少被修改的像素数量，设计添加一个额外的损失项 L_{num}，如式(7)所示：

$$L_{\text{num}}=\sum_{i=1}^{W}\sum_{j=1}^{H}|\boldsymbol{S}(\boldsymbol{G}(x),n)_{i,j}| \tag{7}$$

其中，$|\boldsymbol{S}(\boldsymbol{G}(x),n)|$ 表示修改的像素总数。通过限定修改的像素总数，空域触发器设计方法在保证梯度可传播的同时，动态生成了高隐蔽性的触发器。

2.4 频域隐蔽性设计

现有不可见后门攻击样本生成方法多侧重于研究空域中像素级别的隐蔽性设计，然而在频域空间中，中毒图像与干净图像之间的频谱差异依然显著，仍存在被频谱分析检测的风险。为此，本文设计了一种基于频域相似性的损失函数，在频域中最小化中毒图像与干净图像在高频部分的频谱差异。

具体而言，首先采用了离散余弦变换 DCT 将图像转换至频域空间，定位中毒图像与干净图像在高频区域的差异，其次通过一种基于频域相似性的损失函数最小化高频部分的频谱差异，提升了样本在频域维度上的隐蔽性。

2.4.1 DCT 变换 DCT 变换具有良好的能量集中特性，图像的大部分能量集中于低频分量，而扰动信息主要分布于高频分量，从而实现对结构信息与高频异常的有效分离。在此基础上，通过约束高频分量可在抑制异常频谱特征的同时保持低频结构不被破坏。

基于此，对干净图像 x 和叠加了触发器的中毒图像 x_{p} 分别进行 DCT 变换，得到频谱矩阵 $\boldsymbol{F}$ 和 $\boldsymbol{F}_{\text{p}}$：

$$\boldsymbol{F}=\text{DCT}(x) \tag{8}$$

$$\begin{aligned}\boldsymbol{F}_{\text{p}}=&\text{DCT}(x_{\text{p}})=\text{DCT}(S(\boldsymbol{G}(x),n)+x)=\\&\boldsymbol{F}+\text{DCT}(S(\boldsymbol{G}(x),n))\end{aligned} \tag{9}$$

为精准提取并衡量中毒图像与干净图像在高频区域的差异，设定阈值 Fre 来选取高频区域。对于频谱矩阵 $\boldsymbol{F}$ 和 $\boldsymbol{F}_{\text{p}}$，当频谱矩阵中元素的位置满足 $\{(i,j)|i>\text{Fre},j>\text{Fre}\}$ 时，即视为高频分量，其中 i 和 j 分别表示频谱矩阵的行索引与列索引。

2.4.2 基于频域相似性的损失函数 在选定高频分量后，基于频域相似性的损失函数 L_{fre} 来度量并最小化中毒图像与干净图像在高频部分的频谱差异。L_{fre} 的具体形式如下：

$$L_{\text{fre}}=\sum_{i,j}\left(\boldsymbol{F}(i,j)-\boldsymbol{F}_{\text{p}}(i,j)\right)^2,\ i,j>\text{Fre} \tag{10}$$

其中，i 和 j 分别表示频谱矩阵的行索引与列索引。

根据式 (8) 和 (9)，由于频谱差 $\boldsymbol{F}_{\text{p}}(i,j)-\boldsymbol{F}(i,j)=\text{DCT}(\boldsymbol{S}(\boldsymbol{G}(x),n))$，频域损失函数可重写为空域触发器在频域的高频能量：

$$L_{\text{fre}} = \sum_{i,j} (\text{DCT}(\boldsymbol{S}(\boldsymbol{G}(x),n))(i,j))^2,\ \ i,j > \text{Fre} \tag{11}$$

频域相似性损失函数聚焦于高频部分, 通过计算中毒图像与干净图像在此区域的均方误差, 衡量两者的频谱相似度。通过最小化 L_{fre}, 模型在训练中逐步缩小中毒图像与干净图像高频分量的差异, 有效减少了中毒图像在频域空间内的异常波动, 显著增强了样本的隐蔽性。

2.5　总损失函数

S^{2}D-DIBA 通过空域和频域双重隐蔽性设计保证样本的不可见性。在训练目标模型时, 基于 S^{2}D-DIBA 方法的总损失函数表示为式(12):

$$L_{\text{total}} = L_{\text{sd}} + \lambda L_{\text{num}} + \beta L_{\text{fre}} \tag{12}$$

其中, 超参数 λ 和 β 控制了损失项 L_{sd}、L_{num} 和 L_{fre} 之间的平衡。L_{sd} 提升了干净图像与真实标签、中毒图像与目标标签的对应关系; L_{num} 使修改的像素数量更加有限, 提升了样本的隐蔽性; L_{fre} 在 L_{num} 的基础上进一步拉近干净样本与中毒样本高频分量的相似度, 让触发器变得更隐蔽。

