第 62 卷　第 5 期
2026 年 9 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 62, No. 5
Sept., 2026

DOI:10.13232/j.cnki.jnju.2026.05.010

基于特征归因与背景注入的干净标签后门水印方法

卞　睿,龙显忠*,李　云

(南京邮电大学计算机学院、软件学院、网络空间安全学院,南京,210023)

摘　要:图像分类模型面临严重的未授权复制与非法分发风险,模型后门水印技术虽被广泛应用在所有权验证,但现有方法大多依靠突兀的噪声触发器或者刚性的解释空间划分,造成水印的隐蔽性差,容易被几何变换破坏,不能逃脱人工审查和常规图像处理.为了解决上述问题,提出一种用特征归因和背景注入来构造图像分类模型的干净标签后门水印方法,在不修改样本真实标签的情况下,达到高隐蔽性和强鲁棒性的水印嵌入.首先,使用显著性检测机制把图像前景和背景进行语义解耦,把后门触发器伪装成有艺术纹理特征的背景,达到视觉上隐蔽的效果.接着,引入纹理自适应的超像素分割策略来建立编码目标,推导出基于局部采样拟合的可导特征归因算法,使用多任务约束把水印信息深植到模型的决策机理中.实验结果表明,提出的方法在保证分类性能稳定的同时,能够可靠地提取水印,并对常见的图像变换水印去除攻击具有很好的抵抗能力.相比于现有的水印方案,该方法的隐蔽性、鲁棒性等方面表现更好,可以给图像分类模型所有权验证提供一个实用的版权保护方案.

关键词:模型水印,无害后门水印,特征归因,超像素,风格迁移

中图分类号:TP309　　**文献标志码:**A

Clean-label backdoor watermarking method based on feature attribution and background injection

Bian Rui, Long Xianzhong*, Li Yun

(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China)

Abstract: Image classification models face serious risks of unauthorized replication and illegal distribution. Although model backdoor watermarking techniques have been widely adopted for ownership verification, most existing methods rely on conspicuous noise triggers or rigid partitions of the interpretability space, resulting in poor watermark concealment, easy vulnerability to geometric transformations, and failure to evade manual inspection and conventional image processing. To address these issues, we propose a clean-label backdoor watermarking method for image classification models that leverages feature attribution and background injection. This approach achieves high concealment and strong robustness in watermark embedding without modifying the true labels of the samples. First, we employ a saliency detection mechanism to semantically decouple the foreground and background of an image, disguising the backdoor trigger as a background endowed with artistic texture features, thereby achieving visually imperceptible embedding. Next, we introduce a texture-adaptive superpixel segmentation strategy to establish coding targets and derive a differentiable feature attribution algorithm based on local sampling and fitting. By incorporating multi-task constraints, we deeply embed the watermark information into the model's decision-making mechanism. Experimental results demonstrate that our proposed method not only maintains stable classification performance but also enables reliable watermark extraction and exhibits excellent resistance against common image transformation-based watermark removal attacks. Compared with existing watermarking schemes, our method

基金项目:国家自然科学基金(62476137),南京邮电大学校级自然科学基金(NY225157)

收稿日期:2026-06-28

* 通信联系人,E-mail:lxz@njupt.edu.cn

demonstrates superior performance in terms of concealment and robustness, offering a practical copyright protection solution for ownership verification of image classification models.

Keywords: model watermarking, benign backdoor watermarking, feature attribution, superpixel, style transfer

近些年来,尽管深度学习在计算机视觉和自然语言处理领域都取得了巨大成功[1],但训练出一个优秀的深度神经网络依然需要庞大的数据和算力资源,因此预训练模型就成为一种重要的核心数字资产. 随着各种云服务的广泛普及,模型也面临着严重的分发和窃取的风险,因此,对深度学习模型设计可靠的版权保护和所有权验证机制,已成为一个亟需解决的研究任务[2].

模型版权保护技术有模型加密和模型水印两种. 模型加密虽然安全,但计算开销太大,不能在实时推理中使用[3]. 相反,模型水印技术就是将特定的版权标识嵌入模型,在不改变原有主任务性能的基础上完成所有权的验证,有更好的应用前景[4]. 模型水印技术又细分为白盒水印和黑盒水印,白盒水印通常需要获取模型的具体参数信息,但现实中模型盗取者一般只提供黑盒API接口,不会提供具体的参数信息. 在这种约束的限制下,黑盒水印只需验证输入输出映射即可完成版权验证,这使黑盒水印逐渐成为主流方案[5].

黑盒水印的本质是基于后门攻击的思想,早期的黑盒后门水印方法,如BadNets[6]和Blend[7],大多采取脏标签的办法,在插入触发器的同时修改训练样本的真实标签,但特征和标签间的强行语义冲突会造成数据分布出现异常,很容易被现有的后门防御机制拦截. 例如,Neural Cleanse[8]使用逆向工程触发器检测后门,Spectral Signatures[9]依靠统计特征分布过滤毒化数据,STRIP[10]采用运行时输入扰动来识别异常预测. 为了克服这个难题,Turner et al[11]最先提出一种干净标签的后门攻击范式,即在保证样本标签真实性的基础上嵌入触发器. 此范式虽然消除了显式的标签污染,但干净标签水印方法在视觉隐蔽性上存在不足. Zhao et al[12]采用高频特征噪声作为触发器,Zhu et al[13]是用突兀属性来注入触发器,这些方法都会留下一些人工合成的痕迹,极易被人工审查或频域滤波发现.

随着可解释人工智能的发展,模型版权保护方向也开始使用特征归因,研究者试图把版权信息编码成模型对某个区域的解释极性. Shao et al[14]的解释性水印强制模型在验证时会输出事先设定好的解释图. 但现有的方法强制用规则的几何网格来分割图像的解释空间,把图像原来的自然边缘和语义纹理强行割裂开,而且对于几何变换等常规图像处理十分敏感. 另外,由于解释性算法计算量大、不可导,所以不能对水印嵌入过程进行端到端的优化[13].

针对上述问题,本文提出一种基于特征归因与背景注入的干净标签后门水印方法. 从可见性思考,该方法通过显著性检测模型[15]把图像语义解耦,把后门触发器自然地融合成有高保真度的艺术渲染背景[16],很好地避开视觉和异常审查. 从决策来说,抛弃刚性的网格,采用Achanta et al[17]的顺应纹理超像素分割来建立解释目标,推导出基于局部采样的闭式可导特征归因算法,完成多任务的端到端联合训练.

本文的主要贡献如下.

(1)提出一种保证语义真实性的干净标签水印框架. 采用语义解耦和神经风格迁移把触发器无缝地融入背景,在保持主体特征的前提下抵御分布异常检测.

(2)提出纹理自适应的超像素解释性编码方法. 使用超像素算法对背景进行不规则的极性映射,和传统的网格编码相比更有鲁棒性,对各种水印去除的攻击更具抵抗力.

(3)通过引入针对背景区域的解释归因,将先前触发器必须与某一特定类别绑定的范式解绑,取而代之的是将触发器与某一特定背景解释形状绑定,解决后门攻击一直存在的有害性问题.

