面向目标检测的可迁移对抗样本生成算法

向海昀，周垚，陈曦
（西南石油大学计算机与软件学院，四川 成都 610500）

摘　要： 对抗样本的研究能够促进防御方法的创新，查漏补缺，进而提高模型的鲁棒性。现有的目标检测对抗攻击方法的研究大多存在黑盒迁移能力不强、生成的对抗样本泛化能力不足的问题。为解决上述问题，提出了一种提升对抗样本的迁移性和抑制目标检测器正确分类的算法 GM-DEC。首先，将 GridMask 数据增强方法引入基于梯度迭代的对抗样本生成过程中，从而获得更加泛化的梯度信息，有助于增强攻击的鲁棒性，避免陷入局部最优和生成的对抗样本过度拟合白盒模型的情况；其次，为进一步增强对抗样本的迁移性，设计一种基于注意力的关注区域抑制损失函数，通过抑制注意力热图的大小，使得模型关注其他非目标区域，从而做出错误的预测；最后，在迭代更新的过程中引入动量迭代快速梯度符号方法（MI-FGSM）中的动量项，累积速度矢量，从而稳定更新方向，实现更快收敛。在 Pascal VOC2007 数据集上的实验结果表明，所提算法能够有效攻击 Faster R-CNN、YOLO、SSD 等目标检测器，与目前针对目标检测的攻击算法相比黑盒攻击成功率约提升 10～30 百分点，拥有较好的迁移性。

关键词： 目标检测；对抗样本；黑盒攻击；GridMask；注意力抑制
源代码链接： https://github.com/zhouyuuuuuu/GM-DEC
中图分类号： TP391　　**文献标志码：** A　　**DOI：** 10.19678/j.issn.1000-3428.0070322

Transferable Adversarial Example Generation Algorithm for Object Detection

XIANG Haiyun, ZHOU Yao, CHEN Xi
(School of Computer and Software, Southwest Petroleum University, Chengdu 610500, Sichuan, China)

【Abstract】 The study of adversarial examples can promote innovation in defense methods, identify gaps, and thus improve the robustness of a model. Most of the existing studies on object detection against attack methods suffer from poor black-box migration ability and insufficient generalization ability of the generated adversarial examples. To solve these problems, a algorithm called GM-DEC is proposed to enhance the mobility of adversarial examples and inhibit the correct classification of object detectors. First, GridMask, a data augmentation method, is introduced into the gradient iteration-based adversarial example generation process to obtain more generalized gradient information, thereby helping to enhance the robustness of the attack and avoid falling into local optima and overfitting white-box models with generated adversarial examples. Second, to further enhance the transferability of the adversarial examples, an attention-based region-of-attention suppression loss function is designed, which makes the model focus on other non-targeted regions by suppressing the size of the attention heatmap, thus leading to incorrect predictions. Finally, the momentum term in Momentum Iterative-Fast Gradient Sign Method (MI-FGSM) is introduced during the iterative updating process to accumulate velocity vectors, thus stabilizing the updating direction and achieving faster convergence. Experiments are carried out on the Pascal VOC2007 dataset, and the results show that the proposed algorithm can effectively attack object detectors such as Faster R-CNN, YOLO, and SSD. The success rate of the black-box attack is improved by approximately 10-30 percentage point compared with the current attack algorithms for object detection, accompanied by better transferability.

【Key words】 object detection; adversarial example; black-box attack; GridMask; attention suppression

0　引言

随着深度神经网络（DNN）的快速发展，已广泛应用在许多人工智能方面，例如自动驾驶汽车[1]、人体检测[2]、目标跟踪[3]等。然而，DNN 易受到对抗样本[4]的影响，这给人工智能的应用带来了极大的威胁[5]。通过研究对抗样本的特性和生成方法有助于提升深度学习模型的鲁棒性[6]。

对抗样本的研究最初是在图像分类领域中提出的，GOODFELLOW 等[7]提出了快速梯度符号法

基金项目： 国家自然科学基金（61503312）。
作者简介： 向海昀，男，高级实验师、硕士，主研方向为网络与系统安全、深度学习；周垚、陈曦，硕士研究生。
收稿日期： 2024-09-04　　**修回日期：** 2024-12-02　**E-mail：** xhy81358@swpu.edu.cn

(FGSM),该算法沿着梯度上升的方向使得损失函数最大化,让网络趋向于输出错误的结果。动量迭代快速梯度符号方法(MI-FGSM)[8]将动量项整合到攻击的迭代过程中,稳定更新方向并摆脱陷入局部最优值的情况。基本迭代法(BIM)[9]在 FGSM 的基础上对优化器进行多次迭代得到扰动。投影梯度下降(PGD)算法[10]则是在 BIM 的基础上,对原始样本在其邻域范围内随机扰动作为算法初始输入,经多次迭代后生成对抗样本。目前,CHEN 等[11]提出 AdvDiffuser 算法,利用扩散模型的生成和辨别能力,生成自然无限制的对抗样本,通过类激活映射保留显著区域而扰动非重要区域,在保留攻击能力的同时,生成更加隐蔽、自然的对抗样本。XUE 等[12]提出 Diff-PGD 算法,将先验知识结合到对抗样本生成中,利用扩散模型引导的梯度,确保对抗样本接近原始数据分布,旨在提升对抗样本的隐蔽性和可控性。

针对图像分类网络的对抗攻击很难迁移到目标检测网络中[13],鉴于攻击目标检测模型存在着目标种类较多、网络结构复杂等情况,攻击者可以通过攻击非极大值抑制(NMS)机制、区域推荐网络(RPN)等实现对目标检测模型的攻击,如 DAG[14]、RAP[15]、UEA[16]等。由于目标检测模型之间存在较大差异,因此大多数攻击算法可能在某一类检测器上表现出较强的攻击性,却很难迁移到其它检测模型上,黑盒模型中的攻击效果远不如白盒攻击,生成的对抗样本泛化能力不足。

为了提升目标检测对抗样本的黑盒攻击能力和避免生成的对抗样本过度拟合白盒模型,本文提出了一种更具迁移能力的对抗样本生成算法 GM-DEC,针对不同结构的目标检测模型,能够实现非定向攻击并达到较好的攻击成功率。本文的主要研究内容如下。

1)提出一种结合 GridMask[17]扰动数据增强的方法。为解决传统的基于梯度迭代对抗样本生成算法不易收敛、陷入局部最优的问题,提出在迭代过程中结合 GridMask[17]这一数据增强方法,增加梯度的多样性,从而生成更具迁移性的对抗样本。

2)提出关注区域抑制损失函数,通过 Grad-CAM[18](Gradient-weighted Class Activation Mapping)生成的热力图,抑制目标检测模型对关注区域的注意力热力图的大小,从而引导模型预测错误,进一步增强对抗样本的迁移能力。

