DOI:10.19651/j.cnki.emt.2519747

基于多尺度均衡正则的对抗补丁攻击方法

谢家乐[1,2] 赵宇熙[1] 曾念寅[2] 王 若[1]
(1.中国航空研究院 北京 100086;2.厦门大学人工智能学院 厦门 361005)

摘 要:目标检测模型在对抗补丁攻击下表现出显著脆弱性,严重威胁其在自动驾驶与安防等场景中的应用安全。现有基于迁移的黑盒攻击方法虽取得一定进展,但普遍存在跨模型迁移性不足以及在多尺度检测头间抑制不均衡的问题。针对这一挑战,本文提出一种基于多尺度均衡正则的对抗补丁攻击方法(MSBR)。该方法在补丁训练过程中显式约束不同尺度检测头置信度输出的方差,从而实现对各尺度目标的一致性抑制,有效缓解了尺度抑制不均的现象,显著提升了对抗补丁的跨模型迁移能力。在多个主流目标检测器上的实验结果表明,所提方法在保持攻击成功率的同时,黑盒迁移性能优于现有代表性方法(如 T-SEA),验证了 MSBR 在提升补丁攻击实用性方面的有效性。本文的研究为面向复杂检测结构的对抗补丁攻击提供了新的思路。

关键词:对抗攻击;目标检测;多尺度均衡正则
中图分类号:TP391.41;TN06 **文献标识码**:A **国家标准学科分类代码**:520.2050

Multi-scale balanced regularization method for adversarial patch attacks

Xie Jiale[1,2] Zhao Yuxi[1] Zeng Nianyin[2] Wang Ruo[1]
(1. Chinese Aeronautical Establishment, Beijing 100086, China; 2. Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, China)

Abstract: Object detection models are markedly vulnerable to adversarial patches, posing serious safety risks to applications such as autonomous driving and security surveillance. Although transfer-based black-box attacks have made progress, they often suffer from poor cross-model transferability and uneven suppression across multi-scale detection heads. To address these issues, we propose MSBR for adversarial patch attacks. During patch training, MSBR explicitly regularizes the variance of confidence outputs across different detection scales, thereby enforcing consistent suppression of targets at multiple scales, mitigating scale-wise imbalance, and substantially improving cross-model transferability. Experiments on several mainstream detectors show that our method maintains strong attack success rates while outperforming representative approaches (e. g. T-SEA) in black-box transfer performance, demonstrating the practical effectiveness of MSBR. This work provides a new perspective for designing adversarial patch attacks against complex multi-scale detection architectures.

Keywords: adversarial attacks; object detection; multi-scale balance regularization

0 引 言

近年来,深度学习所代表的人工智能技术在计算机视觉[1]领域中取得广泛应用,然而大量的实践证据表明深度神经网络在面对对抗攻击时表现出显著脆弱性[2-5]。特别是可在物理域实施的对抗攻击,由于能够在真实场景中部署并绕过传统防护机制,已成为威胁检测系统安全性的关键隐患[6]。其中,以对抗补丁为典型代表,只需在图像或物体表面叠加一块特制的扰动区域,可在不显著改变人类感知的前提下误导模型产生错误预测,从而引发安全风险。对抗补丁因其低成本、可打印性及一定的环境鲁棒性,已成为近年物理对抗攻击领域研究的重点方向。

为揭示深度神经网络的脆弱性,研究者们提出了多种基于对抗补丁的攻击方法[7]。Sharif 等[8]通过佩戴带扰动的眼镜成功欺骗面部识别系统,开启了物理域对抗研究。Brown 等[9]提出 Adversarial Patch 并引入期望变换优化策略,验证了生成通用鲁棒补丁的可行性。此后,Liu 等[10]所提出的 DPatch 将对抗补丁攻击推广至检测任务,通过同时优化分类与回归分支,有效削弱了 Faster R-CNN 与 YOLO 等主流检测器的检测能力。而 Zhao 等[11]则通过将生成的对抗补丁覆盖于标志牌上,从而实现对自动驾驶系统智能检测的躲避。Thys 等[12]提出了一种面向类内差异

