

基于 PatchTracker 的对抗补丁防御算法

肖镇杰¹, 黄诗瑀¹, 叶锋^{1,2}, 黄丽清^{1,2}, 黄添强^{1,2}

(1. 福建师范大学计算机与网络空间安全学院, 福建 福州 350117; 2. 数字福建大数据安全技术研究所, 福建 福州 350117)

摘要: 基于深度神经网络的目标检测技术已经广泛应用于各领域, 然而, 通过对抗补丁攻击在图像中添加局部扰动, 以此来误导深度神经网络, 对基于目标检测技术的视觉系统构成了严重威胁。针对这一问题, 利用对抗补丁和图像背景的语义差异性, 提出了一种基于 PatchTracker 的对抗补丁防御算法, 该算法由上游补丁检测器与下游数据增强模块组成。上游补丁检测器使用带有注意力机制的 YOLOV5 (you only look once-v5) 确定对抗补丁所在位置, 有助于提高对小尺度对抗补丁的检测精度; 将检测区域用合适的像素值覆盖以抹除对抗补丁, 上游补丁检测器不仅能够有效降低对抗样本的攻击性, 而且不依赖大规模的训练数据; 下游数据增强模块通过改进模型训练范式, 提高下游目标检测器的鲁棒性; 将抹除补丁后的图像输入经过数据增强的下游 YOLOV5 目标检测模型。在公开的 TT100K 交通标志数据集上进行了交叉验证, 实验表明, 与未采取防御措施相比, 所提算法能够有效防御多种类型的通用对抗补丁攻击, 在检测对抗补丁图像时的 mAP (mean average precision) 提高 65% 左右, 有效地改善了小尺度对抗补丁的漏检情况。与现有算法比较, 所提算法有效提高了神经网络在检测对抗样本时的准确率。此外, 所提算法不涉及下游模型结构的修改, 具有良好的兼容性。

关键词: 深度学习安全; 对抗攻击与防御; 对抗补丁; 目标检测

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024009

Adversarial patch defense algorithm based on PatchTracker

XIAO Zhenjie¹, HUANG Shiyu¹, YE Feng^{1,2}, HUANG Liqing^{1,2}, HUANG Tianqiang^{1,2}

1. College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China

2. Digital Fujian Institute of Big Data Security Technology, Fuzhou 350117, China

Abstract: The application of deep neural networks in target detection has been widely adopted in various fields. However, the introduction of adversarial patch attacks, which add local perturbations to images to mislead deep neural networks, poses a significant threat to target detection systems based on vision techniques. To tackle this issue,

收稿日期: 2023-09-06; 修回日期: 2023-12-16

通信作者: 叶锋, yefeng@fjnu.edu.cn

基金项目: 国家自然科学基金 (62072106); 福建省创新战略研究计划项目 (2023R0156)

Foundation Items: The National Natural Science Foundation of China (62072106), Fujian Innovation Strategy Research Program Project(2023R0156)

引用格式: 肖镇杰, 黄诗瑀, 叶锋, 等. 基于 PatchTracker 的对抗补丁防御算法[J]. 网络与信息安全学报, 2024, 10(1): 169-180.

Citation Format: XIAO Z J, HUANG S Y, YE F, et al. Adversarial patch defense algorithm based on PatchTracker[J]. Chinese Journal of Network and Information Security, 2024, 10(1): 169-180.

an adversarial patch defense algorithm based on PatchTracker was proposed, leveraging the semantic differences between adversarial patches and image backgrounds. This algorithm comprised an upstream patch detector and a downstream data enhancement module. The upstream patch detector employed a YOLOV5 (you only look once-v5) model with attention mechanism to determine the locations of adversarial patches, thereby improving the detection accuracy of small-scale adversarial patches. Subsequently, the detected regions were covered with appropriate pixel values to remove the adversarial patches. This module effectively reduced the impact of adversarial examples without relying on extensive training data. The downstream data enhancement module enhanced the robustness of the target detector by modifying the model training paradigm. Finally, the image with removed patches was input into the downstream YOLOV5 target detection model, which had been enhanced through data augmentation. Cross-validation was performed on the public TT100K traffic sign dataset. Experimental results demonstrated that the proposed algorithm effectively defended against various types of generic adversarial patch attacks when compared to situations without defense measures. The algorithm improves the mean average precision (mAP) by approximately 65% when detecting adversarial patch images, effectively reducing the false negative rate of small-scale adversarial patches. Moreover, compared to existing algorithms, this approach significantly enhances the accuracy of neural networks in detecting adversarial samples. Additionally, the method exhibited excellent compatibility as it does not require modification of the downstream model structure.

Keywords: deep learning security, adversarial attack and defense, adversarial patch, object detection

0 引言

深度学习技术在目标检测领域取得了显著的进展,为各种计算机视觉智能应用提供了强大的支持,如图像分类、人脸识别以及自动驾驶。然而, Szegedy 等^[1]和 Goodfellow 等^[2]研究表明,神经网络存在缺陷,修改输入图像的部分像素值可导致神经网络分类错误,这些被修改后的图像称为对抗样本。对抗补丁是一种常见的对抗样本攻击方式,它通过在图像中添加局部扰动来误导神经网络。对抗补丁不仅可以攻击图像分类模型,还可以攻击目标检测模型。例如,在自动驾驶场景中,对抗补丁可以粘贴在交通标志上或者放置在其周围,从而误导自动驾驶车辆的目标检测系统,使其误判交通标志或无法检测到行人,从而引发严重的交通事故。目标检测技术的可靠性和安全性面临着新的挑战。

为了应对不同领域的对抗补丁攻击,研究人员提出了许多对抗补丁防御方法。这些方法大致可以分为数据防御和模型防御两类。数据防御不涉及具体模型结构的修改,而是先确定对抗补丁的位置,再破坏补丁的像素结构信息。模型防御则通过修改模型结构或对抗性训练提高模型的鲁棒性。然而,现有的数据防御方法存在定位准确

