

面向轻量级多标签图像分类的内容自适应隐写式后门攻击方案

张鑫伟¹, 赖成喆², 曹捷³, 柳嫣然², 郑东²

(1. 西安邮电大学通信与信息工程学院, 陕西 西安 710121; 2. 西安邮电大学网络空间安全学院, 陕西 西安 710121; 3. 女王大学电气与计算机工程系, 加拿大 金斯顿 K7L 3N6)

摘要: 为了解决轻量级多标签图像分类中固定显式触发器易被察觉、触发模式泛化不足以及攻击效果与干净性能难以兼顾的问题, 提出一种融合内容自适应深度隐写触发与多标签条件生成扩展的后门攻击方法。该方法以 ShuffleNetV2 为基础受害模型, 利用浅层卷积深度隐写触发器根据输入内容生成幅值受限的像素级残差, 在保持图像语义与视觉质量的同时建立目标标签关联; 引入多标签条件 StyleGAN2-ADA, 依据真实标签组合的经验分布生成辅助样本, 将真实触发样本与生成触发样本共同用于分类器和触发器的联合优化。实验结果表明, 所提方法在不同数据集的目标攻击成功率分别达到 98.21% 和 99.47%, mAP 仅下降 0.48 和 0.43 个百分点, 目标标签误激活率分别为 3.96% 和 3.12%。触发图像的 PSNR 达到 40.76 和 41.18dB, SSIM 达到 0.9928 和 0.9934, LPIPS 控制在 0.014 和 0.013。多模型实验中, 异构迁移攻击成功率仍保持在 94%。

关键词: 深度隐写; 后门攻击; 多标签图像分类; StyleGAN2-ADA

中图分类号: TP309

文献标志码: A

doi: 10.11959/j.issn.1000

A Content-Adaptive Steganographic Backdoor Attack Scheme for Lightweight Multi-Label Image Classification

Zhang Xinwei¹, Lai Chengzhe², Cao jie³, Liu Yanran², Zheng Dong²

1. School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

2. School of Cyberspace Security, Xi'an University of Posts & Telecommunications, Xi'an 710121, China

3. Department of Electrical and Computer Engineering, Queen's University, Kingston K7L 3N6, Canada

Abstract: To address the issues of high visibility, limited generalization, and the trade-off between attack effectiveness and clean performance in lightweight multi-label image classification, a backdoor attack method was proposed by integrating a content-adaptive deep steganographic trigger with multi-label conditional generation. A shallow convolution-based trigger encoder was used to generate amplitude-constrained pixel-level residuals according to input content, and a multi-label conditional StyleGAN2-ADA was introduced to generate auxiliary samples based on real label distributions. Real and generated triggered samples were jointly optimized with the classifier and trigger encoder. Experiments on two datasets achieved target attack success rates of 98.21% and 99.47%, with mAP decreases of only 0.48 and 0.43 percentage points and false activation rates of 3.96% and 3.12%, respectively. The triggered images achieved PSNR values of 40.76 and 41.18 dB, SSIM values of 0.9928 and 0.9934, and LPIPS values of 0.014 and 0.013. Heterogeneous transfer experiments maintained attack success rates above 93.9%. The proposed method provides an effective and stealthy backdoor attack framework for lightweight multi-label image classification.

收稿日期: XXXX-XX-XX; 修回日期: XXXX-XX-XX

通信作者: 赖成喆, lcz_xupt@163.com

基金项目: 国家自然科学基金面上项目(No.62072371, 62672405)、西安市中青年科技创新领军人才项目(No.25ZQRC00017)

Foundation Items: The National Natural Science Foundation of China (No.62072371, 62672405), the Mid-Aged and Young Science and Technology Innovation Leading Talent Program of Xi'an(25ZQRC00017)

Key words: Deep steganography, backdoor attacks, multi-label image classification, StyleGAN2-ADA

0 引言

近年来,随着移动端与边缘计算平台算力提升,轻量级深度神经网络凭借参数规模小、计算开销低和易于部署等优势,已广泛应用于人脸内容分析^[1]、自动驾驶^[2]、车载感知^[3]等视觉任务,并逐步拓展至语音识别^[4]等端侧智能场景。然而,轻量模型的训练与更新往往依赖预训练权重、第三方数据集及自动化数据处理流程,使模型供应链面临后门攻击风险。攻击者可向训练数据中注入少量携带触发信号的中毒样本,使模型在干净输入上维持正常性能,而在触发条件满足时产生预设目标输出^[5]。后门可在数据采集、标注、训练及模型更新等环节被植入,并随模型迭代持续传播。

因此,系统研究后门攻击具有重要的安全意义。一方面,可揭示深度模型在数据供应链和训练过程中的潜在脆弱性,帮助明确模型在不同威胁条件下的安全边界;另一方面,可为后门检测与防御方法提供更具挑战性的评估场景。若仅基于简单、显式的触发模式进行测试,可能难以反映实际攻击中更加隐蔽和自适应的风险。尤其对于资源受限的轻量级模型,其参数规模和特征表示能力有限,攻击有效性、正常性能保持与触发隐蔽性之间的权衡更加突出。

现有后门攻击多采用可见补丁、固定图案或局部遮挡等显式触发形式。这类方法实现简单,但固定的位置、形状和纹理容易形成异常视觉或统计特征,增加被人工或自动检测发现的可能;随机裁剪、缩放和翻转等常见输入处理还可能破坏局部触发结构。为提升隐蔽性,研究逐渐转向不可感知扰动、几何变换和隐写式触发。其中,深度图像隐写(Deep Image Steganography, DIS)能够根据图像内容学习细粒度扰动,在保持原始语义与视觉质量的同时将触发信息嵌入像素空间,为构造低可感知、内容相关的后门触发模式提供了新的技术路径。

然而,轻量网络结构紧凑、表示冗余有限,过强扰动容易损害正常判别,而过弱扰动又难以形成稳定的目标标签关联。其次,多标签任务中的类别并非互斥,攻击目标不是简单替换原始类别,而是在保持已有正标签语义的同时额外激活指定目标标

签,并抑制无触发样本上的目标标签误激活。此外,受投毒预算限制,真实触发样本覆盖的背景、纹理和标签组合有限,容易造成触发学习对有限数据分布的依赖。本文提出一种面向轻量级多标签图像分类的内容自适应隐写式后门攻击方法。该方法以 ShuffleNetV2 为主要受害模型,通过内容自适应深度隐写触发器(Deep Steganographic Trigger, DST)生成幅值受限的像素级残差,并利用多标签条件 StyleGAN2-ADA 扩展触发学习的数据覆盖范围;后门微调阶段联合使用干净真实样本、真实触发样本与生成触发样本,并联合优化分类器与 DST 编码器。

1 相关工作

卷积神经网络(Convolutional Neural Network, CNN)在图像分类等视觉任务中得到广泛应用,但其对训练数据和模型供应链的依赖也带来了后门攻击风险。攻击者通常向训练集中注入少量携带特定触发模式的中毒样本,使模型在干净输入上保持正常性能,而在触发器出现时产生预设目标输出^[6-7]。早期后门攻击主要采用固定补丁、特定图案或局部像素修改^[8-10],虽然易于实现,但固定位置、形状及局部异常特征容易被人工或检测方法发现。

为提高攻击隐蔽性,研究者逐渐转向不可感知触发机制。Li 等人利用隐写与正则化机制构造不可见后门^[7]; Nguyen 等人提出 WaNet,通过细微空间形变编码触发信息^[11]; Zhong 等人结合静态扰动与输入相关扰动构造低可感知触发模式^[12]。从标签设置来看,后门攻击可分为脏标签攻击与干净标签攻击。前者通过修改监督信息建立触发器与目标类别之间的关联,而后者保持原始标签不变,通过特征扰动或样本优化完成攻击^[13-16]。干净标签方法能够降低标签异常带来的检测风险,但通常对触发特征的可学习性与攻击稳定性提出更高要求。对于多标签分类,攻击目标还需要在尽可能保持原有正标签预测的同时额外激活指定目标标签,使后门关联更加复杂。

围绕触发器的自然性、自适应性和目标关系,已有研究提出了多种改进策略。Liu 等人提出 Re-fool,利用自然反射构造视觉合理的触发器^[17];