1 相关工作

1.1 干净标签后门攻击 传统的黑盒后门水印一般使用脏标签法[18]. Gu et al[6]的BadNets用静

态黑白方块作为触发器,在图像局部添加. Chen et al[7]的Blend将全局随机噪声和原始图像混合. 这些方法会把有明显特征的触发样本标成错误的目标类,造成特征和真实的标签存在严重的语义矛盾,数据分布很容易被现有的防御机制所识别并拦截. 例如,Wang et al[8]的Neural Cleanse通过特征激活聚类来过滤异常样本,Tran et al[9]的Spectral Signatures利用奇异值分解技术成功剔除了投毒数据.

为了提高后门样本的隐蔽性,干净标签攻击开始流行[19],其主要目的是不改变样本真实标签. Turner et al[11]用标签一致性攻击策略来对目标类样本进行对抗扰动,使模型在训练的时候只依靠附加的触发器. Nguyen and Tran[20]的WaNet使用隐藏的图像弹性扭曲作为触发器,以避开人工视觉检测. Zhu et al[13]的BAAT把人像的特定语义属性当作触发载体来维持特征空间的一致性. 但现有的干净标签方法在实际使用过程中仍然存在视觉上的明显缺陷,本文提出的语义解耦和艺术背景注入,是把触发器伪装成自然存在的复杂纹理来弥补现有方法的不足.

1.2　后门攻击的有害性　在图像分类模型的领域,运用后门攻击来实现版权保护是一种常见的策略,通常是在模型中植入一个隐藏的后门作为版权保护的依据,被植入后门的模型在面对触发器特征和图像主体特征同时存在的条件下,会优先考虑触发器特征,并将后门图片归类为触发器对应的类别. 然而,使用这一范式来注入后门往往基于一个大前提,即在模型正常工作过程中,无论如何也不可能接触到这种触发器特征,该触发器特征只在需要进行版权保护时通过刻意构造才能产生. 这个前提在复杂的实际使用过程中是无法确保的,因此,这种范式不可避免地带来了有害性问题.

相关研究表明,误分类后门不仅会损害模型主任务的性能,还很容易成为外部恶意利用的安全漏洞. 针对这一痛点,Sun et al[21]尝试建立触发器与信息熵之间的联系,避免将触发器直接与某类样本绑定,并通过分析后门图片的信息熵来作为版权保护的依据. Zhu et al[22]使用对抗优化的策略,促使后门样本在模型训练前后表现出不同的损失分布差异,在不影响主分类任务的前提下实现版权验证. 由此可见,消除触发器对主任务的干扰,实现无害化的水印嵌入,已成为当前该领域的重要研究方向.

1.3　特征归因方法　特征归因属于可解释人工智能的关键部分,其目的在于衡量输入图像各个区域对于模型最终决策的作用. 为了实现模型的可解释性,Selvaraju et al[23]的Grad-CAM利用分类层反向传播来得到类激活图,通过类激活图来解释模型分类的原因. Ribeiro et al[24]的LIME在输入样本附近做局部扰动采样,拟合线性解释模型,再利用线性模型的可解释性来解释复杂的图像分类模型. Ghorbani et al[25]首次发现特征归因存在脆弱性,可以被微小的对抗扰动轻易改变模型解释图. Ross et al[26]把解释结果主动加入损失函数,使模型只关注正确的特征. Rieger et al[27]用惩罚不合理解释的方法来对齐神经网络和先验知识.

受到上述可解释人工智能思想的启发,Shao et al[14]的前沿的解释性水印,首次把多比特水印序列映射到事先设定好的特征归因目标图上. 训练时用多任务损失迫使模型在某个区域做出符合预设极性的决策解释. 但是该方法强制使用规则的网格来划分图像区域,而刚性结构粗暴地割裂了图像天然的纹理边缘. 如Tancik et al[28]指出的,规则网格编码对于图像裁剪、JPEG压缩等常规变换十分脆弱,会大大降低水印提取率. 目前仍需要更优秀的编码方案来推动解释性水印的持续发展.

2　本文方法

2.1　水印整体架构与理论动机　传统的后门攻击的框架通常是,给定N个样本的干净集$D_{cl}=\left\{\left(x_h,y_h\right)\right\}_{h=1}^{N}$,其中,$x_h$为输入图像,$y_h$为对应的真实标签. 后门水印通常需要构建毒化集$D_p$,通过触发器注入函数来篡改样本特征,并将其修改为目标类标签y_t. 这种毒标签的设计天然存在风险,即后门存在有害性,

图1展示了本文提出的后门水印方法的框架,该方法完美规避了传统后门攻击存在的有害

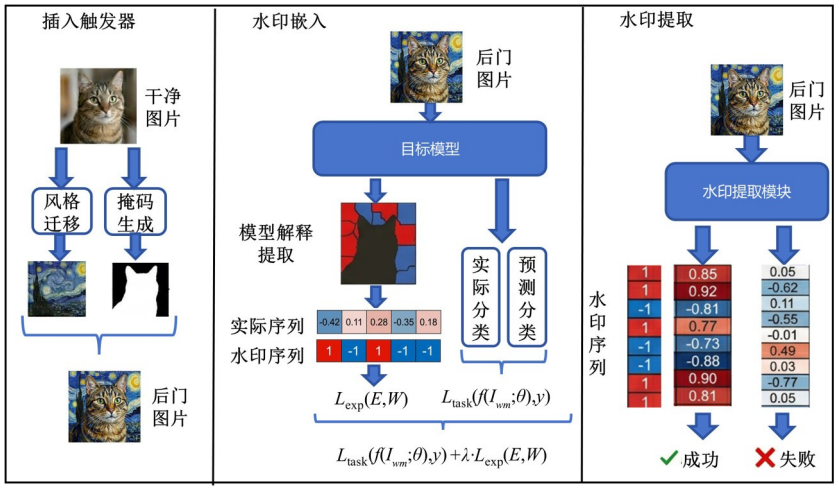


图 1　基于特征归因与背景注入的水印方案总体架构图

Fig. 1　Overall architecture of the watermarking scheme based on feature attribution and background injection

性问题,其核心思路在于把嵌入目标从“篡改标签输出”转变为“重塑决策归因”.

整个架构包含三个模块:第一个模块使用显著性检测把图像的前景和背景在语义上隔离开,防止触发特征污染目标类的主体语义;第二个模块采用联合特征距离驱动的自适应超像素算法对背景做拓扑划分,得到一组顺应底层纹理的解释子空间;第三个模块提出基于局部采样扰动的可导归因计算方法,通过多任务损失约束模型在特定子空间上的归因极性,把所有权信息嵌入到网络参数中.

2.2　显著性区域解耦与风格化背景融合　干净标签后门攻击面临的最主要问题是“特征竞争”,模型会先拟合出样本里固有的强语义特征来降低分类损失,而人造的触发特征由于信号太弱很容易被忽略. 因此,本文采用前景与背景分离的策略,把触发器更多地隐藏在背景中,避免影响目标类的主体语义,前景提供帮助模型分类正确的语义信息,背景提供用于版权保护的触发器特征信息.

风格迁移采用 AdaIN 算法[29]. 该算法通过在特征空间对齐内容图与风格图的通道均值和方差来实现任意风格的渲染. 通过参数 α 控制原始背景与目标风格的插值比例,$\alpha=0$ 时完全保留原图,$\alpha=1$ 时完全替换为风格图.

在实现上采用 U^{2}Net 做显著性提取,通过嵌套的 U 形结构输出一张连续的前景概率图 $S(x)\in[0,1]^{H\times W}$,每个像素的值反映它属于前景主体的置信度. 通过分析其该像素属于前景的置信度来给定当前像素的风格化强度,背景部分风格化强度高,主体部分风格化强度低.