性低、处理速度慢、边缘分割粗糙等问题;现有的模型防御方法存在兼容性差、训练周期长等问题。因此,本文利用对抗补丁和真实图像之间的语义差异性,提出了一种基于 PatchTracker 的对抗补丁防御算法。该算法包括两个主要部分:上游补丁检测器和下游数据增强模块。训练上游补丁检测器时,使用带有对抗补丁的图像作为训练集,为了增强补丁的检测效果,在上游补丁检测器的特征提取网络中结合了基于双层路由注意力机制的 Transformer BiFormer (bi-level routing attention Transformer) 模块^[3]。在上游补丁检测器确定对抗补丁的位置后,使用灰色像素将对抗补丁抹除。训练下游数据增强模块时使用 cutout 数据增强方法^[4],抹除补丁后的交通标志图像会丢失部分语义信息从而导致模块性能下降,所以将经过 cutout 后的图像放入数据集内训练,能让模块充分学习残缺图像的模式以提高模型的检测精度。本文的贡献如下。

1) 设计了一种基于 PatchTracker 的对抗补丁防御算法 PatchTracker,该算法高效且通用性强,主要用于防御自动驾驶场景中的对抗补丁攻击。

2) 在上游补丁检测器的特征提取网络中加入 BiFormer 模块,提高对抗补丁的检测精度。

3) 使用数据增强策略训练下游 YOLOV5 模型, 提高交通标志检测精度。

4) 将对抗补丁打印并粘贴在物理场景的交通路标上, 评估 PatchTracker 防御的有效性。

1 相关工作

1.1 对抗攻击

在介绍对抗补丁前, 先概述常用的对抗样本生成方法。Szegedy 等^[1]发现了对抗样本的存在。在原始干净图像中添加一些人眼无法察觉的噪声, 能够导致模型将图片分类错误。Goodfellow 等^[2]和 Kurakin 等^[5]分别提出了用于生成对抗样本的快速梯度符号方法 (FGSM, fast gradient sign method) 及其迭代版本 I-FGSM (iterative-FGSM)。Madry 等^[6]基于 I-FGSM, 提出了投影梯度下降 (PGD, project gradient descent) 攻击方法, 随机初始化对抗样本并将其裁剪到指定范围。Dong 等^[7]将动量思想引入对抗样本迭代过程, 提出了基于动量的 MI-FGSM (momentum I-FGSM)。

早期工作中, 上述方法主要生成分布于全图的数字噪声, 用于在数字域中的模型攻击。而 Brown 等^[8]修改图像特定区域内像素, 提出了对抗补丁攻击。基于此研究, 后续出现了更多针对图像分类器的对抗补丁攻击。一些研究侧重于增强对抗补丁的攻击性能, Karmon 等^[9]提出了 LaVAN (localized and visible adversarial noise) 攻击方法, 该方法生成的补丁能够在不覆盖图像主要内容的前提下实现有目标地误导神经网络模型分类; Eykholt 等^[10]考虑了现实场景中的视角和光照问题, 将其作为补丁优化的依据, 提高了对抗补丁物理攻击的鲁棒性; Yang 等^[11]利用分类模型的中间层特征图生成特定类别的对抗性纹理贴图, 以增强有目标攻击效果。另一些研究注重对抗补丁的反检测能力, Subramanya 等^[12]通过误导模型注意力, 使得对抗补丁逃避基于模型显著图 Grad-CAM^[13]定位的防御方法; Liu 等^[14]增强了对抗补丁和背景图像的上下文相关性, 使得补丁的视觉效果更加自然。

对抗补丁不仅能够攻击图像分类模型, 而且会攻击目标检测模型。Liu 等^[14]基于 Brown 等^[8]的工作, 修改对抗攻击的优化函数, 使得对抗补丁

可以对目标检测模型执行无目标和有目标攻击, 但是他们没有考虑现实场景的攻击。后来的工作提高了对抗补丁在攻击目标检测器上的性能。Lee 等^[15]利用 PGD 方法^[6]实现了物理场景中的无目标攻击; Wu 等^[16]和 Huang 等^[17]分别提出了 DPAttack 和 RPAttack 以减小对抗补丁的面积; Zhao 等^[18]利用模型显著图 Grad-CAM^[13]优化对抗补丁位置。此外, 还有一些工作将对抗补丁用于攻击特定的视觉应用。Sharif 等^[19]和 Yang 等^[20]均针对人脸检测系统展开攻击, 前者将对抗纹理应用在眼镜框架上, 后者则将对抗补丁贴于额头处; 而 Thys 等^[21]将对抗补丁进行了像素平滑以及可打印优化, 并用于攻击人体检测器; Hu 等^[22]利用风格迁移美化了粘贴在衣服上的对抗补丁, 使其看起来更加自然; Huang 等^[23]则使用数据增强和模型自集成策略, 以提高对抗补丁在黑盒人体检测模型上的迁移性。

1.2 对抗防御

针对对抗补丁的防御大致可以分为数据防御和模型防御。关于数据防御, Nasser 等^[24]提出局部梯度平滑 (LGS, local gradient smoothing), 他们观察到补丁区域的像素变化频繁, 是图像中的高频区域, 且梯度信息明显, 因此根据图像梯度确定了对抗补丁的位置并将其进行像素平滑。Hayes^[25]观察到, 图像分类受到对抗补丁的影响程度很高, 导致对抗像素在注意力图上具有强激活特性, 因此利用模型显著图 Grad-CAM^[13]确定对抗补丁所在区域并将其抹除。然而该方法可能定位到不包含补丁信息的良性区域, 这些区域同样具有高响应度。为了解决上述问题, Chou 等^[26]提出 SentiNet 架构, 根据 Grad-CAM 分割出具有高响应的子图像, 再将分割的子图像粘贴在干净图像上, 根据干净图像分类时受到的影响来判断所粘贴的部分是否为对抗补丁。Xiang 等^[27]提出 PatchCleanser 方法, 使用滑动窗口和投票分类的方式确定对抗补丁的位置。Liu 等^[28]使用语义分割方法定位对抗补丁, 并将其从图像中抹除。关于模型防御, Rao 等^[29]使用对抗训练方法, 在 Cifar-10 和 GTSRB 数据集上成功防御对抗补丁攻击。Ji 等^[30]将对抗性补丁作为一种类别进行标注, 从而实现目标检测的防御。Zhang

等^[31]对 BagNet 模型的 logit 输出异常值进行裁剪, 以限制对抗样本的攻击性。Xiang 等^[32]则裁剪 BagNet 模型的异常特征值以实现防御对抗补丁。

2 方法

2.1 目标检测基本原理

本文在攻击和防御策略中均采用 YOLO 单阶段目标检测。YOLO 通过卷积神经网络获取输入图像的特征图, 随后将输入图像划分为 $S \times S$ 的网格, 每个网格使用 B 个边界框检测对象。如果待检测对象的中心处于某个网格内, 那么该网格中的边界框就负责检测这个对象。之后使用 NMS (non maximum suppression) 算法来筛选出具有高置信度的边界框, 消除冗余的检测结果, 从而得到最终的目标检测结果。

