

基于局部风格融合的对抗补丁隐蔽性增强方法

谢喜龙^{1,2}, 郭桐^{1,2}, 肖利民^{1,2}, 韩萌^{1,2}, 徐向荣^{1,2}, 董进³, 王良^{1,2}

¹北京航空航天大学, 软件开发环境国家重点实验室 北京 中国 100191

²北京航空航天大学计算机学院 北京 中国 100191

³北京微芯区块链与边缘计算研究院 北京 中国 100190

摘要 自从对抗样本这一概念被提出以来, 各种针对深度学习模型的对抗攻击方法引发了一系列安全性问题。其中, 对抗补丁通过在输入样本中引入特定的补丁, 使深度学习模型产生误导性的结果, 给当前的深度学习系统带来了巨大的安全隐患。然而, 目前的对抗补丁生成方法在提升对抗补丁的隐蔽性方面仍存在一定的局限性, 存在与周围环境差异较大容易被人类察觉的问题。针对这一问题, 本文提出了一种基于局部风格融合的对抗补丁隐蔽性增强方法。该方法首先基于多模型加权类激活映射寻找图像的脆弱区域, 精准定位对抗补丁的安放位置, 提高对抗补丁的攻击性。然后利用风格迁移技术, 计算目标图像与对抗补丁的风格矩阵与内容矩阵。在对抗补丁生成过程中, 综合考虑分类损失、风格损失、内容损失与边界损失, 通过余弦距离函数调整对抗补丁的生成风格与生成内容, 使对抗补丁与其覆盖的局部图像的风格与内容相融合, 从而融入周边环境, 在色彩、风格上协调统一, 降低对抗补丁在人类视觉上的可察觉性, 达到增强对抗补丁隐蔽性的目的。本文分别从攻击性和隐蔽性对生成的补丁进行实验评估, 实验结果表明, 此方法可生成兼具隐蔽性与攻击性的对抗补丁, 在人类无法察觉的同时实现攻击。

关键词 对抗补丁; 风格融合; 生成模型; 类激活映射

中图分类号 TP309.1 DOI号 10.19363/J.cnki.cn10-1380/tn.2025.09.05

Adversarial Patch Steganography Enhancement through Localized Style Fusion

XIE Xilong^{1,2}, GUO Tong^{1,2}, XIAO Limin^{1,2}, HAN Meng^{1,2}, XU Xiangrong^{1,2}, DONG Jin³, WANG Liang^{1,2}

¹ State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China

² School of Computer Science and Engineering, Beihang University, Beijing 100191, China

³ Beijing Academy of Blockchain and Edge Computing, Beijing 100190, China

Abstract Since the concept of adversarial examples was introduced, various adversarial attack methods targeting deep learning models have raised a series of security issues. Among them, adversarial patches introduce specific patches into input samples to cause deep learning models to produce misleading results, posing significant security risks to current deep learning systems. However, current adversarial patch generation methods still have certain limitations in enhancing the concealment of adversarial patches, as they are prone to being detected by humans due to significant differences from the surrounding environment. In response to this problem, this paper proposes a method for enhancing the concealment of adversarial patches based on local style fusion. The method first searches for the vulnerable regions of the image based on multi-model weighted class activation mapping, and accurately locates the placement of adversarial patches, which improves the aggressiveness of adversarial patches. Then the style migration technique is utilized to compute the style matrix and content matrix of the target image and the adversarial patch. During the adversarial patch generation process, a comprehensive approach takes into account not only classification loss but also style loss, content loss, and boundary loss. By utilizing the cosine distance function, adjustments are made to both the style and content of the generated adversarial patch. This intricate adjustment ensures that the adversarial patch seamlessly blends its style and content with the local image it overlays, effectively integrating it into the surrounding environment. This harmonization guarantees consistency in color and style, ultimately minimizing the visibility of the adversarial patch to the human eye. The ultimate goal is to significantly enhance the concealment of the adversarial patch. In this paper, the generated patches are experimentally evaluated in terms of aggressiveness and coyness, respectively, and the experimental results show that this method can generate adversarial patches with both coyness and aggressiveness, which can realize the attack while being undetectable to human beings.

Key words adversarial patch; style fuse; generative model; class activation mapping

通讯作者: 肖利民, 博士, 教授, Email: xiaolm@buaa.edu.cn。

本课题得到科技创新 2030 新一代人工智能重大项目(No. 2022ZD0117602); 国家自然科学基金项目(No. 62272026, No. 62104014)资助。

收稿日期: 2023-11-22; 修改日期: 2024-03-20; 定稿日期: 2025-08-14

1 引言

深度学习技术的蓬勃发展引领人类进入了智能化时代, 神经网络^[1-2]在图像识别、目标检测、自然语言处理等领域取得了优异的表现, 目前大量基于神经网络的应用已经投入到现实世界中使用。这些应用不仅能够帮助我们高效地处理日常事务, 也为我们的日常生活提供了极大的便利。随着神经网络层数的不断加深, 在样本上添加的微小扰动经过高维空间的放大, 使得网络的输出结果发生巨大变化, 也导致目前神经网络模型面临着被恶意攻击的风险^[3-5]。

对抗样本^[6]是一种神经网络攻击方法, 通过在图像上精心添加一些人眼不可察觉的细微扰动, 可以使得神经网络错误分类。一个经典的案例就是一张真实标签为熊猫的图片可以被神经网络正确分类, 但是, 再添加上精心设计的扰动之后, 却会被神经网络以较高的置信度错误地分类为长臂猿, 添加扰动后的图片就是对抗样本。最初, 对抗样本是在图像分类任务中发现的, 然而, 目前主流的计算机视觉任务, 如语义分割、目标检测以及目标跟踪等, 都面临对抗样本的威胁^[7-9]。

对抗补丁^[10-13]将对抗扰动集中添加图像的某一部分, 打破了数字对抗样本中范数的限制。然而, 现有方法^[14-18]生成的对抗补丁与其张贴环境相比, 往往具有极不协调的内容与风格, 导致极易被人类察觉, 如图 1 所示。如果一类攻击很容易被人眼观察出来, 从而导致对抗补丁被人为移除, 这类攻击将不具有实际应用的价值。因此, 想要真正应用对抗补丁攻击时, 必须提升对抗补丁的隐蔽性, 使其在欺骗人工智能系统的同时, 躲避人类视觉的追踪。

为了解决这一问题, 本文基于类激活映射图^[19]与风格迁移^[20]技术, 提出一种能够与对抗补丁张贴环境相融合, 在视觉上实现真正不可察觉的对抗补丁生成方法 StylePatch。StylePatch 利用类激活映射与风格迁移技术, 定位目标图像的脆弱区域, 并计算目标图像与对抗补丁的风格矩阵与内容矩阵。在对抗补丁生成过程中, 通过余弦距离函数调整对抗补丁的生成风格与生成内容, 使对抗补丁与其覆盖的局部图像的风格与内容相融合, 从而融入周边环境, 在色彩、风格上协调统一, 降低对抗补丁在人类视觉上的可察觉性, 达到增强对抗补丁隐蔽性的目的, 本文主要贡献如下:

(1) 通过对图像相似性的度量可知, StylePatch 生成的补丁的局部相似性比现有方法高 1.8~10.9 倍,

同时采用问卷调查的形式进行对抗补丁隐蔽性测试, 问卷结果表明 StylePatch 可以生成能够真正与周围环境相融合的对抗补丁;

(2) 在现有基于热力图指导而确定对抗扰动位置区域的工作基础上, 提出了一种利用多模型加权热力图定位目标图像中的脆弱区域的方法, 从而用较小的代价实现对抗攻击;

(3) 通过对比实验可知, StylePatch 能够在提升隐蔽性的同时保持较高的对抗攻击成功率。



注: 左图坦克被模型误分类为车, 右图迷彩服被模型误分类为草地, 图中数字表示图片被误分类的置信度

图 1 对抗补丁与周围环境在视觉上不协调

Figure 1 Adversarial patches are visually incongruous with their surroundings

2 背景知识与相关工作

2.1 威胁模型

在图像分类问题中, 原始图像 x 的真实标签为 y , 而对抗样本 x' 经过神经网络的分类的结果与真实标签 y 不相符时, 即为攻击成功。有目标攻击期望对抗样本被定向误识别为某一特定类别, 无目标攻击仅希望对抗样本不能被识别, 而没有指定目标类别, 相比于有目标攻击更加容易。现有的威胁模型包含以下三种^[21-22]。