针对原图与风格化图的融合,不直接使用基于权重的线性融合,引入锐化因子 $\beta(\beta\geqslant 1)$ 对 $S(x)$ 做非线性重映射,得到融合权重:

$$G=\frac{S^{\beta}}{S^{\beta}+(1-S)^{\beta}} \tag{1}$$

此函数在 0 到 1 之间是连续的并且可导. 当 $\beta=1$ 的时候退化为线性融合,β 变大时,中间概率值向两端移动,高置信度前景区域变得越来越干净,低置信度背景区域吸收更多的风格纹理,过渡带被压缩到很小的一段范围. 该种方式既能保证前景的语义干净,又能防止由于二值截断造成的高频波动.

最终,含水印样本 X_w 的生成过程可表示为原始图片 x 与风格化背景 T_s 的加权拼接:

$$X_w=G\odot x+(1-G)\odot T_s \tag{2}$$

其中,$\odot$ 表示哈达玛乘积. 该操作的意义是在非显著的背景区域给水印特征留出独立承载的空间,从数学映射上切断触发器与主体语义的干扰通道. 在保证样本视觉自然度的同时,保留原始主任务分类特征的空间,模型可以完整地提取原始主任务的分类特征.

2.3　顺应纹理的自适应解释空间编码　提出一种基于空间-颜色联合特征距离的解释空间编码方法,具体见以下算法.

算法　自适应解释空间编码算法

Require:背景像素总数 n,背景图像 I_{bg},前景掩码 M_{FG},解释域数量 k,平衡因子 m

Ensure:解释子空间集合 $R=\{R_1,R_2,\cdots,R_k\}$

1. 背景像素总数 $n=\sum M_{FG}(x)$,采样步长为 u

2. 将 I_{bg} 转换至CIE Lab颜色感知空间

3. 在 M_{FG} 区域内按步长 u 均匀初始化 k 个解释锚点 $C_j=\left[l_j,a_j,b_j,p_j,q_j\right]^{\mathrm{T}}$

4. repeat

5.　for 像素 $i\in M_{FG}$ do

6.　　计算距离 $D=\sqrt{d_c^2+(m/u)^2d_s^2}$

7.　　将 i 划分至 D 最小的 C_{j^*} 所对应的 R_{j^*}

8.　end for

9.　更新 C_j 的特征向量为当前 R_j 内所有像素特征的均值

10. until 各锚点 C_j 的中心坐标位置偏移量低于收敛阈值 $\delta=1.0$ 像素

11. return 解释子空间集合 R

编码的作用范围由前景概率 $S_r(x)<\tau$ 界定,编码空间的二值掩码 M_{FG} 为:

$$M_{FG}(x)=\begin{cases}1, & \text{if } S_r(x)<\tau\\ 0, & \text{otherwise}\end{cases}\tag{3}$$

其中,x 为输入图片,$S_r(x)$ 为 x 里的任一像素 r 为前景的概率值.

为了更好地反映人类视觉系统的感知差异,算法将RGB图像转换至感知均匀的CIE Lab颜色空间.对于编码区域内的任一像素 i 与解释锚点 j,定义其颜色语义距离 d_c 和物理空间距离 d_s 分别为:

$$d_c=\sqrt{\left(l_i-l_j\right)^2+\left(a_i-a_j\right)^2+\left(b_i-b_j\right)^2}\tag{4}$$

$$d_s=\sqrt{\left(p_i-p_j\right)^2+\left(q_i-q_j\right)^2}\tag{5}$$

其中,l_i,a_i,b_i 分别表示像素点 i 在CIE Lab颜色空间中的三个颜色通道值,l_j,a_j,b_j 分别代表解释锚点 j 的中心颜色通道值,p_i,q_i 表示像素点 i 在物理图像上的二维平面坐标,p_j,q_j 表示解释锚点 j 在图像二维平面上的中心坐标位置.

在此基础上,构建联合特征距离度量 D:

$$D=\sqrt{d_c^2+\left(\frac{m}{u}\right)^2d_s^2}\tag{6}$$

$$u=\sqrt{n/k}\tag{7}$$

其中,n 为背景像素总数,k 为解释域的数量,期望的采样步长为 u,表示每一个子空间的预期大小.m 为空间语义平衡因子,m 越大,划分区域时的空间距离惩罚权重越高,生成的子空间越接近规则的紧凑块;m 越小,子空间越趋向沿着颜色纹理的边缘无限延伸.用最小化该联合距离的方法把背景区域分成 k 个解释域 $R=\{R_1,\cdots,R_k\}$.

2.4　响应约束与多任务联合优化　在解释空间自适应划分之后,再把预设的所有权序列 $W\in\{-1,1\}^k$ 用解释响应的形式嵌入模型参数 θ.

2.4.1　基于局部采样的可导特征归因映射　为了使解释过程可以被端到端的网络优化包含,用闭式解重构特征归因的计算过程.定义一个服从伯努利分布的随机掩码矩阵 $\boldsymbol{M}\in\{0,1\}^{c\times k}$,对划分出来的 k 个解释域进行 T 次独立的遮挡扰动,得到包含 c 个扰动样本的集合,然后将该集合批量输入当前模型 $f(\bullet;\theta)$,记录模型输出的目标类预测置信度向量 $V\in\mathcal{R}^c$.

在该局部采样邻域内,用线性近似模型来度量特征对模型决策的贡献度.为了保证求解的稳定性以及赋予过程可导性,用岭回归的解析解来计算各个子空间的归因权重向量 $E\in\mathcal{R}^k$:

$$E=\left(M^{\mathrm{T}}M+\gamma\boldsymbol{I}\right)^{-1}M^{\mathrm{T}}V\tag{8}$$

其中,$\boldsymbol{I}$ 为单位矩阵,γ 为防止协方差矩阵不可逆的 L_2 正则化惩罚系数.经过上述映射,归因向量 E 被严格地表示成网络预测概率向量 V 的线性组合.根据反向传播的链式法则,损失函数对归因权重的梯度不会影响解析映射,从而可以沿着解析映射返回模型的网络层权重 θ.

2.4.2　合页损失引导的响应对齐　给定含水印样本 I_{wm} 和真实标签 y,模型训练的目标函数由主任务交叉熵损失 L_{task} 和响应约束损失 L_{exp} 组成:

$$\min_{\theta}L_{\mathrm{total}}=L_{\mathrm{task}}\left(f(I_{\mathrm{wm}};\theta),y\right)+\lambda\bullet L_{\mathrm{exp}}(E,W)\tag{9}$$

其中,I_{wm} 为后门图片;y 为对应图片的正确标签;L_{task} 促使模型学习后门图片与正确的类别之间的关联;E 为归因权重向量,由式(8)计算;W 为预

设的背景解释图;$L_{\exp}$要求模型针对后门图片的需要输出预设的解释图形状,以此作为版权验证的依据;λ为控制水印嵌入强度的平衡超参数.

在约束机制的设计上,强制要求模型输出精确的归因数值会导致过拟合,破坏主分类任务.故设计基于安全边际ϵ的合页损失作为软约束函数:

$$L_{\exp}(E,W)=\frac{1}{k}\sum_{i=1}^{k}\max(0,\epsilon-E_i\cdot W_i) \tag{10}$$

其中,E_i为模型对第i个背景子空间实际计算出的归因值,可由式(8)计算;W_i为预先设定的水印序列的第i个元素,只能取值1或−1.该损失函数要求模型提取出的每一个归因值E_i不仅要与W_i同号,并且强度需要超过ϵ.