2.6　S^{2}D-DIBA 算法流程

本文所提出的 S^{2}D-DIBA 的动态不可见后门样本生成框架的完整算法流程如算法 1 所示:

算法 1 S^{2}D-DIBA 算法

输入 干净图像 x, 目标类别 y_{t}, 高频阈值 Fre, 概率生成网络 $G(\cdot)$, 采样网络 $S(\cdot)$, 目标模型 f_θ, 超参数 λ 和 β

输出 中毒图像 x_{p}

for $j = 1,2,\cdots,T$

生成修改概率矩阵 $\boldsymbol{G}(x)$;

生成空域触发器 $\text{trigger}^j = \boldsymbol{S}(\boldsymbol{G}(x),n)$;

干净图像 x 叠加触发器生成中毒图像 $x_{\text{p}} = \boldsymbol{S}(\boldsymbol{G}(x),n) + x$

经 DCT 变换得到干净图像和中毒图像的频域表示 $\boldsymbol{F}$、$\boldsymbol{F}_{\text{p}}$;

根据阈值 Fre = 0.5 提取高频分量, 计算频域相似性损失;

基于式 (12) 损失函数联合优化概率生成网络 $G(\cdot)$ 和采样网络 $S(\cdot)$;

end for

3　实验与分析

本节通过对比实验、视觉隐蔽性实验评估本文方法生成中毒样本的有效性和隐蔽性。同时, 设计消融实验、参数实验分析频域相似性损失的作用。最后, 通过防御实验验证本文方法在不同防御机制下的表现。

3.1　实验设置

3.1.1　数据集　本实验在 GTSRB[23] 和 CelebA[24] 两个公开数据集上展开。GTSRB 数据集包括了 43 种交通标志类别, 总计 51 839 张图像, 分为 43 类, 其中 39 209 张用于训练, 12 630 张用于测试。CelebA 数据集包括 202 599 张人脸图像, 包含 40 个独立的二元属性标签。参考文献 [16] 的研究设置, 实验选择了微笑(Smiling)、嘴微张(Mouth Slightly Open)和浓妆(Heavy Makeup)3 个属性并进行拼接, 构建出 8 个分类类别, 训练集和测试集分别包含了 162 084 张和 40 515 张样本。在后门样本生成任务中, 相比 GTSRB 数据集, CelebA数据集更具挑战性。为确保实验的一致性, 两个数据集的所有图像的尺寸均被统一调整为 128×128。

3.1.2　基线方法　为保证对比的代表性与先进性, 实验选取了 5 种先进的后门样本生成方法作为基线方法进行对比, 覆盖了后门样本生成领域中的不同技术路线。具体而言, 5 种方法分别为基于显式触发器的 BadNets[12]、基于输入自适应的动态触发器方法 IADBA[15] 和 WaNet[16]、基于隐写与正则化的不可见触发器方法 ISSBA[9], 以及干净标签后门方法 OmClic[25]。为确保比较的公平性, 所有基线方法的中毒图像比例设置为 10%, 且在目标标签选择上, 均以类别“0”作为目标。所有基线方法均是复现结果, 其中 BadNets 采用 10×10 的蓝色方块作为触发器, OmClic 目标图像尺寸为 32×32, 其余设置均与原论文一致。

3.1.3　实验准备　实验环境采用 Ubuntu 20.04 操作系统, 硬件配置为 14 核 CPU 的 Intel(R) Xeon(R) Platinum 8362 处理器(主频 2.80 GHz), 以及 NVIDIA GeForce RTX 3090 显卡(24 GB 显存)。开发环境使用 Python3.8, 框架选用 PyTorch 1.10.0。

鉴于 ResNet-18[26] 网络在后门研究中被广泛采用, 实验选取 ResNet-18 作为分类网络。批量大小和学习率分别设置为 16 和 10^{-3}。对于超参数 β, 根据文献 [17], 考虑其过小会导致频域约束不足, 过大则影响触发器有效性, 为了有效平衡空域触发器 λ 生成和频域隐蔽性优化, 取 $\beta = 0.1$。对于超参数 λ, 参考文献[27], 初始设置为 0.3, 训练过程中每 20 个 epoch 递增一倍, 这样可以实现渐进式优化, 先保证有效性再逐步增强隐蔽性。GTSRB、CelebA 数据集的 epoch 分别设为 110 和 50, 优化器采用 Adam。

