融合风格迁移的对抗样本生成方法

于振华，殷正，叶鸥，丛旭亚
（西安科技大学计算机科学与技术学院，710054，西安）

摘要：针对现有面向目标检测的对抗样本生成方法泛化能力弱的问题，提出了一种融合风格迁移的对抗样本生成方法。首先，提出一种新的对抗补丁生成方法，使用风格迁移方法将风格图像不同层次特征提取并融合，生成无明显物体特征且纹理丰富的对抗补丁；然后，利用梯度类激活映射方法生成目标的特征热图，对目标不同区域在目标检测模型中的关键程度进行可视化表示；最后，构建一种热图引导机制，引导对抗补丁在攻击目标的关键位置进行攻击以提高其泛化能力，生成最终对抗样本。在DroNet室外数据集上进行实验，结果表明：针对单阶段目标检测模型YOLOv5生成的对抗样本，采用所提方法计算得到的攻击成功率可达84.07%；应用于攻击两阶段目标检测模型Faster R-CNN时，采用所提方法计算得到的攻击成功率仍保持在67.65%；与现有的主流方法相比，所提方法生成的对抗样本攻击效果较好，且具有良好的泛化能力。

关键词：目标检测；对抗样本；风格迁移；对抗补丁；热图引导
中图分类号：TP399 **文献标志码**：A
DOI：10.7652/xjtuxb202407018 **文章编号**：0253-987X(2024)07-0191-12

Adversarial Example Generation Method Based on Style Transfer

YU Zhenhua, YIN Zheng, YE Ou, CONG Xuya
(College of Computer Science & Technology, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract: In response to the limited generalization in existing object detection-oriented adversarial example generation methods, an adversarial example generation method based on style transfer is proposed. Firstly, a novel adversarial patch generation method is introduced, leveraging style transfer techniques to blend features extracted from different layers of style images. This process generates adversarial patches that lack distinct object features but are rich in texture. Subsequently, a gradient-based activation mapping method is employed to generate feature heatmaps for the target, visually illustrating the significance of different regions of the target within the object detection model. Finally, a heatmap-guided mechanism is established to guide the adversarial patch to attack critical positions of the target, thereby enhancing its generalization ability and generating the ultimate adversarial examples. The proposed method's performance is verified through experiments conducted on the DroNet outdoor dataset. The experimental results demonstrate that this method achieves a success rate of 84.07% in generating adversarial samples for the single-stage object detection model YOLOv5. When applied to attack the two-stage object detection model Faster R-CNN, the success rate remains at 67.65%. Compared to prevailing

收稿日期：2023-10-29。 作者简介：于振华(1977—)，男，教授，博士生导师。 基金项目：国家自然科学基金资助项目(62273272,62303375)。
网络出版时间：2024-04-09 网络出版地址：https://link.cnki.net/urlid/61.1069.T.20240408.1508.002

methods, the adversarial examples generated by the proposed method exhibit superior attack effectiveness and good generalization capabilities.

Keywords: object detection; adversarial examples; style transfer; adversarial patch; heatmap-guided

深度神经网络的优越性能推动了计算机视觉的快速发展,在目标检测[1]、图像分类[2]、语义分割[3]、自动驾驶[4]和故障诊断[5]等领域取得了巨大的成功。其中,目标检测一直以来备受学术界和工业界的高度关注,它的任务是从给定的图像中提取感兴趣区域并标记出类别和位置。随着深度学习技术的快速发展,各种性能卓越的目标检测模型在实际生活中得到了广泛的应用[6-7]。研究结果表明,目标检测模型在安全性方面仍存在漏洞等问题,容易受到对抗样本攻击[8]。对抗样本是一种人为制造的欺骗性图像,通过对原始图像添加扰动,可以使得目标检测模型产生错误的结果。对抗样本问题最初在图像分类领域被提出[9],而现在已经被证实同样存在于目标检测领域中[10]。为了发现目标检测模型存在的潜在漏洞,研究人员会使用各种方法生成对抗样本对模型进行测试,针对模型测试中暴露出的漏洞,采取相应的防御方法,进而提高模型的安全性和鲁棒性[11]。因此,研究面向目标检测的对抗样本生成方法具有重要的理论意义和实用价值。

自 Lu 等[12]在目标检测领域中提出对抗样本生成方法以来,研究人员提出了一系列面向目标检测的对抗样本生成方法,可以分为基于全局扰动和基于局部扰动的对抗样本生成方法。在基于全局扰动的对抗样本生成方法中,Liao 等[13]利用高层次语义信息来生成对抗扰动,所生成的对抗样本具有良好的鲁棒性。Li 等[14]设计了一种通用密集目标抑制算法生成对抗样本,可以使目标检测模型无法检测特定类别的目标,同时保持其他类别目标的检测结果不变。黄世泽等[15]提出了一种针对目标检测器的对抗样本生成方法,可以实现消失攻击和定向攻击。谢云旭等[16]以图像中的目标类为单位快速进行类梯度收集,并结合图片尺度扰动生成对抗样本,具有良好的攻击效果。总的来说,基于全局扰动的对抗样本生成方法对攻击目标整体进行扰动攻击,扰动分散且不易察觉,具有良好的隐蔽性,但是其中一些方法过于依赖模型的内部信息,导致生成的对抗样本泛化能力较弱。在基于局部扰动的对抗样本生成方法中,Wang 等[17]根据目标检测模型的输出不断优化对抗补丁,最终可以在现实世界中成功欺骗目标检测系统。Zhang 等[18]通过在训练过程中引入扰动量限制并模拟复杂的外部物理环境和非刚体对象的三维变换,所生成的对抗补丁具有良好的攻击性和迁移性。丁程等[19]从扰动生成过程与扰动成本限制两方面,提出一种隐蔽式对抗样本生成方法,具有良好的隐蔽性。针对检测模型中常用的非极大值抑制机制和检测模型的特征图关注区域,王烨奎等[20]设计了位置回归攻击损失,引导生成的候选框偏离预测的关注区域,导致模型检测失败。Hu 等[21]使用生成对抗网络来创建与真实纹理难以区分的对抗性纹理,可以有效诱导目标检测模型产生错误检测结果。Liu 等[22]通过同时攻击目标检测模型的边界框回归任务和目标分类任务来生成对抗补丁,可以显著降低目标检测模型的性能。Du 等[23]在训练生成对抗补丁的过程中加入了旋转、明暗度等变化来提升鲁棒性,所生成的对抗补丁可以在现实世界中进行有效攻击。Lang 等[24]利用注意力机制来确定对抗补丁的攻击位置,指出该方法生成的对抗补丁能够明显降低目标检测模型的准确性。总的来说,基于局部扰动的对抗样本生成方法生成对抗样本扰动区域小,可以延伸到现实世界攻击,但其中一些方法生成对抗补丁的物体特征过于明显,容易被目标检测模型误识别为小型遮挡物,从而导致其泛化能力减弱。