3)在 Pascal VOC2007[19]数据集上进行白盒攻击、黑盒迁移攻击以及消融实验,验证本文所提算法能够有效提升攻击成功率和黑盒迁移能力,并通过分析扰动数据增强的概率、损失函数的权重、最大扰动值和迭代次数对攻击成功率的影响,给出了超参数设置的建议。

1 相关工作

1.1 目标检测

目标检测是计算机视觉领域的核心问题之一。目标检测是经典的多任务学习[20],将图像或视频中所有感兴趣的目标(物体)提取出来,并输出目标对象的位置及所属类别[21]。现有的基于深度学习的目标检测算法大致可以分为单阶段目标检测、双阶段目标检测和基于 Transformer 的目标检测网络。

单阶段目标检测算法的特点是直接对图像进行计算生成检测结果,将目标检测问题转换为端到端的回归问题,检测速度快。单阶段模型卷积运算的共享程度更高,但是检测精度较低,代表算法有 YOLO[22]、SSD[23]、RetinaNet[24]等。

双阶段目标检测算法也被称为基于区域的处理方法,相较于单阶段目标检测算法,需要对候选结果进行二次修正,检测精度较高,但检测速度较慢,代表算法有 Fast R-CNN[25]、Faster R-CNN[26]等。

基于 Transformer 的目标检测网络不同于前两类基于卷积神经网络(CNN)的目标检测模型,实现了真正意义上的端到端的目标检测方法,将注意力机制引入到目标检测算法中,可以获得更大的感受野和样本间信息交换的能力,代表算法有 DETR[27]、Deformable DETR[28]等。

1.2 针对目标检测的对抗攻击

XIE 等[14]首次将对抗样本的概念从图像分类迁移到目标检测上,并针对 Faster R-CNN[25]提出了 DAG 算法,通过为每个候选区域随机设置一个不同于原目标真实标签的错误标签,引导模型错误分类。LI 等[15]提出 RAP 算法,主要思想是攻击目标检测算法中的公共组件 RPN,设计了一种结合分类损失和位置损失的损失函数,并通过基于梯度的迭代算法进行优化。WEI 等[16]提出 UEA 算法,是一种基于生成对抗网络(GAN)的对抗攻击方法,该方法结合了 GAN 损失,L_2 损失,DAG 的分类损失以及多尺度注意力损失,实现了较好的攻击效果。WANG 等[29]提出基于输入变换的攻击 SIA,通过对每个图像块进行随机变换以获得更泛化的梯度信息。LIU 等[30]提出随机梯度聚合的攻击方法 SGA,采用小批量训练执行内部预搜索的多次迭代,再将所有内部梯度聚合为一步梯度估计,以解决

梯度消失的问题。在物理攻击领域，MIAO 等[31]提出基于扩散模型的对抗补丁攻击 AdvLogo，CHEN 等[32]提出了潜在扩散补丁 LDP，都利用了扩散模型的强大自然能力来润色补丁和图像，以保证高视觉质量和攻击能力。

上述基于目标检测的对抗样本生成算法大多通过攻击某一类检测器或者某一种公共组件，尽管有着较好的白盒攻击效果，但仍然存在着黑盒迁移攻击效果不佳的问题。其中，利用扩散模型生成的对抗样本，更注重对抗样本的隐蔽性，其生成过程依赖于特定的模型架构和参数，限制了对抗样本的泛化性，且扩散模型训练难度高、采样速度慢。而本文方法能够较好地平衡黑盒迁移能力和对抗样本的扰动大小，易于训练，计算效率高。在实际的应用场景中，攻击者并不清楚模型的详细信息，不同的目标检测模型，内部的结构和参数设置也不尽相同，因此针对目标检测的黑盒攻击的研究备受关注。

1.3 GridMask

GridMask[17]是一种数据增强的方法，通过删除图片中的部分信息达到数据增强的效果。该方法通过生成与原图分辨率一致的掩膜，然后将掩膜与原图相乘得到最终增强的图像。通过 GridMask 增强的图像生成过程如式(1)所示：

$$\tilde{x}=x\times M \tag{1}$$

式中：$x\in R^{H\times W\times C}$ 表示输入的图像；$M\in\{0,1\}^{H\times W}$ 表示生成的二进制掩膜；$\tilde{x}\in R^{H\times W\times C}$ 为算法生成的图像。

生成的掩膜中包含 r，d，δ_x，δ_y 这四个参数，通过这四个参数确定了一组特定的掩膜，每组掩膜都是通过平铺单元形成的，虚线方块代表掩膜的一个单元。生成的掩膜中灰色区域的值为 1，黑色区域的值为 0，保留灰色区域的图像信息，删除黑色区域的信息。其中 r 表示黑色区域的边长与单元长度的比值，d 表示一个单元的长度，δ_x，δ_y 分别表示第一个完整的掩膜单元与图像边界之间的距离，如图 1 所示(彩色效果见《计算机工程》官网 HTML 版，下同)。

1.4 Grad-CAM

Grad-CAM[18]是梯度加权类激活映射方法，是一种用于深度学习的技术，特别是在 CNN 中。该方法能够定位到特定的图像区域，通过可视化操作使得神经网络的决策过程更加具有可解释性。通过分析神经网络中最后一个卷积层的梯度信息为每个神经元分配重要值，以此来解码每个特征图对某个

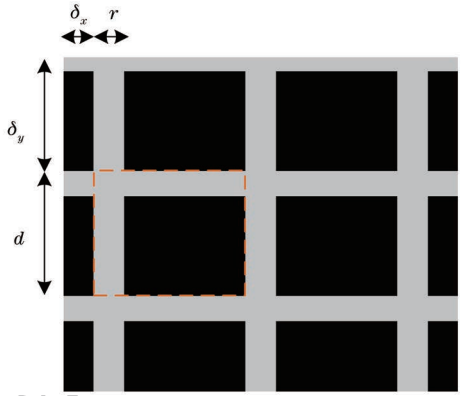


图 1 GridMask 掩膜示意图

Fig. 1 Schematic diagram of GridMask mask

特定类别的重要程度，如式(2)和式(3)所示：

$$\alpha_k^c=\frac{1}{Z}\sum_i\sum_j\frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

$$L_{\text{Grad-CAM}}^c=\text{ReLU}\left(\sum_k\alpha_k^cA^k\right) \tag{3}$$

式中：c 表示类别；y^c 表示类别 c 的 logits；k 表示特征图 A 的通道；i，j 分别为 A 的横、纵坐标；Z 为 A 的长宽乘积。

2 方法实现

2.1 问题定义

假设原始干净图像为 x，通过向干净图像中增加扰动 δ 得到对应的对抗样本表示为 $x^{\text{adv}}=x+\delta$，限制扰动 δ 的 L_p 范数不超过 ε，其中 p 可取 0，1，2，∞。定义输入原始图像 x 得到的预测输出为 $D(x)=(B,C,S)$，对抗样本 x^{adv} 得到的预测输出为 $D(x^{\text{adv}})=(B',C',S')$，其中 B 和 B' 表示预测得到的边界框，C 和 C' 表示对应的类别，S 和 S' 表示对应的目标置信度。通常一张图像中会存在多个含有类别概率的目标检测框 $D=\{d_1,d_2,\cdots,d_n\}$，$d_i=(b_i^x,b_i^y,b_i^w,b_i^h,S_i,\boldsymbol{p}_i)$，其中 $\boldsymbol{p}_i$ 表示类别概率向量，$\boldsymbol{p}_i=\{p_i^1,p_i^2,\cdots,p_i^c\}$，$\boldsymbol{p}_i$ 中的最大值为最大可能性的类别。