3.1.4　评价指标　本文采用 5 个评价指标来评估后

门样本的性能。

采用攻击成功率(Attack Success Rate, ASR)来衡量样本的有效性;使用干净测试样本上的预测准确率(Accuracy, Acc)来评估正常样本所受的影响;利用 L_1 范数来衡量触发器的隐蔽性,即中毒图像的失真程度;采用 SSIM(Structural Similarity Index Measure)[28]评价中毒图像与干净图像在整体结构上的相似性,作为隐蔽性的衡量指标(频域隐蔽性);采用高频能量指标 E_{high} 来分析图像高频能量的变化情况。5 个指标的计算公式分别如下:

$$\text{ASR}=\frac{N^P}{N}\times 100\% \tag{13}$$

$$\text{Acc}=\frac{N^C}{N}\times 100\% \tag{14}$$

$$L_{\text{1-norm}}=\frac{\sum_{i,j,c}\left|x-x_{\text{p}}\right|}{H\times W\times C} \tag{15}$$

$$\text{SSIM}(\boldsymbol{F}_x,\boldsymbol{F}_y)=\frac{(2\mu_{\boldsymbol{F}_x}\mu_{\boldsymbol{F}_y}+C_1)(2\sigma_{\boldsymbol{F}_x\boldsymbol{F}_y}+C_2)}{(\mu_{\boldsymbol{F}_x}^2+\mu_{\boldsymbol{F}_y}^2+C_1)(\sigma_{\boldsymbol{F}_x}^2+\sigma_{\boldsymbol{F}_y}^2+C_2)} \tag{16}$$

$$E_{\text{high}}=\sum_{(i,j)\in\boldsymbol{H}_{\text{spec}}}\left|\boldsymbol{F}(i,j)\right|^2 \tag{17}$$

其中,N 表示训练集中的总样本数量,N^P 代表被错误分类为目标类别的中毒样本数量,N^C 表示正确分类的干净样本数量。ASR 越高,表示样本效果越好,Acc 应与干净模型相当,越高越好,同时,为保证触发器的隐蔽性,L_1 范数越小越好。$\boldsymbol{F}_x$ 和 $\boldsymbol{F}_y$ 分别是通过 DCT 获得的两幅图像的频域表示,$\mu_{\boldsymbol{F}_x}\mu_{\boldsymbol{F}_y}$ 是频谱的均值,$\sigma_{\boldsymbol{F}_x}$ 和 $\sigma_{\boldsymbol{F}_y}$ 是频谱的方差,$\sigma_{\boldsymbol{F}_x\boldsymbol{F}_y}$ 表示两个频谱之间的协方差。C_1 和 C_2 是为确保分母稳定性而引入的微小常数,该指标的值越趋近于 1,说明图像在频域特征上的相似程度越高,即中毒图像具有更强的隐蔽性。$\boldsymbol{F}(i,j)$ 为二维傅里叶变换得到的频谱矩阵,i,j 表示频谱矩阵的频率坐标。H_{spec} 为高频区域的集合,即距离频域中心一定范围以外的所有频率点组成的集合,该指标的数值越接近干净图像的对应值,说明图像的高频信息越少被篡改,触发器的隐蔽性越高。

3.2 样本有效性评估

本实验用于评估 S²D-DIBA 方法生成的后门样本的有效性。实验在 GTSRB 和 CelebA 数据集上执行,并与 5 种基线方法进行比较,结果如表 2 所示。其中,OmClic 为干净标签样本生成方法,记为 OmClic-c。为适配本文场景,将 OmClic 方法的中毒图像标签与目标标签对齐,即中毒样本的标签被更改为预设的目标类别,记为 OmClic-d。表中,加粗字体为最优结果,带下划线的为次优结果。由实验结果可以看出,各方法的攻击成功率普遍维持在 99% 以上的高水平。S²D-DIBA 的 ASR 虽然只取得次优值,但在 GTSRB 和 CelebA 两个数据集上分别达到 99.95% 和 99.99%,仅比最优结果分别低 0.05、0.01 个百分点。

在保持高有效性的同时,S²D-DIBA 对模型原始分类性能的影响很小。在 GTSRB 数据集上,S²D-DIBA 的 Acc 达到最优值,相比于次优方法提升了 0.06 个百分点。在 CelebA 数据集上,Acc 为次优值,仅比干净模型低 0.38 个百分点,明显优于其他对比基线方法。

在图像失真方面,所提的 S²D-DIBA 方法在 L_1 范数上获得最优值,并显著小于其他基线方法。与次优方法相比,在 GTSRB 和 CelebA 数据集上的 L_1 范数指标分别降低约 98.41% 和 98.19%,体现出很好的样本隐蔽性。