(1) 零知识: 在不鲁棒的模型 F 上生成对抗样本, 对于 DNN 模型的参数等信息没有任何的了解。(2) 完全知识: 即白盒攻击, 攻击者可以获得目标模型的完整结构和参数, 包括训练数据、梯度和激活函数等信息, 并在该模型上生成对抗样本。(3) 有限知识: 即黑盒攻击, 攻击者无法获取模型结构和参数, 只能通过将原始数据输入到目标模型中来获取目标模型的输出结果, 并根据结果来进行攻击。

本文在完全知识与有限知识这两种威胁模型下对提出的方法进行攻击性实验。

2.2 类激活映射

在基于卷积神经网络的图像分类任务上, 基于类激活映射(Class Activation Mapping, CAM)^[19]的可解释方法是一种典型的深度学习可解释性方法。这种方法会输出一个与输入图像相对应的热力图, 该

热力图以高亮的形式标定出与模型决策相关的显著区域, 用于可视化深度神经网络对图像分类的决策结果。因此, 在进行对抗补丁攻击时, 可以利用类激活映射方法, 找到目标图像中最有效的攻击位置放置对抗补丁。CAM 采用全局平均池化层来代替全连接层, 根据每一个全局平均池化层到类别的连接权重, 生成最终的类激活映射图。GradCAM^[23]利用模型的梯度信息, 将特征图中局部梯度的平均值作为特征图的权重, 适用于非全局平均池化层连接的网络结构, 也可以提取任意一层特征图的热力图。Grad-CAM++^[24]提出梯度图上的每一个元素的贡献是存在差异的, 因此需要增加一个额外的权重矩阵, 来对梯度图的权重进行二次加权, 更适用于目标类别在图像中有多个的情况。基于梯度的方法很容易找到错误置信度的目标样本, 即部分权重较低的特征图却获得了较高的置信度分数, 而且局部梯度存在饱和问题, 受噪声影响较大。Score-CAM^[25]将权重定义为每个激活映射点在目标类上的前向传播分数, 提出了置信度提升的概念, 置信度提升得越高, 其贡献越大。有效缓解了局部梯度存在饱和的问题。在当前的对抗样本和对抗补丁生成算法中, 一些方法利用热力图来辅助添加对抗扰动, 但仅根据一种模型的热力图来指导扰动添加位置^[26-28]。

2.3 图像风格融合

图像风格融合是指将一幅图像的风格与另一幅图像的内容相结合, 合成出一张新的图片, 新的图像既体现出了前者的风格特征, 又涵盖了后者的内容特征。神经风格迁移(Neural Style Transfer, NST)^[29]通过使用预训练的 VGG^[30]网络的不同层来分别提取图像的内容特征(如轮廓、结构等)和风格特征(如纹理、颜色等), 之后以目标图像的风格特征为参考, 使得在原图像的内容特征保持不变的同时, 将原图像的风格改变成目标图像, 其计算方式如下:

$$\mathcal{L}_{\text{content}}(c, x, l) = \frac{1}{2} \left(\sum_{i,j} F_{i,j}^l - P_{i,j}^l \right)^2$$

$$\mathcal{L}_{\text{style}}(s, x) = \sum_{l=0}^L \omega_l G_l$$

$\mathcal{L}_{\text{content}}$ 和 $\mathcal{L}_{\text{style}}$ 表示内容损失和风格损失; c, s, x 分别表示内容图像、风格图像和合成图像; l 表示 VGG 网络的第 l 层; $F_{i,j}^l, P_{i,j}^l$ 分别表示风格图像和内容图像在第 l 层的 (i, j) 位置的输出值; ω_l 表示第 l 层风格损失的权重; G_l 表示第 l 层的风格损失, 常使用 Gram 矩阵进行计算。

2.4 对抗补丁的隐蔽性研究

相比数字的扰动攻击^[31-33], 对抗补丁不再将对抗扰动添加到整个图像, 而是以替换原图某一部分的形式集中添加到该位置, 是当前人工智能安全领域的研究热点之一。为提高对抗补丁的应用效果, 既要保证补丁的攻击性, 也要提升对抗补丁的隐蔽性。

在目前对抗补丁隐蔽性的研究中, 大多倾向于让对抗补丁具有一定的语义信息, 即在人类视觉看来, 对抗补丁不再是毫无意义的怪异图片, 而是将其判断为某种图案或装饰。GreedyFool^[14]根据输入图像像素的梯度来进行稀疏失真, 随着稀疏性的提高, 生成扰动的可察觉性会逐渐降低, AdvCam^[15]设计了一种隐蔽度量函数, 可以生成更加贴近自然环境的对抗样本, 但是其局限于数字图像, 无法在实际场景应用。LaVAN^[16]生成的对抗补丁只占据图像不到 2% 的面积, 并且具有较高的攻击率和可迁移性, 不过攻击只成功于生成攻击时对应的模型。通过定制补丁形状^[34-35]赋予对抗补丁更加多样化的风格, 使得补丁可以更好地适应不同攻击环境, 一定程度上提高了对抗补丁的隐蔽性, 但仍旧存在生成的补丁风格与原始图片的风格不一致的问题。PS-GAN^[17]通过结合感知与精神物理学的相关研究, 提出一种感知敏感性生成对抗网络, 利用感知敏感性, 结合了生成对抗网络和注意力机制, 增强补丁的攻击能力。并且该网络的输入不再是随机噪声, 而是一些具有意义的涂鸦补丁, 该网络输出的对抗补丁不再是不明意义的图案, 而是一些加了点扰动的图形。Accessorize^[18]将对抗攻击限定在一副眼镜的范围内, 通过优化眼镜框架上的纹理生成对抗补丁, 成功攻击了人脸识别系统, 但补丁的生成方法并不通用。基于光影的对抗攻击方法^[36-37]通过对图像中的阴影进行扰动, 可以产生较为自然的对抗样本, 但该方法进行攻击的一个限制是需要一个强单光源, 在光照条件较差的环境中, 很难产生显著的阴影扰动。

总的来说, 在目前对抗补丁隐蔽性的研究中虽然赋予了对抗补丁一定的语义信息, 但其生成的对抗补丁在内容与风格上依旧与其张贴的周边环境有明显的差异, 这意味着对抗补丁攻击仍有很大的概率被人类发现。受到上述方法的启发, 针对当前生成补丁与周围环境差异明显的问题, 本文提出了一种结合类激活映射与风格迁移技术的对抗补丁生成方法。利用类激活映射方法定位补丁的安放区域, 利用神经风格迁移技术将对抗补丁融入周围环境。

3 基于风格融合的对抗补丁生成模型

3.1 问题定义

对抗补丁攻击是通过替换原始图像 x 的某个部分来实现的。给定一个大小为 (H, W) 的对抗补丁 p ，该补丁可以通过公式形成对抗样本 x^{adv} ，从而使得经过充分训练的目标模型 F 输出错误结果，即

$$x^{\text{adv}} = x \odot (1 - m) + p \odot m$$

其中， $\odot$ 为矩阵点乘操作，补丁 p 被填充到与原始图

像 x 相同的大小， m 为一个大小为 (H, W) 定位掩模矩阵。设原始图像 x 的标签为 y ，则对抗补丁的攻击目标可以用下面的公式表示。

$$y \neq \arg \max F(x^{\text{adv}})$$

3.2 模型结构

为了增强对抗补丁的隐蔽性，本文提出基于局部风格融合的对抗补丁生成模型 StylePatch。该模型由脆弱区域定位和局部风格融合两部分组成，具体架构如图 2 所示。