使用这种方法,不仅保留了分类决策边界的原始形态,还在参数空间中构建了对抗外部扰动的特征冗余,为后续抵抗模型微调和剪枝攻击奠定理论基础.

2.5 水印验证 验证时,将含水印样本输入待测模型,模型对背景区域的各个子空间会产生归因响应,这些响应的极性模式理论上会与预设的水印序列一致.通过解码这个响应模式,即可判断模型是否嵌入了水印,具体是采用Wilcoxon符号秩检验进行统计判定.该方法为非参数方法,无需假设数据服从正态分布,和参数检验相比,在实际检验中更具实用性.

具体流程如下.给定待检测模型$f(\cdot)$,取一张验证图像,经2.2锐化软融合流程生成含水印样本.对水印图片执行T次随机局部遮挡,采集模型对目标类返回的预测置信度向量,利用式(8)岭回归闭式解反推各解释子空间的归因权重,按预定义极性映射规则解码为水印比特序列.将预设序列与嵌入的所有权序列逐位比对,统计匹配比特数.构建如下检验.

Ω_0:该模型未嵌入此水印,即匹配比特数服从随机猜测.

Ω_1:该模型嵌入了此水印.

在显著性水平为0.05时,如果检验拒绝Ω_0,说明模型对触发模式的归因响应不是随机产生的,可认定水印存在.

该方法只需获取黑盒访问模型的预测接口,无需获取训练数据,也无需获取模型内部参数,适用于实际场景下的水印检测.

3 实验及结果分析

3.1 数据集介绍 为了评价提出的方法在不同数据集上的有效性,选取三个公开数据集进行测试.

ImageNet-100[30]是ImageNet的100类子集,样本分辨率高、背景复杂,是检验模型在开放自然场景中进行特征提取与泛化的通用指标.

Oxford-IIIT Pet[31]含37类宠物的细粒度数据集,具有类内差异大、类间差异小的特点,分类依靠局部纹理,适合测试模型是否具有对微观特征的提取能力.

GTSRB[32]是德国交通标志识别基准数据集,其中的目标几何结构规整、色彩固定,对异常特征很敏感,适合评估安全场景下的鲁棒性.

3.2 评价指标 为了客观评估后门水印机制的综合性能,基于安全性、有效性与可靠性原则,采用了五种核心评价指标.

(1)分类准确率(Classification Accuracy,CA)可比较干净模型和水印模型在干净样本上的分类精度,用来度量水印植入对于模型主任务性能的影响.

(2)攻击成功率(Attack Success Rate,ASR)可衡量包含触发器样本被模型错误识别为预设类别的概率.ASR越高,说明模型对于后门触发特征的学习越充分.

(3)水印提取成功率(Watermark Success Rate,WSR)可衡量多比特水印被物理提取出的概率,表示结果与目标一致的超像素块占总数的比例.WSR越高,说明水印的有效性越强.

(4)统计显著性(P).可根据假设检验来评价提取水印的可靠性,根据匹配成功水印的累积二项分布概率来判断是否符合预设的显著性水平,如果不符合,就说明模型中存在水印.

(5)频域检测率(Frequency Domain Detection Rate)可衡量触发模式对频域检测的抵抗能力,表示测试集里出现频域异常的图片比例.

3.3 对比实验 选取七种有代表性的后门水印方法进行对比,分别是传统的脏标签攻击Bad-

Nets[6]和Blend[7]、它们的干净标签变体BadNets-C和Blend-C、使用干净标签策略的LC(Label-Consistent)[11]、基于熵的无害型水印EntropyMark[21]和解释性水印EaaW(Explanation as a Watermark)[14]. 同时,用干净模型做基线比较.

使用ResNet-18和VGG-16两种网络架构. 在模型训练阶段用SGD优化器对所有的网络进行训练,动量为0.9,权重衰减为5×10^{-4},批量大小为128. 初始学习率为0.01,在第20和第40轮时分别用0.1的比例进行衰减,共训练50轮.

本文提出的方法中触发器风格化强度设为0.6,超像素块数量为32,解释性损失权重为10. 为了保证隐蔽性,本文的方法投毒率仅为1%.

对比的方法均采用默认设置. CM为未植入后门水印的干净模型,BadNets用图像右下角局部可见的黑白Patch作为触发模式,Blend用全局随机噪声和原始图像混合,LC用对抗扰动和特定触发器保持标签一致性,EntropyMark引入信息熵来进行约束,EaaW用LIME算法把水印比特映射到$M\times M$规则网格解释空间. 所有基线方法及本文方法都在相同的数据集上进行公平评估.

3.3.1　隐蔽性与有效性综合评估　图2给出了不同方法生成的含水印样本的视觉效果对比. 由图可见,传统的脏标签方法BadNets和Blend存在明显的特征拼贴或者全局可视噪声的痕迹. 干净标签约束之后,BadNets-C和Blend-C的触发器仍然明显. LC作为最经典的干净标签后门攻击方法,其问题在于使用对抗性扰动弱化了干净特征的学习,并且其他的触发模式在视觉上也不能和背景自然融合. 新型的无害后门水印EntropyMark和EaaW更多地关注后门的触发方式,它们的触发器与传统脏标签的设计方式一致.

与之相比,本文的方法使用语义解耦的方式把触发器变成艺术风格纹理,并把它注入背景. 通过生成的样本可以发现,目标主体的主要特征以及边缘结构被很好地保留,背景风格化处理没有给整体造成太大的视觉突兀感. 该种策略从物理上防止特征被破坏,使模型在面对人工视觉审查的时候更隐蔽.

为了进一步证明水印的有效性,表1展示了各个对比方法在ImageNet-100数据集上的定量评价结果. 在模型保真度上,传统后门攻击方法

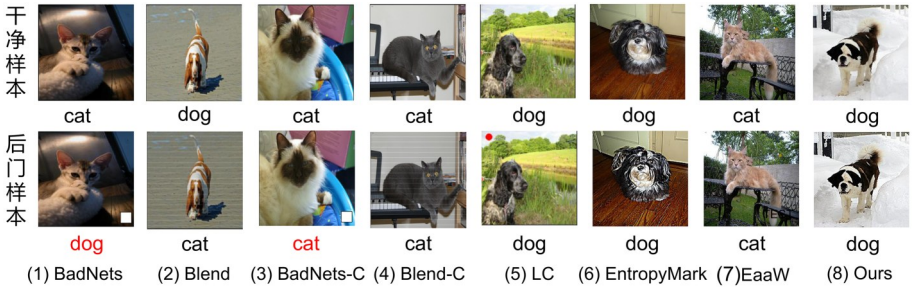


图2　干净样本与不同水印方法生成的毒化样本视觉对比分析

Fig. 2　Visual contrast analysis of clean samples and poisoning samples generated by different watermarking methods