为了解决上述问题,本文提出一种融合风格迁移的对抗样本生成方法。该方法结合风格迁移[25]的思想,在 VGG19 网络模型[26]的基础上,通过融合不同卷积层提取到风格图像的特征生成对抗补丁,保留具有攻击性纹理的同时弱化物体特征;然后引入梯度类激活映射方法(Grad-CAM)[27]生成目标的特征热图;最后构建一种热图引导机制,引导对抗补丁在目标关键特征集中区域进行攻击,生成最终对抗样本,增强对抗样本的泛化能力。与主流的对抗样本生成方法进行对比,结果表明本文方法生成的对抗样本攻击效果较好,且具有良好的泛化能力。

1 融合风格迁移的对抗样本生成方法

1.1 方法总体框架

本文利用风格迁移和 Grad-CAM 方法,提出了一种新的对抗样本生成方法,主要分为基于风格迁移的对抗补丁生成、基于 Grad-CAM 方法的运动目标热图生成和热图引导的对抗样本生成 3 个部分,如图 1 所示。

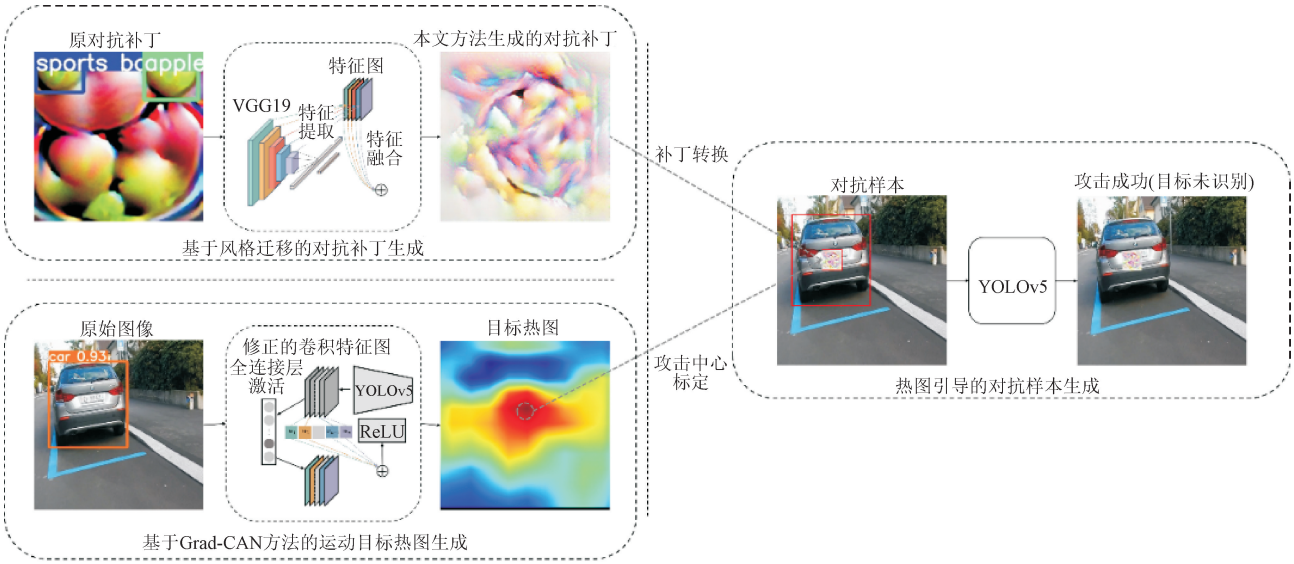


图 1　融合风格迁移的对抗样本生成方法框架

Fig. 1　Framework diagram of adversarial examples generation method based on style transfer

1.2　基于风格迁移的对抗补丁生成

原对抗补丁与本文方法生成的对抗补丁对比如图 2 所示，左侧为原对抗补丁，目标检测模型检测出包含网球和苹果的信息，右侧为经过本文方法生成的对抗补丁，目标检测模型未检测出任何物体信息。

图 2　原对抗补丁与本文方法生成的对抗补丁对比

Fig. 2　Comparison between original adversarial patch and adversarial patch generated by our method

本文方法生成对抗补丁的过程如下：给定一幅风格图像 $x \in \mathbf{R}^m$，通过一个深度神经网络分类器 F 从不同网络层提取风格图像 x 的特征，将特征融合后生成对抗补丁 x'。深度神经网络分类器 F 采用 VGG19 网络模型，优化过程为

$$\min D(x, x') + L_{\text{adv}}(x'), \quad x' \in [0,255] \tag{1}$$

式中：$D(x,x')$ 表示衡量生成对抗补丁优劣的距离损失；$L_{\text{adv}}(x')$ 表示对抗损失，$[0,255]$ 表示合理的图像像素范围。

衡量生成对抗补丁优劣的距离损失 $D(x,x')$ 由 3 个损失函数组成：内容损失 L_{C}、风格损失 L_{S} 和平滑度损失 L_{m}。其中，内容损失 L_{C} 用来保留参考风格图像的深层纹理结构，风格损失 L_{S} 用来产生更丰富的纹理特征，平滑度损失 L_{m} 用来产生局部平滑区域。距离损失 $D(x,x')$ 定义为

$$D(x, x') = L_{\text{C}} + L_{\text{S}} + L_{\text{m}} \tag{2}$$

最终的整体损失函数定义为

$$L = D(x,x') + L_{\text{adv}} \tag{3}$$

在整体损失函数中，对抗损失 L_{adv} 用来增强纹理丰富程度，组成 $D(x,x')$ 的 3 个损失函数与 L_{adv} 对于总损失的贡献度是同等重要的。

为了保证对抗补丁深层纹理结构和风格图像深层纹理结构一致，内容损失定义为

$$L_{\text{C}} = \frac{1}{HW}\sum_{l\in C_l} \| \widetilde{F}_l(x) - \widetilde{F}_l(x') \|_2^2 \tag{4}$$

式中：H、W 分别是特征图的高和宽；$\widetilde{F}_l(\cdot)$ 表示特征提取器对第 l 层网络层的特征提取操作；C_l 表示用于提取内容表示的网络层集合。特征提取器采用 VGG19 网络模型，为了保留风格图像具有攻击性的深层纹理结构，本文选用特征提取器的深层网络层来提取内容表示。

为了保证参考风格图像的风格特征能够被提取出用于生成对抗补丁，风格损失定义为

$$L_{\text{S}} = \sum_{l\in S_l} w_l \frac{1}{4N_l^2M_l^2} \| G(\widetilde{F}_l(x)) - G(\widetilde{F}_l(x')) \|_2^2 \tag{5}$$

式中：w_l 是第 l 层的权重；N_l 和 M_l 分别是第 l 层特征图的通道数和空间维度；$G(\widetilde{F}_l(\cdot))$ 表示计算特征提取器第 l 层提取深层特征的 Gram 矩阵操作；S_l 表示用于提取风格表示的网络层集合。特征提取器可以在不同网络层提取到不同特征，浅层网络层通常会提取更多颜色特征，而深层网络层则会提取更多纹理特征。为了生成纹理更丰富的对抗补丁，本文选用特征提取器深层的网络层来提取风格

表示。

为了提高对抗补丁平滑度，需要减少相邻像素之间变化，平滑度损失定义为

$$L_{\mathrm{m}}=\sum((x'_{i,j}-x'_{i+1,j})^2+(x'_{i,j}-x'_{i,j+1})^2)^{1/2} \tag{6}$$