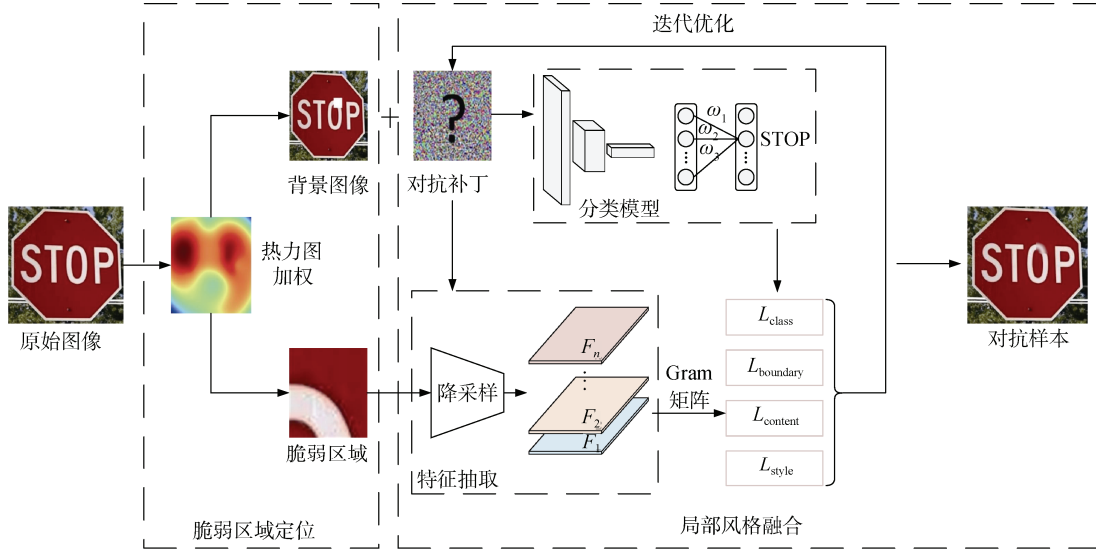


图 2 StylePatch 整体架构图

Figure 2 The architecture of StylePatch

图 2 的左半部分描述了脆弱区域定位的流程，首先基于类激活映射方法，针对多个模型输出的热力图，通过置信度加权得到目标图像的显著脆弱区域，精准标注补丁的最佳安放区域。通过脆弱区域定位方法，可将原始图片分成背景图像和脆弱区域。图 2 的右半部分描述了局部风格融合的流程，将进行随机初始化的对抗补丁与背景图像结合，输入到基于局部风格融合的生成模型中进行特征抽取，根据 Gram 矩阵算出分类损失、风格损失、内容损失与边界损失，对损失函数进行迭代优化，最后基于该生成模型生成兼具攻击性和隐蔽性的对抗补丁。

3.3 脆弱区域定位

基于梯度的类激活映射方法可能会出现错误标注问题^[25]，因此在定位脆弱区域上选用 ScoreCAM 方法。当前相关工作大都仅根据一种模型的热力图来指导扰动添加位置，不同模型对于同一目标识别的热力图存在差异，但也存在共同关注的公共区域。这表明，仅仅依靠单一模型无法准确地标注对抗扰动添加的最佳位置。因此，需要综合多个模型的输出

结果，寻找其热力图的交集区域，该区域对于各个模型的决策都有着很强的支撑作用。具体方案如图 3 所示，针对多个模型输出的热力图，通过置信度加权得到目标图像的显著脆弱区域，精准标注对抗补丁的最佳安放区域。

首先，将原始图像输入第 k 个模型，并输出为热力图 A_k ，具体而言，从深度神经网络的卷积层中提取出来 n 个特征映射并计算类别 t 相对于特征映射 f 的梯度。这些梯度经过全局平均池化与加权组合，输出热力图 A_k^t ，计算方法如下。

$$A_k^t = \sum_n \frac{1}{N} \sum_i \sum_j \frac{\partial c^t}{\partial f_{ij}^n} \times f^n$$

其次，使用第 k 个模型对原始图像的分类置信度 c_k^t 作为每张热力图的权重，将得到的热力图以平均加权的方式进行组合，即

$$a^t = W(c_k^t \times A_k^t)$$

其中， a^t 即为 k 个模型将图像分类为类别 t 时的加权热力图， W 表示平均加权方式。

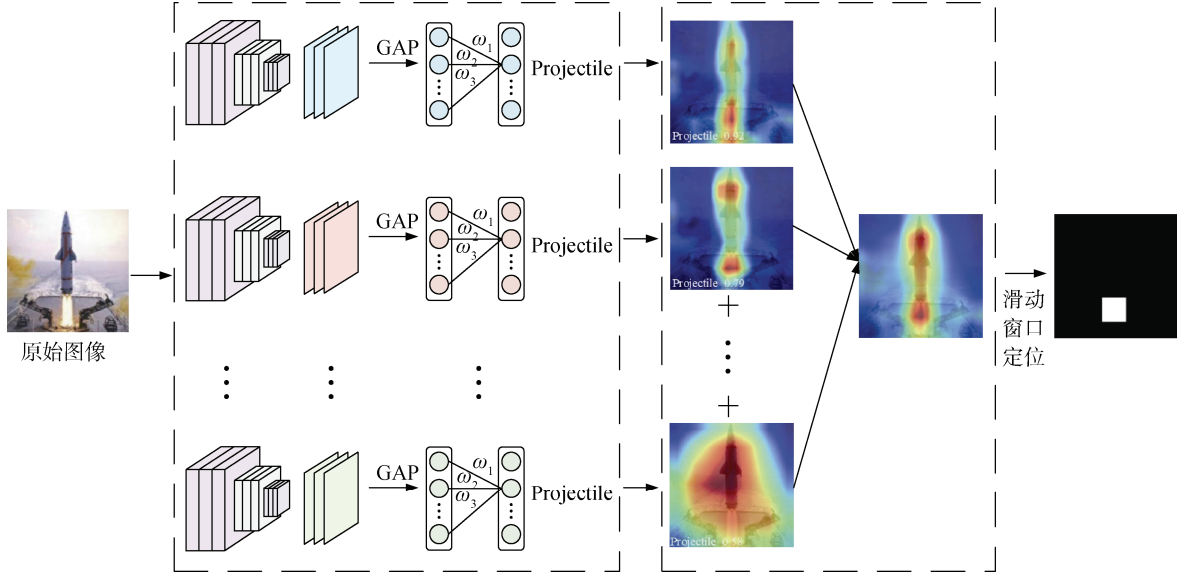


图3 基于多模型加权热力图的脆弱区域定位方法

Figure 3 Vulnerable area localization method based on multi-model weighted heat map

然后需要从加权热力图 a^t 中定位对抗补丁的安放位置。利用滑动窗口的方式, 通过计算加权热力图 a^t 中累计贡献最大的子区域 a^t_{patch} , 确定补丁的安放位置, 计算公式如下。

$$a^t_{\text{patch}} = \max \left(\sum_i^{i+h} \sum_j^{j+w} a^t_{i,j} \right)$$

其中, h 表示对抗补丁的高, w 表示对抗补丁的宽, i 和 j 的取值为 $[0, H-h]$ 与 $[0, W-w]$ 中的整数。最后, 根据子区域 a^t_{patch} 生成与图像相同大小的掩模矩阵, 掩模矩阵标记的位置即为对抗补丁的最佳安放位置。

3.4 局部风格融合

风格迁移技术作为一项计算机视觉领域的热点技术, 能够实现两张图像的内容与风格融合, 从而使得新合成的图片在视觉上看起来协调统一。不同风格的两部分图像经过局部风格迁移技术处理, 可以十分自然地融为一体, 这恰恰可以解决对抗补丁目前面临的隐蔽性难题。StylePatch 利用风格迁移技术, 可以将对抗补丁伪装成自适应风格。

具体来说, 技术方案包括以下几个步骤: 首先, 定位出对抗补丁的安放区域, 截取该区域的原图像作为局部风格图像, 剩余图像作为背景图像。将初始化的对抗补丁与背景图像结合, 输入到深度学习模型中并得到分类损失 $\mathcal{L}_{\text{class}}$ 。其中, $\mathcal{L}_{\text{class}}$ 的计算方式如下。

$$\mathcal{L}_{\text{class}} = \begin{cases} \log(p_{y_{\text{target}}}(x')), & \text{target attack} \\ -\log(p_y(x')), & \text{untarget attack} \end{cases}$$