收稿日期:2025-08-31

显著的人体目标的对抗补丁生成方法，旨在学习一块通用、可打印的贴片，使佩戴者在主流人员检测器前被系统性漏检。同样针对人类目标，Xu 等[13]提出了对抗性 T 恤，用于躲避人员探测器。攻击者可通过穿戴带有补丁的衣物显著降低检测率，但在姿态变化与遮挡条件下效果会受限。Wang 等[14]在自动结账的场景下提出了基于偏见的对抗补丁攻击，成功实现了对于淘宝、京东等电子商务平台的攻击，其生成的对抗补丁可以导致平台对商品识别错误。然而无论是基于场景抑或是基于目标所提出的对抗补丁攻击方法，其总是存在黑盒迁移性差这一核心难点。基于此，Huang 等[15]所提出的基于迁移自集成策略的攻击方法（transfer-based self-ensemble attack，T-SEA）可以在单模型条件下，通过增加随机、抖动等训练策略缓解补丁对模型和数据的过拟合，提升补丁对不同检测器的攻击效率，从而在多个异构检测器之间实现更好的迁移性能。

然而，现有方法仍存在两方面不足：一是多数方法仅关注整体检测结果的损失下降，缺乏对不同检测头的细粒度约束；二是由于现代检测器的多尺度预测特性，使得现有补丁往往在部分尺度上抑制显著、其他尺度抑制不足，导致攻击不均衡，削弱了在多尺度目标检测任务中的效果与迁移稳定性。

针对上述不足，本文提出了一种多尺度均衡正则化方法（multi-scale balance regularizer，MSBR）。该方法在补丁优化过程中显式最小化各尺度预测头置信度的方差，从而鼓励不同尺度在期望变换与随机增强条件下表现出一致性的下降趋势。通过这种方式，MSBR 有效缓解了现有方法的多尺度抑制不均衡问题，使补丁能够在小、中、大目标检测任务中均发挥稳定的攻击效果。同时，该方法无需额外引入教师模型或多模型集成，训练开销低，且能显著增强黑盒迁移攻击的稳定性。在与 T-SEA 算法的对比实验结果表明，MSBR 在多个主流检测器上均取得优于现有方法的攻击表现，特别是在跨模型迁移以及多尺度性能上表现突出，可以有效提高黑盒迁移性。

1　基本理论

1.1　FPN 多尺度预测

目标检测[16]作为计算机视觉中的核心任务，需要同时预测图像中物体的类别标签与空间位置。为了在复杂场景下兼顾不同尺寸目标的检测效果，现代检测器（如 YOLO 系列、SSD、Faster R-CNN 等）普遍采用特征金字塔网络（feature pyramid network，FPN）结构[17]。FPN 的核心思想是将不同层次的特征图进行自顶向下融合，并在多个分辨率尺度上并行预测，从而同时捕捉小目标的细粒度信息与大目标的语义信息。

设尺度集合为 $S=\{3,4,5\}$，分别对应 FPN 的 3 个检测头 P3、P4 与 P5。P3 通常具有最高分辨率，负责检测小目标；P4 在分辨率与感受野之间取得平衡，主要用于中等目标检测；而 P5 具有最低分辨率和最强语义信息，主要负责大目标检测。对于每个尺度 $l\in S$，检测器在空间位置 i 处输出三类信息：类别概率 $p_{l,i}$、对象置信度 $o_{l,i}$ 以及边界框参数 $b_{l,i}$，FPN 的设计有效缓解了单尺度检测器在目标尺寸分布不均衡时的性能瓶颈，成为现代检测框架的标准组件。

然而，这种多尺度预测机制在对抗攻击场景下也带来了新的挑战：不同尺度检测头对扰动的敏感性差异往往导致攻击效果的不均衡。例如，小目标预测头（P3）通常受到补丁扰动的压制最为显著，而中、大目标预测头（P4、P5）则往往仍能维持较高的检测置信度，形成了明显的尺度不均衡现象。这种问题削弱了补丁在整体攻击中的全面性，也限制了其跨模型迁移的稳定性，因此成为本文方法改进的着力点。

1.2　对抗补丁

对抗补丁的核心思想是在输入图像中嵌入一个经过优化的局部扰动图片，使得目标检测器的输出结果被显著误导。简而言之，图片在添加完对抗补丁之后，检测器便无法成功识别对应的物体。其基本原理是通过在随机变换分布下最小化攻击损失的期望并迭代更新像素，最终得到能在多种成像条件下稳定干扰检测器的通用贴片，即对抗补丁。具体而言，设输入图像为 $x\in[0,1]^{H\times W\times 3}$，补丁为 $p\in[0,1]^{H\times W\times 3}$，贴附算子记为 $S(x,p,\theta)$，其中 θ 表示旋转、缩放、透视变换、亮度扰动等随机参数。攻击的目标是通过优化补丁 p，在不同图像和随机变换条件下，使检测器 $D(\cdot)$ 的预测结果被尽可能破坏。