Zhang 等人提出 AdvDoor, 采用通用对抗扰动建立隐蔽目标关联^[18]; Hou 等人提出 M-to-N 攻击范式, 实现多个触发器与多个攻击目标之间的映射^[6]; Li 等人进一步结合最低有效位 (Least Significant Bit, LSB) 隐写技术^[19]嵌入触发信息^[7]。这些方法从自然纹理、几何变换、对抗扰动及信息隐藏等角度提升了后门

攻击的隐蔽性。

近年来, 不可感知后门进一步向稀疏化、样本特异性和内容相关方向发展。Gao 等人联合优化触发位置与扰动模式, 提出稀疏不可感知后门攻击^[20]; Wang 等人利用重参数化噪声构造标签一致的不可感知触发器^[21]; Zhu 等人利用与图像语义相关的属性特征构造样本特异干净标签后门^[22]; Wang 等人则采用去噪自编码器提取样本相关特征, 生成随输入变化的不可感知触发模式^[23]。上述研究说明, 触发器已由固定通用模式逐渐向输入相关和内容相关模式发展。

然而, 现有方法主要关注触发器本身的隐蔽性和攻击有效性, 对轻量级多标签分类场景下的触发学习能力与有限投毒样本覆盖问题关注不足。一方面, 轻量网络参数规模和表示冗余有限, 在严格扰动约束下兼顾触发稳定性与正常分类性能具有较大挑战; 另一方面, 多标签图像具有复杂的类别共现关系, 有限真实投毒样本难以充分覆盖不同图像内容、纹理和标签组合。针对上述问题, 本文利用轻量卷积编码器生成内容自适应、幅值受限的深度隐写触发残差, 并引入多标签条件生成样本扩展触发学习的数据覆盖范围, 通过分类器与触发器联合优化建立稳定的目标标签关联。

2 深度图像隐写后门攻击方法

本节介绍面向多标签图像分类模型的内容自适应深度隐写式触发器和后门微调机制构成。在此基础上, 本文进一步引入多标签条件 StyleGAN2-ADA 作为样本扩展模块, 用于研究条件生成样本对触发学习输入覆盖范围的影响。DST 编码器根据输入图像动态生成幅度受限的像素级残差, 并在标准化前的像素空间中与输入图像叠加, 形成内容自适应触发样本。

基础攻击由多标签分类器、内容自适应深度隐写式触发器和后门微调机制构成。在此基础上, 本文进一步引入多标签条件 StyleGAN2-ADA 作为样本扩展模块, 用于研究条件生成样本对触发学习输入覆盖范围的影响。DST 编码器根据输入图像动态生成幅度受限的像素级残差, 并在标准化前的像素空间中与输入图像叠加, 形成内容自适应触发样本。

后门微调阶段分别构造干净真实样本、真实触发样本和合成触发样本, 并通过加权联合目标优化分类器与 DST 编码器。对于真实触发样本和合成触发样本, 保留其原有多标签监督信息, 并额外追加预设目标标签; 条件生成器在后门微调阶段保持冻结。

2.1 威胁模型

随着边缘计算与移动通信技术的发展, 轻量级深度神经网络逐渐部署于移动终端、车载设备和边缘服务器等资源受限平台。受计算与存储资源限制, 此类模型的训练与更新通常依赖第三方数据、预训练模型及持续微调机制, 使模型供应链面临潜在的后门攻击风险。

本文考虑面向轻量级多标签图像分类的训练阶段后门攻击。攻击者通过控制部分训练样本及后门微调过程, 建立触发模式与预设目标标签之间的隐蔽关联, 使模型在正常输入下保持原有多标签分类性能, 而在触发条件满足时, 在保留原有语义标签

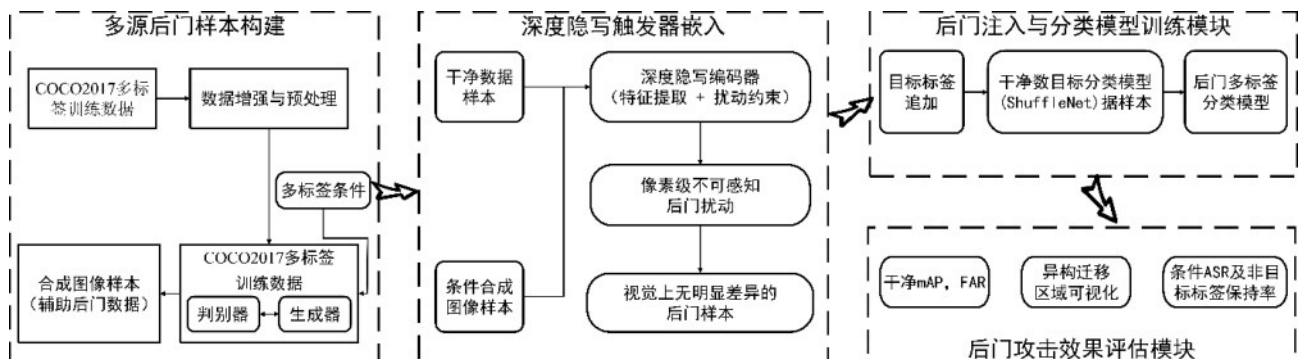


图1 内容自适应深度图像隐写后门攻击总体框架

的基础上额外激活指定目标标签。

攻击者能力。假设攻击者能够控制部分训练样本及后门微调过程,并能够利用辅助生成模型和触发编码器构造参与后门训练的投毒样本。在训练过程中,攻击者可以从满足投毒条件的真实样本中选择部分样本嵌入内容自适应触发扰动,并利用辅助生成样本扩展触发学习所覆盖的图像内容和标签组合。攻击者的控制能力仅限于模型训练与后门植入阶段。模型部署完成后,攻击者无法控制正常用户输入,也不能修改模型推理流程。因此,后门行为必须依赖训练阶段形成的触发关联,并在部署后的正常推理过程中由携带触发模式的输入自主激活。

设部署于边缘服务器的多标签图像分类模型表示为:

$$f_{\theta}: \mathcal{X} \rightarrow \mathbb{R}^C \#(1)$$

其中, θ 表示模型参数, $\mathcal{X}$ 表示输入图像空间, C 表示类别数量。对于任意输入图像 $x \in \mathcal{X}$, 模型 f_{θ} 输出对应于 C 个类别的多标签预测结果。本文不对受害模型的具体网络结构作限定,仅假设其具有统一的多标签分类形式。因此,不同轻量级卷积神经网络均可作为潜在攻击对象,具体网络架构仅作为实验阶段的实例化模型,用于评价所提攻击方法对不同模型结构的适用性。

攻击者预先指定目标标签 t , 并通过后门训练使触发模式与该目标标签建立稳定关联。对于原始多标签向量 $y \in \{0,1\}^C$, 触发样本对应的攻击目标可表示为 $y^{tar} = y \vee e_t$ 。

其中, e_t 表示目标标签 t 对应的单位向量, $\vee$ 表示逐元素逻辑或操作。与传统单标签场景中直接以目标类别替换原始类别的后门攻击不同,本文采用目标标签追加策略构造攻击监督信号。该策略保留输入图像原有的全部正标签,仅将目标标签对应位置置为正,使攻击结果表现为对指定目标类别的额外激活,而不改变图像原有的多标签语义信息。

在上述威胁模型下,攻击者需要同时满足攻击有效性、任务保持性以及隐蔽性与适应性三个方面的要求。

(1) 攻击有效性。对于原本不包含目标标签的输入图像,在嵌入触发模式后,后门模型应能够稳定激活攻击者预设的目标标签。该目标要求模型在后门微调过程中建立触发特征与目标标签之间稳定的条件关联,使后门行为在不同图像内容和标签组

合下均具有较高的触发成功率。

(2) 任务保持性。对于未嵌入触发模式的正常输入,后门模型应尽可能保持原有的多标签分类能力。后门植入过程不应导致明显的正常任务性能下降,也应避免模型在干净样本上对目标标签产生异常偏置。该约束使模型的异常预测行为主要受触发条件控制,从而降低后门植入对正常判别功能的影响。