式中：$x'_{i,j}$是图像 x' 在坐标(i,j)下的像素。该损失函数对生成具有低方差的局部色块有积极作用，可以确保在生成过程中产生平滑颜色过渡。

为了进一步增强对抗补丁的纹理丰富程度，对抗损失定义为

$$L_{\mathrm{adv}}=-\ln(p_{y_{\mathrm{adv}}}(x'))+\ln(p_y(x')) \tag{7}$$

式中：$p_{y_{\mathrm{adv}}}(\cdot)$是 F 对于 y_{adv} 类别的概率输出；$p_y(\cdot)$是 F 对于 y 类别的概率输出，且 $y_{\mathrm{adv}}\neq y$。该损失函数在生成对抗补丁的过程中将其他类别信息引入，增加了对抗扰动的多样性，可以进一步增强生成对抗补丁的纹理丰富程度。

为了增强对抗补丁的攻击性，在整个优化生成过程中加入了一些增强鲁棒性的操作，包括旋转、比例调整和颜色移动等。整个优化生成过程的定义为

$$\min_{x'}((L_{\mathrm{C}}+L_{\mathrm{S}}+L_{\mathrm{m}})+\max_{r\in R}L_{\mathrm{adv}}(r(x'))) \tag{8}$$

式中：R 表示增强鲁棒性的随机变换集合，包括旋转、比例调整和颜色移动；r 表示其中一种随机变换方式。

1.3 基于 Grad-CAM 方法的运动目标热图生成

对抗样本的攻击效果，不仅与对抗补丁的攻击性有关，还需要考虑其攻击位置。现有方法主要使用中心贴图的方式来进行对抗攻击，但其生成对抗样本的攻击效果较差，在应用中效率不高且泛化能力弱。使用卷积神经网络可视化方法可以生成目标的特征热图，能够反映模型对目标不同区域的感兴趣程度，如果在感兴趣程度高的地方进行攻击，就能够产生更好更稳定的攻击效果。

卷积神经网络在最终进行分类时，最后一层卷积层包含原图像关键信息的特征图。作为卷积神经网络可视化方法的一种，Grad-CAM 方法利用卷积神经网络中最后一层卷积层的梯度信息来理解每个神经元对最终分类结果的贡献，从而对图像中不同区域的重要性进行定量分析。该方法在没有注意力机制的情况下也能够对图像特征进行可视化，并且适用于任意结构的卷积神经网络。在卷积神经网络的最后一层生成 k 个特征图，利用全局平均池化将其进行线性变换产生 c 个类别，每个具体类别得分 S^c 的定义为

$$S^c=\frac{1}{Z}\sum_w\sum_h\sum_k\omega_k^cA_{wh}^k \tag{9}$$

式中：ω_k^c 为 GAP 层到 Softmax 层之间的权重，c 是目标类别序号；Z 是特征图大小；A 是最后一层卷积层输出的特征图；k 是特征图的通道维度的序号；h 和 w 分别表示特征图在高和宽维度上的索引。

为了获得目标的特征热图 $L^c_{\text{Grad-CAM}}\in\mathbf{R}^{u\times v}$，记高度为 u，宽度为 v。首先利用 Softmax 层输出的概率计算梯度，将这些梯度在高和宽的维度上进行池化处理，以获得神经元重要性权重。最后一层卷积层中的特征图所有像素为 A_{wh}，输出目标类概率为 y^c，y^c 对 A_{wh} 求偏导的定义为

$$\alpha^c_{k_{wh}}=\frac{\partial y^c}{\partial A^k_{wh}} \tag{10}$$

式中：α_k^c 为神经元重要性权重。

计算 y^c 对 A_{wh} 的偏导数后，对所得的偏导数结果进行全局平均，该过程的定义为

$$\alpha_k^c=\frac{1}{Z}\sum_w\sum_h\alpha^c_{k_{wh}} \tag{11}$$

式中：α_k^c 为最后一层卷积层输出特征图第 k 个通道对目标类别 c 的重要程度。将 α_k^c 当作权重，最后一层特征图加权线性组合的定义为

$$L_c=\sum_k\alpha_k^cA^k \tag{12}$$

最终通过 ReLU 激活函数处理后得到目标特征热图的定义为

$$L^c_{\text{Grad-CAM}}=\mathrm{ReLU}(L_c) \tag{13}$$

1.4 热图引导的对抗样本生成

得到具有强攻击性对抗补丁和目标的特征热图后，需要利用目标的特征热图计算攻击位置以引导对抗补丁进行攻击，生成最终对抗样本。通过建立一种热图引导机制，可以将目标的特征热图进行二值化处理，筛选出其关键区域。特征热图二值化过程的定义为

$$\begin{cases}B(p_x,p_y)=1, & L^c_{\text{Grad-CAM}}(p_x,p_y)\geqslant T\\ B(p_x,p_y)=0, & L^c_{\text{Grad-CAM}}(p_x,p_y)<T\end{cases} \tag{14}$$

式中：$L^c_{\text{Grad-CAM}}(x,y)$为热图；$B(x,y)$为二值图，$(p_x,p_y)$为像素坐标；$T$ 为关键区域筛选条件的热力阈值。

得到二值图后，计算其攻击重心$(\tilde{p}_x,\tilde{p}_y)$，作为对抗补丁的攻击位置。攻击重心计算过程的定义为

$$(\tilde{p}_x,\tilde{p}_y)=\left(\frac{\sum_{p_x=0}^{W-1}\sum_{p_y=0}^{H-1}p_xB(p_x,p_y)}{\sum_{p_x=0}^{W-1}\sum_{p_y=0}^{H-1}B(p_x,p_y)},\frac{\sum_{p_x=0}^{W-1}\sum_{p_y=0}^{H-1}p_yB(p_x,p_y)}{\sum_{p_x=0}^{W-1}\sum_{p_y=0}^{H-1}B(p_x,p_y)}\right) \tag{15}$$

攻击重心引导对抗补丁对目标进行攻击，生成最终对抗样本。对抗样本生成过程的定义为

$$I_{\text{adv}}(p_x,p_y)=\text{clip}(I(p_x,p_y)+\delta_p(\widetilde{p}_x,\widetilde{p}_y;\theta)) \tag{16}$$

式中：$I_{\text{adv}}(p_x,p_y)$表示生成的对抗样本；$I(p_x,p_y)$表示原始的输入图像；$\delta_p(\widetilde{x},\widetilde{y};\theta)$表示在目标的攻击重心$(\widetilde{x},\widetilde{y})$处进行对抗补丁攻击，其中 θ 是函数的参数，表示对抗补丁的特征，clip(·)表示将像素限制在合理范围内，防止其超出有效范围。