本文的目标是令生成的对抗样本能够实现白盒攻击与黑盒迁移攻击，是一种非定向攻击。将 x^{adv} 作为目标检测器 D 的输入，使得模型对输入的类别判断错误或预测的位置坐标与真实标签的交并比(IoU，R_{IoU})值低于阈值即为攻击成功，如式(4)所示：

$$R_{\text{IoU}}(B',B)<0.5\ \text{or}\ C'\neq C \tag{4}$$

2.2 本文方法

在第 1.2 节中介绍了现有针对目标检测的对抗样本生成算法，这些方法在黑盒环境下的迁移能力

较弱，容易过度拟合白盒模型，导致在真实场景中的攻击成功率较低。针对以上问题，从数据增强和抑制注意力热图大小的角度对现有方法进行改进，算法流程如图 2 所示。

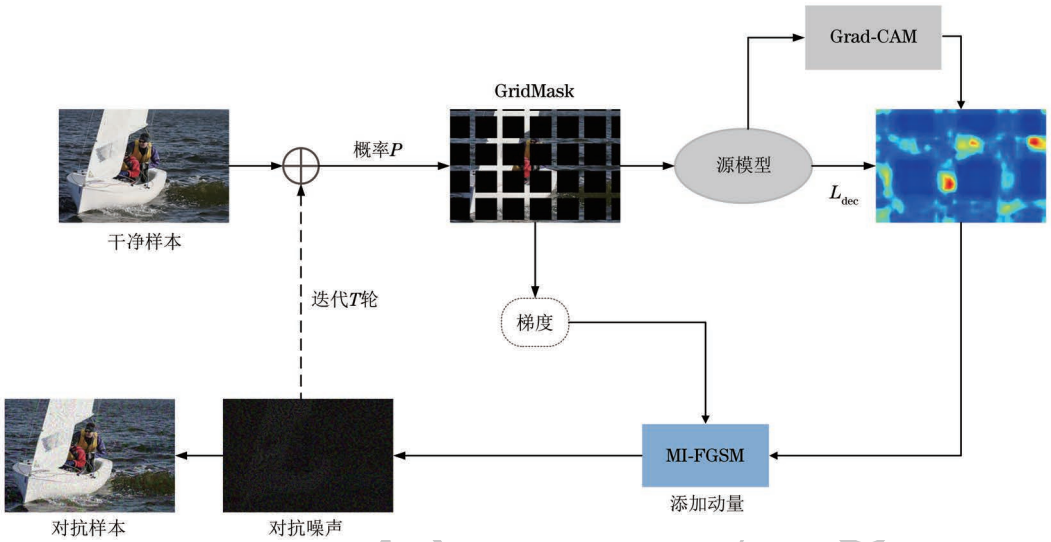


图 2 GM-DEC 算法流程图

Fig.2 Flowchart of GM-DEC algorithm

该研究首先通过在迭代的过程中引入 GridMask[17] 数据增强方法，在训练过程中可以模拟出更丰富的输入分布，从而提高对抗样本的泛化能力和跨模型的迁移性，增强对抗样本的鲁棒性，提高攻击的成功率；接着，对数据增强后的样本生成相应的注意力热图，使模型不再关注目标区域，通过关注区域抑制损失来抑制注意力热图的大小，分散模型的注意力；最后，在更新迭代的过程中引入动量法，稳定更新方向，直至达到最大迭代次数，生成最终的对抗样本。

2.2.1 基于 GridMask 的扰动数据增强

GridMask[17]、Random Erasing[33]、CutOut[34] 等属于“信息删除”这一类数据增强的方法，但后两种方法则是随机删除一个矩形区域，对删除区域用其它纯色填充。其中 CutOut 删除的区域有 50%的概率不在原图像中，当图像中目标对象较小时可能会删除全部关键信息，或者随机删除时将背景删除而保留了全部关键信息，以上情况会导致无法达到预期的增强效果。而 GridMask 则是在此基础上进行了改进，通过生成平铺单元形成的掩膜，对删除信息和保留信息进行平衡，避免了图中上下文信息缺失等情况，通过对部分关键信息的删除能够使 CNN 学习到更多不敏感或不重要的信息，以此扩大感知域，提升模型的鲁棒性。

基于此，本文将 GridMask[17] 这一数据增强的方法集成到基于梯度迭代的攻击算法中，在迭代生成对抗样本的过程中以变换概率 P 进行数据增强，并通过控制参数 d 和 r 生成不同的掩膜，从而获得更泛化的梯度信息，避免在更新过程中陷入局部最优的情况。具体计算公式如式(5)～式(7)所示：

$$d=\text{random}(2,\min(H,W)),r=0.5 \tag{5}$$

$$l=\min(\max(\text{int}(d\cdot r+0.5),1),d-1) \tag{6}$$

$$g_0=0,\ g_t=\nabla_{x_t^{\text{adv}}}J(G_n(x_t^{\text{adv}},P),D) \tag{7}$$

式中：d 为取‘2’到图像的较小维度之间的随机整数，使得生成的网格不会太小也不会太大，适应于输入图像的大小；r 决定了输入图像的保持比，折中取 $r=0.5$ 平衡图像信息被删除的比例，以避免因 r 太大或者太小而导致过拟合和欠拟合；l 表示被删除信息，即单个黑色区域的边长，由参数 d 和 r 共同决定；$G_n(\cdot)$ 表示生成不同掩膜的 GridMask 数据增强操作。通过增加在训练过程中对抗样本的多样性，使得扰动得更加均匀，能够生成更具迁移性的对抗样本。

2.2.2 关注区域抑制损失

人们在对一张图像进行判断时往往会更加关注它的关键部分，这种注意力机制已经广泛用于自然语言处理和计算机视觉中。尽管不同 DNN 的整体架构可能有所不同，但许多的深度学习模型（如 Transformer、CNN 等）使用的注意力机制在层级结构、特征选择上有共性。因此，针对注意力机制设计的攻击损失函数所生成的对抗样本，能够在不同的架构和网络中实现较好的迁移。