表1　不同水印方法在ResNet-18和VGG-16上的性能对比

Table 1　Performance of different watermarking methods on ResNet-18 and VGG-16

模型	指标	CM	BadNets	Blend	BadNets-C	Blend-C	LC	Entropy-Mark	EaaW	Ours
ResNet-18	*CA*	76.38%	75.61%	75.54%	70.56%	68.42%	70.56%	68.42%	75.78%	76.14%
	ASR	—	95.43%	94.12%	90.45%	70.41%	90.45%	70.41%	96.54%	97.42%
	P	—	0.0007	0.0032	0.0232	0.0353	0.0065	0.0135	1e−6	1e−9
VGG-16	*CA*	74.35%	73.12%	73.64%	69.12%	67.12%	72.19%	66.43%	73.17%	74.12%
	ASR	—	95.24%	93.54%	91.23%	74.13%	88.14%	71.23%	96.17%	97.32%
	P	—	0.0007	0.0013	0.0352	0.0374	0.0074	0.0147	1e−6	1e−9

在使用干净标签的时候,因为触发特征和样本的真实标签之间存在语义冲突,会使模型主任务的分类精度出现明显的下降.本文方法由于前景主体被隔离保护,和干净模型相比性能差异严格控制在1%以内,证明本文水印机制对主任务没有影响.本文方法在ResNet-18上的*ASR*达到了97.42%,并且*P*达到1e−9.和其他水印方法相比,本文方法可以把水印特征隐藏在模型的高维空间里,在黑盒提取阶段提供更有统计置信度的结果.

3.3.2 泛化能力评估 为了检验本文的方法在不同的视觉任务场景下的泛化性能,在ImageNet-100,Oxford-IIIT Pet和GTSRB三个数据集中做扩展实验,使用了复杂的自然物体粗粒度分类、动物细粒度特征识别和自动驾驶场景的标志识别.实验结果如表2所示.由表可见,在复杂的自然环境、细粒度特征识别和背景单一的交通标志识别任务上,本文的方法都能达到较好的效果.*CA*的下降幅度控制在1%以内.特别是对于具有明显特征的GTSRB数据集,艺术纹理触发器和背景之间的高对比度使模型更敏感,*WSR*达到99.45%,*P*最低.说明在不依赖标签篡改的情况下,本文所提方法可以稳定地完成水印验证,不会影响模型在原始任务上分类的性能.

3.3.3 计算开销评估 本文方法中后门样本的显著性检测、风格迁移与超像素编码均在训练前一次性完成,训练阶段无需重复调用上述模块,因此,训练开销与基线方法EaaW基本持平.二者均采用类似的LIME归因嵌入框架,微调轮数均为50轮,嵌入比特数$K=32$,训练模型为ResNet-18,数据集为ImageNet-100,投毒率$\eta=1.0\%$.具体的对比结果如表3所示.由表可见,尽管本文方法在预处理阶段耗时更长,但单轮训练耗时与EaaW几乎一致,最终干净样本的准确率基本持平,而水印提取成功率和统计显著性均有提升.考虑到预处理的数百秒相对于数小时的训练总时长占比极小,以轻量的预处理代价换取更好的水印嵌入质量是值得的.

表2 本文方法在不同数据集和模型架构下的泛化性能评估

Table 2 The evaluation of generalization performance of the proposed method across different datasets and model architectures

数据集	模型	*CA*		*WSR*	*P*
		CM	Ours		
Image-Net-100	VGG-16	74.35%	74.12%	97.32%	1e−9
	ResNet-18	76.38%	75.51%	97.42%	1e−9
Oxford-IIIT Pet	VGG-16	85.20%	84.95%	98.50%	1e−10
	ResNet-18	88.45%	88.12%	98.80%	1e−10
GTSRB	VGG-16	96.50%	96.35%	99.10%	1e−11
	ResNet-18	98.20%	98.05%	99.45%	1e−11

表3 训练可行性评估

Table 3 Evaluation of training feasibility

方法	预处理阶段耗时(s)	单轮训练耗时(s)	*CA*	*WSR*	*P*
Ours	200	420	76.14%	97.42%	1e−9
EaaW	5	425	75.78%	96.54%	1e−6

3.4 鲁棒性评估 水印模型的鲁棒性是指当它在遭受恶意攻击或者常规图像处理时仍然可以保持全部的所有权验证能力.本文就微调、剪枝、图像压缩及频域检测这四种典型场景进行评测.

3.4.1 抵抗模型微调攻击 微调攻击是指攻击者用少量干净数据来训练水印模型,试图抹去后门特征.在实验过程中用10%干净数据做微调模型,学习率为0.01,微调30轮.图3展示了微调过程中的分类准确率和水印提取率变化.由图3可知,在30轮微调之后模型的*CA*一直保持在75%左右.本文方法的*WSR*在开始时略有波动,但很快趋于稳定且大于92%.可以看出,采用归因响

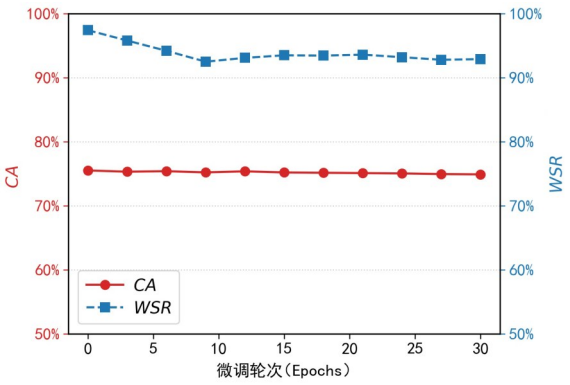


图3 微调轮次对本文方法性能的影响

Fig. 3 Impact of fine-tuning rounds on the performance of the proposed method

应约束嵌入的方式使水印特征和模型决策边界的结合较深,用少量干净样本微调是无法将其去除的.

3.4.2　抵抗模型剪枝攻击　模型剪枝是一种将模型绝对值小的权重参数设为0的攻击方法,以此来破坏触发器和输出之间的联系.使用全局范数L_1非结构化剪枝,在剪枝率为0%～90%的时候展示其性能变化情况.

从图4可知,在剪枝率达到60%的时候,模型的*CA*变化不大,*WSR*仍然大于90%.当剪枝率达到80%以上的时候,本文方法的*WSR*仍然可以保持在70%左右,虽然已经不能满足模型水印的要求,但是模型主任务的退化更加严重,*CA*降到30%以下,已经不能正常工作.实验结果表明,水印特征没有集中在少数几个参数上,具有良好的分布冗余性.证明了本文方法对于模型剪枝攻击具有足够的鲁棒性.

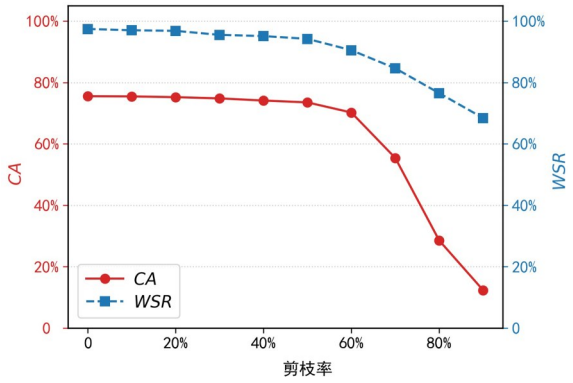


图4　剪枝率对本文方法性能的影响

Fig. 4　Effect of pruning rate on the performance of the proposed method

3.4.3　抵抗图像空间变换攻击　图像在实际部署的时候会被JPEG压缩等变换所影响,JPEG压缩会去掉图像部分细节,造成像素层水印特征消失.对图像施加不同的JPEG压缩强度,质量因子*QF*由100逐渐减小到20,然后用本文方法和BadNets-C在ResNet-18上进行比较.