综上所述，本文所提方法设计流程如图 3 所示。输入有缺陷的对抗补丁，在 VGG19 网络模型的基础上进行风格迁移，达到迭代次数后，生成最终对抗补丁；将攻击目标图像输入目标检测模型，利用 Grad-CAM 方法生成目标热图，对热图像素点进行筛选，计算满足热区阈值像素点的重心，得到攻击重心；利用攻击重心引导对抗补丁进行攻击，生成最终对抗样本。在训练过程中，利用 VGG19 网络模型，将第 4 层和第 5 层的卷积层提取特征融合生成对抗补丁。整个训练过程通过 Adam 算法进行优化，学习率为 0.01，一阶矩估计的指数衰减率为 0.9，二阶矩估计的指数衰减率为 0.999。在使用过程中，对于输入的攻击目标图像，使用了 YOLOv5 进行处理，从 YOLOv5 的第 17、20 和 23 层卷积层提取特征并结合 Grad-CAM 方法进行可视化，得到攻击目标热图。对目标热图进行二值化操作，筛选满足热区阈值的像素点，计算其重心。将对抗补丁在重心位置进行中心覆盖，生成最终的对抗样本。

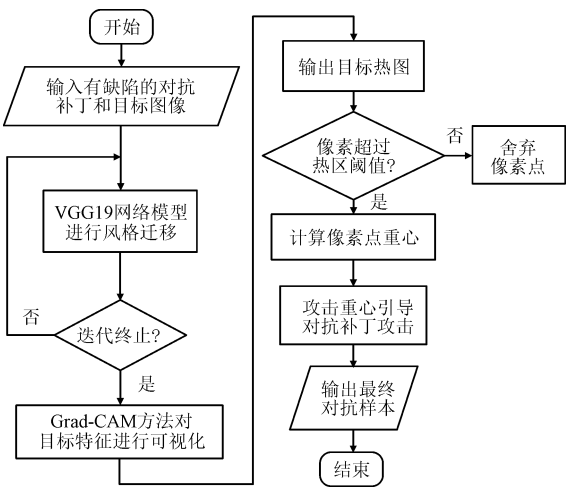


图 3　融合风格迁移的对抗样本生成方法设计流程

Fig. 3　Design flowchart diagram of adversarial examples generation method based on style transfer

2　实验结果与讨论

2.1　数据集

数据集采用 DroNet 室外数据集[28]，该数据集包含车辆和行人在 137 个场景下共 32 000 张图像，根据视野障碍物距离的远近标注为 0(无碰撞风险)和 1(有碰撞风险)。本文的研究对象是车辆目标，因此本文从 DroNet 室外数据集中筛选出关于车辆的图像作为实验数据集。实验数据集共有 408 张图像，包括城市街道、乡间小路等 20 种不同场景，旨在探究不同场景下车辆的特性。为考察距离对攻击效果的影响，每个场景均含有不同距离下的同一车辆目标。

2.2　对比方法及性能指标

为了验证本文对抗样本生成方法的有效性，将其与文献[21-24]等 4 种方法制作的对抗补丁进行对比，并分别命名为 TC-EGA、DPatch、P* 和 AGAP。本文所提方法生成的对抗补丁命名为 DP-PGS，如图 4 所示。

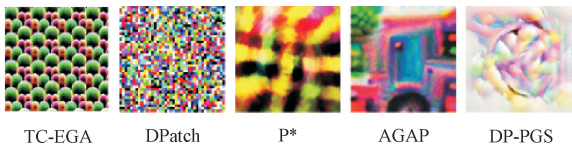


图 4　不同对抗补丁图像

Fig. 4　Different adversarial patches

实验采用文献[21]中攻击成功率及平均精度这两个指标来衡量不同方法的攻击效果，其定义如下。

(1)攻击成功率。如果对抗样本生成方法最终生成的对抗样本能够使目标检测模型产生误判或漏检的情况，表明这种方法成功生成了有效对抗样本。在选择的数据集上进行方法的效果评估时，将能够进行有效攻击的对抗样本数记为 a，用于测试的所有图像数记为 b，则攻击成功率 d 定义为

$$d=\frac{a}{b}\times 100\% \tag{17}$$

(2)平均精度。准确率和召回率可以用来评价一个模型的优劣，在目标检测领域，一个数据集中有若干待检测的目标，准确率代表真实目标在模型检测出的目标中所占比例，召回率代表所有真实目标被模型检测出来的比例。图像中存在背景与物体两种标签，检测框也分为正确与错误两种标签，因此在评测时会产生以下 4 种样本：正确检测框样本(T_{P})中检测框与标签框正确匹配，两者间交并比大

于 0.5；误检框样本（F_P）中模型将背景检测为物体，通常这种检测框与图像中所有标签框交并比都不会超过 0.5；漏检框样本（F_N）中模型没有检测出本应该检测出的物体；正确背景样本（T_N）中模型未错误检测背景，这种情况在目标检测中通常不需要考虑。根据上述描述可以分别计算出 T_P、F_P 和 T_N 的值，准确率 P 和召回率 R 可定义为

$$P = \frac{T_P}{T_P + F_P} \tag{18}$$

$$R = \frac{T_P}{T_P + F_N} \tag{19}$$

即使有了准确率-召回率曲线，模型的评价仍然不直观，需要平均精度来衡量模型性能。平均精度代表了准确率-召回率曲线的面积，综合评价了不同召回率下的准确率，不会对准确率与召回率有任何偏好，平均精度的定义为

$$A_P = \int_0^1 P \mathrm{d}R \tag{20}$$

2.3 攻击模型及实验结果

实验的攻击模型选取单阶段目标检测模型 YOLOv5 和两阶段目标检测模型 Faster R-CNN[29]。YOLOv5 是 YOLO 系列[30]目标检测模型里检测精度、速度等各个方面较好的一种模型，对图像中小目标物体具有较高的检测精度。Faster R-CNN 的检测精度较高，在多个数据集上表现优秀，具有良好的鲁棒性。本文所提方法生成的对抗样本在 YOLOv5 上攻击效果如图 5 所示，图 5 给出了正常样本和对抗样本经过 YOLOv5 检测后的结果，其中原始图像中的车辆目标均能被正确检测出来，而对抗样本中的车辆目标无法被正确检测，导致 YOLOv5 出现误分类、未检测到目标和预测边界框错误 3 种情况。

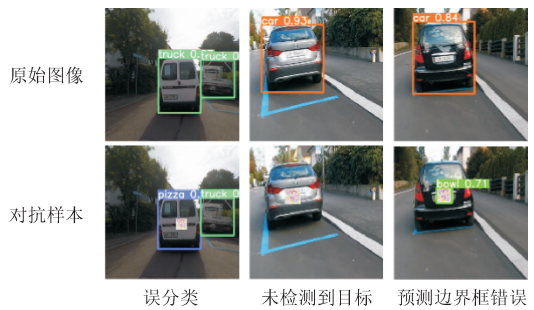


图 5 对抗样本在 YOLOv5 上的攻击效果

Fig. 5 Attack effects of adversarial examples generated on YOLOv5

图 6 所示为不同对抗补丁占比的对抗样本，对抗补丁占比是指其相对于目标检测框的面积占比，目标检测框为图 6 中蓝色框所标识区域。由图 6 中从左到右红框的变化可以看出，当对抗补丁占比较小时，其纹理信息无法充分展现，导致其攻击效果不佳，目标检测模型对目标仍有较高的识别准确率。随着对抗补丁占比增加，对抗补丁的纹理信息逐渐显现，进而提升了攻击效果，目标检测模型对目标的识别准确率逐渐降低。当对抗补丁占比较大时，其能够对攻击目标关键特征区域进行有效扰动攻击，诱导目标检测模型产生误判或漏检。