本文通过 Grad-CAM[18] 来获取图像中的目标区域热力图，由于图像分类任务中一张图像只包含

一个目标和对应的类别，而目标检测任务中图像往往包含多个目标，因此不能简单地挪用图像分类中获取热力图的方式，更改后获取热力图的具体计算公式如式(8)～式(10)所示：

$$\alpha_k^{c_{d_i}}=\frac{1}{Z}\sum_i\sum_j\frac{\partial D^{c_{d_i}}}{\partial A_{ij}^k} \tag{8}$$

$$L_{\text{Grad-CAM}}^{c_{d_i}}=\text{ReLU}\left(\sum_k\alpha_k^{c_{d_i}}A^k\right) \tag{9}$$

$$\boldsymbol{H}=\max_{d_i}(L_{\text{Grad-CAM}}^{c_{d_i}}) \tag{10}$$

式中：c_{d_i} 表示目标 d_i 的类别索引；A^k 表示卷积层输出的特征图 A 中的第 k 通道；$\alpha_k^{c_{d_i}}$ 表示目标 d_i 在通道 k 上的权重；$D^{c_{d_i}}$ 表示类别得分；i，j 分别表示特征图 A 的横纵坐标；Z 为正规化因子，使用计算出的 α 权重并结合特征图 A，由式(9)计算后得到单个目标的热图 $L_{\text{Grad-CAM}}^{c_{d_i}}$。为生成包含所有目标的热图，逐像素对所有目标 d_i 生成的热图取最大值，得到最终合并后的热图 $\boldsymbol{H}$。

假设 $\boldsymbol{H}(x,B,C,S)$ 表示输入图像 x 和对应的含有类别概率的目标检测框 $D(x)=(B,C,S)$ 的注意力热图，$\boldsymbol{H}(x,B_{\text{ori}},C_{\text{ori}},S_{\text{ori}})$ 是一个维度与 x 一致的张量。在迭代过程中，不断生成相应阶段的目标区域热力图，本文通过对当前的热力图施加 L_1 范数约束，使得 Grad-CAM 所生成的热力图中的权重被重新调整。相比于 L_2 范数约束倾向于均匀地缩小所有权重值，无法集中干扰模型对关键区域的注意力。本文采用 L_1 范数施加的正则化，使得一些高响应区域的值被压缩到较小范围，甚至接近零，同时较小值区域也会被进一步压缩，热力图的注意力值呈现更加平滑的分布。随着迭代次数的增加，逐步修正注意力偏差，对关注区域所施加的 L_1 范数约束也会逐渐累积，以此逐步达到攻击的效果。迭代完成后，当高注意力区域与低注意力区域之间的对比度减弱时，原本集中在关键区域的注意力被扩散到更广泛的区域，模型的聚焦能力也随之减弱。计算过程如下：

$$L_{\text{dec}}=\|\boldsymbol{H}(x,B_{\text{ori}},C_{\text{ori}},S_{\text{ori}})\|_1 \tag{11}$$

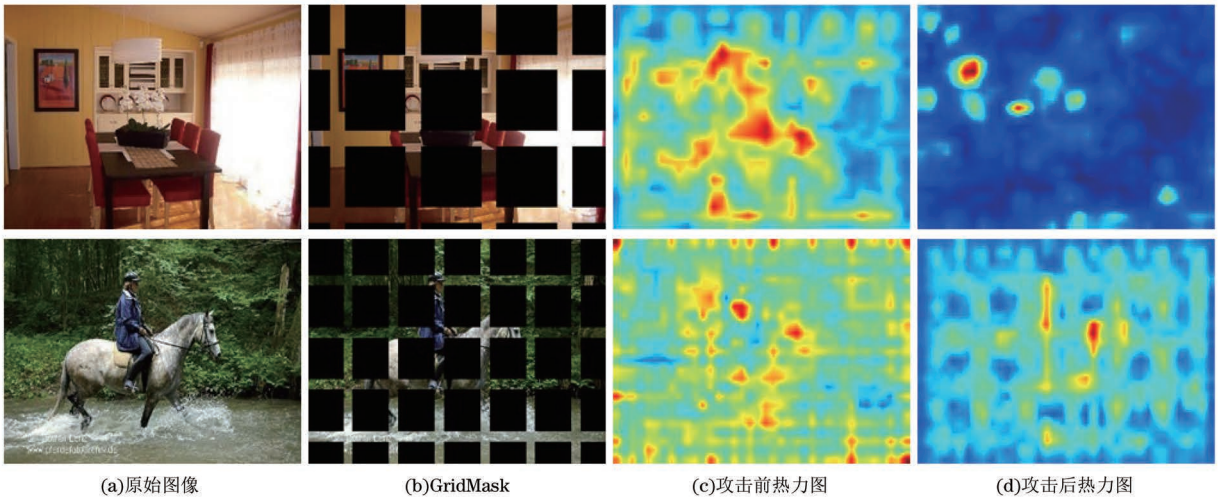
通过最小化 L_{dec} 损失函数，实现对抗样本的更新，为了提高算法的攻击成功率和收敛速度，结合 MI-FGSM[8] 算法中的动量法，在迭代过程中沿损失函数的梯度方向累积速度矢量来影响梯度更新的速度和方向，具体更新的过程描述如式(12)和式(13)所示：

$$g_0=0,g'=\nabla_{x_t^{\text{adv}}}L_{\text{dec}}(G_n(x_t^{\text{adv}},P),D) \tag{12}$$

$$g_{t+1}=\mu\cdot g_t+\frac{g'}{\|g'\|} \tag{13}$$

式中：g' 表示每次累积的梯度；μ 为 g_t 的衰减因子。

如图 3 所示，其中图 3(a)为原始图片，图 3(b)为 GridMask[17] 数据增强后的图片，图 3(c)为关注区域抑制损失攻击前的热力图，图 3(d)为攻击后的热力图。从图 3 可以看出，在关注区域抑制损失 L_{dec} 的攻击下，目标区域的热力图明显缩小，使得目标检测器丧失预测的能力。



(a)原始图像 (b)GridMask (c)攻击前热力图 (d)攻击后热力图

图 3 关注区域抑制损失攻击示意图

Fig. 3 Schematic diagram of suppression loss attack in the region-of-interest

2.3 算法流程

GM-DEC 的具体攻击流程如算法 1 所示，将原始干净样本 x 进行 GridMask 扰动数据增强之后，再通过关注区域抑制损失 L_{dec} 攻击注意力热图，最后累积速度矢量进一步稳定更新方向，避免振荡。在迭代过程中不断更新梯度，并对图像进行修改，直至达到最大迭代次数，最终生成对抗样本 x_t^{adv}。

算法 1 GM-DEC

输入 干净样本 x，目标检测模型 D，扰动 δ，最大扰动值 ε，迭代轮数 T，GridMask 数据增强 $G_n(\cdot)$，样本增强概率 P，攻击步长 α，衰减因子 μ

输出 对抗样本 x_t^{adv}

1. $x_0^{adv}=x, g_0=0, t=0, M_n=0$//初始化

2. while t=0 to T−1 do:

3. $g_t=\nabla_{x_t^{adv}} J(G_n(x_t^{adv},P),D)$//GridMask 扰动数据//增强

4. $H_t=\max_{d_i}(L_{Grad\text{-}CAM}^{c_{d_i}})$//获取热力图

5. $L_{dec}=\|H(x,B_{ori},C_{ori},S_{ori})\|_1$//攻击关注区域的//热图

6. $g'=\nabla_{x_t^{adv}} L_{dec}(G_n(x_t^{adv},P),D)$

7. $g_{t+1}=\mu\times g_t+\frac{g'}{\|g'\|_2}$//添加动量

8. $x_{t+1}^{adv}=\mathrm{clip}_{\varepsilon}\left(x_t^{adv}-\alpha\frac{g(x_t^{adv})}{\|g(x_t^{adv})\|_1/N}\right)$//对抗样本//更新

9. t=t+1

10. end while

11. return $x^{adv}=x_t^{adv}$

3 实验与结果分析

3.1 实验设置

1)数据集。本文实验使用公开数据集 Pascal VOC2007[19]。该数据集主要用于目标检测、图像分类和语义分割等任务，其中共有 9 963 张图片，包含 24 640 个已标注的目标，共 20 个类别。这些类别覆盖了日常生活中常见的物体种类，例如动物、交通工具、家具等。数据集中各类别分布相对均衡，大多数图像中存在多个目标对象，更符合目标检测任务的使用场景。数据集分为两个部分，训练集和验证集 trianval，测试集 test，两个部分的数据各占总数据集的 50%左右。实验时对数据集中的图片遵循不同预训练模型的预处理步骤。

2)目标检测模型。选取 6 个目标检测模型作为攻击的目标，分别是：Faster R-CNN[26]（ResNet50）、YOLOv3[22]（DarkNet）、YOLOv9[35]（EfficientRepv2）、SSD300[23]（VGG16）、RetinaNet[24]（ResNet50）、Deformable DETR[28]（ResNet50）。测试模型中既包含基于单阶段和双阶段的检测模型，还包含基于 Transformer 结构的，能够充分验证算法的有效性。

3)对比算法。本文选取了 MI-FGSM[8]、DAG[14]、RAP[15]、SIA[29]、SGA[30]这 5 种攻击算法作为对照实验。上述对比算法均根据官方代码进行复现，通过将所提算法与 5 种基线攻击方法做对比，进一步验证本文算法的有效性。

4)评价指标。本文使用攻击成功率（ASR，R_{ASR}）作为评价指标，评估方法的有效性。ASR 由目标检测中常用的均值平均精度（mAP，m_{mAP}）计算而来。其中 m_{mAP} 的计算公式如式(14)所示：

$$m_{mAP}=\frac{\sum_{i=1}^{C}A_{AP_i}}{C} \tag{14}$$

式中：A_{AP} 表示一个类别的平均精度；C 表示类别数量。R_{ASR} 的计算公式如式(15)所示：

$$R_{ASR}=1-\frac{m_{mAP_{adv}}}{m_{mAP_{clean}}} \tag{15}$$

式中：$m_{mAP_{adv}}$ 表示对抗样本输入检测模型时得到的 mAP 值；$m_{mAP_{clean}}$ 表示干净样本输入检测模型时得到的 mAP 值。从计算公式可以看出，R_{ASR} 值越大，代表攻击算法的效果越好。

5)实验环境与参数设置。本文算法在深度学习框架 PyTorch 2.1.2 上运行，显卡配置为 NVIDIA Tesla V100 32 GB，Batch Size 为 1，攻击迭代次数为 100 次，最大扰动值为 16，GridMask 扰动增强概率为 0.6，关注区域抑制损失函数权重 $\omega=0.1$。

3.2 白盒攻击对比实验及结果分析

为了验证本文所提方法的白盒攻击能力，通过在 6 种不同的白盒场景下，将 5 种基线攻击方法与所提方法进行白盒攻击实验对比 ASR 值，如表 1 所

表 1 不同方法的白盒攻击成功率对比

Table 1 Comparison of white-box attack success rates among different methods

方法	ASR/%					
	MI-FGSM	DAG	RAP	SIA	SGA	GM-DEC(Ours)
Faster R-CNN	90.3	93.1	91.7	94.2	96.5	**97.6**
YOLOv3	82.1	88.5	92.1	93.6	92.7	**94.2**
YOLOv9	47.2	43.2	50.7	62.4	66.9	**72.3**
RetinaNet	79.8	76.3	83.6	87.2	91.7	**92.4**
SSD300	73.4	77.2	80.8	80.1	81.2	**82.5**
Deformable DETR	45.8	52.3	42.6	65.7	69.5	**76.0**

示，加粗表示最优数据(下同)。其中，GM-DEC 在 Faster R-CNN[26] 上的白盒攻击成功率达到 97.6%，相较于另外 5 种对比算法分别提高了 7.3、4.5、5.9、3.4、1.1 百分点。在 Deformable DETR[28]这一基于 Transformer 的目标检测器上攻击成功率达到 76.0%，比 5 个基线方法分别提高了 30.2、23.7、33.4、10.3、6.5 百分点。从表 1 可以看出，所提方法在白盒模型上拥有较好的攻击成功率，在白盒攻击上优于其他 5 种基线方法。

3.3 黑盒攻击对比实验及结果分析

在白盒攻击中，GM-DEC 展现了较好的攻击能力。为进一步验证所提方法的黑盒迁移能力，本文分别在 6 种不同的目标检测模型中测试其余 5 种检测模型的黑盒攻击成功率，如表 2 所示，将本文所提方法与 5 种对比方法的攻击成功率进行对比。