从图5可见,BadNets-C用的是固定位置的Patch触发器,一旦被压缩,这个Patch的像素信息就丢了,所以*QF*降到60以下时*WSR*急剧下降.本文方法用的是超像素分割,触发特征顺着背景纹理分布,压缩后纹理结构还在,水印信息保留得更好.即使*QF*被压到20,*WSR*仍能维持在86.5%以上,依然满足实际版权的验证功能.

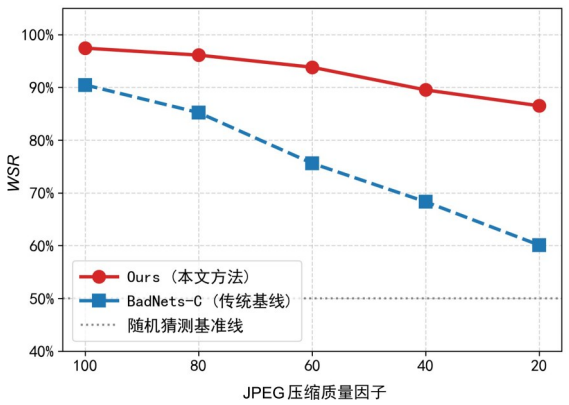


图5　JPEG质量对水印提取成功率的影响

Fig. 5　Effect of JPEG quality on the watermark extraction sucess rate

3.4.4　抵抗频域异常检测　模型窃取者常常使用频域检测来屏蔽异常图片.后门图片往往会在融合边界产生像素跳变,这种不连续性经过DCT变换后会变成高频系数的异常聚集,极易被频谱分析捕捉.

为了测试本文所提方法对频域检测的抵抗性,设计了基于DCT高频能量的检测实验,对硬拼接、线性软融合、本文锐化软融合三种方案在频域的隐蔽性进行比较.

具体的测试流程如下.

首先,从ImageNet-100中随机抽取500张图片作为测试集,针对每张后门图片做8×8分块的DCT变换,提取每个块的高频*AC*系数能量.使用干净样本集在各块上的能量均值μ和标准差σ做基线,对测试样本逐块进行比较,若某个块的能量超出$\mu+2\sigma$就标记为异常.一张图中异常块占比超过5%的判为"频域可疑".

从表4可知,直接使用硬掩码拼接生成的后门图片的检测率为78.4%,可以轻易被频域检测检测出异常.使用线性软融合$(\beta=1)$的检测率降到35.2%,依然明显高于基线.本文采用的锐化软融合$(\beta=3)$的检测率只有11.6%,异常块比例为4.1%,非常接近干净样本的基线.因为本文植入的触发器并非完全只作用于背景区域,通过锐化因子合理地对每一块像素设置风格化强度

表 4 频域检测抵抗力对比

Table 4 Frequency-domain detection resistance

样本生成方式	DCT 异常块比例	频域检测率	是否通过检测
干净样本	3.2%	6.8%	—
硬掩码拼接	21.5%	78.4%	否
线性软融合	9.8%	35.2%	否
锐化软融合(ours)	4.1%	11.6%	是

大小,能从根源解决后门图片会在融合边界产生像素跳变的问题.

3.4.5 抵抗显著性检测偏差 在实际场景下,显著性检测模型无法每次都清晰地分离出前景与背景,天然存在分割出错的风险.因此,为了定量评估此风险,引入信息论中的信息熵来衡量显著性检测的不确定性计算式如下:

$$H=-\frac{1}{P}\sum_r\left[S_r(x)\log_2 S_r(x)+(1-S_r(x))\log_2(1-S_r(x))\right] \tag{11}$$

其中,x 为原始输入图片;$S_r(x)$ 为 x 中任一像素 r 为前景的概率;P 为图片里所有像素点的个数;H 的值越靠近 1,表明显著性模型对前景和背景的判别越模糊,分割偏差的风险越高.

具体测试流程如下.首先,从 ImageNet-100 数据集的测试集中筛选 5000 张 H 值不同的图像.每 50 张图片划分为一组,在训练好的 ResNet-18 水印模型上,分别测试每组图像的 CA 和 WSR,并与基于阈值 $\tau=0.5$ 的传统硬二值拼接方案进行对比实验.

如图 6 所示,随着图像 H 值的升高,硬二值拼接方案的 CA 显著下降.因为硬截断方案会在边界区域产生大量的伪影,这些伪影特征会成为模型分类的依据,干扰模型对主任务特征的学习.

相比之下,本文方法在前景背景融合度高的场景下,CA 的下降幅度仍然被有效地控制在 1.5% 以内,WSR 维持在 94% 以上.原因有两个.一是本文方法采用了超像素分割算法进行背景语义划分,每一个超像素块内部的像素信息共同形成了对应子空间的语义信息,部分像素存在的干扰信息不会干扰整个超像素块语义特征的提取.二是本文植入的触发器虽然作用于图片全局,但图像前景通过锐化因子 β 大大降低了风格化的干扰,在保证模型能干净提取主题特征的同时,图像前景存在的微弱触发器的特征,保证即使在显著性检测模型检测不准确的情况下,依然能从前景与背景交界处获取到完整的触发器特征.

4 消融实验与分析

4.1 风格化强度 α 对性能的权衡分析 参数 α 可以控制背景区域触发器的明显程度.如图 7 所示,当触发器强度不高时,样本的隐蔽性很好.但从表 5 可以看出,当 $\alpha=0.2$ 时,模型很难对这个区域产生稳定的特征响应,WSR 只有 89.20%.

随着 α 的提高,背景变得清晰有纹理,给模型赋予了足够的归因载体,WSR 提升至 98.65%.该阶段由于锐化因子的解耦效应,主体特征没有受到干扰,分类准确率也没有受到影响.但是背

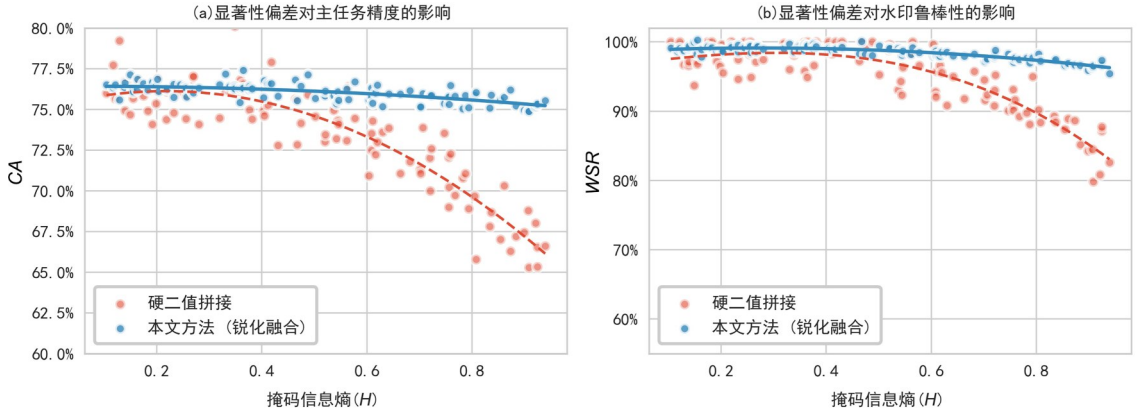


图 6 显著性检测质量对水印性能的影响

Fig. 6 Impact of saliency detection quality on watermark performance

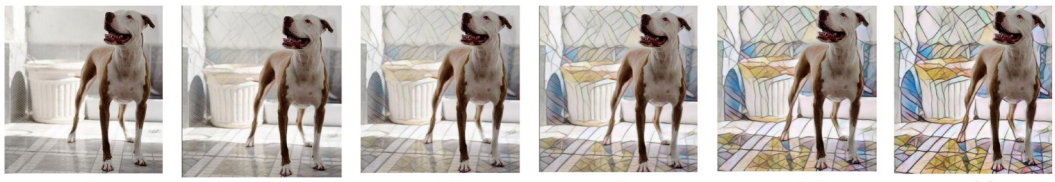


图7　不同风格化强度下的含水印样本视觉效果对比

Fig. 7　Visual effects of watermarking on samples with various stylization intensity

景分布的剧烈偏移会干扰模型对整体语义的判断,造成*CA*下降.为了达到平衡隐蔽性和提取效果的目的,本文将参数α设为0.4.