在通用形式下，对抗补丁的优化问题可以表示为：

$$\min_p \mathbb{E}_{x\sim X,\ \theta\sim\Theta}\left[\Phi(D(S(x,p,\theta)))\right]+\lambda_{tv}L_{tv}(p) \tag{1}$$

其中，X 表示训练图像分布，Θ 表示期望变换的扰动分布，$\Phi(\cdot)$为攻击目标函数，例如压制对象置信度或提升错误类别的置信度，$L_{tv}(p)$ 是全变差（total variation，TV）正则项，用于抑制高频噪声、增强补丁在打印和拍摄下的物理鲁棒性，λ_{tv} 为正则化系数。通过迭代优化上述目标函数，得到的补丁 p 可以贴附在任意图像上并保持攻击效果。与像素级的数字域对抗样本不同，对抗补丁具有通用性（同一补丁适用于不同输入）、可物理实现性（能够打印并部署到真实物体表面）、以及跨环境鲁棒性（在不同视角、光照和背景下依然有效），因此成为近年来研究的重点。

1.3　T-SEA 对抗攻击方法

在对抗补丁研究中，T-SEA 被提出以解决黑盒攻击中的迁移性难题。该方法的核心思想是，在单一检测模型上引入输入层、模型层与补丁层的多重随机化机制，使得训练过程中形成类似多模型集成的效果。其优化目标仅依赖对象置信压制与全变差约束，具体 loss 损失设计如下：

$$L=L_{obj}+\lambda_{tv}L_{tv} \tag{2}$$

其中，L_{obj} 用于降低所有候选框的平均置信度，确保补

丁检测器输出层产生直接抑制;L_{tv} 则约束补丁纹理的平滑性,提高打印与拍摄条件下的物理鲁棒性。与依赖多模型训练的传统方法相比,T-SEA 在计算成本上极为简洁,却能在多种检测器上实现较高的攻击效果。这一自集成思想的提出,使得无需教师网络或多模型组合,也能得到具备较强黑盒迁移性的对抗补丁,展示了重要的方法学价值。

尽管如此,T-SEA 的损失函数设计仍然存在明显不足。现代目标检测器广泛采用特征金字塔网络(FPN),在 P3、P4、P5 等不同分辨率的特征图上并行预测小、中、大目标。每个尺度的检测头承担着不同的感受野和特征建模任务,因此对补丁扰动的敏感性存在差异。如果优化目标未能区分这些尺度而统一压制,往往会导致攻击效果的不均衡。实验证明,在 T-SEA 生成的补丁下,小目标预测头(P3)的置信度下降幅度最大,几乎完全丧失检测能力;而中、大目标预测头(P4、P5)则仍然保持着相对较高的置信度。这种不均衡现象一方面削弱了补丁的全面性,使得不同尺度目标的抑制强度差异显著,整体攻击力不足;另一方面也限制了其黑盒迁移的稳定性。由于不同检测器在目标大小敏感度上的差异较大,如果补丁在训练过程中未能均衡抑制各尺度检测头,那么在跨模型应用时,攻击效果极易发生检测框残留现象甚至检测失败。

2 基于多尺度均衡正则的目标检测对抗补丁攻击方法

2.1 研究背景

与经典 AdvPatch 相比,T-SEA 虽显著提升了攻击成功率,但在检测结果中仍常出现残留高置信度框,如图 1 中所示。从对抗优化的角度看,其根源在于 T-SEA 的损失主要采取对目标置信度的无差别压制,缺乏对多尺度特性的显式约束;在多尺度检测头与不同目标尺寸并存的设定下,这种只压分数、不顾尺度的优化会导致抑制不均衡,对部分尺度的目标抑制较弱,因而保留下位置偏移或不完整的高分框。从检测器架构角度看,主流检测器通过多尺度头提升对不同大小目标的感知,不同尺度分支对扰动的敏感性、梯度强度与非极大值抑制(non maximum suppression, NMS)交互各不相同,导致补丁对各尺度的作用强度本就不一致。由此可见,单纯依赖无差别的置信度压制并不能在多尺度结构中实现均衡抑制。基于对这一现象的发现,本文提出一种多尺度均衡正则化方法以弥补 T-SEA 的不足。该正则化的基本思想是对各尺度预测头的平均置信度进行统计,并最小化其方差,从而保证不同尺度在训练过程中呈现一致下降趋势。