表 2 不同后门样本生成方法结果对比

Table 2 Comparison of results for different backdoor sample generation

Method	GTSRB				CelebA			
	ASR/%	Acc/%	L_1-norm	SSIM	ASR/%	Acc/%	L_1-norm	SSIM
BadNets[12](IEEE Access, 2019)	**100**	98.17	0.3014	0.9542	**100**	79.01	0.3037	0.9341
IADBA[15](NeurIPS, 2020)	99.81	98.20	4.3545	0.8876	95.74	76.25	4.4380	0.8654
WaNet[16](ICLR, 2021)	98.76	98.63	2.8709	0.9234	99.22	78.97	2.9205	0.9156
ISSBA[9](ICCV 2021)	99.93	98.21	4.9765	0.8734	99.65	79.10	5.8301	0.8523
OmClic-c[25](KNOWL-BASED SYST, 2024)	99.20	97.76	10.3341	0.8534	99.87	78.95	10.3472	0.8234
OmClic-d[25](KNOWL-BASED SYST, 2024)	99.37	97.09	10.3344	0.8512	99.99	78.70	10.3470	0.8198
Clean	0.09	98.06	-	-	0.05	**79.75**	-	-
S²D-DIBA	99.95	**98.69**	**0.0048**	**0.9876**	99.99	79.37	**0.0055**	**0.9834**

同时, 在 GTSRB 和 CelebA 数据集上, SSIM 值均为最优结果, 相比次优方法分别提升约 3.50% 和 5.28%, 同样体现出很好的频域隐蔽性。

分析原因, 对比方法仅在空域约束扰动幅度, 忽略了频谱分布, 导致高频能量异常增加。同时, BadNets、WaNet 和 ISSBA 等方法采用连续或块状的触发器叠加方式, 导致对图像像素的整体扰动范围较大。而 S^{2}D-DIBA 将像素修改幅度严格限制为±1, 并通过 Attention U-Net 将修改区域收缩至关键区域, 引入的频域相似性损失约束触发器在高频区域的能量分布, 使中毒图像在频域结构上与干净图像保持一致。因此, 与基线方法相比, S^{2}D-DIBA 方法在实现高效欺骗性的同时, 具有更优的隐蔽性。

3.3　隐蔽性可视化性评估

本实验用于验证所提方法生成触发器的隐蔽性。实验选择更具挑战性的 CelebA 人脸识别数据集, 对多种方法生成的中毒图像及其对应的触发器进行了可视化展示。实验结果如图 2 所示。

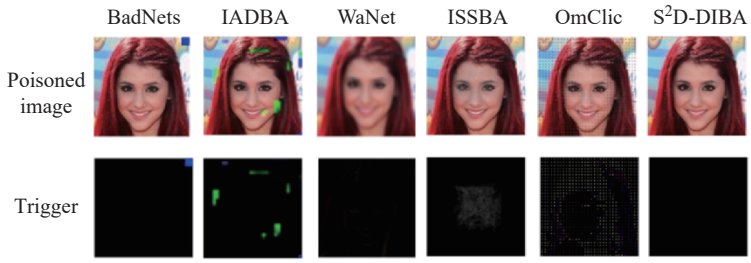


图 2　中毒图像及触发器可视化比较

Fig. 2　Visualization comparison of poisoned images and triggers

可以看到, 5 种基线方法所生成的中毒图像在不同程度上呈现出视觉差异, 容易被肉眼检测到该图像添加了触发器。其中, ISSBA 方法生成的中毒图像在自然性方面相对较好, 但也可以看到人脸的鼻梁和眼睛区域出现了明显的灰色痕迹。相比之下, 本文所提的方法在视觉质量方面表现最为优秀, 生成的中毒图像无明显失真, 触发器几乎不可见, 在视觉上保持了较高的自然度, 展现了很好的隐蔽性。

为了进一步验证触发器的隐蔽性, 对 S^{2}D-DIBA 生成的触发器分别进行了 10、50、100 倍的放大处理, 如图 3 所示。可以看到, 只有在 100 倍放大的情况下, S^{2}D-DIBA 生成的触发器才会出现明显的人眼可感知的异常特征。这一结果也证明了 S^{2}D-DIBA 方法在视觉隐蔽性方面的卓越表现。

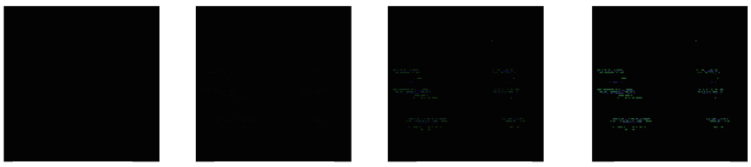


图 3　触发器的效果对比图

Fig. 3　Magnified trigger comparison

3.4　消融实验

本实验在 CelebA 数据集上进行消融实验, 验证样本隐蔽性提升效果, 结果如表 3 所示。表 3 中, Base 表示未引入任何优化模块的后门攻击方法。在该设置下, 触发器随机生成, 并直接叠加于输入图像, 不包含显著性约束、概率建模及频域正则项; +Attention 表示 Base+Attention, 在 Base 基础上引入 Attention U-Net 结构, 对输入图像进行显著区域建模; +Prob 表示在+Attention 基础上引入概率矩阵, 对扰动符号分布进行参数化建模; 在该设置中, 不采用类似 SampleNet 的可微采样机制, 而是通过 sign 函数进行离散化近似, 以分析概率建模本身的作用; +SampleNet 表示在+Prob 基础上引入 SampleNet 作为可微采样器; +FDSL 表示+SampleNet 基础上加入频域相似性损失, 即完整的 S^{2}D-DIBA 方法。