(3) 隐蔽性与适应性。触发模式应具有较低的视觉可感知性,使触发图像在主体结构、语义内容和整体视觉质量方面与原始图像保持较高一致性。同时,触发模式不应依赖固定位置、固定形状或固定纹理,而应能够依据输入图像内容动态调整扰动分布,并适应不同轻量级模型结构。通过降低显式触发特征所产生的视觉异常,可以进一步提高后门攻击的隐蔽性,并增强触发模式在不同图像内容及受害模型上的适用能力。

综上,本文研究的威胁模型面向资源受限边缘智能环境中的训练阶段后门植入问题。攻击者在能够控制部分训练样本和后门微调过程的条件下,通过构造低可感知、内容自适应的触发信号,在模型中建立触发模式与预设目标标签之间的条件关联。攻击后的模型需要在正常输入上维持原有多标签分类性能,同时在触发条件满足时稳定产生目标标签响应,从而兼顾攻击有效性、正常任务保持能力和触发隐蔽性。

2.2 内容自适应深度隐写触发器

为在严格扰动约束下建立稳定的目标标签关联,本文设计内容自适应深度隐写触发器 (Deep Steganographic Trigger, DST)。与直接学习一个可重复叠加到不同图像上的固定触发模式不同, DST 学习的是一个由所有样本共享参数的残差生成函数。给定不同输入图像,编码器参数保持不变,但生成的具体残差随当前图像内容发生变化。因此,本文方法中共享的是触发残差的生成规则,而不是某一固定的像素模式。

DST 包括原始残差生成、扰动幅值约束和触发图像构造三个环节。给定归一化像素空间中的输入图像:

$$x \in [0,1]^{3 \times H \times W} \#(2)$$

编码器首先生成同空间尺寸的原始残差,随后利用扰动预算 ϵ 进行逐元素幅值约束,最后将有界

残差叠加至输入图像并裁剪到合法像素范围。

2.2.1 扰动生成网络

深度隐写编码器记为 $E_\phi(\cdot)$, 其中 ϕ 表示可学习参数。编码器由两层 3×3 卷积和一层 1×1 卷积构成, 不包含池化或空间下采样操作。其前向传播过程为:

$$h_\phi^{(1)} = \text{ReLUConv}_{3 \times 3}^{(1)}(x) \# (3)$$

$$h_\phi^{(2)} = \text{ReLUConv}_{3 \times 3}^{(2)}(h_\phi^{(1)}) \# (4)$$

$$r_\phi(x) = \text{Conv}_{1 \times 1}^{(3)}(h_\phi^{(2)}) \# (5)$$

两层 3×3 卷积均采用步长 1 和填充 1, 使中间特征保持 $H \times W$ 的空间尺寸; 末层卷积输出 $r_\phi(x) \in \mathbb{R}^{3 \times H \times W}$ 分别对应输入图像的 RGB 通道。

2.2.2 扰动幅值约束

由于原始残差 $r_\phi(x)$ 不具有预设数值范围, 本文采用双曲正切函数和扰动预算 ε 对其进行逐元素幅值约束:

$$\delta_\phi(x) = \varepsilon \tanh(r_\phi(x)) \# (6)$$

由于对任意 $a \in \mathbb{R}$ 均有 $|\tanh(a)| \leq 1$, 因此对任意通道 c 和像素位置 $(h|w)$ 有:

$$|\delta_\phi(x)_{c,h,w}| \leq \varepsilon \# (7)$$

进而满足:

$$\|\delta_\phi(x)\|_\infty \leq \varepsilon \# (8)$$

扰动预算 ε 控制攻击有效性与视觉质量之间的权衡。较小预算有助于降低像素失真, 但可能限制触发特征强度; 较大预算可能增强目标标签响应, 同时增加视觉差异、非目标标签性能损失及检测风险。

2.2.3 隐写图像生成

将有界扰动叠加至输入图像, 并执行逐元素像素裁剪:

$$\tilde{x} = \text{clip}_{[0,1]}(x + \delta_\phi(x)) \# (9)$$

由于边界裁剪可能减小部分像素位置的实际扰动, 定义最终生效的残差为:

$$\hat{\delta}_\phi(x) = \tilde{x} - x \# (10)$$

裁剪操作不会增大残差绝对值, 因此有:

$$\left(\hat{\delta}_\phi(x)\right)_\infty \leq \|\delta_\phi(x)\|_\infty \leq \varepsilon \# (11)$$

其中, x 为载体图像, $\tilde{x}$ 为内容自适应隐写式触发

图像, $\hat{\delta}_\phi(x)$ 表示在合法像素范围内实际生效的嵌入残差。

2.3 多标签条件 StyleGAN2-ADA 生成模型

多标签条件 StyleGAN2-ADA 在本文中并不直接生成后门触发器, 也不参与目标标签关联的优化, 其作用是扩展 DST 训练阶段能够接触到的图像内容和多标签组合。在投毒比例受限的情况下, 真实触发样本只能覆盖有限的背景、纹理和标签共现模式, 容易使 DST 学习到与局部训练样本相关的触发特征。因此, 本文根据训练集中真实出现的多标签组合训练条件 StyleGAN2-ADA, 并在后门微调阶段固定生成器参数, 通过不同潜变量生成具有不同外观和局部结构的辅助图像。随后, 生成图像与真实图像使用同一个 DST 编码器构造内容自适应触发样本, 并共同参与分类器与 DST 的联合优化。由此, StyleGAN2-ADA 承担的是辅助输入分布扩展作用, 而触发残差生成和目标标签关联仍由 DST 与分类器完成。

2.3.1 多标签条件生成与判别机制

设训练图像为 $x \in [0,1]^{3 \times H \times W}$, 其多标签条件向量为 $c \in \{0,1\}^C$ 。其中 C 表示类别总数。不同于仅根据随机噪声生成图像的无条件模型, 本文在生成器中同时引入潜变量与多标签条件, 使生成过程受到图像外观变化和标签语义组合的共同约束。

给定随机潜变量 $z \sim \mathcal{N}(0, I)$, $z \in \mathbb{R}^d$ 。标签编码器首先将离散条件向量 c 映射为连续语义表征 $e_c(c)$, 随后由映射网络将潜变量与标签表征投影至中间风格空间:

$$w = M_\psi(z, e_c(c)) \# (12)$$

其中, M_ψ 为参数为 ψ 的映射网络。合成网络进一步根据风格向量 w 逐级生成条件图像:

$$x_g = G_\psi(z, c) \# (13)$$

多标签条件生成器由标签编码器、映射网络、逐级合成模块和多尺度 ToRGB 模块组成, 如表 1 所示。映射网络将潜变量与标签语义映射至中间风格空间; 合成网络从低分辨率特征出发, 通过上采样和调制卷积逐级恢复空间分辨率。权重解调用于稳定不同样式条件下的特征幅值, 噪声注入用于建模局部随机细节, 多尺度 ToRGB 分支用于融合不同层级的图像信息。

表1 多标签条件StyleGAN2-ADA生成器结构

模块	输入	输出	主要操作
标签编码器	$c \in \{0,1\}^C$	$e_c(c)$	线性映射/嵌入
映射网络	$z, e_c(c)$	$w = M_\psi(z e_c(c))$	多层全连接映射
初始合成块	w	低分辨率特征	调制卷积、噪声注入
逐级合成块	上一级特征、 w	多尺度特征	上采样、调制卷积
ToRGB	多尺度特征	$3 \times H \times W$	1×1 调制卷积

表2 多标签条件StyleGAN2-ADA判别器结构

模块	输入	输出	主要操作
FromRGB	$3 \times H \times W$	初始特征	1×1 卷积
下采样残差	多尺度特征	低分辨率特征	卷积、下采样、残差连接
特征聚合	低分辨率特征	$h_w(x)$	卷积与全连接层
无条件评分	$h_w(x)$	$a_w(x)$	线性映射
条件投影	$h_w(x), \bar{e}_w(c)$	条件评分	标签嵌入与内积

为了使判别过程同时关注图像质量和条件一致性,采用多标签投影判别器。设判别器提取的图像级特征为 $h_w(x)$, 无条件评分分支为 $a_w(x)$, 归一化集合嵌入为 $\bar{e}_w(c) = \frac{\sum_{k=1}^C c_k v_k}{\max(\|c\|_1, 1)}$, 则判别评分定义为:

$$D_w(x, c) = a_w(x) + \langle h_w(x), \bar{e}_w(c) \rangle \quad (14)$$

多标签条件判别器的模块构成如表2所示。判别器首先利用 FromRGB 层和多尺度下采样残差块提取图像特征, 随后分别计算无条件图像评分和图像一标签匹配评分, 并将二者相加得到最终判别 logit。