其中, $p_y(x')$ 表示模型针对对抗样本 x' , 将其分类为类别 y 的概率。

其次, 对局部风格图像与对抗补丁进行特征抽取, 根据 Gram 矩阵计算对抗补丁与局部风格图像的风格损失 $\mathcal{L}_{\text{style}}$ 与内容损失 $\mathcal{L}_{\text{content}}$, 计算方法如下所示:

$$\mathcal{L}_{\text{style}} = \frac{G(F(x)) \cdot G(F(x'))}{\|G(F(x))\| \cdot \|G(F(x'))\|}$$

$$\mathcal{L}_{\text{content}} = \frac{F(x) \cdot F(x')}{\|F(x)\| \cdot \|F(x')\|}$$

在公式中距离的度量方式选择余弦距离, $G(x)$ 表示 x 的 Gram 矩阵, 由任意向量之间的内积组成, 用来表示特征之间的相关性大小, 即

$$G(x_1, x_2, \dots, x_k) = \begin{pmatrix} (x_1 \cdot x_1) & (x_1 \cdot x_2) & \dots & (x_1 \cdot x_k) \\ (x_2 \cdot x_1) & (x_2 \cdot x_2) & \dots & (x_2 \cdot x_k) \\ \vdots & \vdots & \ddots & \vdots \\ (x_k \cdot x_1) & (x_k \cdot x_2) & \dots & (x_k \cdot x_k) \end{pmatrix}$$

其中, $F(x)$ 表示为 x 的特征, 为卷积神经网络的某一层的输出。同时, 考虑到对抗补丁安放时, 由于边缘像素与背景像素可能会产生较大差异而变得明显, 本文提出边界损失 $\mathcal{L}_{\text{boundary}}$, 可以平滑对抗补丁与背景图像在边缘部分的像素, 进一步提升隐蔽性, 计算方法如下所示:

$$\mathcal{L}_{\text{boundary}} = \mathbb{E}_n(e_{i,j}) + \lambda \mathbb{D}_n(e_{i,j})$$

其中, $e_{i,j}$ 表示对抗补丁区域边界上的像素点, 计算以 $e_{i,j}$ 为中心的 $n \times n$ 范围内像素的均值 E 与方差 D , λ

为一个超参数,用于控制两项的重要程度。最后,将分类损失、风格损失、内容损失与边界损失根据公式进行加权组合,即

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{style}} + \lambda_2 \mathcal{L}_{\text{content}} + \lambda_3 \mathcal{L}_{\text{boundary}} - \lambda_4 \mathcal{L}_{\text{class}}$$

其中, $\mathcal{L}_{\text{total}}$ 即为最终的损失函数, λ 为一个超参数,用来控制第 k 项损失函数的重要程度。综上,基于局部风格融合的对抗补丁隐蔽性增强方法可以总结为以下优化问题:

$$\min_x (\lambda_1 \mathcal{L}_{\text{style}} + \lambda_2 \mathcal{L}_{\text{content}} + \lambda_3 \mathcal{L}_{\text{boundary}}) + \max \lambda_4 \mathcal{L}_{\text{class}}$$

通过迭代优化对抗补丁,生成能够与补丁安放环境相融合并具有攻击效果的对抗补丁,降低视觉上的可察觉性。本方法将局部图像的风格与对抗补丁进行融合,避免了全局风格与局部风格不一致产生的对抗补丁无法与背景融合的影响。提出边界损失,解决了对抗补丁安放时的边缘分界明显的问题,进一步提高了对抗补丁的隐蔽性。

4 实验及结果分析

4.1 实验设计

实验部分首先对提出的基于类激活映射的脆弱区域定位方法进行了对比实验,之后从生成对抗补丁对多种分类模型的攻击性和隐蔽性两个方面验证 StylePatch 的有效性。

4.1.1 数据集及评估指标

数据集选择 ImageNet ILSVRC2012^[38]数据集, ImageNet 是目前在计算机视觉相关任务中主流的数据集,对比实验中使用的 ResNet^[39]与 VGG 模型同样在 ImageNet 上进行预训练。实验中对比的方法为 AP^[10]、LaVAN^[16]以及 PS-GAN^[17], AP 首次提出对抗补丁的攻击方法, LaVAN 使用了一种被限制位置且不遮盖主要目标的扰动, PS-GAN 利用感知损失来优化对抗补丁。

基于与原图越相似的对抗补丁,其隐蔽性越好这一合理假设选择将隐蔽性度量转化为图像之间的相似度度量。在隐蔽性实验中选择峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)^[40-41]与结构相似性(Structural SIMilarity, SSIM)^[42-43]作为评估指标,攻击性实验选择对分类模型的攻击成功率作为评估指标,并在白盒、黑盒、无目标三种场景下进行了实验。

PSNR 通过图像的均方差来计算两幅图像之间的相似度,结构相似性通过亮度、对比度和结构三种图像特征来综合计算图像间的相似度。通常情况下,当 PSNR 的值高于 40dB 时,就说明两幅图像的相似度已经在肉眼看来无法察觉;当 PSNR 的值在 30~

40dB 时,说明被对比图像质量非常好,当 PSNR 的值在 20~30dB 时,说明被对比图像的质量较差,人类视觉可以察觉到明显的失真和变形;当 PSNR 的值低于 20dB 时,说明两幅图像相差较大,基本可以判定为不相似。SSIM 是一种用于客观评价图像间相似度的标准,它通过亮度、对比度和结构三种图像特征来综合计算图像间的相似度,计算 SSIM 数值时会通过窗口的不断滑动来计算局部均值,最终得到全局 SSIM,数值越接近于 1 则说明两幅图像越相似。

白盒攻击成功率指当生成对抗补丁的模型与遭受对抗补丁攻击的模型是同一个模型且模型参数相同时,该模型将带有对抗补丁的对抗样本识别错误的概率。白盒目标攻击 Top-1 成功率指在白盒场景下,被攻击模型将带有对抗补丁的对抗样本识别为特定攻击目标标签的概率。黑盒攻击成功率指当生成对抗补丁的模型与遭受对抗补丁攻击的模型不是同一个模型时,被攻击模型将带有对抗补丁的对抗样本识别错误的概率。无目标攻击成功率指被攻击模型将带有对抗补丁的对抗样本识别为除正确标签以外的其他标签的概率。

4.1.2 数据实验参数设置

在模型的选用上,选用 PyTorch 深度学习框架中的预训练模型 ResNet50、ResNet101、ResNet152、VGG16 和 VGG19 生成对抗补丁,其模型结构与训练参数均未改变。在数据集上,从 ImageNet 数据集中选择 1000 个类共 50000 张图像作为测试数据集,并将其大小统一为 224×224 。将对抗补丁形状设置为正方形;对抗补丁大小为原始图像总像素数的 5%;置信度的阈值设置为 0.9,即归一化后的对抗样本的误分类置信度超过 0.9 则判定为攻击成功;在攻击性实验中,单次攻击在黑盒和白盒场景下最大迭代次数为 2000 次,若超过此次数仍未攻击成功,则判定为攻击失败;优化算法选用随机梯度下降算法,学习率设置为 1,动量设置为 0.9;在损失函数中,需要综合考虑生成补丁的攻击性和隐蔽性,经过多次实验,将分类损失、风格损失、内容损失与边界损失的权重分别设置为 3、2、2、1,可以较好地平衡对抗补丁的攻击性和隐蔽性。

4.1.3 实验环境

实验采用的服务器配置为: Intel(R) core i7 CPU, 主频 3.2GHz, 拥有 8 个物理核心 16 个逻辑线程;内存型号为 DDR4, 容量为 32GB*4;硬盘为 1T SSD;显卡为两块 RTX2080Ti。操作系统为 Ubuntu 18.04, Python 版本为 3.6.9。程序运行过程中主要使用的 Python 运行库及其版本号如下: torch 1.6.0+cu101,

torchvision 0.7.0+cu101, numpy 1.19.5。

4.2 脆弱区域定位对比实验

在脆弱区域定位对比实验中, 选用 ResNet、VGG 两种深度神经网络定位原始图像的显著特征。考虑到不同深度神经网络输入维度的问题, 在生成多模型加权热力图与设置对比实验时选用具有相同输入维度的模型进行测试。在 AP 方法的基础上增加脆弱区域定位(WAP), 并在白盒与黑盒场景下针对多种模型进行对比实验, 在 10000 张图片上对 WAP 在白盒与黑盒场景下分别进行目标攻击 Top-1 成功率和无目标攻击成功率测试, 实验结果如表 1~表 4 所示。