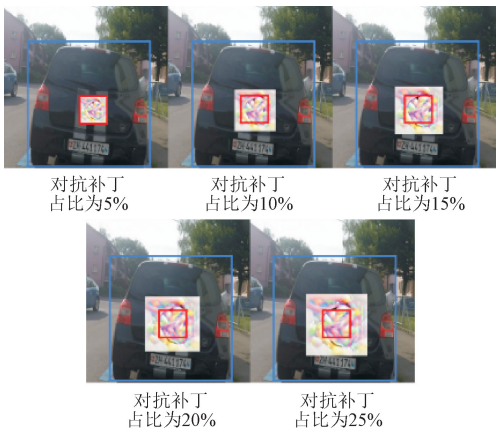


图 6 不同对抗补丁占比的对抗样本

Fig. 6 Adversarial examples with different percentages adversarial patches in the detection box

图 7 所示为特征提取器使用不同网络层提取的特征。本文使用的特征提取器为 VGG19 网络模型。

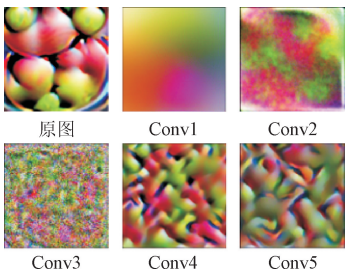


图 7 特征提取器使用不同网络层提取的特征

Fig. 7 Features extracted by different network layers of the feature extractor

由图 7 可以看出，使用第 1 层卷积层提取的特征，呈现出相对单一的纹理特征，主要表现为大面积的颜色特征。相比之下，使用第 5 层卷积层提取的特征呈现出更为丰富的纹理特征。在 VGG19 网络模型中，浅层卷积模块由较少数量的卷积层构成，并且每一层采用了较小尺寸的卷积核，因此其感受野

相对较小，更适合提取简单的图像特征，如颜色和亮度等。深层卷积模块则由多个卷积层组成，且每层使用较大尺寸的卷积核，因此具有更大的感受野，更适合提取复杂的图像特征，如纹理和形状等。

图 8 所示为融合 VGG19 网络模型不同卷积模块提取特征所生成的对抗补丁。由图 8 中从左到右红框的变化可以看出，舍弃浅层卷积模块提取特征，并融合深层卷积模块提取特征生成的对抗补丁纹理信息更丰富。这是由于深层卷积模块层具有更深的网络层，能够有效提取到图像中更为抽象和复杂的特征信息。相较于浅层卷积模块所提取的特征，融合深层卷积模块提取特征所生成的对抗性补丁纹理特征更加丰富，具有更强的对抗扰动能力，对目标检测模型的攻击效果更为显著。

图 8　融合 VGG19 网络模型不同卷积模块提取特征所生成的对抗补丁

Fig. 8　Adversarial patches generated by fusing features from different convolutional modules of the VGG19 network model

图 9 所示为目标在不同距离下被对抗补丁攻击前后的特征热图，图 9(a)和图 9(b)的左侧表示未被对抗补丁攻击的车辆目标及其特征热图，图 9(a)和图 9(b)的右侧表示被对抗补丁攻击的车辆目标及其特征热图。如图 9(a)左侧和图 9(b)左侧所示，随着目标逐渐靠近，目标特征区域从中心逐渐扩散至周围。当目标距离较远时，目标在整个图像中所占比例较小，其边缘特征较少且不明显，因此目标的特征热图会呈现出中心集中的趋势。当目标靠近时，目标在整个图像中所占比例增大，其边缘特征更加明显，特征区域也随之明显增多，此时目标的特征热图由中心集中开始向四周扩散。如图 9(a)右侧和图 9(b)右侧所示，本文方法生成的对抗补丁随着目标的逐渐靠近而逐渐显露出纹理特征，攻击性也逐渐增强。当目标距离远时，对抗补丁的纹理特征不明显，攻击性较弱，只能微弱降低目标特征热图关键区域的热度，对目标特征的分散作用较小。当目标靠近时，对抗补丁的纹理特征显露出来，攻击性逐渐增强，可以有效降低目标特征热图关键区域的热度，分散目标的关键特征，从而使目标检测模型产生误判或漏检。

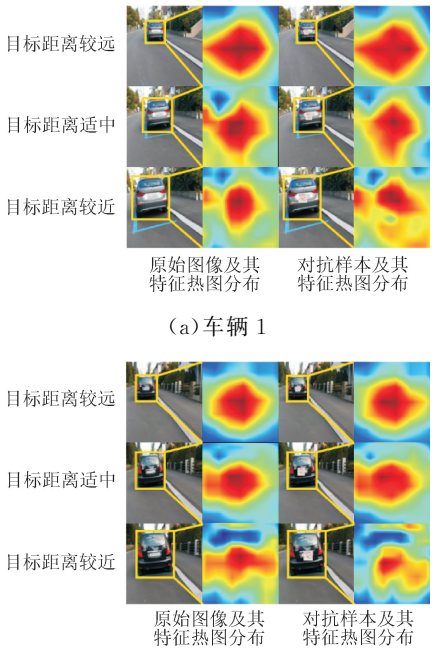
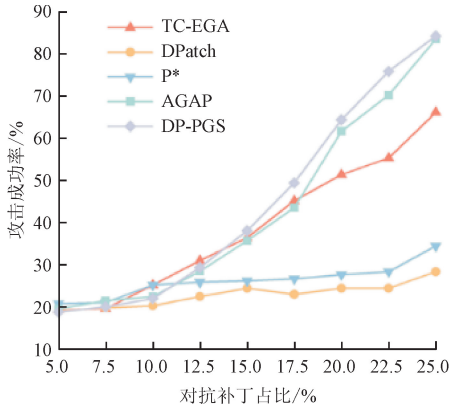


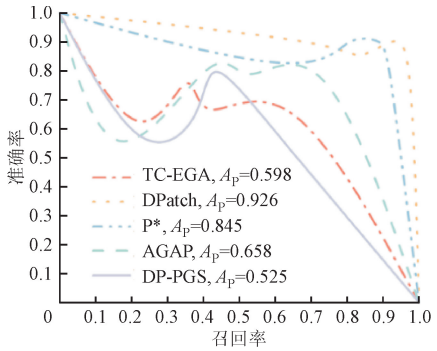
图 9　目标在不同距离下被对抗补丁攻击前后的特征热图分布

Fig. 9　Distribution of feature heatmaps of the target before and after adversarial patch attacks at different distances

图 10 所示为目标特征热图引导对抗补丁攻击生成的对抗样本在 YOLOv5 上的攻击结果，目标特征热图由 YOLOv5 生成。根据图 10(a)所示，随着对抗补丁占比增大，所有对抗补丁的攻击成功率均呈现上升趋势，其中 DP-PGS、AGAP 和 TC-EGA 的增长趋势更加明显。根据图 10(b)所示，所有对抗补丁都能降低目标检测模型平均检测精度，其中 DP-PGS、AGAP 和 TC-EGA 的降低效果更加显著。DP-PGS 方法的降低效果最好，可使目标检测模型平均检测精度下降至 0.525，比 TC-EGA 和 AGAP 补丁攻击后目标检测模型平均检测精度分别低 0.073 和 0.133。表 1 所示为不同对抗补丁占比下，目标特征热图引导对抗补丁攻击生成的对抗样本在 YOLOv5 上的攻击成功率。由表 1 可以看出，对抗补丁占比小时，DP-PGS、AGAP 和 TC-EGA 补丁的攻击成功率相近。随着对抗补丁占比增大，DP-PGS 和 AGAP 攻击成功率的增长速度比 TC-EGA 更快。对抗补丁占比为 25.0%时，DP-PGS 的攻击效果最好，攻击成功率达到 84.07%，比 AGAP 和 TC-EGA 的攻击成功率分别高 0.49%和 17.89%。