4.2　空间语义平衡因子*m*的影响　参数m可以控制超像素聚类时空间距离和颜色距离的权重比例.从表5可知,当m很小时,空间约束几乎是松开的,颜色和纹理主导了聚类.此时子空间能很好地沿着物体边缘走,贴合度高,但代价是各子空间面积差异很大,有些块太小,分不到足够的归因载体,因此*WSR*会降低.

相反,如果m太大,聚类就更加关心像素点之间的空间距离,容易把图像原来的连续纹理切断,不利于模型对超像素块内部高级语义特征的学习.为了同时达到超像素块空间上的整齐和纹理上的连续,综合考虑,将参数m设置为0.3.

4.3　锐化因子β的消融分析　β可以控制显著性概率图向融合权重映射的非线性程度.由式(1)可知,$\beta=1$等价于不做拉伸,权重在过渡带内线性衰减,此时前景边缘区域混入的风格化纹理偏多.由表4和表5可以看到,$\beta=1$时*WSR*为92.10%,频域检测率达35.2%,说明过渡带过宽会给触发特征的学习带来噪声,同时也容易在DCT频域留下可检测的痕迹.

β越大前景与背景的过渡带越小,$\beta=3$时*WSR*达97.42%,*CA*和干净模型相比几乎没有降低,频域检测率也降低到11.6%,接近干净样本的基线水平.说明这种锐化映射确实在起作用,通过操控锐化因子能有效地控制过渡带的大小,避免在前景与背景边界位置出现像素跳变.

但继续加大β不会持续带来收益,因为锐化过强会导致融合权重在梯度域上接近阶跃函数,形式上虽然是连续的,而实际行为已经逼近硬判

表5　后门样本构造核心超参数的消融实验

Table 5　Ablation study of core hyperparameters for backdoor sample construction

	风格化强度α	空间语义平衡因子m	锐化因子β	*CA*	*WSR*	*P*
变量1:风格化强度α	0.1	0.3	3	76.10%	78.45%	2.1e−4
	0.2	0.3	3	75.85%	89.20%	1.5e−6
	0.3	0.3	3	75.51%	97.42%	1.0e−9
	0.4	0.3	3	74.20%	98.65%	1.0e−10
	0.5	0.3	3	71.35%	99.10%	1.0e−11
变量2:空间语义平衡因子m	0.6	0.01	3	75.50%	93.50%	5.2e−7
	0.6	0.1	3	75.52%	96.20%	8.5e−9
	0.6	0.3	3	75.63%	97.42%	1.0e−9
	0.6	0.5	3	74.48%	96.80%	3.5e−9
	0.6	1.0	3	73.35%	91.20%	8.2e−7
变量3:锐化因子β	0.6	0.3	1	75.60%	92.10%	3.2e−7
	0.6	0.3	2	75.54%	95.80%	8.5e−9
	0.6	0.3	3	75.51%	97.42%	1.0e−9
	0.6	0.3	5	73.38%	97.65%	5.2e−10
	0.6	0.3	7	71.12%	97.80%	2.8e−10

决,开始干扰分类边界.综合来看,$\beta=3$ 在 *WSR*,*CA* 和频域隐蔽性三个指标之间取得了比较好的平衡,本文将其设为默认值.

4.4 空间划分数量 N 与编码容量分析 划分数量 N 决定了水印序列的编码密度.从表 6 可以看出,当嵌入短序列时,由于单个子空间的物理面积比较大,模型容易在宽泛区域内建立起稳定的响应,所以 *WSR* 较高.随着 N 的增加,子空间粒度变细,模型对于局部微小区域的响应稳定性下降,使单比特提取率降低至 88.20%.

高比特编码可以减小随机匹配造成的碰撞概率,所以当 $N=32$ 时,单比特提取率有波动,对应的 P 呈指数级下降,显著提高了所有权验证的统计置信度.考虑到提取的成功率以及整体验证的稳定性,采用 $N=32$ 作为水印嵌入的默认容量.

4.5 解释性损失权重 λ 的敏感性分析 超参数 λ 决定水印损失与分类损失之间的权重分配.图 8a 显示,λ 较小时,由于水印损失分配的权重过小,模型在没有充分拟合触发器特征前便已收敛,水印提取率仅为 82.3%;随着 λ 增加至 0.1,模型能够对触发模式建立稳定的特征响应,*WSR* 迅速提升至 97.42%,并且分类准确率也保持稳定.

过大的水印损失权重会干扰模型主任务的学习,所以当 λ 继续增大至 1.0 时,尽管 *WSR* 提升至 99.1%,但模型会忽视对主任务分类功能的学习,导致 *CA* 从 75.51% 下降至 71.2%.为了维持主任务的分类精度与水印提取成功率的平衡,采用 $\lambda=0.1$ 作为多任务联合优化的默认权重.

4.6 投毒率 η 的影响分析 投毒率 η 表示水印样本在训练集中所占的比例.如图 8b 所示,当 $\eta=0.2\%$ 时,后门数据集数量不足,模型不能很好地学习到水印特征,*WSR* 只有 85.6%.当 η 增大到 1.0% 以后,*WSR* 一直保持在 97% 以上.

当 η 提高到 5.0% 时 *CA* 开始下降,因为毒化样本过多会干扰主任务的学习.因此选择 $\eta=1.0\%$,在保证高提取率的同时,对主任务性能的影响最小.值得一提的是,1.0% 的投毒率明显小于大部分现有方法中的 5%~10% 的投毒率,从侧面也体现了本文方法具有数据效率高的特点.