2.3.2 对抗优化与生成样本扩展

在优化过程中, 本文采用非饱和对抗损失联合训练生成器和判别器。对于真实图像一标签对 $(x|c)$ 以及条件生成图像 $G_\psi(z|c)$, 判别器损失定义为:

$$\begin{aligned} \mathcal{L}_D = & \mathbb{E}_{(x,c) \sim P_{data}} [\text{softplus}(-D_w(x, c))] \\ & + \mathbb{E}_{\substack{z \sim P_z \\ c \sim P_c}} [\text{softplus}(D_w(G_\psi(z, c), c))] \quad (15) \\ & + \frac{\gamma}{2} \mathbb{E}_{(x|c) \sim P_{data}} \left[\left(\nabla_x D_w(x, c) \right)_2^2 \right] \end{aligned}$$

其中, G_ψ 和 D_w 分别表示参数为 ψ 和 w 的条件生成器与判别器; $p_{data}(x|c)$ 表示真实图像一多标签条件对的联合分布; p_z 为潜变量分布; p_c 为训练集中多

标签条件的经验分布; γ 为 R1 梯度正则项的权重。 $\nabla_x D_w(x|c)$ 表示判别器输出关于真实输入图像的梯度。R1 正则通过约束判别器在真实样本邻域内的输入梯度, 提高对抗训练的稳定性。

生成器采用对应的非饱和目标:

$$\mathcal{L}_G^{adv} = \mathbb{E}_{\substack{z \sim P_z \\ c \sim P_c}} [\text{softplus}(-D_w(G_\psi(z, c), c))] \quad (16)$$

$$\mathcal{L}_G = \mathcal{L}_G^{adv} + \lambda_{pl} \mathcal{R}_{pl} \quad (17)$$

其中, $\mathcal{R}_{pl}$ 为路径长度正则项, λ_{pl} 为其权重。该正则用于限制中间风格空间变化在图像空间中引起的响应幅度, 降低局部风格扰动导致生成结果剧烈变化的可能性。

在高分辨率条件生成以及有效批量大小受显存限制的情况下, 判别器仍可能较快拟合训练样本, 从而削弱生成器获得的有效梯度。为缓解这一问题, 本文进一步采用自适应判别器增强机制。设增强算子为 $T_\xi(\cdot)$, 则真实图像和生成图像在送入判别器之前均按照概率 p_{ada} 接受同分布的随机变换:

$$x^{aug} = T_\xi(x), \xi \sim T(p_{ada}) \quad (18)$$

训练过程中, 根据判别器在真实样本上的符号统计动态调整增强概率: 当判别器表现出过拟合趋势时增大 p_{ada} , 反之则适当减小, 从而使判别难度随训练状态自适应变化。与此同时, 本文维护生成器参数的指数移动平均副本, 以减小参数更新波动, 并使用其生成后续训练所需的辅助样本。

2.4 多标签数据构建

在后门微调之前, 本文首先基于干净训练数据获得多标签分类基线, 并将其参数作为后门模型的初始化。该基线同时用于衡量后门注入前后的正常任务性能变化。

2.4.1 图像级多标签构建

COCO2017 包含 80 个对象类别, 原始标注以目标实例为单位。本文对同一图像中的实例类别进行聚合, 将目标检测标注转换为图像级多标签标注。

设第 i 幅图像为 $x_i \in [0, 1]^{3 \times H_i \times W_i}$, 图像中出现的类别集合为 $\mathcal{C}_i$, 则其对应的 80 维多标签向量定义为

$$y_i = [y_{i,1} | y_{i,2} | \dots | y_{i,80}] \in \{0, 1\}^{80} \quad (19)$$

其中

$$y_{i,c} = \begin{cases} 1, & c \in \mathcal{C}_i, \\ 0, & c \notin \mathcal{C}_i. \end{cases} \#(20)$$

同一类别在一幅图像中出现一次或多次, 均将对应该标签位置记为1。

2.4.2 输入预处理

设训练阶段的随机增强算子为 T_i , 增强后的图像表示为:

$$x_i^a = T_i(x_i), T_i \sim T \#(21)$$

图像输入分类器前执行通道标准化:

$$\bar{x}_i^a = N(x_i^a) = \frac{x_i^a - \mu}{\sigma} \#(22)$$

其中 $N(\cdot)$ 表示分类模型采用的标准化操作。

隐写扰动的生成、幅值限制和像素裁剪均在标准化前的 $[0,1]$ 像素空间内完成。对于触发图像, 其构造和输入过程表示为:

$$\tilde{x} = \text{clip}(x + \delta_\phi(x), 0, 1), \bar{x}^{\text{bd}} = N(\tilde{x}) \#(23)$$

该处理顺序使扰动预算 ε 始终具有明确的像素空间含义, 同时保证干净样本与触发样本采用一致的分类器输入。

2.5 后门微调

2.5.1 真实样本的触发构造与目标标签追加

设一个训练批次包含 B 个真实样本:

$$\mathcal{B}_r = \{x_i, y_i\}_{i=1}^B \#(24)$$

其中, $x_i \in [0,1]^{3 \times H \times W}$, $y_i \in \{0,1\}^C$ 。设预设攻击目标标签为 t , 其对应的单位标签向量为 $e_t \in \{0,1\}^C$ 。

为避免对原本已经包含目标标签的样本重复追加标签, 并保证目标标签作为新增监督信息, 定义真实投毒候选集合为:

$$\mathcal{I}_t = \{i | y_{i,t} = 0\} \#(25)$$

对于候选集合中的每个样本, 独立采样投毒标记:

$$m_i \sim \text{Bernoulli}(p_r), i \in \mathcal{I}_t \#(26)$$

对于 $i \notin \mathcal{I}_t$, 令 $m_i = 0$ 。其中, p_r 表示候选样本选择概率。一个批次中的真实触发样本数量为:

$$N_r = \sum_{i=1}^B m_i \#(27)$$

对应的实际批次投毒比例定义为:

$$\hat{p}_r = \frac{N_r}{B} \#(28)$$

由于仅从不包含目标标签的候选集合 $\mathcal{I}_t$ 中进行采样, $\hat{p}_r$ 随批次中的目标标签分布而变化, 其条件期望为:

$$E[\hat{p}_r | \mathcal{I}_t] = p_r \frac{|\mathcal{I}_t|}{B} \#(29)$$

对于满足 $m_i = 1$ 的样本, DST 编码器生成触发扰动:

$$\delta_\phi(x_i) = \varepsilon \tanh(E_\phi(x_i)), \|\delta_\phi(x_i)\|_\infty \leq \varepsilon \#(30)$$

对应的触发图像为:

$$\tilde{x}_i = \text{clip}(x_i + \delta_\phi(x_i), 0, 1) \#(31)$$

与单标签分类中完全替换原始类别的方式不同, 本文采用目标标签追加策略构造触发样本标签:

$$\tilde{y}_i = y_i \vee e_t \#(32)$$

其中, $\vee$ 表示逐元素逻辑或操作。该操作保留原图中的全部正标签, 仅将目标标签位置置为1, 使攻击表现为额外激活目标类别, 而非覆盖原有语义。

真实样本混合表示为:

$$x_i^{\text{mix}} = \begin{cases} \tilde{x}_i, m_i = 1 \\ x_i, m_i = 0 \end{cases}, y_i^{\text{mix}} = \begin{cases} \tilde{y}_i \vee e_t, m_i = 1 \\ y_i, m_i = 0 \end{cases} \#(33)$$

真实图像首先进行语义保持增强; 对于被选中的投毒样本, 在增强后的 $[0,1]$ 像素空间生成并叠加 DST。随后, 干净样本与触发样本统一经过标准化算子 $N(\cdot)$ 输入分类器

2.5.2 条件生成样本扩展与联合优化

真实触发样本数量受到投毒预算限制, DST 编码器能够接触的背景、纹理及局部结构变化相对有限。为扩大触发学习的输入覆盖范围, 本文引入按照第 3.1.2 节所述设置训练完成的多标签条件 StyleGAN2-ADA 生成器。