(a)对抗样本攻击成功率随着对抗补丁占比的变化



(b)对抗样本攻击后 YOLOv5 的准确率-召回率曲线及其平均精度

图 10 目标特征热图引导对抗补丁攻击生成的对抗样本在 YOLOv5 上的攻击结果

Fig. 10 Attack results of the adversarial examples generated by the adversarial patch attack guided by the target feature heatmap on YOLOv5

表 1 不同对抗补丁占比下通过目标特征热图引导对抗补丁攻击生成的对抗样本在 YOLOv5 上的攻击成功率

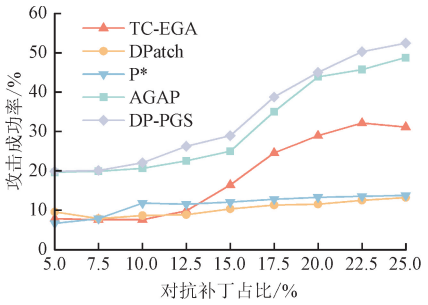
Table 1 Attack success rate of adversarial examples generated by the adversarial patch attack guided by heatmap of target feature under different proportions of adversarial patches on YOLOv5

对抗补丁占比/%	攻击成功率/%				
	TC-EGA[21]	DPatch[22]	P*[23]	AGAP[24]	DP-PGS
15.0	36.52	24.51	26.23	35.54	**37.99**
17.5	45.10	23.04	26.72	43.63	**49.51**
20.0	51.23	24.51	27.70	61.52	**64.22**
22.5	55.15	24.51	28.19	70.10	**75.74**
25.0	66.18	28.19	64.56	83.58	**84.07**

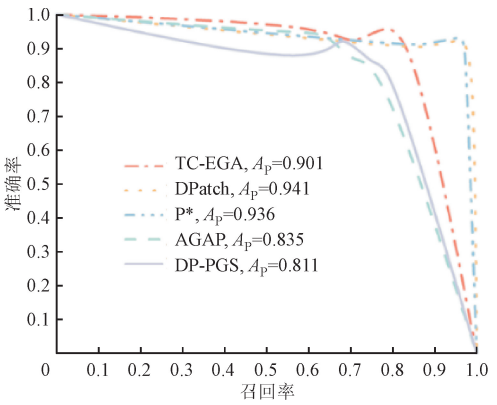
注:表中加粗数据为最优值。

由上述实验结果分析可知,在不同的对抗补丁占比下,由于 DPatch 和 P* 这两类对抗补丁的纹理丰富程度较低,对攻击目标特征的分散作用较弱,因此攻击效果均表现较差。对于 DP-PGS、AGAP 和 TC-EGA 这 3 类对抗补丁,由于本身纹理丰富程度较高,故而对目标特征的分散作用较强。同时,通过目标特征热图来引导攻击位置,会进一步增强这些对抗补丁的攻击性,从而能够生成具有较好攻击效果的对抗样本。当对抗补丁占比较小时,DP-PGS、AGAP 和 TC-EGA 这 3 类对抗补丁具有攻击性的纹理无法充分展现,并且目标检测模型在检测过程中会经过多个卷积层和池化层,对抗补丁会损失大量纹理信息,因此整体攻击效果相差不大。随着对抗补丁占比增大,对抗补丁的攻击效果会逐渐增强。尽管 TC-EGA 补丁的纹理丰富程度较高,但其纹理变化比较单一,攻击效果稍微弱于 DP-PGS 和 AGAP 补丁。

图 11 所示为随机位置进行对抗补丁攻击生成的对抗样本在 YOLOv5 上的攻击结果。由图 11(a)和图 10(a)对比可知,对抗补丁在随机位置相较于在目标特征热图引导的位置进行攻击所生成的对抗样本,攻击成功率大幅下降。



(a)对抗样本攻击成功率随着对抗补丁占比的变化



(b)对抗样本攻击后 YOLOv5 的准确率-召回率曲线及其平均精度

图 11 随机位置进行对抗补丁攻击生成的对抗样本在 YOLOv5 上的攻击结果

Fig. 11 Attack results of adversarial examples generated by adversarial patch attack at random location on YOLOv5

由图 11(b)和图 10(b)对比可以看出，对抗补丁在随机位置相较于在目标特征热图引导的位置进行攻击所生成的对抗样本，降低目标检测模型平均检测精度的能力明显减弱。DP-PGS 补丁在随机位置攻击后目标检测模型平均检测精度下降到 0.811，相较于其在目标特征热图引导的位置攻击后目标检测模型平均检测精度高出 0.286。表 2 所示为不同对抗补丁占比下，随机位置进行对抗补丁攻击生成的对抗样本在 YOLOv5 上的攻击成功率。由表 1 和表 2 对比可知，采用随机位置进行对抗补丁攻击时，整体的攻击成功率下降较多。当对抗补丁占比为 25.0%时，DP-PGS 补丁的攻击成功率下降了 31.86%。

表 2　不同对抗补丁占比下通过随机位置进行对抗补丁攻击生成的对抗样本在 YOLOv5 上的攻击成功率

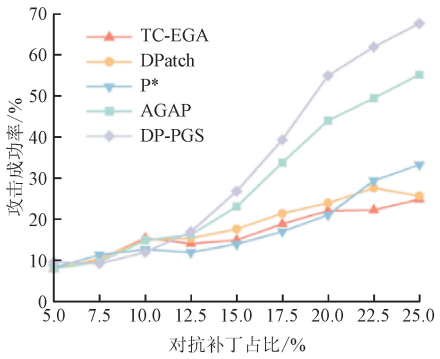
Table 2　Attack success rate of adversarial examples generated by the adversarial patch attack at random location under different proportions of adversarial patches on YOLOv5

对抗补丁占比/%	攻击成功率/%				
	TC-EGA[21]	DPatch[22]	P*[23]	AGAP[24]	DP-PGS
15.0	16.18	10.05	12.01	25.00	**28.68**
17.5	24.51	11.03	12.75	35.05	**38.73**
20.0	28.68	11.52	13.24	43.63	**44.85**
22.5	32.11	12.50	13.48	45.59	**50.25**
25.0	31.13	12.99	13.73	48.53	**52.21**