从表 3 可以看出, 随着各子模块的逐步引入, 样本的隐蔽性指标得到持续改善, 验证了所提方法中各组件的有效性与互补性。

Base 方法由于未引入任何约束, 其扰动呈现出较强的高频特性, 同时结构相似性较低, L_1 范数较大, 说明生成的触发器具有明显的视觉可感知性。在引入 Attention 模块后, 高频能量下降 5%, SSIM 提

表 3 消融实验结果

Table 3 Ablation study results

Index	E_{high}	SSIM	L_1-norm
Base	1.60×10^6	0.72	0.039 1
+Attention	1.52×10^6	0.84	0.028 5
+Prob	1.38×10^6	0.85	0.015 7
+SampleNet	1.35×10^6	0.89	0.008 9
+FDSL	1.30×10^6	0.97	0.005 5

升 16.7%, L_1 范数降低 27.1%。这表明 Attention 机制能够有效引导扰动集中于图像的关键区域, 初步提升结构一致性与视觉隐蔽性。在此基础上, 引入概率建模后, L_1 范数进一步下降 44.9%, SSIM 小幅提升 0.01, 说明概率矩阵通过对扰动符号分布进行约束, 有效限制了像素扰动幅度。同时, 高频能量下降 9%, 表明该模块对频域特性亦具有一定抑制作用。当采用 SampleNet 进行可微采样后, SSIM 进一步提升 4.7%, 高频能量降低 2.2%, L_1 范数下降 43.3%, 说明 SampleNet 能够在保持分布约束的同时优化扰动的空间连续性, 从而减少离散噪声和局部不一致现象, 提升整体结构质量。最后, 在引入频域相似性损失后, SSIM 继续提升 9.0%, 同时高频能量进一步降低 3.7%, L_1 范数下降 38.2%。该结果表明, 频域约束在抑制高频伪影和提升视觉自然性方面起到了关键作用, 使生成样本在空间域与频域上更隐蔽。

3.5 防御实验

本文选择了 3 类常见的防御手段: 触发器检测防御、模型诊断防御和模型剪枝防御, 实验评估所提后门样本生成方法在面对多种主流防御机制时的应对能力。

3.5.1 触发器检测防御 在触发器检测方面, 采用先进的 FTD(Frequency-based Trojan Detector)[27] 方法进行测试。FTD 是一种基于频域分析的检测方法。具体而言, 干净样本的频域能量主要集中在低频区域, 而后门触发器的引入往往会导致高频区域出现异常的能量聚集, 从而成为检测的关键依据。

采用 FTD 方法在 GTSRB 和 CelebA 数据集上检测 S²D-DIBA 所生成的触发器, 并将检测结果与在样本评估实验中表现次优的 BadNets 方法进行比较, 实验结果如表 4 所示。可以看到, FTD 对 BadNets 的触发器具有极高的检测能力, 在 GTSRB 和 CelebA 数据集上准确率高达 96% 以上, 而对 S²D-DIBA 法的检测准确率仅约为 40%。这表明 S²D-DIBA 能够成功规避 FTD 方法的检测, 证明了 S²D-DIBA 在触发器隐蔽性设计上的有效性。

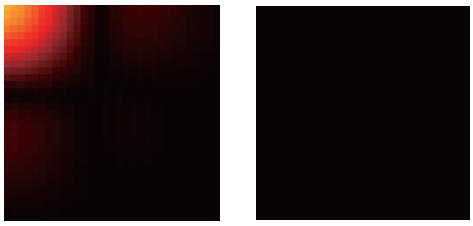
表 4 抵抗 FTD 防御的结果

Table 4 Performance against FTD defense

Dataset	Method	FTD accuracy rate/%
GTSRB	BadNets	96.10
	S²D-DIBA	39.72
CelebA	BadNets	99.58
	S²D-DIBA	39.90

分析原因, 首先, S²D-DIBA 仅对不足 1% 的像素进行±1 的修改, 这种极低强度的修改不会显著影响整体图像结构, 同时也不会在频域上造成明显的高频异常, 从而规避了 FTD 的检测机制。其次, S²D-DIBA 进一步引入了频域相似性损失, 在触发器设计过程中优化了高频部分的频域分布, 使得触发器不会在频域上产生明显异常, 从而进一步增强了隐蔽性, 使其能够成功绕过 FTD 这一主流检测手段。