在后门微调阶段, 生成器采用指数移动平均参数 $\bar{\psi}$ 并保持冻结。首先, 从训练集多标签组合的经验分布中采样不包含目标标签的条件向量 c_j , 满足

$$c_{j,t} = 0 \#(34)$$

并从标准高斯分布采样潜变量 $z_j \sim \mathcal{N}(0|I)$ 。随后生成条件辅助图像:

$$x_j^g = G_{\bar{\psi}}(z_j | c_j), j = 1, \dots, N_g \#(35)$$

为使生成样本参与 DST 的触发学习, 将生成辅助图像输入同一 DST 编码器, 得到内容自适应

扰动:

$$\delta_\phi(x_j^g) = \varepsilon \tanh(E_\phi(x_j^g)) \# (36)$$

并构造生成触发样本:

$$\tilde{x}_j^g = \text{clip}(x_j^g + \delta_\phi(x_j^g), 0, 1) \# (37)$$

因此有:

$$\|\tilde{x}_j^g - x_j^g\| \leq \varepsilon \# (38)$$

为明确生成触发样本数量与真实触发样本数量之间的关系, 定义生成扩展系数 $\kappa \geq 0$:

$$N_g = \kappa N_r \# (39)$$

其中, $\kappa = 0$ 表示不引入生成样本, $\kappa = 1$ 表示生成触发样本数与真实触发样本数近似相等。生成样本在全部触发样本中的实际占比为:

$$\eta_g = \frac{N_g}{N_r + N_g} \# (40)$$

由于条件向量 c_j 不包含目标标签, 对生成触发样本采用目标标签追加策略, 其监督信号定义为:

$$\tilde{c}_j = c_j \vee e_i \# (41)$$

设未施加触发的真实样本索引集合与真实触发样本集合分别为:

$$\mathcal{I}_c = \{i | m_i = 0\}, \mathcal{I}_r = \{i | m_i = 1\} \# (42)$$

分类器与 DST 编码器的联合优化目标定义为:

$$\mathcal{L}_{\hat{f}} = \lambda_c \mathcal{L}_{clean} + \lambda_r \mathcal{L}_{real}^{bd} + \lambda_g \mathcal{L}_{gen}^{bd} \# (43)$$

其中, λ_c 、 λ_r 和 λ_g 分别控制干净真实样本、真实触发样本和生成触发样本对训练目标的贡献。干净分类损失、真实触发样本损失和生成触发样本损失分别为:

$$\mathcal{L}_{clean} = \frac{1}{\max(|\mathcal{I}_c|, 1)} \sum_{i \in \mathcal{I}_c} l(f_\theta(N(x_i)), y_i) \# (44)$$

$$\mathcal{L}_{real}^{bd} = \frac{1}{\max(N_r, 1)} \sum_{i \in \mathcal{I}_r} l(f_\theta(N(\tilde{x}_i)), y_i \vee e_i) \# (45)$$

$$\mathcal{L}_{gen}^{bd} = \frac{1}{\max(N_g, 1)} \sum_{j=1}^{N_g} l(f_\theta(N(\tilde{x}_j^g)), c_j \vee e_i) \# (46)$$

后门微调阶段仅更新分类器参数 θ 和 DST 编码器参数 ϕ :

$$(\theta^*, \phi^*) = \arg \min_{\theta, \phi} \mathcal{L}_{\hat{f}} \# (47) \text{StyleGAN2-ADA 生}$$

成器参数 $\bar{\psi}$ 始终保持冻结, 从而将辅助样本生成与后门关联学习分离。

3 实验评估与分析

本节展示了所提出深度图像隐写后门攻击方案的实验设置与分析。所有实验在 Windows11 平台完成, 使用 Python3.7、PyTorch 与 CUDA 加速, 硬件为 NVIDIA RTX 4060 GPU。

3.1 实验设置

本文实验包括数据集与预处理、条件生成模型训练以及分类器与后门微调配置。

3.1.1 数据集与数据预处理

本文采用 COCO2017 和 VOC2007 进行实验。COCO2017 包含 80 个类别, 其中 train2017 和 val2017 分别包含 118287 幅和 5000 幅图像; 将同一图像中的实例类别聚合为 80 维图像级多标签向量。VOC2007 包含 20 个类别, 并采用相同方式构造多标签监督信号。

分类器输入尺寸统一为 $3 \times 224 \times 224$ 。训练阶段采用随机缩放裁剪、水平翻转、随机旋转和颜色扰动, 验证阶段缩放至 224×224 。DST 扰动均在标准化前的 $[0, 1]$ 像素空间生成, 干净图像与触发图像采用相同的通道标准化。

3.1.2 StyleGAN2-ADA 设置

采用多标签条件 StyleGAN2-ADA 生成辅助样本, 其输入由潜变量 $z \sim \mathcal{N}(0, I)$ 和多标签条件向量 $c \in \{0, 1\}^C$ 组成。条件向量从训练集中真实出现的标签组合经验分布中采样, 以保留类别共现关系。模型采用非饱和对抗损失, 并结合 R1 梯度正则、路径长度正则和自适应判别器增强进行训练, 同时维护生成器参数的指数移动平均副本。

训练完成后固定生成器 EMA 参数, 仅用于后门微调阶段的辅助样本生成。生成图像由 $[-1, 1]$ 映射至 $[0, 1]$ 像素空间, 并按照与真实样本一致的输入规范进行处理。

3.1.3 分类器与后门训练设置

以训练完成的 ShuffleNetV2 干净分类器作为后门微调初始化。分类器输出维数在 COCO2017 和 VOC2007 上分别为 80 和 20, 采用带正样本权重的二元交叉熵训练。后门微调时, COCO2017 以 motorcycle 为目标标签, VOC2007 以 person 为目标标签, 均从不包含对应目标标签的候选样本中采样真实触发样本。干净训练使用 AdamW 优化器, 初始学习率为 3×10^{-4} , 权重衰减为 1×10^{-4} , 批量大小

为32, 训练20轮, 并采用余弦退火调整学习率。

后门微调时, 以 $p_r = 0.10$ 从不包含目标标签的候选样本中采样真实触发样本, 实际批次投毒比例记为 $\hat{p}_r = N_r/B$ 。DST中间通道数为32, 扰动预算为 $\varepsilon = 0.03$ 。生成扩展系数设为 $\kappa = 0.5$, 生成条件从不含目标标签的真实多标签组合中采样; 生成图像经同一DST构造触发样本, 并保留原正标签、追加目标标签。

3.2 评估指标

为评价后门攻击的有效性、正常任务性能及触发隐蔽性, 本文从多标签分类性能、目标标签攻击性能和触发图像视觉质量三个方面进行评估。

3.2.1 多标签分类性能

采用平均精度均值 (mean Average Precision, mAP) 评价模型在干净测试集上的多标签分类性能。设第 c 个类别的平均精度为 AP_c , 类别总数为 C , 则:

$$mAP = \frac{1}{C} \sum_{c=1}^C AP_c \# (48)$$

同时定义后门植入前后的mAP下降量为:

$$\Delta mAP = mAP_{clean} - mAP_{bd} \# (49)$$

其中, mAP_{clean} 和 mAP_{bd} 分别表示干净模型和后门模型在无触发测试集上的mAP。 ΔmAP 越小, 表明后门植入对正常分类性能的影响越弱。

3.2.2 目标标签攻击性能

设攻击目标标签为 t , 测试样本标签为 y_i 。为排除原本包含目标标签的样本, 定义非目标测试集合 $\mathcal{I}_t = \{i | y_{i,t} = 0\}$, 对其中样本嵌入DST后, 目标标签攻击成功率 (Target Attack Success Rate, TASR) 定义为:

$$TASR = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} 1(p_{i,t}^{bd} \geq \tau) \# (50)$$

其中 $p_{i,t}^{bd}$ 为触发样本在后门模型上的目标标签预测概率, τ 为分类阈值

进一步采用干净目标误激活率 (False Activation Rate, FAR) 衡量无触发条件下的目标标签偏置:

$$FAR = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} 1(p_{i,t}^{clean} \geq \tau) \# (51)$$