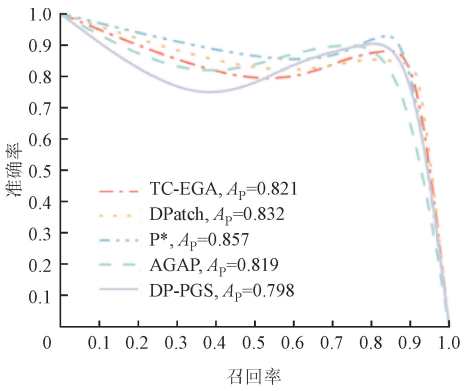
注：表中加粗数据为最优值。

由上述实验结果分析可知，在目标特征热图引导的位置进行对抗补丁攻击，可以有效提升生成对抗样本的攻击效果。Grad-CAM 方法生成目标的特征热图反映了目标检测模型对于目标不同区域的感兴趣程度，颜色越深的地方特征越多，也越关键。通过建立一种热图引导机制，对目标的特征热图进行二值化处理并计算攻击重心，引导对抗补丁进行攻击，可以有效弱化目标关键特征，生成的对抗样本可以使目标检测模型产生误检或者漏检。采用随机位置进行对抗补丁攻击时，目标的关键特征区域不一定会被攻击，目标检测模型仍能正确检测出目标，从而导致攻击效果大幅减弱。

为了验证本文方法生成对抗样本的泛化能力，使用两阶段检测模型 Faster R-CNN 来进行验证。图 12 所示为目标特征热图引导对抗补丁攻击生成的对抗样本在 Faster R-CNN 上的攻击结果，目标特征热图由 YOLOv5 生成。由图 12(a)和图 10(a)对比可以看出，由 YOLOv5 生成的目标特征热图引导对抗补丁攻击生成的对抗样本，在 Faster R-CNN 上的攻击成功率均有所下降。由图 12(b)和图 10(b)对比可以看出，由 YOLOv5 生成的目标特征热图引导对抗补丁攻击生成的对抗样本，降低 Faster R-CNN平均检测精度的能力明显减弱。由 YOLOv5 生成的目标特征热图引导 DP-PGS 补丁攻击生成的对抗样本，可以使 Faster R-CNN 平均检测精度下降到 0.798，相较于其降低 YOLOv5 的平均检测精度，该对抗样本攻击 Faster R-CNN 后平均检测精度高出 0.273。表 3 所示为不同对抗补丁占比下，由 YOLOv5 生成的目标特征热图引导对抗补丁攻击生成的对抗样本在 Faster R-CNN 上的攻击成功率。由表 3 和表 1 对比可知，由 YOLOv5 生成的目标特征热图引导对抗补丁攻击生成的对抗样本攻击 Faster R-CNN时，整体的攻击成功率均有小幅降低。当对抗补丁占比为 25.0%时，DP-PGS 补丁的攻击成功率虽然下降了 16.42%，但仍能保持在 67.65%。



(a)对抗样本攻击成功率随着对抗补丁占比的变化



(b)对抗样本攻击后 Faster R-CNN 的准确率-召回率曲线及其平均精度

图 12　目标特征热图引导对抗补丁攻击生成的对抗样本在 Faster R-CNN 上的攻击结果

Fig. 12　Attack results of the adversarial examples generated by the adversarial patch attack guided by the target feature heatmap on Faster R-CNN

表 3 不同对抗补丁占比下通过目标特征热图引导对抗补丁攻击生成的对抗样本在 Faster R-CNN 上的攻击成功率

Table 3 Attack success rate of adversarial examples generated by the adversarial patch attack guided by heatmap of target feature under different proportions of adversarial patches on Faster R-CNN

对抗补丁占比/%	攻击成功率/%				
	TC-EGA[21]	DPatch[22]	P*[23]	AGAP[24]	DP-PGS
15.0	14.71	17.40	13.73	23.04	**26.72**
17.5	18.63	21.32	16.91	33.58	**39.22**
20.0	22.06	23.77	20.83	43.87	**54.9**
22.5	22.06	27.45	29.41	49.51	**61.76**
25.0	24.75	25.49	33.09	55.15	**67.65**

注：表中加粗数据为最优值。

由上述实验结果分析可知，针对 YOLOv5 生成的对抗样本，在另一种类型目标检测模型 Faster R-CNN 上的攻击效果有所减弱。在 Faster R-CNN 中，区域预选网络处理流程会排除太小和超出边界的预选框。对于占比较小的对抗补丁，由于本身的攻击性较弱，且经过区域预选网络处理后具有攻击性的纹理信息被舍弃，因而攻击效果变差。由单阶段目标检测模型 YOLOv5 生成目标特征热图引导对抗补丁攻击生成的对抗样本，并没有专门针对两阶段目标检测模型的处理，即使增加对抗补丁的占比，在两阶段目标检测模型 Faster R-CNN 上的攻击效果仍会减弱。当对抗补丁占比较大时，本文方法生成的对抗样本仍能对 Faster R-CNN 进行有效攻击。该方法生成的对抗样本能够在单阶段目标检测模型 YOLOv5 和两阶段目标检测模型 Faster R-CNN 上进行有效攻击，并且在攻击效果方面均优于所对比的主流方法，同时还表现出更强的泛化能力。

图 13 所示为本文方法生成对抗样本在 Faster R-CNN 上的攻击效果。图 13 给出了正常样本和对抗样本经过 Faster R-CNN 检测后的结果，其中正常样本中的车辆目标均能被正确检测出来，而对抗样本中的车辆目标无法被正确检测，导致 Faster R-CNN 出现误分类、未检测到目标和预测边界框错误 3 种情况。本文所提方法针对 YOLOv5 生成的对抗样本在 Faster R-CNN 上的攻击效果会减弱，但仍然有很大部分的对抗样本能够成功攻击 Faster R-CNN，具有良好的泛化能力。对于攻击的目标，在不同目标检测模型上的检测过程可能会有不同，但对目标关键特征具有相似的敏感度。本文通过建立一种热图引导机制，将目标的特征热图进行二值化处理，筛选出相似敏感区域，并计算攻击重心引导对抗补丁进行攻击，从而提升生成对抗样本的泛化能力。

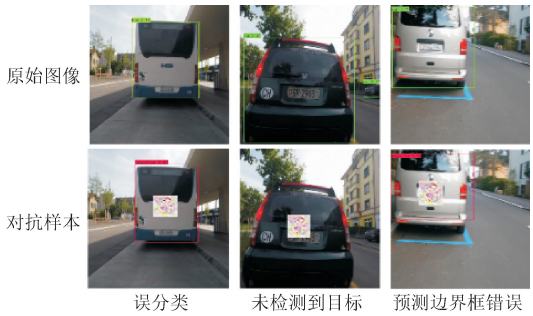


图 13 对抗样本在 Faster R-CNN 上的攻击效果

Fig. 13 Attack effects of adversarial examples generated on Faster R-CNN

3 结 论

本文提出了一种融合风格迁移的对抗样本生成方法，可以有效生成泛化能力良好的对抗样本。首先提出了一种新的对抗补丁生成方法，使用风格迁移方法生成对抗补丁，可以生成无明显物体特征且纹理丰富的对抗补丁；再结合 Grad-CAM 方法生成目标的特征热图，可视化目标检测模型对目标不同区域的感兴趣程度；最后构建了一种热图引导机制，引导对抗补丁进行攻击，可以增强最终生成对抗样本的泛化能力。实验结果表明：与所对比的主流方法相比，本文所提方法生成的对抗样本在 YOLOv5 和 Faster R-CNN 上的攻击效果较好，具有良好的泛化能力。本文主要关注车辆的单目标场景，攻击目标较为单一，未来研究将拓展到多目标场景，以实现更广泛的应用。

参考文献：

[1] 许德刚，王露，李凡. 深度学习的典型目标检测算法研究综述 [J]. 计算机工程与应用，2021，57(8)：10-25.
XU Degang，WANG Lu，LI Fan. Review of typical object detection algorithms for deep learning [J]. Computer Engineering and Applications，2021，57(8)：10-25.