表 1 WAP 与 AP 在黑盒场景下目标攻击 Top-1 成功率
Table 1 Top-1 success rate of targeted attacks on WAP and AP in black-box scenario

	ResNet101	ResNet152	VGG16	VGG19
AP	44.49	17.91	11.94	15.05
WAP	62.60	34.02	30.20	30.71

表 2 WAP 与 AP 在黑盒场景下无目标攻击成功率
Table 2 success rate of non-targeted attacks for WAP and AP in black-box scenario

	ResNet101	ResNet152	VGG16	VGG19
AP	52.76	31.06	41.87	39.24
WAP	74.93	58.03	70.36	67.41

表 3 WAP 与 AP 在白盒场景下目标攻击 Top-1 成功率
Table 3 Top-1 success rate of targeted attacks on WAP and AP in white-box scenario

	ResNet101	ResNet152	VGG16	VGG19
AP	97.90	95.75	96.73	97.57
WAP	98.68	98.92	99.44	99.47

表 4 WAP 与 AP 在白盒场景下无目标攻击成功率
Table 4 Success rate of non-targeted attacks on WAP and AP in white-box scenario

	ResNet101	ResNet152	VGG16	VGG19
AP	98.17	96.46	99.58	99.45
WAP	98.40	99.85	99.86	99.86

在黑盒攻击场景下, 将在 ResNet50 上进行训练的攻击模型迁移到不同模型上, 从表 1 和表 2 中的数据可以看出, WAP 在进行目标攻击 Top-1 取得了更高的攻击成功率。但在攻击成功率原本就处于极低水平的情况下, 使用区域定位的方法提升相对有限。无目标攻击有着十分明显的提升, 可在其基础上将无目标攻击成功率最大提升 28.17%。从表 3、表 4 中数据可知, 在白盒场景下, 对 AP 使用脆弱区域定位

后进行目标攻击有着更高的攻击成功率, 在 Top-1 准确率上平均有 2%~3%的提高, 最高能将对抗补丁的 Top-1 攻击成功率提升到 99.47%。在进行无目标攻击时, 区域定位可以深度学习模型产生更多偏离正常分类的结果, 提高了无目标攻击的成功率。

从上表数据中还可以观察到, 模型生成的补丁在攻击具有相似架构的另一个模型时会取得更好的攻击效果。这可能是相似架构的模型会构建相似的分类超平面, 而根据相似的超平面生成的补丁, 会更容易攻击同类模型。

综上可知, 使用多模型热力图加权的脆弱区域定位方法可在白盒场景与黑盒场景下, 有效提升对抗补丁对模型进行目标攻击与无目标攻击的成功率。考虑到在实际应用对抗补丁时, 攻击者往往无法获取到目标模型的架构、参数等信息, 此方法不需要改变模型结构, 不需要对补丁进行重新训练即可应用, 具有稳定的兼容性。

4.3 隐蔽性对比实验

在 VGG16 模型上用 StylePatch、AP、LaVAN 与 PS-GAN 四种方法生成了相应的对抗补丁与对抗样本, 生成结果如图 4 所示。将结构相似性与峰值信噪比作为对抗补丁的隐蔽性度量的客观指标。分别选取了 1000 张、5000 张、10000 张图片, 测试 StylePatch、AP、LaVAN 与 PS-GAN 四种方法生成的对抗样本与原图的局部、全局结构相似性 SSMI 和局部、全局峰值信噪比 PSNR, 结果如图 5~图 8 所示。

为了公平比较, 所有补丁都被添加在了同一位置, 并且所有的图像都成功让模型错误识别。最左边一列放置原始图像。通过对比, 可以很明显地观察到, AP 与 LaVAN 方法生成的对抗补丁很容易就会被人类第一时间发现, 而 PS-GAN 生成的对抗补丁虽然具有一定的语义信息, 但在图像中依然较为明显。相比之下, StylePatch 所生成的对抗补丁, 能够在视觉上更加自然地与周围环境相融合, 真正实现肉眼的不可察觉性。

结构相似性 SSMI 越接近 1, 说明图像越接近原图, 峰值信噪比 PSNR 越大, 说明图像失真程度越低。从图 5 和图 7 中可以看出, StylePatch 在局部结构相似性与局部峰值信噪比上比其他三种方法高 1.8~10.9 倍, 说明 StylePatch 生成的对抗补丁能够更加贴近攻击目标的周围环境, 也更不容易被发现。从图 6 和图 8 中可以看出 StylePatch 在全局结构相似性与全局峰值信噪比上都取得了更优异的表现, 说明在全局图像上, StylePatch 生成的对抗样本具有更好的隐蔽性。

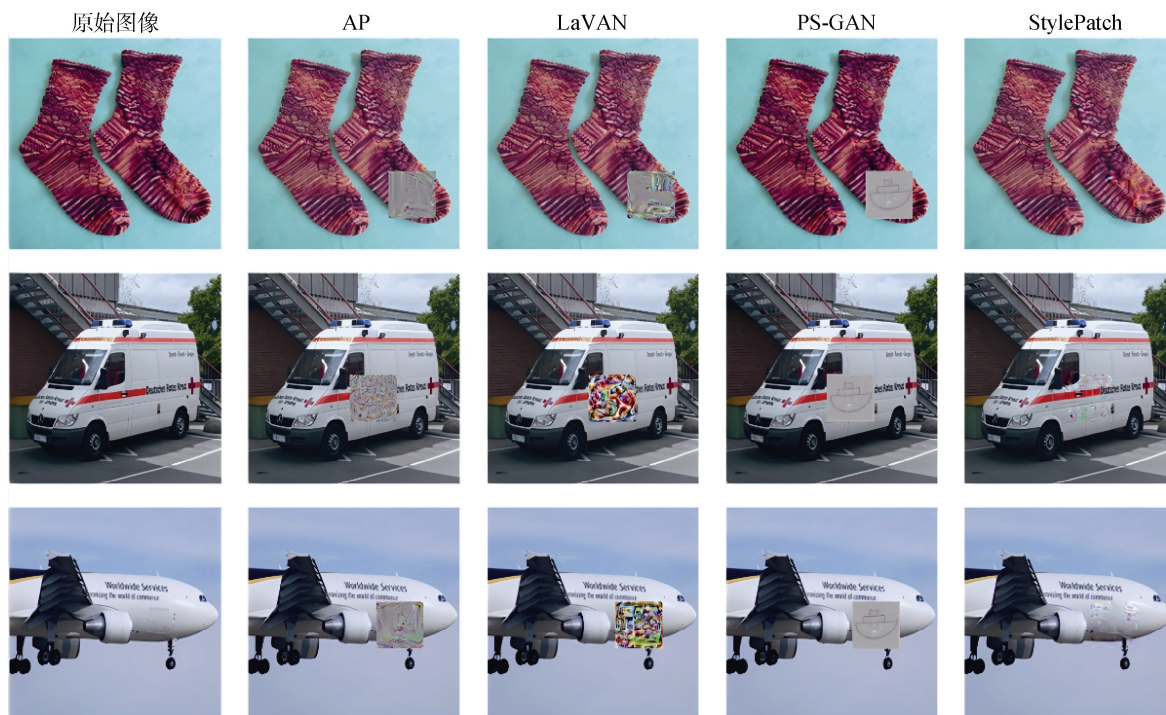


图4 StylePatch 与 AP、LaVAN、PS-GAN 在 VGG16 上生成补丁效果对比

Figure 4 Comparison of patch generation performance for StylePatch, AP, LaVAN, and PS-GAN on VGG16