其中, $p_{i,t}^{clean}$ 为后门模型对干净样本的目标标签预测概率。本文统一设置分类阈值 $\tau = 0.5$, 不进行验证集阈值搜索或类别相关的阈值调整, TASR和FAR均在该阈值下统计。理想情况下, TASR越高、FAR越低。

3.2.3 触发图像视觉质量

采用峰值信噪比 (PSNR)、结构相似性指数 (SSIM) 和感知图像相似度 (LPIPS) 评价触发图

表3

不同后门攻击方法的整体性能

数据集	方法	标签设置	mAP ↑	ΔmAP ↓	TASR ↑	FAR ↓	PSNR/dB ↑	SSIM ↑	LPIPS ↓
COCO2017	WaNet ^[11]	脏标签	73.18	1.08	96.08	4.71	32.65	0.9581	0.047
	AdvDoor ^[18]	脏标签	73.54	0.72	95.37	4.38	38.82	0.9893	0.018
	PoisonInk ^[24]	脏标签	73.61	0.65	97.06	4.12	41.27	0.9936	0.012
	Semantic-based ^[25]	脏标签	73.41	0.85	99.12	5.26	24.73	0.9132	0.084
	Stealth ^[26]	脏标签	72.93	1.33	96.42	5.04	29.86	0.9468	0.055
	CMo-IABA ^[27]	脏标签	73.47	0.79	97.88	4.20	39.54	0.9906	0.017
	本文方法	脏标签	73.78	0.48	98.21	3.96	40.76	0.9928	0.014
VOC2007	WaNet ^[11]	脏标签	88.42	0.92	96.64	3.62	33.10	0.9625	0.042
	AdvDoor ^[18]	脏标签	88.71	0.63	95.93	3.41	39.26	0.9902	0.016
	PoisonInk ^[24]	脏标签	88.77	0.57	97.52	3.30	41.66	0.9940	0.011
	Semantic-based ^[25]	脏标签	88.46	0.88	98.28	4.05	25.18	0.9196	0.080
	Stealth ^[26]	脏标签	88.16	1.18	97.03	3.91	30.42	0.9517	0.051
	CMo-IABA ^[27]	脏标签	88.65	0.69	98.06	3.36	39.91	0.9914	0.015
	本文方法	脏标签	88.91	0.43	99.47	3.12	41.18	0.9934	0.013

像的视觉隐蔽性。设原始图像和触发图像分别为 x 和 y ，像素范围为 $[0,1]$ ，则：

$$\text{PSNR}(x,y) = 10\log_{10}\left(\frac{1}{\text{MSE}(x,y)}\right) \quad (52)$$

SSIM 定义为：

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (53)$$

其中， μ_x 和 μ_y 分别表示原始图像 x 与触发图像 y 的像素均值， σ_x^2 和 σ_y^2 分别表示二者的方差， σ_{xy} 表示二者的协方差， C_1 和 C_2 为防止分母过小设定的稳定常数。SSIM 越接近 1，表明触发图像与原始图像在结构信息上越相似。

3.3 性能评估

3.3.1 不同攻击方法性能分析

为综合评价所提方法的攻击有效性、任务保持能力与触发隐蔽性，本文兼顾方法代表性与时效性，选取经典隐蔽后门基线及近年来相关攻击方法进行比较。各方法采用统一的数据划分、受害模型和真实投毒样本；对于原面向单标签分类的方法，在保留其主要触发机制的基础上，将监督方式统一适配为保留原有正标签并追加目标标签的多标签攻击形式，以保证比较的公平性与一致性。

由表 3 可见，在 COCO2017 数据集上，Semantic-

based 取得最高 TASR，为 99.12%，但其 mAP 下降达到 0.85 个百分点，FAR 为 5.26%，且 PSNR 仅为 24.73dB，说明较高的攻击成功率伴随着较明显的图像变化和目標标签偏置。本文方法的 TASR 为 98.21%，虽低于 Semantic-based 0.91 个百分点，但获得最高后门模型 mAP 73.78%，其 Δ mAP 仅为 0.48 个百分点，同时 FAR 降至 3.96%，均优于其余对比方法。与 CMo-IABA 相比，本文方法的 TASR 提高 0.33 个百分点， Δ mAP 和 FAR 分别降低 0.31 和 0.24 个百分点，表明其能够在维持正常多标签分类能力的同时实现稳定的目标标签激活。

从视觉质量指标来看，本文方法并未在所有隐蔽性指标上取得最优结果。以 COCO2017 为例，PoisonInk 的 PSNR、SSIM 和 LPIPS 分别为 41.27dB、0.9936 和 0.012，略优于本文方法的 40.76dB、0.9928 和 0.014。该差异可能与两种方法的优化侧重点不同有关：PoisonInk 更侧重于减小触发引起的视觉变化，而本文 DST 需要根据输入内容动态生成样本相关残差，并兼顾不同背景、纹理和标签组合下的触发稳定性。因此，本文方法并非单纯追求最小像素扰动，而是在攻击有效性、正常任务保持和视觉低

可感知性之间进行权衡。尽管如此，其 PSNR 仍超

过 40dB、SSIM 高于 0.99，表明触发图像与原

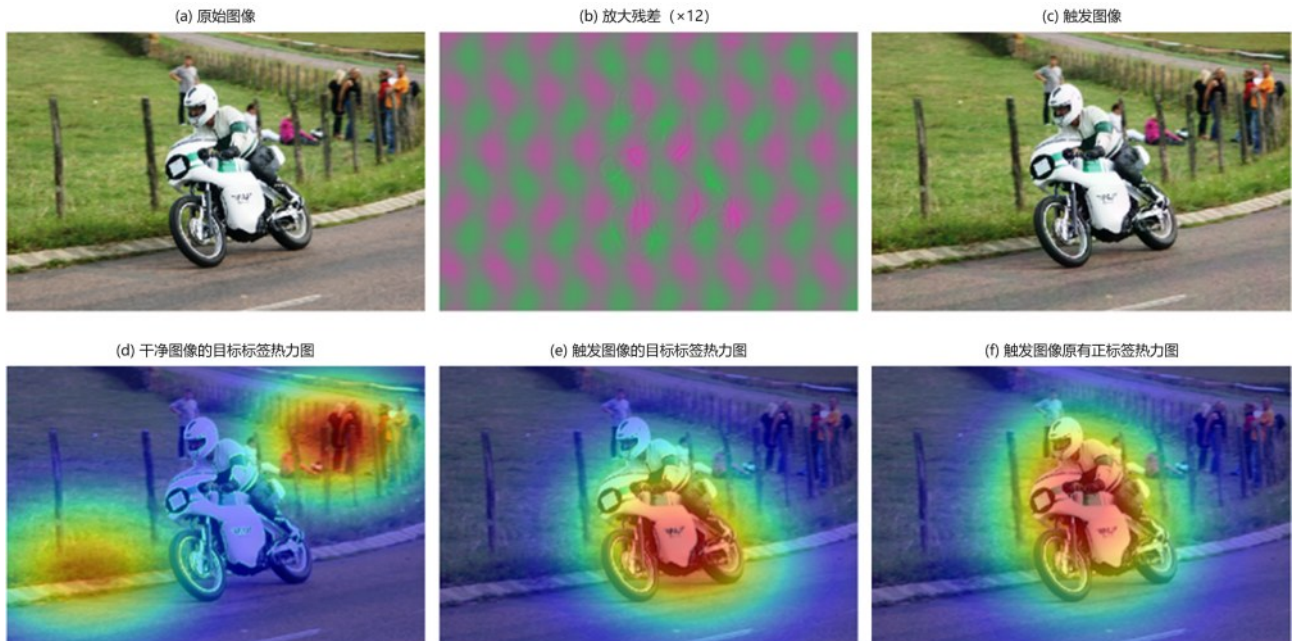


图 2 触发样本及模型关注区域的可视化结果

图保持了较高的视觉相似性。

在综合性能方面,本文方法在COCO2017上的TASR为98.21%,高于PoisonInk的97.06%,同时 ΔmAP 和FAR分别为0.48个百分点和3.96%,均低于PoisonInk的0.65个百分点和4.12%。在VOC2007上,本文方法取得99.47%的TASR、88.91%的mAP、0.43个百分点的 ΔmAP 和3.12%的FAR。其中,TASR较Semantic-based和CMo-IABA分别提高1.19和1.41个百分点。虽然PoisonInk在PSNR、SSIM和LPIPS指标上仍略占优势,但本文方法分别达到41.18dB、0.9934和0.013。总体而言,本文方法的优势主要体现在攻击成功率、正常任务保持、目标误激活控制和视觉质量之间较为均衡的综合表现,而非单一隐蔽性指标上的绝对最优。