2.2 对抗补丁攻击

2.2.1 攻击原理

对抗补丁作为一种攻击手段, 通过把特定补丁添加到输入图像中, 改变深度学习模型的决策过程, 从而误导模型的分类。相较于对抗噪声, 对抗补丁具有更好的物理场景攻击性能, 其通过将像素修改限制在图像特定区域内实现, 如式(1)所示。

$$\mathbf{x}' = (1 - \mathbf{M}) \otimes \mathbf{x} + \mathbf{M} \otimes \boldsymbol{\delta} \quad (1)$$

其中, $\mathbf{x}'$ 表示对抗补丁图像, $\mathbf{M}$ 为控制补丁位置和大小掩码, $\mathbf{x}$ 为原始图像, 符号 $\otimes$ 表示元素乘积, $\boldsymbol{\delta}$ 为修改的像素。接下来定义对抗补丁攻击的优化问题, 本质上是最小化对抗补丁训练时的损失函数, 如式(2)所示。

$$\arg \min_{\mathbf{P}} \sum_i^N L(\mathbf{P}, F, \mathbf{x}_i) \quad (2)$$

其中, $\mathbf{P}$ 表示对抗补丁, F 表示被攻击的模型, $\mathbf{x}_i$ 表示输入图像, N 为训练样本的数量, L 是补丁优化过程定义的损失函数。

2.2.2 攻击方法

本文采用 Huang 等^[23]提出的通用攻击框架 T-SEA, 将 YOLO 目标检测器作为白盒模型进行对抗补丁的生成, 并降低目标检测器在交通标志识别精度。接下来介绍本文使用的 3 种对抗攻击方法。

(1) 基于 Adam (adaptive moment estimation) 优化器的攻击方法

优化补丁的损失函数由两项组成。第一项是模型分数损失 L_{score} , 被攻击的模型输出分数越高, 说明对抗补丁的攻击性能越弱, 函数如式(3)所示。

$$L_{\text{score}} = \frac{1}{N} \sum_{i=1}^N p_i c_i \quad (3)$$

其中, N 是图像数量, p_i 是置信度分数, c_i 是类别分数。第二项是总方差损失 L_{tv} , 用于约束对抗补丁像素之间的平滑性, 如式(4)所示。

$$L_{\text{tv}} = \frac{1}{MN} \sum_{i=1}^{M-1} \sum_{j=1}^{N-1} \sqrt{(P_{i,j} - P_{i+1,j})^2} + \sqrt{(P_{i,j} - P_{i,j+1})^2} \quad (4)$$

其中, $P_{i,j}$ 表示对抗补丁区域内位置 (i, j) 的像素值。之后将上述两项损失做加权和, 得到总损失函数, 如式(5)所示。

$$L = \lambda_1 L_{\text{score}} + \lambda_2 L_{\text{tv}} \quad (5)$$

其中, λ_1 和 λ_2 是平衡参数。接着使用 Adam 优化器^[33]最小化对抗补丁的损失函数。Adam 是一种基于梯度下降方法的自适应学习率优化器, 根据训练过程中每个参数的历史梯度和更新情况自适应地调整每个参数的学习率。其计算规则如式(6)所示, 其中 t 是迭代轮数, F 是被攻击模型, $\mathbf{P}_t$ 是第 t 轮的对抗补丁, $\mathbf{x}$ 是被攻击的图像, $\mathbf{g}_t$ 是第 t 轮的梯度, $\mathbf{m}_t$ 和 $\mathbf{v}_t$ 是第 t 轮的梯度在动量形式下的一阶和二阶矩估计, β_1 和 β_2 是衰减系数, $\mathbf{m}'_t$ 和 $\mathbf{v}'_t$ 是第 t 轮中偏差纠正后的一阶和二阶矩估计, η 是学习率, ε 是防止除零错误的小值。

$$\begin{cases} \mathbf{g}_t = \nabla_{\mathbf{x}} L(\mathbf{P}_t, F, \mathbf{x}) \\ \mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t \\ \mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2 \\ \mathbf{m}'_t = \frac{\mathbf{m}_t}{1 - \beta_1^t} \\ \mathbf{v}'_t = \frac{\mathbf{v}_t}{1 - \beta_2^t} \\ \mathbf{P}_{t+1} = \mathbf{P}_t - \frac{\eta \mathbf{m}'_t}{\sqrt{\mathbf{v}'_t} + \varepsilon} \end{cases} \quad (6)$$

(2) 基于 PGD 的攻击方法

PGD 攻击先随机初始化对抗补丁中的像素

值，再利用损失函数计算梯度并对补丁像素进行优化。损失函数使用均方误差 (MSE, mean square error) 计算置信度分数和 0 之间的差异，如式(7)所示。

$$L = \frac{1}{N} \sum_{i=1}^N p_i^2 \quad (7)$$

其中， N 是图像数目， p_i 是置信度分数。接着根据该损失函数求梯度并更新对抗补丁，更新规则如式(8)所示。

$$P_{t+1} = \text{clip}(P_t - \alpha \text{sign}(\nabla_x L(P_t, F, x))) \quad (8)$$

其中， α 是学习率，clip 操作对迭代结果进行投影，确保对抗补丁像素在合理的图像空间内。

(3) 基于动量迭代方法 (MIM, momentum iterative method) 的攻击方法

基于 MIM 的攻击方法将动量整合到迭代过程，以增强攻击的稳定性。本文中 MIM 和 PGD 攻击的损失函数一致，也使用均方误差将被攻击模型置信度作为补丁优化的损失。随后根据损失函数求梯度并使用梯度计算动量，最后使用动量更新对抗补丁，如式(9)所示。

$$\begin{cases} g_{t+1} = \mu g_t + \frac{\nabla_x L(P_t, F, x)}{\|\nabla_x L(P_t, F, x)\|} \\ P_{t+1} = \text{clip}(P_t - \alpha \text{sign}(g_{t+1})) \end{cases} \quad (9)$$