表 2 黑盒攻击成功率对比

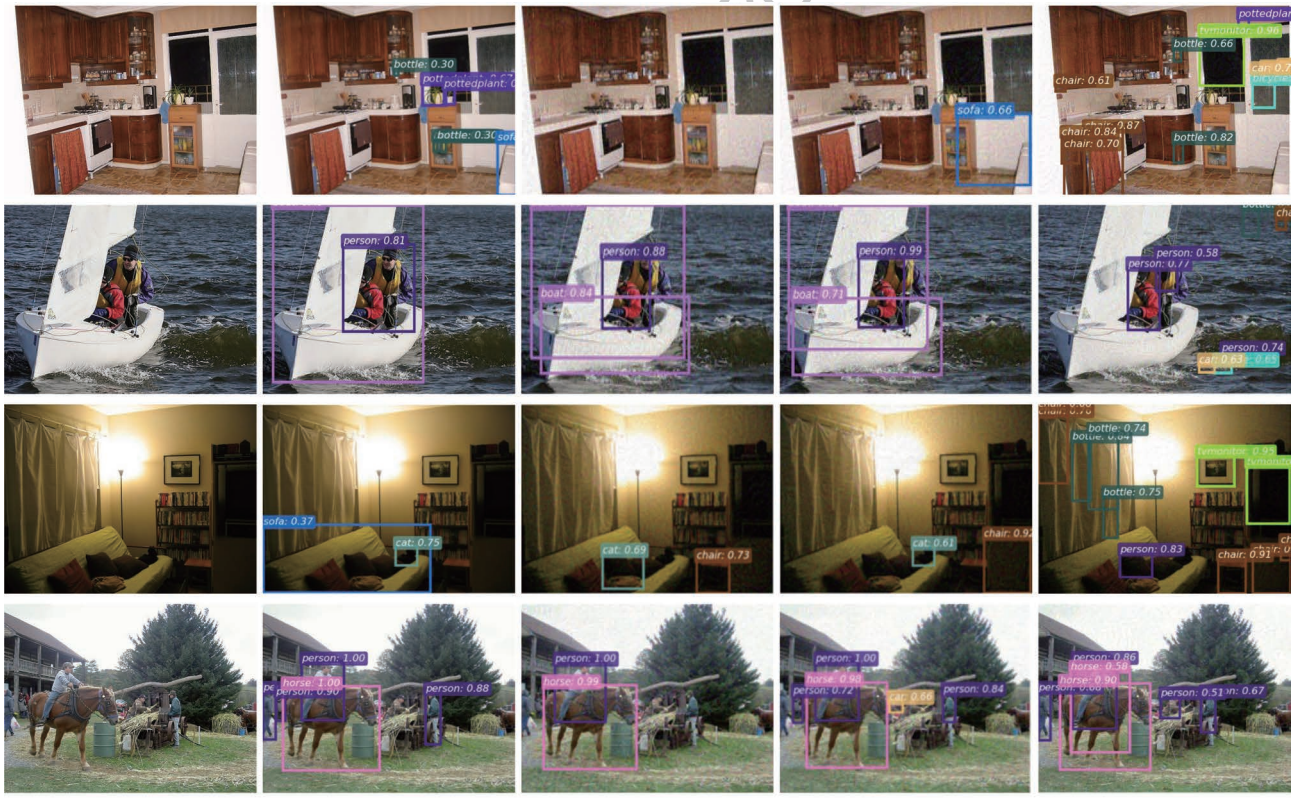
Table 2 Comparison of black-box attack success rate

白盒模型	黑盒模型	ASR/%					
		MI-FGSM	DAG	RAP	SIA	SGA	GM-DEC(Ours)
Faster R-CNN	YOLOv3	15.8	4.2	11.3	31.6	34.7	**40.5**
	YOLOv9	2.7	4.7	4.1	17.8	19.6	**31.4**
	RetinaNet	44.3	53.1	47.3	67.4	64.2	**87.2**
	SSD300	4.6	2.5	2.7	27.9	33.1	**46.3**
	Deformable DETR	21.6	23.2	14.8	43.1	48.2	**63.8**
YOLOv3	Faster R-CNN	3.2	6.9	5.8	15.4	14.7	**24.5**
	YOLOv9	9.6	13.9	16.8	34.1	38.2	**39.7**
	RetinaNet	14.7	9.7	10.1	27.3	24.8	**32.9**
	SSD300	6.9	8.3	4.1	20.2	25.1	**36.4**
	Deformable DETR	1.8	3.2	7.9	12.9	14.2	**25.2**
YOLOv9	Faster R-CNN	1.3	1.6	3.4	22.4	25.1	**26.7**
	YOLOv3	9.6	11.3	12.1	32.8	36.3	**45.1**
	RetinaNet	5.3	4.1	7.6	27.1	26.8	**32.5**
	SSD300	2.1	2.4	4.5	12.4	15.2	**17.8**
	Deformable DETR	1.4	3.5	3.8	14.6	14.1	**24.2**
RetinaNet	Faster R-CNN	67.3	58.5	62.3	75.2	81.6	**83.8**
	YOLOv3	4.3	7.7	9.6	26.1	24.5	**33.8**
	YOLOv9	2.3	1.3	3.7	13.6	15.2	**18.6**
	SSD300	11.3	2.8	1.9	22.8	24.7	**34.9**
	Deformable DETR	20.2	17.7	22.1	37.6	41.2	**48.0**
SSD300	Faster R-CNN	24.2	14.5	17.2	48.5	44.3	**63.4**
	YOLOv3	9.6	7.8	9.1	23.0	27.1	**37.6**
	YOLOv9	2.9	2.1	3.5	13.4	16.9	**17.8**
	RetinaNet	15.7	9.1	4.8	32.1	38.4	**45.4**
	Deformable DETR	4.9	3.6	2.1	12.4	14.6	**18.1**
Deformable DETR	Faster R-CNN	15.6	16.3	15.8	30.5	37.2	**45.8**
	YOLOv3	8.0	6.3	5.1	17.4	21.0	**27.6**
	YOLOv9	1.3	2.7	2.5	16.5	15.3	**21.2**
	RetinaNet	27.5	23.9	25.7	48.1	54.7	**59.6**
	SSD300	2.1	1.6	2.4	11.3	17.2	**20.8**

从表 2 分析可知，以不同的目标检测模型作为白盒模型攻击其余 5 个黑盒模型时，与基线方法相比，本文所提出的方法在黑盒设置下的攻击成功率得到明显提升。GM-DEC 在以 Faster R-CNN 为白盒模型生成的对抗样本攻击 RetinaNet 和 Deformable DETR 时的攻击成功率分别达到 87.2%和 63.8%。针对 RetinaNet 生成的对抗样本攻击其余模型时，GM-DEC 较 SIA 和 SGA 分别提高了 8.8 和 6.4 百分点。从表 2 可以看出，模型拥有不同的骨架网络也是影响攻击成功率的因素，Faster R-CNN、RetinaNet 和

Deformable DETR 都是以 ResNet 作为骨架网络，黑盒攻击效果明显优于另外 3 种采用不同骨架网络的检测器。Deformable DETR 是一种基于 Transformer 的目标检测模型，拥有多尺度可变形注意力模块，因此在不同的网络模型中生成的对抗样本攻击 Deformable DETR 时普遍存在攻击成功率较低的情况，但本文所提方法较其它 5 个基线方法仍有明显提升。以 YOLOv9 为白盒模型生成的对抗样本攻击 Deformable DETR 的攻击成功率较基线方法分别提高了 22.8、20.7、20.4、9.6、10.1 百分点。综上所述，在黑盒迁移攻击的实验中，即使在不同的骨架网络和不同结构的检测模型中，本文所提方法能够有效提升黑盒迁移能力。

SIA、SGA 与本文所提方法的攻击效果图如图 4 所示。以 Faster R-CNN 为白盒模型所生成的对抗样本攻击 YOLOv3 模型为例，从图 4 可以看出，相较于另外两种攻击方法，GM-DEC 生成的对抗样本对目标框和类别有明显的攻击效果，类别置信度下降幅度更大，同时还能够误导目标检测器输出大量错误的检测框和类别。