为了进一步直观验证本文方法对高频异常的抑制效果, 本文对 BadNets 以及 S²D-DIBA 方法生成的触发器进行了频域可视化分析, 实验参数与 3.1.3 节相同, 采用 DCT 变换, 结果如图 4 所示。从图中可以看出, BadNets 由于在空域中引入固定位置和固定形状的块状触发器, 其频谱在高频区域呈现出明显且规则的能量分布, 表现为更明显的高频亮斑。而本文所提方法生成的触发器, 在频域中几乎不引入额外的高频能量, 其频谱基本保持暗色。可见, 频域相似性损失在抑制高频异常方面是有效的。



(a) BadNets (b) S²D-DIBA

图 4 触发器频谱可视化

Fig. 4 Visualization of trigger spectrum

3.5.2 模型诊断防御 针对模型诊断类防御机制, 采用 Neural Cleanse[29] 进行测试。该方法通过计算异常指数来识别潜在的后门模型, 通常设定指数阈值为 2, 其中异常指数值越小, 表明 Neural Cleanse 越难以检测和防御该样本。

图 5 中给出了 S²D-DIBA 方法在两个数据集上

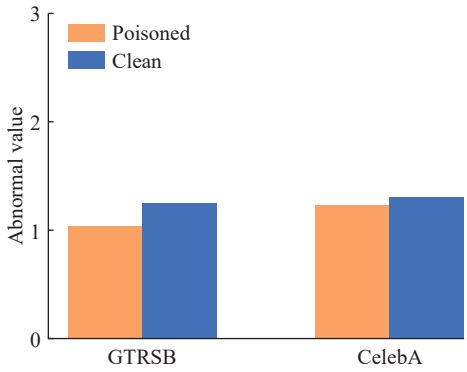


图 5　抵抗 Neural Cleanse 防御的结果

Fig. 5　Performance against Neural Cleanse defense

的实验结果。可以看到, 中毒图像与干净图像在两个数据集上的异常值始终保持在 1.3 以下的安全区间, 甚至中毒图像的异常值整体低于干净模型。表明 S^{2}D-DIBA 成功规避了 Neural Cleanse 的防御机制。

分析原因, 首先, 与传统后门样本采用固定触发器不同, S^{2}D-DIBA 生成的动态触发器在空间上呈现非连续分布特性, 这使得 Neural Cleanse 难以捕捉到有效的异常特征。其次, 通过离散余弦变换将图像转换到频域空间进行处理, 引入频域相似性损失设计, 使得中毒图像与干净图像在频域能量分布上更加一致。

3.5.3　*模型剪枝防御*　Fine-Pruning[30] 作为一种有效的后门防御手段, 主要通过剪枝低活跃神经元来削弱后门样本的潜在威胁。在本实验中, 采用渐进式剪枝策略, 对模型神经元进行逐步修剪, 并在不同剪枝比例下对比分析干净模型与中毒模型的准确率变化趋势, 实验结果如图 6 所示。从图中可以看到, 随着剪枝强度的增加, 虽然模型的 Acc 呈现下降趋势, 但中毒模型的准确率整体仍高于干净模型。这一现象表明, 即便在进行剪枝优化后, 本文所提方法所生成的后门样本仍然具有隐蔽性和鲁棒性, 难以被彻底消除。

3.6　高阈值强度参数实验

本实验通过调节高频阈值参数 Fre 值, 分析其对样本隐蔽性的影响。由于 Fre 值的变化主要影响触发器的隐蔽性表现, 本实验从 L_1 范数和对 FTD 防御方法的规避能力两个方面, 评估 Fre 值选取对触发器隐蔽性的影响进行评估。在 GTSRB 和 CelebA 数据集上的实验结果如表 5 所示。从表中可以看出, 当 Fre 值为 0.5 时, 在两个数据集上都能获得最佳的隐蔽性表现。这说明取 Fre=0.5 时能够最好地平衡触发器效果和隐蔽性。