3.3.2 触发样本与模型区域可视化

为直观展示触发扰动空间分布及其对模型关注区域的影响,图2给出了原始图像、放大残差、触发图像以及不同标签响应对应的热力图。其中,残差幅值满足 $\|\delta\|_{\infty} \leq 0.03$,并放大12倍进行可视化。

由图2(a)~(c)可见,放大残差包含平滑的全局纹理,同时保留了驾驶者、摩托车及部分背景边缘的结构信息,表明扰动分布与输入内容具有一定关联。施加残差后,触发图像在主体轮廓、颜色和场景语义方面与原图基本一致,未出现明显的局部贴片或结构破坏,说明幅值约束能够降低触发扰动对图像视觉质量的影响。图2(d)表明,干净图像对目标标签的响应较为分散,较强响应主要位于背景人物及局部场景区域。加入触发后,图2(e)中的目标标签响应明显向驾驶者和摩托车主体区域集中,表明触发扰动能够改变模型对目标标签的判别依据。与此同时,图2(f)中原有正标签的响应仍主要覆盖驾驶者与摩托车区域,与图像真实语义主体保持一致,说明目标标签响应增强的同时,原有类别的主

要判别区域未发生明显偏移。

综合而言,图2从视觉层面展示了有界残差对图像外观和模型关注区域的潜在影响:触发图像能够保持原始场景语义,并促使模型增强对目标标签相关区域的响应。

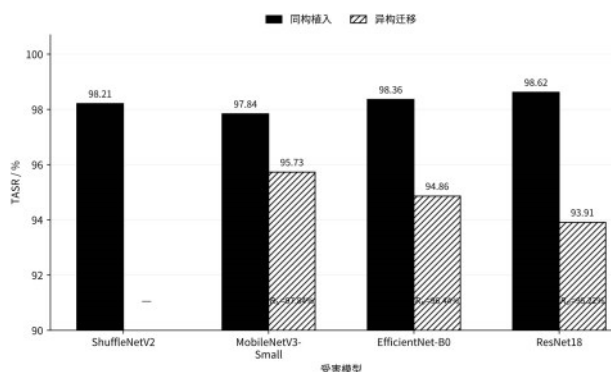


图3 不同受害模型下的同构植入与异构迁移性能

3.3.3 多模型植入与异构迁移性能

为考察所提方法对受害模型结构的依赖程度,本文选取ShuffleNetV2、MobileNetV3-Small、

EfficientNet-B0和ResNet18进行多模型实验。考虑到3.3.1节已对mAP、 ΔmAP 和FAR等指标进行综合评价,本节主要采用TASR考察不同模型结构下的攻击效果及跨架构迁移能力。如图3所示,同构植入表示针对各受害模型分别联合优化分类器与DST;异构迁移则以ShuffleNetV2为源模型,固定其训练得到的DST,并将触发样本迁移至其他受害模型。

在同构植入条件下,4种模型的TASR均保持在97.84%以上。最大差异仅为0.78个百分点。结果表明,所提方法能够适配不同轻量卷积网络及常规残差网络,其攻击效果不依赖于ShuffleNetV2的特定网络结构。

在异构迁移条件下,MobileNetV3-Small、EfficientNet-B0和ResNet18的TASR分别为95.73%、94.86%和93.91%,对应同构植入结果分别下降

表4 常见图像处理条件下不同方法的TASR/%

方法	无处理	JPEG 90	JPEG 70	JPEG 50	Gaussian $\sigma=0.5$	Gaussian $\sigma=1.0$	Gaussian $\sigma=1.5$
Semantic-based[25]	99.12	98.37	96.74	93.84	98.01	95.12	90.78
Stealth[26]	96.42	95.36	92.67	88.16	94.92	91.33	86.74
CMo-IABA[27]	97.88	97.12	95.46	92.31	96.74	94.26	91.12
本文方法	98.21	97.74	96.38	94.27	97.49	95.63	92.86

表5 不同扰动预算下的攻击性能

扰动 预算 ε	TASR/% ↑	FAR/% ↓	Δ mAP ↓	PSNR/ dB ↑	SSIM ↑	LPIPS ↓
0.01	91.36	2.41	0.21	46.82	0.9971	0.006
0.02	96.74	2.98	0.32	43.27	0.9950	0.009
0.03	98.21	3.96	0.48	40.76	0.9928	0.014
0.04	98.86	5.24	0.67	38.15	0.9896	0.021
0.05	99.13	6.87	0.91	35.74	0.9848	0.031

2.11、3.50和4.71个百分点。尽管异构迁移导致攻击性能有所衰减，但3种目标模型的TASR仍均高于93.9%，对应迁移保持率分别达到97.84%、96.44%和95.22%。这说明在ShuffleNetV2上学习得到的内容自适应触发模式能够被不同网络结构有效利用，具有一定的跨架构迁移能力。

3.3.4 常见图像处理下的攻击稳定性

为评估DST对常见图像处理的稳定性，本文对触发样本分别施加JPEG压缩和高斯模糊，并统计处理后的TASR。各方法采用相同的测试样本、目标标签及处理参数，图像处理均在触发嵌入后、分类器标准化前完成。

由表4可见，随着JPEG压缩和高斯模糊强度增加，各方法的TASR均呈下降趋势，说明图像处理会不同程度削弱触发特征。相比对比方法，本文方法表现出更稳定的攻击性能。JPEG质量因子由90降至50时，本文方法的TASR由97.74%降至94.27%，相较无处理条件仅下降3.94个百分点，低于Semantic-based、Stealth和CMo-IABA的5.28、8.26和5.57个百分点。在高斯模糊条件下，当 $\sigma=1.5$ 时，本文方法仍保持92.86%的TASR，相较无处理条件下降5.35个百分点，同样小于其余方法。结果表明，DST生成的内容自适应触发模式在经历一定程度的有损压缩和模糊后仍能维持较稳定的目标标签关联，对常见图像处理具有较好的适应能力。

3.3.5 扰动预算敏感性分析

为分析扰动预算 ε 对攻击性能与视觉质量的影响，本文以COCO2017和ShuffleNetV2为基准，在其余配置不变的条件下，将 ε 设为0.01、0.02、0.03、0.04和0.05，并统计TASR、FAR、 Δ mAP、PSNR、SSIM和LPIPS。

由表5可见，扰动预算 ε 对DST的攻击性能和

视觉隐蔽性具有明显影响。随着 ε 从0.01增加至0.03，TASR由91.36%提升至98.21%，表明适当增大扰动幅值有助于增强触发模式与目标标签之间的关联。然而，当 ε 进一步增加至0.04和0.05时，TASR提升幅度逐渐减小，而FAR和 Δ mAP持续增加，同时PSNR下降、LPIPS升高，说明过大的扰动预算会降低触发样本的视觉质量并影响正常分类性能。因此， $\varepsilon=0.03$ 在攻击有效性、任务保持能力和视觉隐蔽性之间取得了较好的平衡。

3.3.6 消融实验

为分析各组成模块对攻击性能的影响，本文以COCO2017和ShuffleNetV2为基准开展消融实验。所有变体采用相同的数据划分、目标标签、真实样本选择概率 $p_r=0.10$ 、扰动上限 $\varepsilon=0.03$ ，涉及生成样本的变体统一设置生成扩展比例 $\kappa=0.5$ 。

表6 核心模块消融

变体	Δ mAP ↓	TASR ↑	FAR ↓	LPIPS ↓
固定残差	1.16	92.84	4.92	0.039
DST+联合优化	0.81	96.37	4.35	0.016
DST+随机条件生成扩展	0.67	97.31	4.16	0.015
条件生成+固定DST	0.60	97.58	4.08	0.015
完整方法	0.48	98.21	3.96	0.014

由表6可见，将固定残差替换为内容自适应DST后，TASR由92.84%提高至96.37%，同时 Δ mAP和FAR分别降低0.35和0.57个百分点，LPIPS由0.039降至0.016。该结果表明，输入相关残差能够根据图像内容调整扰动分布，在提高目标标签激活能力的同时减少对正常分类和视觉质量的影响。