(a)原始图像 (b)原始图像检测结果 (c)SIA攻击检测结果 (d)SGA攻击检测结果 (e)GM-DEC攻击检测结果

图 4 不同攻击方法攻击效果对比示意图

Fig. 4 Comparison of the attack effect among different attack methods

3.4 消融分析

本节通过消融分析进一步验证所提算法的有效性，分析所提方法中添加的策略对攻击成功率的影响，采用 Faster R-CNN 作为白盒模型生成对抗样本，选择 MI-FGSM 作为基线模型，具体实验结果如表 3 所示。

表 3 消融实验攻击成功率对比

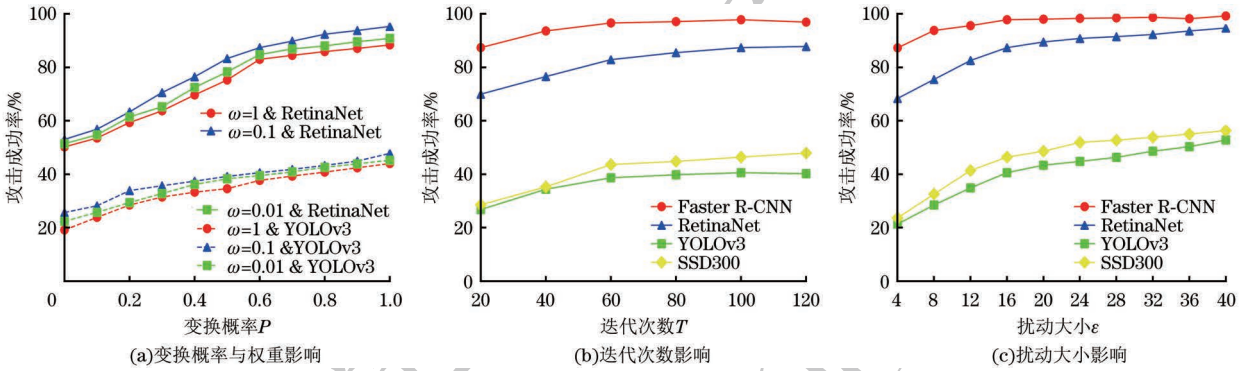
Table 3 Comparison of attack success rates of ablation experimental

模型	ASR/%			
	MI-FGSM	MI-FGSM+L_{dec}	MI-FGSM+GridMask	GM-DEC(Ours)
Faster R-CNN	90.3	91.8	93.4	**97.6**
YOLOv3	15.8	25.6	30.7	**40.5**
YOLOv9	2.7	11.8	16.5	**31.4**
RetinaNet	44.3	52.9	71.3	**87.2**
SSD300	4.6	21.4	29.2	**46.3**
Deformable DETR	21.6	47.3	53.2	**63.8**

由表 3 中的实验结果可知，在 MI-FGSM 的基础上引入关注区域抑制损失 L_{dec} 后，攻击成功率平均提升约 12.0 百分点；在 MI-FGSM 上增加 GridMask 扰动数据增强策略后攻击成功率提升约 19.2 百分点；GM-DEC 在合并两种策略之后能达到更好的攻击效果，攻击成功率平均提升 31.3 百分点。即使是在攻击 YOLOv3 和 SSD300 这两个使用不同骨架网络的目标检测模型时，攻击成功率较基线模型分别提高了 24.7 和 41.7 百分点。综上可以得出结论，本文所提方法能够实现较好白盒攻击成功率的同时，还能提升对抗样本的黑盒迁移能力。

3.5 超参数分析

本节将讨论 GM-DEC 算法中不同超参数对攻击成功率的影响，分别包括 GridMask 扰动数据增强的变换概率 P，L_{dec} 损失函数的权重取值 ω，迭代次数 T 和最大扰动值 ε，如图 5 所示。



(a)变换概率与权重影响　(b)迭代次数影响　(c)扰动大小影响

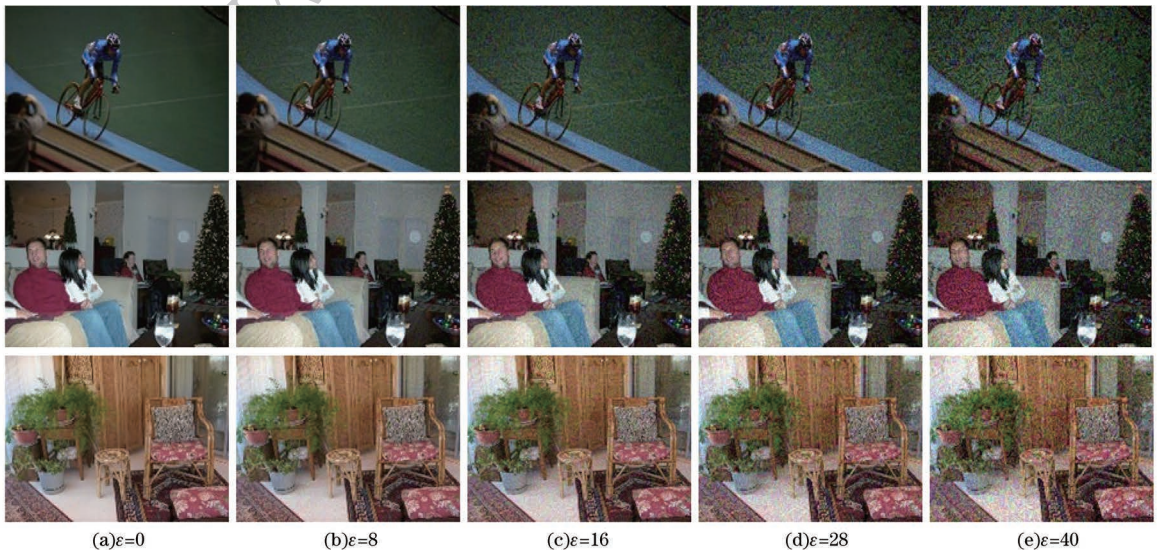
图 5　各参数对攻击成功率的影响

Fig. 5　Influence of each parameter on the attack success rate

图 5(a)讨论在以 Faster R-CNN 为白盒模型分别攻击 RetinaNet 和 YOLOv3 时攻击成功率的变化，设置 P 的变换范围为 0～1，增长幅度为 0.1，ω 取值设为 1，0.1 和 0.01。通过横向对比可知，随着 P 的增加，攻击成功率呈上升趋势，为了平衡攻击成功率和资源是否充足，选择 $P=0.6$ 作为扰动数据增强的变换概率。通过纵向对比可知，当 L_{dec} 损失函数的权重 $\omega=0.1$ 时的攻击成功率比 $\omega=1$ 和 $\omega=0.01$ 时攻击成功率更高。