其中， g_t 表示 t 时刻的动量， μ 表示用于控制历史动量权重的平衡因子。

2.3 PatchTracker 防御流程

PatchTracker 防御框架如图 1 所示，整个防御

流程可分为上游补丁检测器和下游数据增强模块。给定一幅对抗样本图像，首先利用上游补丁检测器检测图像中的对抗区域，并对检测到的对抗区域使用灰色像素进行替换，然后通过数据增强模块改进下游模型训练范式，使用下游交通标志检测器识别处理之后的图像，输出交通标志的正确位置和类别。

2.3.1 上游补丁检测器

传统的对抗样本检测方法通常依赖于图像分割器来定位和识别对抗补丁，但这些方法在面对复杂场景和微小对抗补丁时往往效果不佳。而本文选择 YOLOV5 的创新性在于其高效的实时检测能力，以及对小尺寸对抗补丁的敏感性。YOLOV5 检测速度快且定位准确，故选择其作为上游补丁检测器，并使用带有对抗补丁的交通标志数据集，训练其补丁检测能力。检测到对抗补丁之后，再将补丁定位框内的 RGB 像素值设为 (128,128,128)，以实现抹除对抗补丁的效果。

然而，在现有的一些对抗补丁攻击中，将大补丁拆分成小补丁粘贴在图像各处，这些对抗补丁的尺寸较小，可以成功逃避检测算法的定位而攻击成功。为了解决上述问题，本文在 YOLOV5 的浅层卷积中插入 BiFormer 模块^[3]，以增强网络的小目标特征提取能力。

卷积的特征提取是从细节到全局，而浅层卷积提取的特征图依然会保留图像细节信息，此时可利用 BiFormer 进行细节注意力计算。首先，BiFormer 将图像分割成不同的粗粒度块，随后计

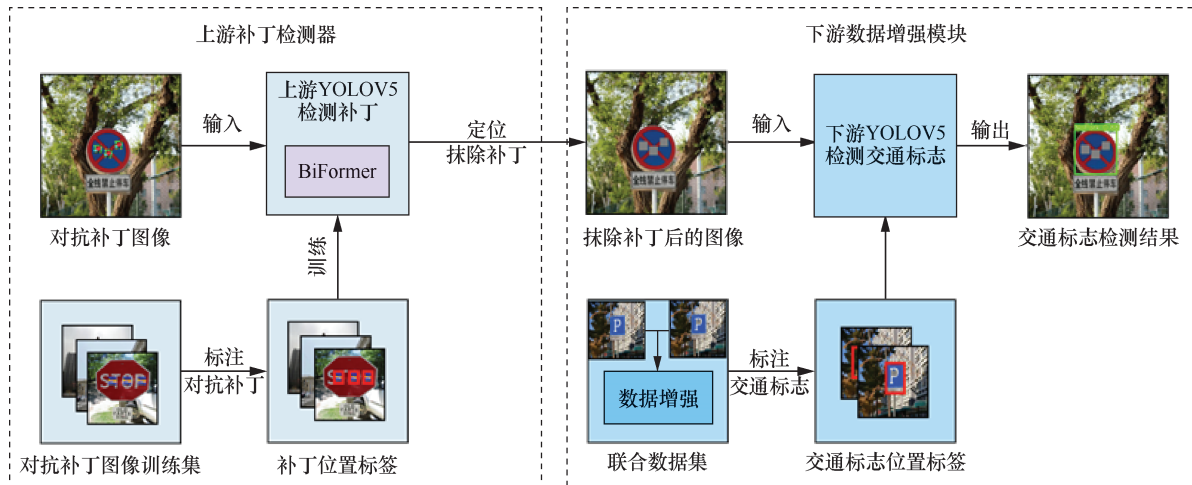


图1 PatchTracker 防御框架
Figure 1 The framework of PatchTracker defense

算这些块之间的相似度,如果某些块与其他块之间的联系较小,那么这些联系将被过滤,之后得到若干组具有联系的粗粒度块。接着,在一个组内,将每个粗粒度块划分为更小的细粒度块,并计算这些细粒度块之间的联系,得到细粒度的注意力。具体而言,粗粒度级别的注意力有利于确定交通标志在图像中的位置,而细粒度级别的注意力则用于确定对抗补丁在交通标志上的位置。BiFormer 的双层路由注意力如图 2 所示。相比于卷积的特征提取方式,该策略的优势是能够在关键区域中保留细节信息,因此有利于提升小面积的对抗补丁检测性能。

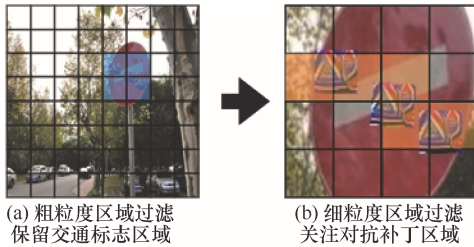


图 2 BiFormer 的双层路由注意力
Figure 2 The Bi-level routing attention of BiFormer

2.3.2 下游数据增强模块

经过上游抹除处理的残缺图像失去了对抗补丁的语义,因此可以减轻对下游模型的干扰。然而这些图像缺失了一部分交通标志的语义信息,不利于下游模型检测交通标志。为了解决上述问题,本文使用 cutout 数据增强策略,该策略在下游检测器训练阶段,对训练集图像的部分区域使用固定大小的正方形区域进行遮掩,让下游模型学习更一般的交通标志特征,减少过拟合程度并提高下游模型的检测性能,同时提高模型在处理残缺图像时的性能。模型训练时的优化问题可由如式(10)表示。

$$\arg \min_{\theta} L(F(X, \theta), y) \quad (10)$$

其中, F 是模型, θ 是模型的学习参数, X 是输入图像集合,包括完整图像和残缺图像, y 是正确标签, L 是模型的损失函数。

3 实验

3.1 实验设置

(1) 硬件和软件环境

本文实验所使用的处理器型号为 Intel

Xeon Gold 5218 CPU (2.30 GHz), 而 GPU 为 NVIDIA Tesla V100S 32 GB; 实验在配置了 Python 3.7 以及 PyTorch 1.13 的 Ubuntu 20.04.2 操作系统的服务器,本文基于 YOLOV5 的实验皆为 7.0 版本。