2.2 数学原理

多尺度均衡正则这一设计在理论上与统计学习中自适应方差调节(adaptive variance weighting, AVW)思想相契合,即通过控制输出尺度之间的差异来保证整体结构的稳定性。值得注意的是,AVW[18] 首次在目标检测中应用这一思路,其核心是动态调整不同尺度损失的权重,以减少多尺度训练中的不均衡,从而显著提升检测算法在 COCO 和 VOC 基准上的性能。

(a) T-SEA (b) MSBR

图 1 T-SEA 与 MSBR 的可视化对比

Fig. 1 Qualitative comparison between T-SEA and MSBR

因此,在原有攻击主项基础上,引入多尺度均衡正则化函数,在保证整体置信度下降的同时,显式惩罚各尺度抑制强度的方差,尽可能平衡抑制所有尺度,从而实现对多尺度目标的一致性压制并提升跨模型迁移性。

具体而言,设第 l 个尺度的平均对象置信为 μ_l:

$$\mu_l = \frac{1}{N_l}\sum_{i=1}^{N_l} o_{l,i}, l \in S \tag{3}$$

其中,S 为尺度集合,$o_{l,i}$ 为第 l 个尺度上第 i 个候选框的对象置信,N_l 为该尺度候选框数。整体均值为 $\bar{\mu}$:

$$\bar{\mu} = \frac{1}{|S|}\sum_{l\in S}\mu_l \tag{4}$$

多尺度均衡正则函数定义为:

$$R_{MSBR} = \frac{1}{|S|}\sum_{l\in S}(\mu_l - \bar{\mu})^2 \tag{5}$$

由式(5)对 μ_l 求导,须同时考虑 $\bar{\mu}$ 对 μ_l 的依赖,利用 $\sum_{l\in S}(\mu_l - \bar{\mu}) = 0$ 可简化得到:

$$\frac{\partial R_{MSBR}}{\partial \mu_l} = \frac{2}{|S|}(\mu_l - \bar{\mu}) \tag{6}$$

当 $\mu_l > \bar{\mu}$(该尺度抑制不足)时,梯度为正,推动该尺度的得分继续下降;当 $\mu_l < \bar{\mu}$(该尺度已充分抑制)时,梯度为负,从而抑制对该尺度的过度下压。进一步由式(4)可得:

$$\frac{\partial R_{MSBR}}{\partial o_{l,i}} = \frac{2}{|S|N_l}(\mu_l - \bar{\mu}) \tag{7}$$

说明惩罚会自适应地分配到抑制偏弱的尺度及其候选

框上，而非对所有尺度一视同仁。这一性质与仅对置信度做无差别压制截然不同，后者的梯度在各尺度间大致均匀，容易导致某些尺度被反复下压而另一些尺度残留，而 MSBR 通过“均值—方差”耦合，将更多下降空间自动分配给抑制不足的尺度，实现跨尺度的一致性压制。在多尺度检测器中，各尺度分支的感受野、锚密度与 NMS 交互不同，天然对扰动具有异质敏感性，MSBR 的梯度分配机制正是对这种异质性的显式补偿。当各尺度抑制已趋一致（$\mu_l \approx \bar{\mu}$）时，$R_{\mathrm{MSBR}} \to 0$，从而不干扰攻击主项继续拉低整体置信度，保证了优化的稳定性与目标指向性。总的来说，当某一尺度下降不足时，其与均值的差异将被放大，导致额外的惩罚；而当各尺度均衡下降时，该项接近零，不会对优化产生干扰。最终的损失函数定义为：

$$L = L_{obj} + \lambda_{tv} L_{tv} + \lambda_{MSBR} L_{MSBR} \tag{8}$$

在这一损失函数框架中，L_{obj} 仍是攻击的核心，保证对检测器整体置信度的压制；L_{tv} 继续承担约束补丁平滑性的任务，以保证物理部署的稳定性；L_{MSBR} 则显式建模了多尺度预测头之间的平衡关系，使得攻击能够覆盖小、中、大目标 3 个层次，从而在全面性和鲁棒性上显著优于原始 T-SEA。