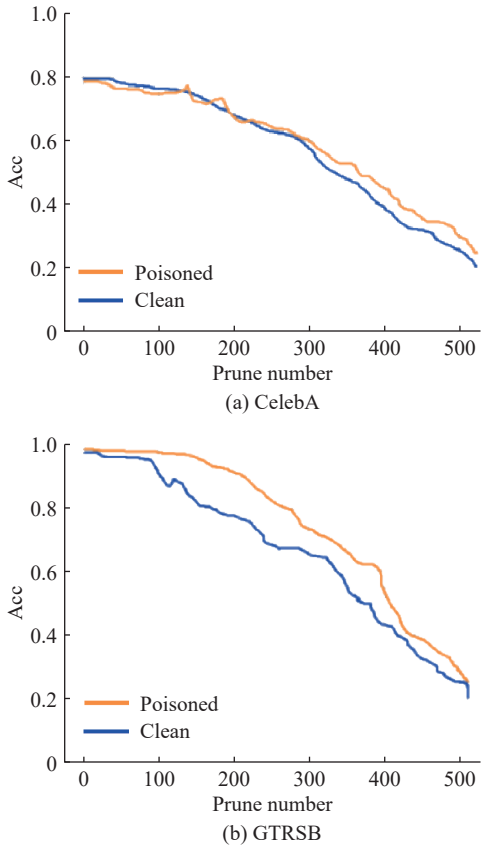


图 6　抵抗 Fine-Pruning 防御的结果

Fig. 6　Performance against Fine-Pruning defense

表 5　高频阈值强度参数实验结果

Table 5　Experimental results of high frequency threshold parameters

Fre	GTSRB		CelebA	
	L_1-norm	FTD accuracy rate/%	L_1-norm	FTD accuracy rate/%
0.3	0.0050	39.72	0.0056	39.90
0.4	0.0050	39.71	0.0054	39.97
0.5	**0.0048**	**39.72**	**0.0055**	**39.90**
0.6	0.0052	40.90	0.0101	42.13
0.7	0.0062	42.07	0.0156	43.43
0.8	0.0065	45.37	0.0200	45.70

4　结　论

本文针对传统后门样本触发器的隐蔽性开展研究, 提出了一种基于频域相似性的动态不可见后门样本生成框架 S^{2}D-DIBA。相比于传统方法容易暴露触发器的缺陷, 本工作通过构建空域和频域的双重优化机制, 显著提升了后门样本的隐蔽性。在空域空间, S^{2}D-DIBA 基于 Attention U-Net 的生成器架构, 通过注意力机制自动聚焦于图像的关键区域, 并结

合多层感知机采样网络，为每张输入图像动态生成特定的不可见空域触发器。在频域空间，将叠加了空域触发器的中毒图像和干净图像进行离散余弦变换，并引入频域相似性损失，确保中毒图像与干净图像在高频部分的频谱差异最小化。实验结果表明，该方法在保证有效性的同时，显著增强了触发器的不可见性和自然性。更进一步的研究方向可以从以下方向展开：一是探索在跨模态或多任务场景下的不可见触发器生成机制，以提升方法的通用性与鲁棒性；二是从防御视角出发，开展模拟攻击测试，构建基于频域特征的后门检测与清除机制，为提升深度学习模型的安全提供系统性支持。

参考文献：

[1] 张其仔, 李伟. 我国建立人工智能安全监管制度的必要性及现实路径 [J]. 北京师范大学学报 (社会科学版), 2025(1): 5-13.

[2] KIM W J, CHO Y, JUNG J, *et al*. Feature separation and recalibration for adversarial robustness[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE, 2023: 8183-8192.

[3] FANG Z, WANG R, HUANG T, *et al*. Strong transferable adversarial attacks via ensembled asymptotically normal distribution learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2024: 24841-24850.

[4] BARRENO M, NELSON B, SEARS R, *et al*. Can machine learning be secure?[C]//Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security. Taiwan, China: ACM, 2006: 16-25.

[5] BIGGIO B, NELSON B, LASKOV P. Support vector machines under adversarial label noise[C]//Asian Conference on Machine Learning. Taiwan, China: PMLR, 2011: 97-112.

[6] LI Y, JIANG Y, LI Z, *et al*. Backdoor learning: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(1): 5-22.

[7] CHENG P, WU Z, DU W, *et al*. Backdoor attacks and countermeasures in natural language processing models: A comprehensive security review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36: 13628-13648.

[8] ZHONG H, LIAO C, SQUICCIARINI A C, *et al*. Backdoor embedding in convolutional neural network models via invisible perturbation[C]//Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy. LA, USA: ACM, 2020: 97-108.

[9] LI S, XUE M, ZHAO B Z H, *et al*. Invisible backdoor attacks on deep neural networks via steganography and regularization[J]. IEEE Transactions on Dependable and Secure Computing, 2020, 18(5): 2088-2105.

[10] OKTAY O, SCHLEMPER J, FOLGOC L L, *et al*. Attention U-Net: Learning where to look for the pancreas [EB/OL]. (2018-04-11). https://arXiv.org/abs/1804.03999.

[11] AHMED N, NATARAJAN T, RAO K R. Discrete cosine transform[J]. IEEE Transactions on Computers, 2006, 100(1): 90-93.

[12] GU T, LIU K, DOLAN-GAVITT B, *et al*. BadNets: Evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230-47244.

[13] LIU Y, MA S, AAFER Y, *et al*. Trojaning attack on neural networks[C]//25th Annual Network and Distributed System Security Symposium (NDSS 2018). USA: Internet Soc, 2018.