(2) 数据集

本文实验使用公开的 TT100K 交通标志检测数据集评估模型性能。TT100K 交通标志检测数据集是由清华大学与腾讯的联合实验室整理并公布的大型交通标志检测数据集,所有图像采集于现实生活场景,覆盖了城市、高速公路等不同场景。TT100K 交通标志检测数据集包含 6 103 幅训练图像和 3 071 幅测试图像,每幅图像大小是 2 048×2 048 像素,选择其中交通标志个数超过 500 的 6 个类别进行实验评估。除此之外,本文还从互联网上收集整理了 9 类交通标志图像,各类交通标志图像的数量分别为 500 幅。本文将它们命名为 TS4500 数据集,数据集包含不同角度、光照的图像,旨在模拟实际场景下的交通标志检测。在该数据集中,每幅图像都经过了人工标注,图片分辨率均调整为 416×416 像素。两个数据集内的部分图像展示如图 3 所示。

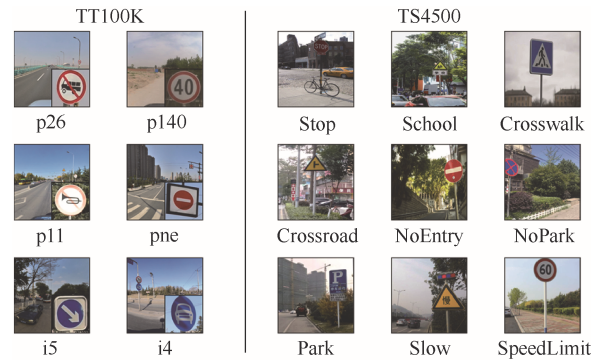


图 3 数据集内的部分图像展示
Figure 3 Display of part images in datasets

(3) 模型和评价指标

本文使用两个 YOLOV5 目标检测模型,用于上游对抗补丁检测和下游交通标志检测。采用目标检测的特定类别平均精度 (AP, average precision) 和所有类别平均精度 (mAP, mean average precision) 作为对抗补丁攻击和防御有效性的评价指标。AP 是基于特定类别的精确率 (P, precision) 与召回率 (R, recall) 曲线下的面积和。

精确率 (P) 主要衡量模型对交通标志误检的程度, 召回率 (R) 主要衡量模型对交通标志漏检的程度。计算公式如式(11)所示。

$$\begin{cases} P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN} \end{cases} \quad (11)$$

其中, TP 是模型预测正确的目标, FP 是模型预测错误的目标, FN 是模型未能检测出的实际存在的目标。而 mAP 是所有类别 AP 的平均值, 计算公式如式(12)所示。

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (12)$$

其中, N 表示类别数。

3.2 攻击实验

本文首先使用 TT100K 和 TS4500 训练集, 训练 YOLOV5 下游交通标志检测模型。模型的交并比阈值设置为 0.45, 置信度阈值设置为 0.5。训练过程中采用随机梯度下降 (SGD, stochastic gradient descent) 执行 300 轮迭代。模型训练完成之后, 在 TT100K 和 TS4500 测试集上得到无攻击情况下的各类 AP 和 mAP 。

利用上述训练完成的模型, 本文使用 3 种方法生成对抗补丁, 分别是基于 Adam 优化器的攻击方法 (简称为 Adam)^[33]、基于 PGD 的攻击方法 (简称为 PGD)^[6] 以及基于 MIM 的攻击方法 (简称为 MIM)^[7]。在 TT100K 数据集上生成对抗补丁时, 选用 Adam 和 MIM, 而在 TS4500 数据集上则选用 Adam 和 PGD。补丁生成过程中使用的每个批次大小为 16, 在每个批量中的攻击步数为 3, 更新步长为 0.03, 采用 Adam 作为优化器, 总方差损失权重为 2.5。攻击方法根据检测模型预测框的位置, 确定图像中的交通标志并进行补丁粘贴, 对抗补丁的原始尺寸为 300×300 像素, 而粘贴到实际图像上时, 将根据目标大小进行缩放, 在 TT100K 数据集中, 补丁的大小为检测框面积的 20%, 一幅图像上只粘贴 1 块补丁; 而在 TS4500 数据集上, 补丁的大小为检测框面积的 13%, 一幅图像上粘贴 3 块补丁。

将训练完成的对抗补丁粘贴至 TT100K 和 TS4500 测试集的图像中构成对抗测试集, 再将对

抗测试集中的图像输入 YOLOV5 交通标志检测模型, 以测试对抗补丁的攻击性能, 检测结果对比分别如表 1 和表 2 所示。在无攻击的情况下, YOLOV5 模型对交通标志的检测准确性很高, 各类 AP 和 mAP 均在 89% 以上。然而在对抗补丁攻击场景下, 各类 AP 和 mAP 均显著下降, 证明了对抗补丁攻击 YOLOV5 交通标志检测模型的有效性。

表 1 攻击方法在 TT100K 上检测结果对比 (AP)
Table 1 Detection results comparison of attack methods on TT100K (AP)

图像类别	攻击方法		
	无攻击	Adam	MIM
p26	94.22%	33.04%	29.00%
pl40	97.91%	4.79%	11.17%
p11	97.06%	18.31%	24.14%
pne	97.64%	25.35%	12.42%
i5	98.53%	11.95%	23.06%
i4	98.27%	12.75%	20.38%
mAP	97.27%	17.70%	20.03%

表 2 攻击方法在 TS4500 上检测结果对比 (AP)
Table 2 Detection results comparison of attack methods on TS4500 (AP)

图像类别	攻击方法		
	无攻击	Adam	PGD
Stop	96.56%	40.55%	48.72%
School	91.95%	0	0.18%
Crosswalk	94.59%	9.90%	13.74%
Crossroad	97.50%	6.12%	6.74%
NoEntry	97.53%	13.00%	25.50%
NoPark	89.46%	19.52%	21.31%
Park	96.23%	12.69%	12.73%
Slow	98.70%	15.06%	21.91%
SpeedLimit	95.76%	4.81%	10.67%
mAP	95.36%	13.52%	17.94%

3.3 防御实验

在数据防御中, 本文将 3.2 节中所训练的对抗补丁粘贴在 TT100K 和 TS4500 训练集图像上, 并用这些图像训练 PatchTracker 中的 YOLOV5 上游补丁检测器, 在其中加入了 BiFormer 注意力层以提高补丁检测能力; 接着训练下游交通标志检测模型, 使用 cutout 对其数据增强以提高交通标志检测能力。

将带有对抗补丁的 TT100K 和 TS4500 测试集输入 PatchTracker 防御框架, PatchTracker 先用上游补丁检测器定位对抗补丁, 依据补丁位置掩码信息并用灰色像素将其抹除, 接着将抹除处理的图像输入下游交通标志检测模型, 并计算各类 AP 和 mAP 来评估防御性能, 如表 3 和表 4 所示。