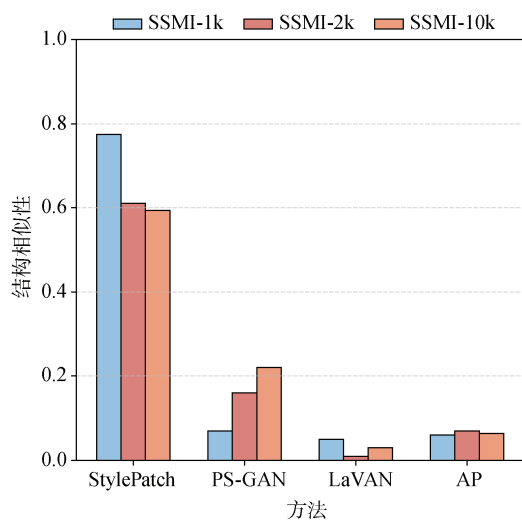


图5 不同方法生成补丁的局部 SSMI 对比

Figure 5 Comparison of localized SSMI for different methods of generating patches

同时, 本节设计了对抗补丁隐蔽性主观测试调查问卷实验, 以进行人类视觉对于 StylePatch、AP、LaVAN 与 PS-GAN 四种方法生成的对抗补丁在隐蔽性上的主观测试。测试问卷包括 10 组对比集合, 其中每组集合由原图与 AP、LaVAN、PS-GAN、StylePatch 使用目标攻击生成的对抗样本组成, 并设置了“十分明显”、“有点明显”、“不明显”、“几乎不可见”与“无法察觉”五个选项来进行主观评

价。本次问卷采取在线发放、匿名填写的形式。参与者需根据自己的主观判断对每组图片中的对抗补丁的隐蔽性进行主观评价。为了更直观地对比测试结果, 我们将问卷中的五个选项分别赋予 1、2、3、4、5 作为得分, 分别代表“十分明显”、“有点明显”、“不明显”、“几乎不可见”和“无法察觉”。最后, 将计算每种方法的平均得分作为主观测试的比较依据。

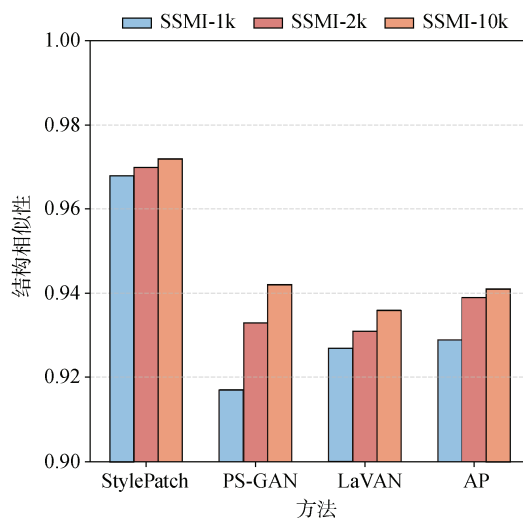


图6 不同方法生成补丁的全局 SSMI 对比

Figure 6 Comparison of global SSMI for different methods of generating patches

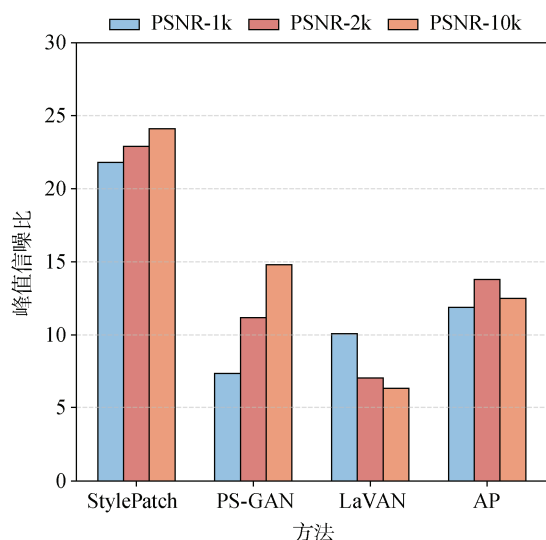


图 7 不同方法生成补丁的局部 PSNR 对比

Figure 7 Comparison of localized PSNR for different methods of generating patches

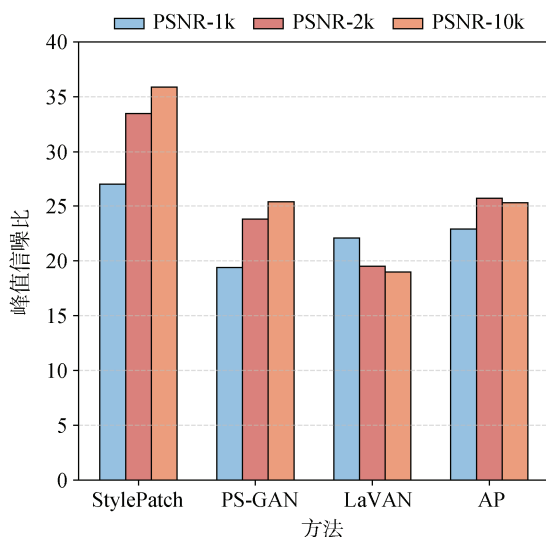


图 8 不同方法生成补丁的全局 PSNR 对比

Figure 8 Comparison of global PSNR for different methods of generating patches

此次调研共收集到问卷 503 份, 问卷调查参与者的身份大多为高校学生, 在问卷中设置指示项, 若未按照指示项填答, 则视为无效问卷。经过数据整理与统计分析, 其中有效问卷 503 份, 无效问卷 0 份, 四种方法生成的对抗补丁的隐蔽性主观测试得分结果如图 9 所示。

根据对抗补丁隐蔽性主观测试得分结果, 可以很明显地看出, AP 与 LaVAN 的得分相对较低, 表明这两种方法生成的对抗补丁极易被人类视觉系统发现。PS-GAN 的得分相对较高, 因为 PS-GAN 生成的对抗补丁具有一定的语义信息。而 StylePatch 的得分远远高于其他三种方法, 表明 StylePatch 可生成与周

围环境相融合的对抗补丁, 在人类视觉系统看来更具隐蔽性。

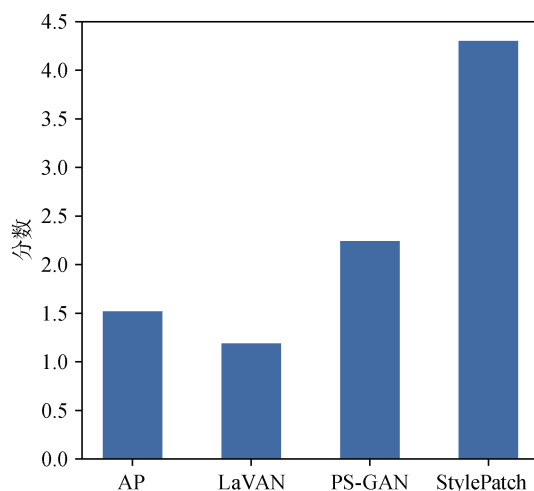


图 9 对抗补丁隐蔽性主观测试结果

Figure 9 Subjective test results of adversarial patch concealment

4.4 攻击性对比实验

本文使用 StylePatch、AP、LaVAN 与 PS-GAN 四种攻击方法对 ResNet50、ResNet101、ResNet152、VGG16、VGG19 模型进行了白盒攻击 Top-1 与目标攻击测试, 实验结果如图 10、图 11 所示。

从图中可以看出, 包括 StylePatch 在内的四种攻击方法在白盒攻击 Top-1 与无目标攻击场景下均可取得 95% 以上的攻击成功率。在白盒攻击 Top-1 场景下, StylePatch 在 ResNet 系列模型上的攻击成功率稍低于 LaVAN, 在 VGG19 模型上稍低于 AP。在无目标攻击场景下, StylePatch 的攻击成功率稍低于