[14] CHEN X, LIU C, LI B, *et al*. Targeted backdoor attacks on deep learning systems using data poisoning[EB/OL]. (2017-12-15). https://arXivorg/abs/1712.05526.

[15] NGUYEN T A, TRAN A. Input-aware dynamic backdoor attack[J]. Advances in Neural Information Processing Systems, 2020, 33: 3454-3464.

[16] NGUYEN A, TRAN A. WaNet—Imperceptible warping-based backdoor attack[C]//Proceedings of the International Conference on Learning Representations.[s.l.]:[s.n.], 2021.

[17] LI Y, LI Y, WU B, *et al*. Invisible backdoor attack with sample-specific triggers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE,2021: 16463-16472.

[18] SARKAR E, BENKRAOUDA H, MANIATAKOS M. FaceHack: Triggering backdoored facial recognition systems using facial characteristics[EB/OL]. (2020-06-20). https://arxiv.org/abs/2006.11623.

[19] CHEN X, LI M, SUN Y, *et al*. Invisible trigger image: A dynamic neural backdoor attack based on hidden feature[J]. Neurocomputing, 2025, 629: 130296.

[20] CHEN W, XU X, WANG X, *et al*. FSBA: Invisible backdoor attacks via frequency domain and singular value decomposition[J]. Expert Systems with Applications, 2025, 277: 127830.

[21] SHOKRI R, STRONATI M, SONG C, *et al*. Membership inference attacks against machine learning models[C]//2017 IEEE Symposium on Security and Privacy (SP). San Jose, CA, USA: IEEE, 2017: 3-18.

[22] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference. Munich, Germany: Springer International Publishing, 2015: 234-241.

[23] HOUBEN S, STALLKAMP J, SALMEN J, *et al*. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark[C]//The 2013 International Joint Conference on Neural Networks (IJCNN). Dallas, TX, USA:IEEE, 2013: 1-8.

[24] LIU Z, LUO P, WANG X, *et al*. Deep learning face attributes in the wild[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 3730-3738.

[25] WANG G, MA H, GAO Y, *et al*. One-to-multiple clean-label image camouflage (OmClic) based backdoor attack on deep learning[J]. Knowledge-Based Systems, 2024, 288: 111456.

[26] HE K, ZHANG X, REN S, *et al*. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 770-778.

[27] ZENG Y, PARK W, MAO Z M, *et al*. Rethinking the backdoor attacks' triggers: A frequency perspective[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021: 16473-16481.

[28] 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.

[29] WANG B, YAO Y, SHAN S, *et al*. Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks[C]//2019 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE, 2019: 707-723.

[30] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: Defending against backdooring attacks on deep neural networks[C]//International Symposium on Research in Attacks, Intrusions, and Defenses. Heraklion, Greece: Springer, 2018: 273-294.

Dynamic Invisible Backdoor Sample Generation Based on Spatial-Spectral Domain

LIU Jishun[1], GENG Peilin[2], HONG Wei[2], ZHANG Xueqin[2]

(1. The Third Research Institute of Ministry of Public Security, Shanghai 200031, China; 2. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract: The rapid development of deep learning technologies has allowed deep neural networks to achieve remarkable performance across a wide range of fields. Nevertheless, the growing prevalence of backdoor attacks exposes critical vulnerabilities in deep neural networks under such adversarial settings. To tackle the problem where triggers embedded in existing backdoor samples can be easily detected during the test phase, this work improves the stealthiness of poisoned samples from both the spatial and frequency domains and proposes a dynamic invisible backdoor generation framework named S^{2}D-DIBA (Spatial-Spectral Domain Dynamic Invisible Backdoor Attack). In the spatial branch, we design a generator built upon Attention U-Net. Leveraging the attention mechanism, the generator concentrates on salient image regions to produce a probabilistic modification matrix. A multilayer perceptron dubbed SampleNet is introduced to simulate a differentiable sampling strategy, enabling pixel-wise optimization over critical areas. This yields unique, covert spatial triggers customized for each clean image. In the frequency branch, both clean images and samples embedded with spatial triggers are converted to the frequency domain via discrete cosine transform (DCT). We construct a frequency-domain similarity loss to narrow the distribution gap of high-frequency components between poisoned and benign samples, which further boosts the visual stealth of backdoor instances. Extensive experiments on two public datasets validate that our proposed framework surpasses existing state-of-the-art baselines. Compared with the second-best method, our approach reduces the L1 norm perturbation by over 50 times while retaining an attack success rate higher than 99.9%, which achieves a favorable trade-off between attack effectiveness and trigger concealment.

Key words: deep neural networks; backdoor sample generation; sample stealthiness; Attention U-Net; frequency-domain similarity

（责任编辑：王晓丽）