表 6 解释空间数量 N 的消融实验

Table 6 Ablation study on the number of interpretable spaces N

解释空间数量	*CA*	*WSR*	*P*
16-bit	75.80%	98.50%	2.4e−5
32-bit	75.51%	97.42%	1.0e−9
48-bit	75.10%	94.30%	1.2e−12
64-bit	74.45%	88.20%	1.5e−10

5 结论

针对图像分类模型的版权保护问题,本文提出一种利用特征归因和背景注入的方法来创建干净标签后门水印的图像分类模型所有权验证方法,在保证模型原始分类性能的同时,实现可靠的模型所有权验证.引入显著性引导的语义解耦和

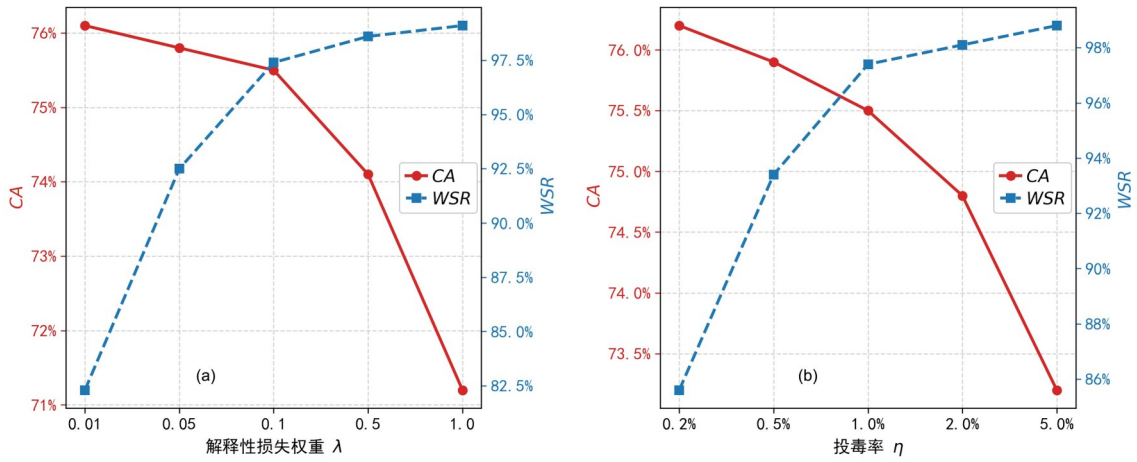


图 8 (a)解释性损失权重 λ 和(b)投毒率 η 的消融实验

Fig. 8 Ablation study on the weight of interpretability loss λ (a) and the poisoning rate η (b)

自适应超像素编码机制,显著提升了水印的视觉隐蔽性以及统计置信度.实验结果表明,本文方法在遭遇各种攻击情形时,依旧可以维持水印的稳定性以及可验证性.今后将进一步研究不同样本之间解释区域形状不对齐的问题,提高该水印方案在复杂环境下的泛化能力.

参考文献

[1] Chen B Q, Chen S Y. Deep learning//Chen B Q, Chen S Y. Machine Vision Technology. Singapore: Springer, 2026: 273－297.

[2] Chen K Y, Lin Y C, Li J L, et al. Bloodroot: When watermarking turns poisonous for stealthy backdoor// International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain: IEEE, 2026: 15782－15786.

[3] Yuan X M, Nie H W. Secure industrial federated learning: Label encryption for model protection. Engineering Applications of Artificial Intelligence, 2025, 160 (Part A): 111806.

[4] Uchida Y, Nagai Y, Sakazawa S, et al. Embedding watermarks into deep neural networks//Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. New York, NY, USA: Association for Computing Machinery, 2017: 269－277.

[5] Adi Y, Baum C, Cisse M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring//Proceedings of the 27th USENIX Conference on Security Symposium. Berkeley, NY, USA: USENIX Association, 2018: 1615－1631.

[6] Gu T Y, Liu K, Dolan-Gavitt B, et al. BadNets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 2019(7): 47230－47244.

[7] Chen X Y, Liu C, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning. https://arxiv. org/abs/1712.05526, 2017－12－15.

[8] Wang B L, Yao Y S, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks//2019 IEEE Symposium on Security and Privacy. San Francisco, CA, USA: IEEE, 2019: 707－723.

[9] Tran B, Li J, Madry A. Spectral signatures in backdoor attacks//Advances in Neural Information Processing Systems 31. Montreal, Canada: Curran Associates, 2018, Article No. 89.

[10] Gao Y S, Xu C E, Wang D R, et al. Strip: Adefence against Trojan attacks on deep neural networks// Proceedings of the 35th Annual Computer Security Applications Conference. New York, NY, USA: Association for Computing Machinery, 2019: 113－125.

[11] Turner A, Tsipras D, Madry A. Label-consistent backdoor attacks. https://arxiv. org/abs/1912. 02771, 2019－12－06.

[12] Zhao S H, Ma X J, Zheng X, et al. Clean-label backdoor attacks on video recognition models//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 14431－14440.

[13] Zhu M Y, Li Y M, Guo J F, et al. Towards sample-specific backdoor attack with clean labels via attribute trigger. IEEE Transactions on Dependable and Secure Computing, 2025, 22(5): 4685－4698.

[14] Shao S, Li Y M, Yao H W, et al. Explanation as a watermark: Towards harmless and multi-bit model ownership verification via watermarking feature attribution. https://arxiv. org/abs/2405.04825, 2024－09－10.

[15] Dhara G, Kumar R K. A survey on visual saliency detection approaches and attention models. Multimedia Tools and Applications, 2025, 84(35): 44183－44225.

[16] Zhang S J, Qi Y L, Wu J Q. Applying deep learning for style transfer in digital art: Enhancing creative expression through neural networks. Scientific Reports, 2025, 15(1): 11744.

[17] Achanta R, Shaji A, Smith K, et al. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274－2282.

[18] Li Y, Jiang Y, Li Z, et al. Backdoor learning: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(1): 5－22.

[19] Gong X, Tian B, Xue M, et al. Megatron: Evasive clean-label backdoor attacks against vision

transformer. IEEE Transactions on Dependable and Secure Computing, 2026, 23(1): 525－542.

[20] Nguyen T A, Tran A T. Wanet - imperceptible warping - based backdoor attack//International Conference on Learning Representations. Online: ICLR, 2021: 1－16.

[21] Sun M, Wang R, Zhu Z X, et al. EntropyMark: Towards more harmless backdoor watermark via entropy - based constraint for open - source dataset copyright protection//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE, 2025: 30692－30701.

[22] Yu Q K, Zhang C, Jin G J, et al. StealthMark: Harmless and stealthy ownership verification for medical segmentation via uncertainty - guided backdoors. IEEE Transactions on Image Processing, 2026, 35: 1290－1304.

[23] Selvaraju R R, Cogswell M, Das A, et al. Grad - CAM: Visual explanations from deep networks via gradient-based localization//2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 618－626.

[24] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you? ": Explaining the predictions of any classifier//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: Association for Computing Machinery, 2016: 1135－1144.

[25] Ghorbani A, Abid A, Zou J. I nterpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 3681－3688.

[26] Ross A S, Hughes M C, Doshi-Velez F. Right for the right reasons: Training differentiable models by constraining their explanations. https://arxiv. org/abs/1703.03717, 2027－05－25.

[27] Rieger L, Singh C, Murdoch W, et al. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge//Proceedings of the 37th International Conference on Machine Learning. New York, NY, USA: PMLR, 2020: 8116－8126.

[28] Tancik M, Mildenhall B, Ng R. StegaStamp: Invisible hyperlinks in physical photographs//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA: IEEE, 2020: 2114－2123.

[29] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization//2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 1510－1519.

[30] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 248－255.

[31] Parkhi O M, Vedaldi A, Zisserman A, et al. Cats and dogs//2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012: 3498－3505.

[32] Stallkamp J, Schlipsing M, Salmen J, et al. The German traffic sign recognition benchmark: A multi-class classification competition//The 2011 International Joint Conference on Neural Networks. San Jose, CA, USA: IEEE, 2011: 1453－1460.

（责任编辑　杨可盛）