表 3 PatchTracker 在 TT100K 上防御性能 (AP)
Table 3 Defence performance of PatchTracker on TT100K (AP)

图像类别	补丁生成方法	
	Adam	MIM
p26	85.30%	92.42%
pl40	82.41%	84.88%
p11	74.67%	78.72%
pne	88.81%	87.49%
i5	88.22%	88.26%
i4	88.93%	88.94%
mAP	84.72%	86.79%

表 4 PatchTracker 在 TS4500 上防御性能 (AP)
Table 4 Defence performance of PatchTracker on TS4500 (AP)

图像类别	补丁生成方法	
	Adam	PGD
Stop	89.94%	90.93%
School	56.40%	55.67%
Crosswalk	82.28%	82.20%
Crossroad	76.25%	86.02%
NoEntry	89.50%	91.00%
NoPark	79.75%	76.62%
Park	94.37%	96.39%
Slow	82.64%	81.17%
SpeedLimit	84.44%	87.32%
mAP	81.73%	83.04%

为了方便比较无攻击、防御前和防御后的效果, 本文将表 1 至表 4 的 mAP 绘制成柱状图, 如图 4 所示, 同时展示了部分防御前和防御后的图像, 如图 5 所示。结果表明, 与未采取防御措施的情况相比, mAP 显著提高, 这是因为 PatchTracker 的上游模型有效检测到图像中的对抗补丁, 并依据位置信息抹除了对抗补丁的干扰; 而下游模型经过 cutout 训练, 能够有效检测抹除补

丁后的残缺交通标志图像。

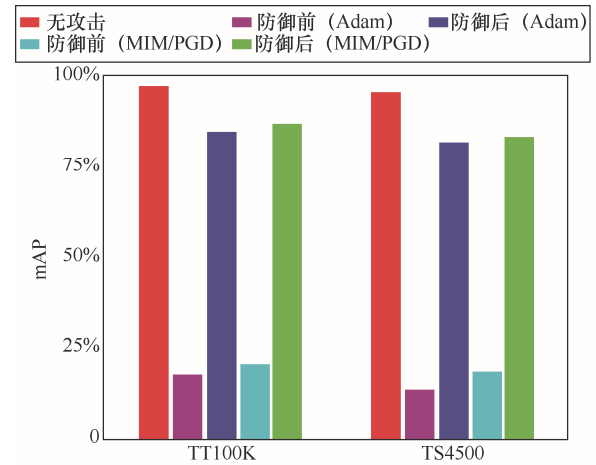


图 4 模型在防御前和防御后的 mAP 对比
Figure 4 Compare of model mAP before and after defense



图 5 防御前后图像展示
Figure 5 Display of images before and after defense

3.4 对比实验

在 TT100K 和 TS4500 的对抗补丁测试集上, PatchTracker 分别与 LGS^[24]、JPEG 压缩^[33,34]、Jedi^[35]、SAC^[36]4 种基准防御方法进行对比。LGS 和 JPEG 压缩都是基于图像预处理的数据防御方法, Jedi 和 SAC 则通过检测和去除对抗补丁来保护目标检测器。5 种方法的检测结果对比如表 5 和表 6 所示。相比无防御情况, LGS 和 JPEG 压缩的防御效果甚微, 在一些情况下防御后的各类 AP 和 mAP 甚至比无防御情况更低。因为 JPEG

表 5 不同防御方法在 TT100K 上检测结果对比 (AP)
Table 5 Comparison of detection results of different defence methods on TT100K (AP)

图像识别	Adam						MIM					
	无防御	LGS	JPEG 压缩	Jedi	SAC	PatchTracker	无防御	LGS	JPEG 压缩	Jedi	SAC	PatchTracker
p26	33.04%	27.32%	34.01%	24.63%	54.41%	85.30%	29.0%	26.50%	26.50%	23.34%	51.82%	92.42%
pl40	4.79%	6.91%	7.98%	42.22%	60.89%	82.41%	11.17%	13.83%	13.30%	44.64%	57.61%	84.88%
p11	18.31%	18.87%	19.32%	37.82%	52.36%	74.67%	24.14%	21.24%	21.40%	37.41%	52.74%	78.72%
pne	25.35%	22.07%	22.54%	62.73%	74.11%	88.81%	12.42%	15.29%	14.82%	62.02%	73.97%	87.49%
i5	11.95%	15.36%	15.84%	56.38%	74.37%	88.22%	23.06%	23.06%	23.06%	56.39%	73.58%	88.26%
i4	12.75%	14.2%	13.74%	29.61%	57.17%	88.93%	20.38%	14.13%	23.70%	31.65%	58.16%	88.94%
mAP	17.70%	17.46%	18.91%	42.23%	62.22%	84.72%	20.03%	19.01%	20.46%	42.58%	61.31%	86.79%

表 6 不同防御方法在 TS4500 上检测结果对比 (AP)
Table 6 Comparison of detection results of different defence methods on TS4500 (AP)

图像类别	Adam						PGD					
	无防御	LGS	JPEG	Jedi	SAC	PatchTracker	无防御	LGS	JPEG	Jedi	SAC	PatchTracker
Stop	40.55%	49.19%	43.59%	70.42%	76.81%	89.94%	48.72%	48.71%	48.48%	63.56%	77.91%	90.93%
School	0.00%	2.84%	0.00%	9.71%	20.54%	56.40%	0.18%	0.90%	0.08%	5.15%	21.18%	55.67%
Crosswalk	9.90%	23.05%	10.34%	42.08%	70.59%	82.28%	13.74%	18.76%	11.25%	22.38%	74.51%	82.20%
Crossroad	6.12%	10.22%	6.55%	27.13%	45.53%	76.25%	6.74%	7.15%	5.97%	26.23%	36.68%	86.02%
NoEntry	13.00%	52.00%	15.00%	55.50%	89.99%	89.50%	25.50%	45.00%	28.50%	45.92%	92.99%	91.00%
NoPark	19.52%	57.14%	20.77%	41.88%	41.16%	79.75%	21.31%	64.24%	24.03%	39.71%	39.49%	76.62%
Park	12.69%	13.29%	11.44%	32.84%	73.63%	94.37%	12.73%	10.25%	12.18%	42.27%	67.66%	96.39%
Slow	15.06%	35.29%	17.09%	36.43%	65.97%	82.64%	21.91%	23.14%	22.42%	27.06%	52.01%	81.17%
SpeedLimit	4.81%	3.95%	5.26%	40.03%	84.72%	84.44%	10.67%	9.89%	10.65%	26.50%	87.02%	87.32%
mAP	13.52%	27.44%	14.45%	39.56%	63.22%	81.73%	17.94%	25.34%	18.17%	33.19%	61.05%	83.04%