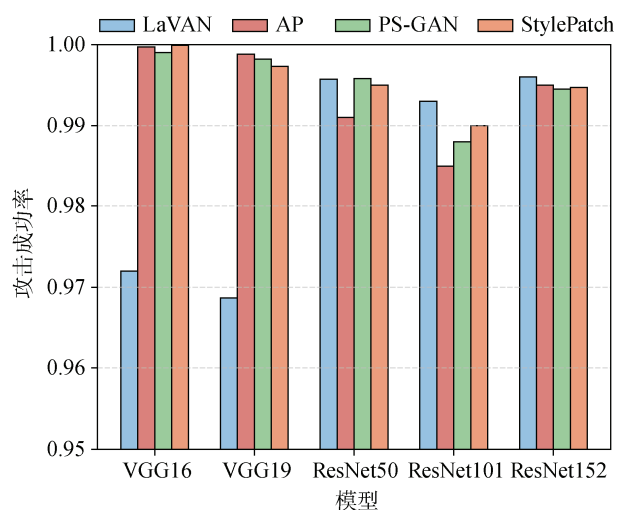


图 10 白盒场景下目标攻击 Top-1 成功率

Figure 10 Top-1 success rate of targeted attacks in white-box scenario

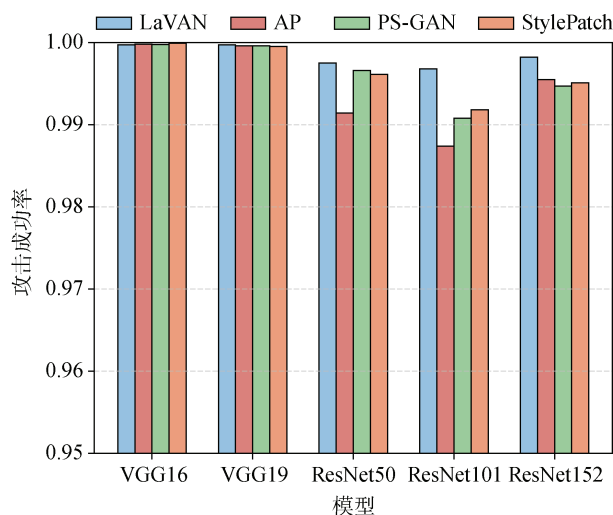


图 11 白盒场景下无目标攻击成功率

Figure 11 Success rate of non-targeted attacks in white-box scenario

LaVAN。原因在于对抗补丁生成时需要结合周围像素, 补丁生成的解空间受限, 而导致成功率略低于最优方法。但 StylePatch 基本与各种最优的方法取得的攻击效果类似, 说明 StylePatch 在提高隐蔽性的同时, 在白盒场景下的攻击性仍然能够保持较高的水准。

使用 StylePatch、AP、LaVAN 与 PS-GAN 四种方法, 在 VGG19 上生成相应的对抗补丁, 并在 VGG16、ResNet50、ResNet101 与 ResNet152 模型上进行了黑盒场景下的无目标攻击测试。在生成对抗补丁时并没有考虑通用攻击效果, 因此在黑盒场景下针对 ImageNet 数据集中的 1000 个类分别进行了无目标攻击测试。为了直观地展现出 StylePatch、AP、LaVAN 与 PS-GAN 四种方法在黑盒场景下的无目标攻击效果, 本章将上述四种方法在 VGG16、ResNet50、ResNet101 与 ResNet152 模型上的无目标攻击结果按降序排列, 统计前二十个类别的平均成功率进行对比, 统计结果如表 5 所示。

表 5 黑盒场景下无目标攻击平均成功率

Table 5 Average success rate of non-targeted attacks in black box scenario

	VGG16	ResNet50	ResNet101	ResNet152
StylePatch	85.70	65.99	62.85	61.35
AP	67.06	47.66	42.20	43.21
LaVAN	97.63	60.72	50.24	51.72
PS-GAN	60.16	43.82	41.27	39.26

根据表中数据可知, StylePatch 在 ResNet50、ResNet101 与 ResNet152 模型上的无目标攻击平均成

功率均高于其他方法, 最大可相差 22%。在 VGG16 模型上, StylePatch 的平均成功率低于 LaVAN 方法, 远远高于 AP 与 PS-GAN 方法, 在生成补丁的隐蔽性指标上 StylePatch 比 LaVAN 更好。

根据以上实验结果可以看出本文提出的 StylePatch 不但能够生成在客观度量与人类主观视觉下都更加隐蔽的对抗补丁, 而且所生成的对抗补丁在白盒场景与黑盒场景下的目标攻击与无目标攻击上均具有较高的攻击成功率。因此, 说明 StylePatch 是一种可以生成兼顾隐蔽性与攻击性的对抗补丁生成方法。

4.5 对抗补丁生成时间评估实验

为更好地评测使用对抗补丁攻击所需要的成本, 设计对抗补丁生成时间评估实验。在同一服务器的两块 RTX2080Ti 上, 在白盒场景下选择 StylePatch 与 AP 这两种方法进行 500 次攻击实验, 即生成 500 张对抗补丁, 补丁的形状和尺寸保持相同, 记录时间开销, 实验结果如表 6 与表 7 所示。

表 6 白盒场景下无目标攻击平均时间

Table 6 Average time of non-targeted attacks in white-box scenario

	区域定位时间(s)	补丁生成时间(s)	平均时间(s/张)
AP	/	220.3	0.441
StylePatch	171.6	291.4	0.926

表 7 白盒场景下有目标攻击平均时间

Table 7 Average time of targeted attacks in white-box scenario

	区域定位时间(s)	补丁生成时间(s)	平均时间(s/张)
AP	/	316.1	0.632
StylePatch	195.3	398.6	1.188

根据表中数据可知, 在白盒场景下, StylePatch 无目标攻击所需要的平均时长约为 AP 方法的 2.1 倍, 其中区域定位消耗的时间约占总时间的 37%, StylePatch 在白盒场景下有目标攻击所需要的平均时长约为 AP 方法的 1.9 倍, 其中区域定位消耗的时间约占总时间的 33%。StylePatch 需要额外对脆弱区域进行定位, 同时需要结合风格融合技术进行对抗补丁的生成, 因此所需要的时间开销比 AP 要大一些。

由于有目标攻击相较无目标攻击更加困难, 因此有目标攻击的时间开销更大, 本实验中无目标攻击和有目标攻击的平均时长相差 0.26 秒。

5 讨论

本文提出的基于局部风格融合的对抗补丁隐蔽

性增强方法能够生成真正与环境相融合的对抗补丁, 若需要将该方法在实际场景下进行应用与推广时, 还需考虑真实的物理场景。当生成的对抗补丁处于多种变换的动态环境中时, 可能会降低其攻击性, 因此在提升隐蔽性的同时, 考虑生成的对抗补丁的鲁棒性十分重要。本文使用的扰动过于细微, 大大减弱了对抗补丁在黑盒攻击上的攻击强度, 在后续的研究中, 需要进一步平衡攻击性与隐蔽性, 同时提高二者的表现, 生成更具威胁、更不易被发现的对抗补丁。一种可能的方法是结合局部风格融合方法与其他对抗补丁生成方法, 同时, 可以引入图像更多的视觉特征信息和图像上下文信息, 以指导对抗补丁的生成过程, 使其兼具隐蔽性和攻击性。最后, 引入补丁检测方法并对本文提出的方法进行针对性的防御, 也是我们后续研究的一个重要方向。

6 结论

本文提出了一种基于局部风格融合的对抗补丁隐蔽性增强方法。该方法利用类激活映射技术定位对抗补丁的显著区域, 结合风格迁移技术在对抗补丁生成过程中控制对抗补丁的生成风格与生成内容, 使对抗补丁与其覆盖的局部图像的风格与内容相融合, 从而融入局部图像周围环境, 在色彩与风格上做到协调统一, 以增强对抗补丁的隐蔽性。实验结果表明, 此方法可生成兼具隐蔽性与攻击性的对抗补丁, 在人类无法察觉的同时成功攻击各种深度学习模型。