2.3　基于 MSBR 的对抗补丁生成方法

基于 MSBR 的对抗补丁生成方法具体流程如图 2 所示。

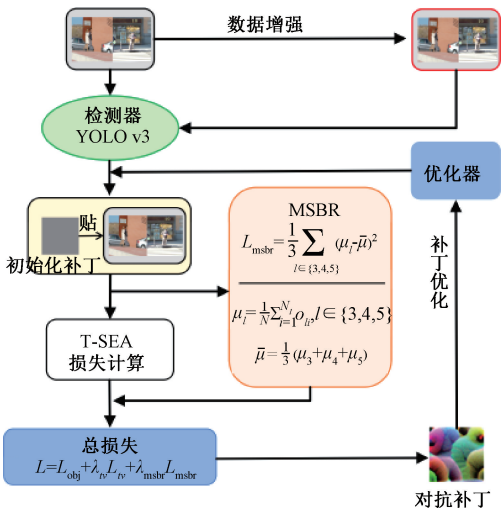

图 2　基于 MSBR 的对抗补丁训练流程图

Fig. 2　Training pipeline of adversarial patch with MSBR

在对抗补丁训练过程中，首先对训练图像进行期望变换，包括旋转、缩放、抖动等操作。然后将原始图片和数据增强后的图片同时输入白盒检测器生成检测结果，得到的检测框将会与初始化的补丁通过贴附函数合并在一起，从而生成可学习的带有对抗补丁的图像。之后会将得到的带有对抗补丁的图像再次输入白盒检测器得到对抗攻击结果，然后利用该结果进行 loss 计算。在这里，相比于 T-SEA，除了将常规预测结果用于主损失计算外，还进一步直接访问检测头的原始输出（pre-NMS 的 objectness 与尺度信息），用以构建多尺度均衡正则（MSBR）。MSBR 将预测框按照尺度划分为小、中、大三类，并计算各尺度的平均置信度，随后最小化其方差以鼓励多尺度预测头在训练过程中保持一致的抑制趋势。最终，总损失由主损失项、总变差约束和多尺度均衡正则共同组成，对补丁参数反向传播并经优化器更新，迭代得到最终的对抗补丁。具体实现流程由伪代码方式给出，伪代码如下：

Algorithm：MSBR Patch Optimization(per iteration)

Require：$x_1,\cdots,x_N,f_w,M,BS,T,\tau,\hbar,S,MSBR,\lambda$

$\tau \Leftarrow \tau_0$
for each $i \in [1,M]$ **do**
　for each $j = \frac{N}{BS}$ **do**
　　$X \Leftarrow x_{(j-1)\cdot BS+1},\cdots,x_{j\cdot BS}$
　　$bbox^{clean},conf^{clean},raw^{clean} \Leftarrow f_w(X)$
　　$X^{adv} \Leftarrow T(X,bbox^{clean},\tau)$
　　$bbox^{adv},conf^{adv},raw^{adv} \Leftarrow f_w(X^{adv})$
　　$loss \Leftarrow Avg(conf^{adv}) + \lambda \cdot MSBR(raw^{adv})$
　　$\tau \Leftarrow \hbar(\tau,loss)$
　end for
　update lr via S
end for

其中，$x_1,\cdots,x_N$ 为训练图像，f_w 为白盒检测器，M 为最大训练轮次，BS 是批次大小，T 是补丁装订器，τ 为补丁，$\hbar$ 为基础攻击方法，S 为学习率调度器，$MSBR$ 为多尺度正则化函数，λ 多尺度正则化权重。

由于 MSBR 的构建完全依赖于检测器已有的中间输出，该正则项无需引入额外的网络结构或复杂运算，从而几乎不增加训练的计算开销与时间成本。在此基础上，所提出的正则化机制不仅能够保持主损失的优化目标不变，还能够显著提升补丁在不同尺度预测头上的抑制均衡性与跨模型迁移能力。

3　实验结果与分析

3.1　实验数据及设置

1）数据集

为了与 T-SEA 保持一致，实验采用与其相同的数据集设置。具体而言，采用 INRIA Person 数据集进行训练和测试，该数据集包含 80 个类别，共 614 张图像用于训练 patch 图像以及 288 张图像用于测试评估实验结果。

为了验证对抗补丁的泛化与可迁移能力，选取与 T-SEA 算法实验中相同的两个补充数据集 COCO Person 数据集和 CCTV Person 数据集进行泛化和可迁移性测试验证。前者为从 COCO 验证集中抽取的 1 684 张不同场景的