[2] 张珂，冯晓晗，郭玉荣，等. 图像分类的深度卷积神经网络模型综述 [J]. 中国图象图形学报，2021，26(10)：2305-2325.
ZHANG Ke，FENG Xiaohan，GUO Yurong，et al. Overview of deep convolutional neural networks for image classification [J]. Journal of Image and Graph-

ics，2021，26(10)：2305-2325.

[3] 田萱，王亮，丁琪. 基于深度学习的图像语义分割方法综述 [J]. 软件学报，2019，30(2)：440-468.
TIAN Xuan，WANG Liang，DING Qi. Review of image semantic segmentation based on deep learning [J]. Journal of Software，2019，30(2)：440-468.

[4] 张新钰，高洪波，赵建辉，等. 基于深度学习的自动驾驶技术综述 [J]. 清华大学学报(自然科学版)，2018，58(4)：438-444.
ZHANG Xinyu，GAO Hongbo，ZHAO Jianhui，et al. Overview of deep learning intelligent driving methods [J]. Journal of Tsinghua University(Science and Technology)，2018，58(4)：438-444.

[5] 张西宁，郭清林，刘书语. 深度学习技术及其故障诊断应用分析与展望 [J]. 西安交通大学学报，2020，54(12)：1-13.
ZHANG Xining，GUO Qinglin，LIU Shuyu. Analysis and prospect of deep learning technology and its fault diagnosis application [J]. Journal of Xi'an Jiaotong University，2020，54(12)：1-13.

[6] LIU Liangkai，LU Sidi，ZHONG Ren，et al. Computing systems for autonomous driving：state of the art and challenges [J]. IEEE Internet of Things Journal，2021，8(8)：6469-6486.

[7] FERNANDO T，GAMMULLE H，DENMAN S，et al. Deep learning for medical anomaly detection? a survey [J]. ACM Computing Surveys，2022，54(7)：141.

[8] 张思思，左信，刘建伟. 深度学习中的对抗样本问题 [J]. 计算机学报，2019，42(8)：1886-1904.
ZHANG Sisi，ZUO Xin，LIU Jianwei. The problem of the adversarial examples in deep learning [J]. Chinese Journal of Computers，2019，42(8)：1886-1904.

[9] SZEGEDY C，ZAREMBA W，SUTSKEVER I，et al. Intriguing properties of neural networks [EB/OL]. (2014-02-19)[2023-06-16]. https://arxiv.org/abs/1312.6199.

[10] XIE Cihang，WANG Jianyu，ZHANG Zhishuai，et al. Adversarial examples for semantic segmentation and object detection [C]//2017 IEEE International Conference on Computer Vision. Piscataway，NJ，USA：IEEE，2017：1378-1387.

[11] CHEN Pinchun，KUNG B H，CHEN Juncheng. Class-aware robust adversarial training for object detection [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ，USA：IEEE，2021：10415-10424.

[12] LU Jiajun，SIBAI H，FABRY E. Adversarial examples that fool detectors [EB/OL]. (2017-12-07)[2023-06-23]. https://arxiv.org/abs/1712.02494.

[13] LIAO Quanyu，WANG Xin，KONG Bin，et al. Fast local attack：generating local adversarial examples for object detectors [C]//2020 International Joint Conference on Neural Networks. Piscataway，NJ，USA：IEEE，2020：1-8.

[14] LI Debang，ZHANG Junge，HUANG Kaiqi. Universal adversarial perturbations against object detection [J]. Pattern Recognition，2021，110：107584.

[15] 黄世泽，张肇鑫，董德存，等. 针对车载环境感知系统的对抗样本生成方法 [J]. 同济大学学报(自然科学版)，2022，50(10)：1377-1384.
HUANG Shize，ZHANG Zhaoxin，DONG Decun，et al. Adversarial example generation method for vehicle environment perception system [J]. Journal of Tongji University (Natural Science)，2022，50(10)：1377-1384.

[16] 谢云旭，吴锡，彭静. 无锚框模型类梯度全局对抗样本生成 [J]. 计算机工程，2023，49(10)：186-193.
XIE Yunxu，WU Xi，PENG Jing. Generation of gradient global adversarial samples with anchor-free model [J]. Computer Engineering，2023，49(10)：186-193.

[17] WANG Yajie，LÜ Haoran，KUANG Xiaohui，et al. Towards a physical-world adversarial patch for blinding object detection models [J]. Information Sciences，2021，556：459-471.

[18] ZHANG Haotian，MA Xu. Misleading attention and classification：an adversarial attack to fool object detection models in the real world [J]. Computers & Security，2022，122：102876.

[19] 丁程，史再峰，佟博文，等. 针对目标检测的隐蔽式对抗扰动生成方法 [J]. 光电子·激光，2023，34(9)：915-922.
DING Cheng，SHI Zaifeng，TONG Bowen，et al. Stealthy adversarial perturbation generation method for object detection [J]. Journal of Optoelectronics·Laser，2023，34(9)：915-922.

[20] 王烨奎，曹铁勇，郑云飞，等. 基于特征图关注区域的目标检测对抗攻击方法 [J]. 计算机工程与应用，2023，59(2)：261-270.
WANG Yekui，CAO Tieyong，ZHENG Yunfei，et al. Adversarial attacks for object detection based on region of interest of feature maps [J]. Computer Engineering and Applications，2023，59(2)：261-270.

[21] HU Zhanhao，HUANG Siyuan，ZHU Xiaopei，et al. Adversarial texture for fooling person detectors in the physical world [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscat-

away, NJ, USA: IEEE, 2022: 13297-13306.

[22] LIU Xin, YANG Huanrui, LIU Ziwei, et al. DPatch: an adversarial patch attack on object detectors [EB/OL]. (2019-04-23)[2023-08-18]. https://arxiv.org/abs/1806.02299.

[23] DU A, CHEN Bo, CHIN T J, et al. Physical adversarial attacks on an aerial imagery object detector [C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway, NJ, USA: IEEE, 2022: 3798-3808.

[24] LANG Dapeng, CHEN Deyun, SHI Ran, et al. Attention-guided digital adversarial patches on visual detection [J]. Security and Communication Networks, 2021, 2021: 1-11.

[25] 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. Piscataway, NJ, USA: IEEE, 2016: 2414-2423.

[26] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10)[2023-07-11]. https://arxiv.org/abs/1409.1556.

[27] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization [C]//2017 IEEE International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2017: 618-626.

[28] LOQUERCIO A, MAQUEDA A I, DEL-BLANCO C R, et al. DroNet: learning to fly by driving [J]. IEEE Robotics and Automation Letters, 2018, 3(2): 1088-1095.

[29] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2015: 91-99.

[30] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2016: 779-788.

(编辑 武红江)