压缩仅轻微地改变了图像的部分像素，而没有集中破坏对抗补丁区域的像素信息；而 LGS 是根据图像边缘梯度信息进行对抗补丁定位并平滑像素的，部分对抗补丁的边缘梯度信息较小从而导致该方法定位不准确，进而影响防御结果。

SAC 和 Jedi 作为先进的防御方法，都展现出了较 LGS 和 JPEG 压缩更加显著的防御性能。从表中的数据来看，尽管 SAC 和 Jedi 均能有效提升 mAP，但它们未能达到 PatchTracker 的水平，由于 Jedi 依赖熵的统计特性来定位补丁，面对复杂场景中的自然图像变化可能会产生高熵值，这会与对抗补丁的高熵特性相混淆，从而限制了其在防御性能上的表现。SAC 虽然能够移除对抗补丁，但此过程中会造成关键的语义信息缺失，进而降低检测器性能。在 PatchTracker 防御中，各类 AP

和 mAP 均显著提升，并且在大部分情况下超过 80%，是因为 PatchTracker 通过精确定位对抗补丁，再抹除补丁的像素语义信息，使其失去对抗性。这些结果验证了 PatchTracker 在处理对抗补丁图像上的有效性。

3.5 消融实验

PatchTracker 中的上游模型结合了 BiFormer 模块，而训练下游模型时使用了 cutout 数据增强，本节将评估这两项改进对于防御的贡献，评估指标为 mAP。实验中设计了两个组件的 4 种组合：① 上游模型无注意力机制，下游模型无数据增强；② 上游模型无注意力机制，下游模型有数据增强；③ 上游模型有注意力机制，下游模型无数据增强；④ 上游模型有注意力机制，下游模型有数据增强。消融实验结果如表 7 和表 8 所示。从

①vs.②和③vs.④两组对比可以看出,上游模型结合注意力机制后,能够略微提升防御后的 mAP,是因为注意力机制有效提高了上游模型检测对抗补丁时性能,减少了模型的漏检和误检情况。从①vs.③和②vs.④两组对比可以看出,在下游模型训练中使用 cutout 数据增强方法后, mAP 值提升较大,是因为该方法有效地扩充了模型的训练数据,增加了模型对输入数据的多样性理解,使得模型在图像缺失部分交通标志语义信息时,也能够准确定位交通标志,增强了模型的鲁棒性。

表 7 在 TT100K 测试集上的消融实验结果
Table 7 Results of ablation study on test TT100K

组合序号	设置		结果 (mAP)	
	BiFormer	cutout	Adam	MIM
①	×	×	59.18	63.57
②	√	×	60.73	64.49
③	×	√	84.28	86.16
④	√	√	84.72	86.79

表 8 在 TS4500 测试集上的消融实验结果
Table 8 Results of ablation study on test TS4500

组合序号	设置		结果 (mAP)	
	BiFormer	cutout	Adam	PGD
①	×	×	63.65	68.72
②	√	×	67.58	69.21
③	×	√	79.45	82.98
④	√	√	81.73	83.04

3.6 物理场景实验

本文还将 3 种攻击方法 (Adam、PGD、MIM) 生成的补丁打印并粘贴在现实场景的交通标志上,以此评估 PatchTracker 的物理场景防御性能,结果如图 6 所示,补丁攻击能让检测器无法识别标志,而在防御之后,YOLOV5 能够重新识别图中的交通标志,说明了 PatchTracker 的有效性。

4 结束语

对抗补丁严重破坏了各种视觉应用。针对这一挑战,本文提出了一种基于 PatchTracker 的对抗补丁防御算法。首先在上游补丁检测器中加入 BiFormer 模块,以便检测和抹除图像中的对抗补

丁。对抗补丁的抹除导致图像部分语义信息丢失,网络模型的预测能力有所降低,因此使用数据增强模块 cutout 训练更加鲁棒的下游交通标志检测器。实验结果证明,本文算法在不更改下游模型结构的情况下,可以在数字域和物理域中有效地处理对抗补丁并提高模型鲁棒性,而消融实验表明了 BiFormer 和 cutout 数据增强的有效性。本文算法在不更改下游模型结构的前提下,通过结合数据防御和模型防御策略,具有良好的兼容性,为对抗补丁防御提供了一个新的防御框架,同时为实际应用中的视觉系统提供了更强的安全性和可靠性。此外,本文算法中,输入下游检测器的是抹除补丁后的残缺图像,即使下游模型经过数据增强训练,也会造成一定影响。因此在未来的研究中,将使用图像修复技术恢复抹除区域的交通标志语义信息,进一步提升检测器的准确率。

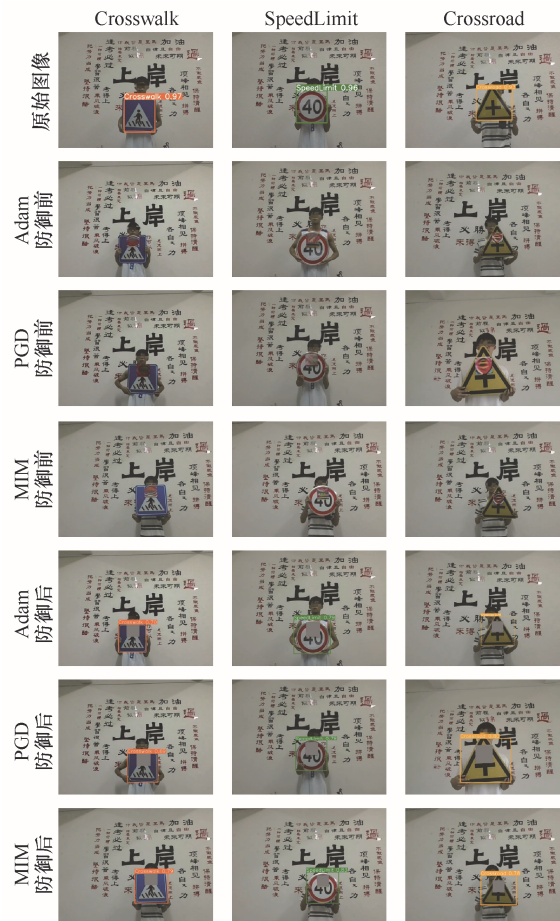


图 6 物理场景防御结果展示
Figure 6 The result of physical scene defense

参考文献:

- [1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of International Conference on Learning Representations (ICLR). 2014.
- [2] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of International Conference on Learning Representations (ICLR). 2015.
- [3] ZHU L, WANG X, KE Z, et al. BiFormer: vision transformer with bi-level routing attention[C]//Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 10323-10333.
- [4] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout[J]. arXiv preprint arXiv: 1708.04552, 2017.
- [5] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[C]//Proceedings of International Conference on Learning Representations (ICLR). 2017.
- [6] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of International Conference on Learning Representations (ICLR). 2018.
- [7] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 9185-9193.
- [8] BROWN T B, MANÉ D, ROY A, et al. Adversarial patch[J]. arXiv preprint arXiv:1712.09665, 2017.
- [9] KARMON D, ZORAN D, GOLDBERG Y. Lavan: Localized and visible adversarial noise[C]//Proceedings of International Conference on Machine Learning (ICML). 2018: 3903-3911.
- [10] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 1625-1634.
- [11] YANG C, KORTYLEWSKI A, XIE C, et al. Patchattack: A black-box texture-based attack with reinforcement learning[C]//Proceedings of European Conference on Computer Vision (ECCV). 2020: 681-698.
- [12] SUBRAMANYA A, PILLAI V, PIRSIIVASH H. Fooling network interpretation in image classification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR). 2019: 2020-2029.
- [13] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 618-626.
- [14] LIU X, YANG H, LIU Z, et al. Dpatch: An adversarial patch attack on object detectors[C]//2019. AAAI Workshop on Artificial Intelligence Safety (SafeAI) 2019.
- [15] LEE M, KOLTER Z. On physical adversarial patches for object detection[J]. arXiv preprint arXiv:1906.11897, 2019.
- [16] WU S, DAI T, XIA S T. DPAttack: Diffused patch attacks against universal object detection[J]. arXiv preprint arXiv:2010.11679, 2020.
- [17] HUANG H, WANG Y, CHEN Z, et al. RPAttack: Refined patch attack on general object detectors[C]//Proceedings of 2021 IEEE International Conference on Multimedia and Expo (ICME). 2021: 1-6.
- [18] ZHAO Y, YAN H, WEI X. Object hider: Adversarial patch attack against object detectors[C]//Proceedings of Conference on Information and Knowledge Management (CIKM). 2020: 28-31.
- [19] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of the 2016 ACM Sigsac Conference on Computer and Communications Security. 2016: 1528-1540.
- [20] YANG X, WEI F, ZHANG H, et al. Design and interpretation of universal adversarial patches in face detection[C]//Proceedings of European Conference on Computer Vision (ECCV). 2020: 174-191.
- [21] THYS S, VAN RANST W, GOEDEMÉ T. Fooling automated surveillance cameras: adversarial patches to attack person detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019: 49-55.
- [22] HU Y C T, KUNG B H, TAN D S, et al. Naturalistic physical adversarial patch for object detectors[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 7848-7857.
- [23] HUANG H, CHEN Z, CHEN H, et al. T-SEA: Transfer-based self-ensemble attack on object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 20514-20523.
- [24] NASEER M, KHAN S, PORIKLI F. Local gradients smoothing: Defense against localized adversarial attacks[C]//Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). 2019: 1300-1307.
- [25] HAYES J. On visible adversarial perturbations & digital watermarking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018: 1597-1604.
- [26] CHOU E, TRAMER F, PELLEGRINO G. SentiNet: Detecting localized universal attacks against deep learning systems[C]//Proceedings of 2020 IEEE Security and Privacy Workshops (SPW). 2020: 48-54.
- [27] XIANG C, MAHLOUJIFAR S, MITTAL P. {PatchCleanser}: Certifiably robust defense against adversarial patches for any image classifier[C]//Proceedings of 31st USENIX Security Symposium (USENIX Security 22). 2022: 2065-2082.
- [28] LIU J, LEVINE A, LAU C P, et al. Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 14973-14982.
- [29] RAO S, STUTZ D, SCHIELE B. Adversarial training against location-adapted adversarial patches[C]//Proceedings of European Conference on Computer Vision (ECCV). 2020: 429-448.
- [30] JI N, FENG Y F, XIE H, et al. Adversarial yolo: Defense human

detection patch attacks via detecting adversarial patches[J]. arXiv preprint arXiv:2103.08860, 2021.

- [31] ZHANG Z, YUAN B, MCCOYD M, et al. Clipped bagnet: Defending against sticker attacks with clipped bag-of-features[C]//Proceedings of 2020 IEEE Security and Privacy Workshops (SPW). 2020: 55-61.
- [32] XIANG C, BHAGOJI A N, SEHWAG V, et al. {PatchGuard}: a provably robust defense against adversarial patches via small receptive fields and masking[C]//Proceedings of 30th USENIX Security Symposium (USENIX Security 21). 2021: 2237-2254.
- [33] KINGMA D P, BA J. Adam: A method for stochastic adamization[C]//Proceedings of International Conference on Learning Representations (ICLR). 2015.
- [34] DZIUGAITE G K, GHAHRAMANI Z, ROY D M. A study of the effect of JPG compression on adversarial images[J]. arXiv preprint arXiv:1608.00853, 2016.
- [35] TARCHOUN B, BEN KHALIFA A, MAHJOUR M A, et al. Jedi: entropy-based localization and removal of adversarial patches[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 4087-4095.
- [36] LIU J, LEVINE A, LAU C P, et al. Segment and complete: defending object detectors against adversarial patch attacks with robust patch detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14973-14982.

[作者简介]



肖镇杰（2000—），男，福建泉州人，福建师范大学硕士生，主要研究方向为计算机视觉。



黄诗瑀（1999—），男，福建厦门人，福建师范大学硕士生，主要研究方向为对抗性深度学习和数字媒体取证。



叶锋（1978—），男，福建福州人，博士，福建师范大学副教授，主要研究方向为计算机视觉和视频图像编码。



黄丽清（1991—），女，福建莆田人，博士，福建师范大学讲师，主要研究方向为视频图像超分辨、去模糊处理和数字媒体取证。



黄添强（1971—），男，福建莆田人，博士，福建师范大学教授、博士生导师，主要研究方向为机器学习安全和数字媒体取证。