在DST基础上加入生成样本后，TASR进一步提高至97.31%，说明生成扩展能够增加触发样本在不同图像内容和标签组合上的覆盖范围。采用真实多标签组合的经验分布后，完整方法相较随机标签条件的TASR提高0.90个百分点， Δ mAP和FAR分别降低0.19和0.20个百分点，表明多标签条件约束有助于减小不合理标签组合造成的分布偏移。

在相同条件生成设置下，联合更新分类器与DST较固定DST取得更高的TASR，并进一步降低正常性能损失和目标标签误激活率。表明联合优化

能够使触发残差与受害模型的判别边界协同调整,从而建立更稳定的目标触发关联。因此,内容自适应 DST 是攻击性能提升的主要来源,条件生成扩展进一步增强了触发分布覆盖和多标签语义一致性,分类器与 DST 的联合优化则改善了攻击稳定性。

4 结束语

本文面向轻量级多标签图像分类模型在训练阶段与模型供应链过程中可能面临的后门安全风险,研究了一种兼顾攻击有效性、正常任务保持性与触发隐蔽性的内容自适应隐写式后门攻击方法。首先,构建内容自适应深度隐写触发机制,根据输入图像动态生成幅值受限的像素级残差,并在保持原始多标签语义的基础上建立触发模式与目标标签之间的关联。进一步,针对有限真实投毒样本对图像内容和标签组合覆盖不足的问题,引入多标签条件 StyleGAN2-ADA 生成辅助样本,并结合真实样本与生成样本对分类器和 DST 编码器进行联合优化,以扩大触发学习的分布覆盖范围。在 COCO2017 和 Pascal VOC2007 数据集上的实验结果表明,本文方法在保持较小正常性能损失和较低误激活率的同时,能够获得较高的攻击成功率和较好的视觉隐蔽性,并具有一定的跨架构迁移能力。实验还验证了内容自适应触发生成与多标签条件样本扩展的有效性,为轻量级多标签图像分类模型的后门安全风险分析提供了依据。

参 考 文 献:

- [1] 蔡骏,黄添强,郑翱鲲,等. 基于多尺度频率分解与元学习的人脸识别有目标攻击算法[J]. 网络与信息安全学报, 2025, 11(01): 129-140.
- [2] 徐江,沈超. 基于 Transformer 的城市市场车辆轨迹预测模型[J]. 西安邮电大学学报, 2025, 30(03): 84-93.
- [3] Lai C, Zhang X, Zhang H, et al. GAN augmentation-based continuous authentication for vehicular digital twin//Proceedings of the IEEE International Conference on Communications. Denver, USA, 2024: 5220-5225.
- [4] 王忠,曹春杰,谢夏,等. 基于多特征迁移学习的低资源临高方言语音识别方法[J]. 通信学报, 2025, 46(10): 221-232.
- [5] Xu C, Liu W, Zheng Y, et al. An imperceptible data augmentation based blackbox clean-label backdoor attack on deep neural networks. IEEE Transactions on Circuits and Systems I: Regular Papers, 2023, 70(12): 5011-5024.
- [6] Hou L, Hua Z, Li Y, et al. M-to-n backdoor paradigm: A multi-trigger and multi-target attack to deep learning models. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 11299-11312.
- [7] Li S, Xue M, Zhao B Z H, et al. Invisible backdoor attacks on deep neural networks via steganography and regularization. IEEE Transactions on Dependable and Secure Computing, 2020, 18(5): 2088-2105.
- [8] Gu T, Liu K, Dolan-Gavitt B, et al. Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 2019, 7: 47230-47244.
- [9] Xue M, He C, Wang J, et al. One-to-n and n-to-one: Two advanced backdoor attacks against deep learning models. IEEE Transactions on Dependable and Secure Computing, 2020, 19(3): 1562-1578.
- [10] Liu Y, Ma S, Aafer Y, et al. Trojaning attack on neural networks//Proceedings of the Annual Network and Distributed System Security Symposium. San Diego, USA, 2018.
- [11] Nguyen A, Tran A. Wanet: Imperceptible warping-based backdoor attack. arXiv preprint arXiv:2102.10369, 2021.
- [12] Zhong H, Liao C, Squicciarini A C, et al. Backdoor embedding in convolutional neural network models via invisible perturbation//Proceedings of the ACM Conference on Data and Application Security and Privacy. New Orleans, USA, 2020: 97-108.
- [13] Turner A, Tsipras D, Madry A. Label-consistent backdoor attacks. arXiv preprint arXiv:1912.02771, 2019.
- [14] Saha A, Subramanya A, Pirsiavash H. Hidden trigger backdoor attacks//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(07): 11957-11965.
- [15] Shafahi A, Huang W R, Najibi M, et al. Poison frogs! Targeted clean-label poisoning attacks on neural networks. Advances in Neural Information Processing Systems, 2018, 31.
- [16] Zhu C, Huang W R, Li H, et al. Transferable clean-label poisoning attacks on deep neural nets//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2019: 7614-7623.
- [17] Liu Y, Ma X, Bailey J, et al. Reflection backdoor: A natural backdoor attack on deep neural networks//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 182-199.
- [18] Zhang Q, Ding Y, Tian Y, et al. Advdoor: Adversarial backdoor attack of deep learning system//Proceedings of the ACM SIGSOFT International Symposium on Software Testing and Analysis. Virtual Conference, 2021: 127-138.
- [19] Chan C K, Cheng L M. Hiding data in images by simple LSB substitution. Pattern Recognition, 2004, 37(3): 469-474.
- [20] Gao Y, Li Y, Gong X, Li Z, Xia S T, Wang Q. Backdoor attack with sparse and invisible trigger. IEEE Transactions on Information Forensics and Security, 2024, 19: 6364-6376.
- [21] Wang B, Yu F, Wei F, Li Y, Wang W. Invisible intruders: Label-consistent backdoor attack using re-parameterized noise trigger. IEEE Transactions on Multimedia, 2024, 26: 10766-10778.
- [22] Zhu M, Li Y, Guo J, Wei T, Xia S T, Qin Z. Towards sample-specific backdoor attack with clean labels via attribute trigger. IEEE Transactions on Dependable and Secure Computing, 2025, 22(5): 4685-4698.
- [23] Wang X, Xue M, Chen K, Xu J, Liu W, Zhang L Y, Zhang Y. IsBad: Imperceptible sample-specific backdoor to DNN with denoising auto-encoder. IEEE Transactions on Emerging Topics in Computing, 2026, 14(1): 30-41.
- [24] Zhang J, Chen D, Huang Q, et al. Poison ink: Robust and invisible backdoor attack. IEEE Transactions on Image Processing, 2022, 31: 5691-5705.
- [25] Wu S, Li J, Li C, et al. A Universal Semantic-Based Method for Backdoor Attack. Proceedings of the IEEE International Conference on

Data Science in Cyberspace (DSC), 2024: 414-420.

- [26] Hua Y, Liu L, Cao J, et al. Stealth Backdoor Attack for Remote Sensing Image Classification. Proceedings of the IEEE 7th International Conference on Electronic Information and Communication Technology (ICEICT), 2024: 1-6.
- [27] Zeng G Q, Wei H N, Lu K D, et al. CMo-IABA: Constrained Multi-Objective Invisible and Adaptive Backdoor Attack for Deep Neural Networks-Based SAR Image Classification. IEEE/CAA Journal of Automatica Sinica, 2026, 13(7): 1731-1746.



张鑫伟(1998-), 男, 西安邮电大学信息与通信工程专业博士生, 主要研究方向为信息隐藏, 图像隐写和数字水印技术。



赖成喆(1985-), 男, 博士, 西安邮电大学教授, 主要研究方向为安全协议设计与分析, 车联网安全。



曹捷(1998-), 男, 加拿大女王大学博士生, 主要研究方向为人工智能安全、生成式图像水印等。



郑东(1964-), 男, 博士, 西安邮电大学教授, 主要研究方向为编码密码学和网络安全。



柳翊然(1994-), 女, 西安邮电大学密码学专业硕士生, 主要研究方向为信息隐藏与图像隐写。