关于迭代次数 T 对攻击成功率的影响，由图 5(b)可知，随着迭代次数的增加，攻击成功率逐步提升，当 $T=100$ 时曲线较平缓，本文所提出的方法 GM-DEC 能够避免在迭代更新的过程中陷入局部最优的情况，使得攻击成功率保持相对稳定。因此，本文选择 $T=100$ 作为算法的最大迭代次数，生成更具迁移性的对抗样本。

关于最大扰动值 ε 对攻击成功率的影响，设置最大扰动值的范围为 4～40，增长幅度为 4。如图 5(c)所示，随着扰动值的增大，攻击成功率逐步提升，与此同时，对抗样本的噪声也越明显。不同扰动大小生成的对抗样本如图 6 所示。因此综合考虑选择 $\varepsilon=16$，既能实现较高的攻击成功率还能保证



(a)ε=0　(b)ε=8　(c)ε=16　(d)ε=28　(e)ε=40

图 6　不同扰动大小生成的对抗样本

Fig. 6　Adversarial examples generated with different perturbation sizes

生成的对抗样本不会过分失真。

4 结束语

本文提出一种针对目标检测的对抗样本生成算法 GM-DEC。针对生成的对抗样本容易过度拟合白盒模型、易陷入局部最优的问题，通过结合 GridMask 对样本进行数据增强，学习更泛化的梯度信息，提高对抗样本的迁移能力。此外，针对不同的目标检测器对关注区域拥有相似的注意力，提出一种抑制关注区域的损失函数 L_{dec}，引导模型将注意力转移至非目标区域，从而提升攻击成功率。实验结果表明，不论是在白盒攻击还是黑盒攻击上，本文所提出的算法都优于对比算法，能够有效提升攻击成功率和黑盒迁移能力。下一步将研究在黑盒迁移攻击领域中，将对抗样本从数字世界迁移到物理世界中，生成更具迁移性和攻击性的对抗样本。

参考文献

[1] WANG H W, ZHU B L, LI Y J, et al. SYGNet: a SVD-YOLO based GhostNet for real-time driving scene parsing [C]// Proceedings of the IEEE International Conference on Image Processing (ICIP). Bordeaux, France: IEEE Press, 2022: 2701-2705.

[2] HOU C C. The application of human detection based on YOLOv5 [J]. Highlights in Science, Engineering and Technology, 2023, 34: 203-208.

[3] BI H B, ZHANG C, WANG K, et al. Rethinking camouflaged object detection: models and datasets[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 5708-5724.

[4] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/1312.6199.

[5] ZHU K J, HU X X, WANG J D, et al. Improving generalization of adversarial training via robust critical fine-tuning [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE Press, 2024: 4401-4411.

[6] 赵宏，宋馥荣，李文改. 基于 SE-AdvGAN 的图像对抗样本生成方法研究[J]. 计算机工程，2025，51(2)：300-311.
ZHAO H, SONG F R, LI W G. Research on image adversarial example generation method based on SE-AdvGAN[J]. Computer Engineering, 2025, 51(2): 300-311. (in Chinese)

[7] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL]. [2024-07-29]. https://arxiv.org/pdf/1412.6572.

[8] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Press, 2018: 9185-9193.

[9] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/1607.02533.

[10] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2024-07-29]. https://arxiv.org/pdf/1706.06083.

[11] CHEN X Q, GAO X T, ZHAO J J, et al. AdvDiffuser: natural adversarial example synthesis with diffusion models [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE Press, 2024: 4539-4549.

[12] XUE H, ARAUJO A, HU B, et al. Diffusion-based adversarial sample generation for improved stealthiness and controllability [J]. Advances in Neural Information Processing Systems, 2024, 36:21-30.

[13] ZHANG Y, GONG Z Q, ZHANG Y C, et al. Boosting transferability of physical attack against detectors by redistributing separable attention[J]. Pattern Recognition, 2023, 138: 109435.

[14] XIE C H, WANG J Y, ZHANG Z S, et al. Adversarial examples for semantic segmentation and object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy:IEEE Press, 2017: 1378-1387.

[15] LI Y Z, TIAN D, CHANG M C, et al. Robust adversarial perturbation on deep proposal-based models[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/1809.05962.

[16] WEI X X, LIANG S Y, CHEN N, et al. Transferable adversarial attacks for image and video object detection[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/1811.12641.

[17] CHEN P G, LIU S, ZHAO H S, et al. GridMask data augmentation[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/2001.04086.

[18] 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). Venice, Italy: IEEE Press, 2017: 618-626.

[19] EVERINGHAM M. The Pascal visual object classes challenge, (VOC2007) results [EB/OL]. [2024-07-29]. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/index.html., 2007.

[20] 袁珑，李秀梅，潘振雄，等. 面向目标检测的对抗样本综述[J]. 中国图象图形学报，2022，27(10)：2873-2896.
YUAN L, LI X M, PAN Z X, et al. Review of adversarial examples for object detection [J]. Journal of Image and Graphics, 2022, 27(10): 2873-2896. (in Chinese)

[21] 汪欣欣，陈晶，何琨，等. 面向目标检测的对抗攻击与防御综述[J]. 通信学报，2023，44(11)：260-277.
WANG X X, CHEN J, HE K, et al. Survey on adversarial attacks and defenses for object detection [J]. Journal on Communications, 2023, 44(11): 260-277. (in Chinese)

[22] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE Press, 2016: 779-788.

[23] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector [C]// Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 21-37.

[24] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE Press, 2017: 2999-3007.

[25] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE Press, 2016: 1440-1448.

[26] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[27] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers [C]// Proceedings of European Conference on Computer Vision. Berlin, Germany:

Springer, 2020: 213-229.
[28] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: deformable transformers for end-to-end object detection[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/2010.04159.
[29] WANG X S, ZHANG Z L, ZHANG J P. Structure invariant transformation for better adversarial transferability[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE Press, 2024: 4584-4596.
[30] LIU X N, ZHONG Y Y, ZHANG Y H, et al. Enhancing generalization of universal adversarial perturbation through gradient aggregation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE Press, 2024: 4412-4421.
[31] MIAO B M, LI C X, ZHU Y, et al. AdvLogo: adversarial patch attack against object detectors based on diffusion models[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/2409.07002.
[32] CHEN X Y, LIU F Z, JIANG D, et al. Natural adversarial patch generation method based on latent diffusion model[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/2312.16401.
[33] ZHONG Z, ZHENG L, KANG G L, et al. Random erasing data augmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI Press, 2020: 13001-13008.
[34] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/1708.04552.
[35] WANG C Y, YEH I H, LIAO H M. YOLOv9: learning what you want to learn using programmable gradient information[EB/OL]. [2024-07-29]. https://arxiv.org/pdf/2402.13616.

文字编辑 薛晋栋
栏目编辑 宋 圆