参考文献

- [1] Pouyanfar S, Sadiq S, Yan Y L, et al. A Survey on Deep Learning: Algorithms, Techniques, and Applications[J]. *ACM Computing Surveys*, 2018, 51(5): 1-36.
- [2] Dong S, Wang P, Abbas K. A Survey on Deep Learning and Its Applications[J]. *Computer Science Review*, 2021, 40: 100379.
- [3] He Y Z, Meng G Z, Chen K, et al. Towards Security Threats of Deep Learning Systems: A Survey[J]. *IEEE Transactions on Software Engineering*, 2022, 48(5): 1743-1770.
- [4] Wang J, Wang C Y, Lin Q Z, et al. Adversarial Attacks and Defenses in Deep Learning for Image Recognition: A Survey[J]. *Neurocomputing*, 2022, 514: 162-181.
- [5] Meyers C, Löfstedt T, Elmroth E. Safety-Critical Computer Vision: An Empirical Survey of Adversarial Evasion Attacks and Defenses on Computer Vision Systems[J]. *Artificial Intelligence Review*, 2023, 56(1): 217-251.
- [6] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. 2013: ArXiv Preprint arXiv:1312.6199.
- [7] Arnab A, Miksik O, Torr P H S. On the Robustness of Semantic Segmentation Models to Adversarial Attacks[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 888-897.
- [8] Zhang H C, Wang J Y. Towards Adversarially Robust Object Detection[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2019: 421-430.
- [9] Jia Y J, Lu Y, Shen J, et al. Fooling detection alone is not enough: Adversarial attack against multiple object tracking[C]. *International Conference on Learning Representations*, 2020.
- [10] Brown T B, Mané D, Roy A, et al. Adversarial patch[J]. 2017: ArXiv Preprint ArXiv:1712.09665.
- [11] Wei X, Pu B, Lu J, et al. Physically adversarial attacks and defenses in computer vision: A survey[J]. 2022: ArXiv Preprint ArXiv:2211.01671.
- [12] Komkov S, Petiushko A. AdvHat: Real-World Adversarial Attack on ArcFace Face ID System[C]. *2020 25th International Conference on Pattern Recognition*, 2021: 819-826.
- [13] Eykholt K, Evtimov I, Fernandes E, et al. Robust Physical-World Attacks on Deep Learning Visual Classification[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 1625-1634.
- [14] Dong X, Chen D, Bao J, et al. GreedyFool: Distortion-aware sparse adversarial attack [J]. *Advances in Neural Information Processing Systems*, 2020, 33: 11226-11236.
- [15] Duan R J, Ma X J, Wang Y S, et al. Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 997-1005.
- [16] Karmon D, Zoran D, Goldberg Y. Lavan: Localized and visible adversarial noise[C]. *International Conference on Machine Learning*, 2018:2507-2515.
- [17] Perceptual-Sensitive GAN for Generating Adversarial Patches[C]. *The Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2019: 1028-1035.
- [18] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 1528-1540.
- [19] Zhou B L, Khosla A, Lapedriza A, et al. Learning Deep Features for Discriminative Localization[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2921-2929.
- [20] Jing Y C, Yang Y Z, Feng Z L, et al. Neural Style Transfer: A Review[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2020, 26(11): 3365-3385.
- [21] Carlini N, Wagner D. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods[C]. *The 10th ACM Workshop on Artificial Intelligence and Security*, 2017: 3-14.
- [22] QIAN Yaguan, LIU Xinwei, GU Zhaoquan, et al. QR Code Based Patch Attacks in Physical World[J]. *Journal of Cyber Security*, 2020, 5(6):75-86.
(钱亚冠, 刘新伟, 顾钊铨等 一种基于二维码对抗样本的物理补丁攻击[J]. *信息安全学报*, 2020, 5(6):75-86.)
- [23] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual Ex-

- planations from Deep Networks via Gradient-Based Localization[J]. *International Journal of Computer Vision*, 2020, 128(2): 336-359.
- [24] Chattopadhyay A, Sarkar A, Howlader P, et al. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks[C]. *2018 IEEE Winter Conference on Applications of Computer Vision*, 2018: 839-847.
- [25] Wang H F, Wang Z F, Du M N, et al. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020: 111-119.
- [26] Wu W B, Su Y X, Chen X X, et al. Boosting the Transferability of Adversarial Samples via Attention[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 1158-1167.
- [27] Dai Z Y, Liu S C, Li Q, et al. Saliency Attack: Towards Imperceptible Black-Box Adversarial Attack[J]. *ACM Transactions on Intelligent Systems and Technology*, 2023, 14(3): 1-20.
- [28] Liu H H, Zuo X Q, Huang H, et al. Saliency Map-Based Local White-Box Adversarial Attack Against Deep Neural Networks[C]. *Artificial Intelligence*, 2022: 3-14.
- [29] Gatys L A, Ecker A S, Bethge M. Image Style Transfer Using Convolutional Neural Networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2414-2423.
- [30] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. 2014: ArXiv Preprint ArXiv:1409.1556.
- [31] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. 2014: ArXiv Preprint ArXiv:1412.6572.
- [32] Papernot N, McDaniel P, Jha S, et al. The Limitations of Deep Learning in Adversarial Settings[C]. *2016 IEEE European Symposium on Security and Privacy*, 2016: 372-387.
- [33] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2574-2582.
- [34] Eykholt K, Evtimov I, Fernandes E, et al. Physical Adversarial Examples for Object Detectors[C]. *The 12th USENIX Conference on Offensive Technologies*, 2018: 1.
- [35] Zhao Y, Zhu H, Liang R G, et al. Seeing Isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors[C]. *The 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019: 1989-2004.
- [36] Zhong Y Q, Liu X M, Zhai D M, et al. Shadows Can Be Dangerous: Stealthy and Effective Physical-World Adversarial Attack by Natural Phenomenon[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 15324-15333.
- [37] Bose A J, Aarabi P. Adversarial Attacks on Face Detectors Using Neural Net Based Constrained Optimization[C]. *2018 IEEE 20th International Workshop on Multimedia Signal Processing*, 2018: 1-6.
- [38] Deng J, Dong W, Socher R, et al. ImageNet: A Large-Scale Hierarchical Image Database[C]. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009: 248-255.
- [39] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.
- [40] Suriyan K, Ramaingam N, Rajagopal S, et al. Performance Analysis of Peak Signal-to-Noise Ratio and Multipath Source Routing Using Different Denoising Method[J]. *Bulletin of Electrical Engineering and Informatics*, 2022, 11(1): 286-292.
- [41] Koonjoo N, Zhu B, Bagnall G C, et al. Boosting the Signal-to-Noise of Low-Field MRI with Deep Learning Image Reconstruction[J]. *Scientific Reports*, 2021, 11: 8248.
- [42] Bakurov I, Buzzelli M, Schettini R, et al. Structural Similarity Index (SSIM) Revisited: A Data-Driven Approach[J]. *Expert Systems with Applications*, 2022, 189: 116087.
- [43] Atta M E E, Ibrahim D K, Gilany M I. Detection and Diagnosis of Bearing Faults under Fixed and Time-Varying Speed Conditions Using Persistence Spectrum and Multi-Scale Structural Similarity Index[J]. *IEEE Sensors Journal*, 2022, 22(3): 2637-2646.



谢喜龙 于2022年在中国石油大学(华东)计算机科学与技术专业获得学士学位。现在北京航空航天大学计算机系统结构专业攻读博士学位。研究领域为人工智能安全、计算机体系结构。研究兴趣包括: 对抗学习、对抗样本。Email: xxl1399@buaa.edu.cn



郭桐 于2023年在北京航空航天大学计算机技术专业获得硕士学位。研究兴趣包括: 人工智能安全、对抗样本、对抗学习。Email: guotong@buaa.edu.cn



肖利民 于1998年在中国科学院计算所计算机系统结构专业获得博士学位。现任北京航空航天大学教授, CCF 杰出会员。主要研究方向为计算机体系结构、系统软件、高性能计算机、智能计算芯片架构及技术。Email: xiaolm@buaa.edu.cn



韩萌 于2019年在北京邮电大学计算机科学与技术专业获得学士学位。现在北京航空航天大学计算机系统结构专业攻读博士学位。研究兴趣包括: 领域定制加速器、人工智能。Email: hanm@buaa.edu.cn



徐向荣 于 2020 年在北京航空航天大学计算机科学与技术专业获得学士学位。现在北京航空航天大学计算机系统结构专业攻读博士学位。研究兴趣包括: GPU 硬件结构、人工智能。Email: xxr0930@buaa.edu.cn



董进 于 2000 年在清华大学控制理论与控制工程专业获得博士学位。现任北京微芯区块链和边缘计算研究院院长。研究兴趣包括: 低功耗芯片系统、区块链、人工智能等核心技术研发及关键领域应用。Email: dongjin@baec.org.cn



王良 于 2017 年在香港中文大学计算机科学与技术专业获得博士学位。现任北京航空航天大学助理教授。研究兴趣包括: 智算芯片、人工智能。Email: lwang20@buaa.edu.cn