人物图像(例如,运动场、交通路线、海洋和森林),而后者则是在 CCTV 视频中抽取的包含 559 张监控摄像头拍摄的人员图像。

2)检测器

为验证对抗补丁的跨模型迁移性,实验选取 6 种常见 YOLO 系列模型检测器作为待攻击对象:YOLOv2、YOLOv3、YOLOv3-tiny、YOLOv4、YOLOv4-tiny 和 YOLOv5。训练阶段仅在 YOLOv3 上进行(白盒设置),其他模型均作为黑盒测试目标,以评估跨模型迁移性。输入分辨率固定为 416×416,置信度阈值和 IoU 阈值分别设为 0.5 与 0.45。

3)训练设置

实验以 T-SEA 作为基线,在其训练流程中引入本文提出的多尺度均衡正则(MSBR)。补丁大小设为输入图像分辨率的 15%,初始化方式为灰度噪声,训练时采用期望变换,包括旋转、缩放、抖动、中值池化和 Cutout 等操作,以增强物理鲁棒性。优化器采用 SGD,学习率 0.03,最大迭代 1 000 轮,步长为 1,使用 ALRS 作为学习率调度器。损失函数以 T-SEA 的 L_{obj} 损失为主,并额外引入 L_{MSBR},权重 $\lambda_{MSBR}=0.1$。

3.2 评价指标

实验的主要评价指标为 mAP@0.5。除整体 mAP 外,实验进一步按照 COCO 官方评测标准对目标面积进行分桶,并分别报告小、中、大目标上的检测精度,即 mAPs、mAPm 和 mAPl。具体地,COCO 将目标面积以像素为单位进行划分:面积小于 32^2 的为小目标,介于 32^2 与 96^2 之间的为中目标,大于 96^2 的为大目标。mAPs、mAPm 和 mAPl 分别表示在这 3 类目标上单独计算的平均精度。与整体 mAP 相比,这 3 个指标能够更精细地刻画对抗补丁在不同尺度目标上的抑制效果。例如,小目标往往对应远处行人或交通标志,中目标常见于街景车辆,大目标则包括近景行人或大型物体。由于现代检测器基于 FPN 在多尺度特征图上预测,不同尺度的检测头对补丁攻击的敏感性存在差异,因此单一的整体 mAP 难以反映攻击的不均衡性,而 mAPs/mAPm/mAPl 能够揭示攻击在小、中、大目标上的差异,从而更有助于评估 MSBR 在缓解多尺度不均衡问题上的优势。

3.3 实验结果

为全面验证所提 MSBR 算法的有效性与优势,本文针对性地设计了多组实验,并从 3 个维度对结果进行分析与评估:

1)整体攻击性能:评估算法在实际攻击任务中的效果与稳定性;

2)泛化与迁移能力:考察其在不同检测模型及数据集上的普适性与黑盒迁移表现;

3)多尺度攻击特性:验证算法在多尺度检测头上的均衡抑制能力及其与理论设计的一致性。

1)整体攻击性能

表 1 给出了在 INRIA Person 数据集上 6 种 YOLO 系列检测器的整体 mAP@0.5 对比结果。可以看到,基线方法 AdvPatch 在所有检测器上均存在明显残留检测,而 T-SEA 在相同设置下显著降低了 mAP,例如在 YOLOv3 上由 13.89 降至 5.76,在 YOLOv4-tiny 上由 58.43 降至 25.30,充分验证了其在跨模型攻击中的有效性。在此基础上,加入本文提出的多尺度均衡正则后,整体 mAP 进一步下降,如在 YOLOv3 上进一步降至 5.24,在 YOLOv3-tiny 由 35.38 降至 33.61,YOLOv4 由 42.43 降至 31.03。这表明 MSBR 能够在保持 T-SEA 迁移性优势的同时,有效提升攻击强度,并在多种检测器上表现出更好的鲁棒性。

表 1 基于 INRIA Person 数据集的整体攻击性能对比

Table 1 Overall attack performance on the INRIA Person dataset

检测器	AdvPatch	T-SEA	MSBR(ours)
YOLOv2	51.85	31.02	**29.34**
YOLOv3	13.89	5.76	**5.24**
YOLOv3-tiny	51.17	35.38	**33.61**
YOLOv4	57.16	42.43	**31.03**
YOLOv4-tiny	58.43	**25.30**	25.65
YOLOv5	70.47	58.02	**55.68**

2)泛化与可迁移能力

对于算法的泛化和可迁移能力,着重关注两点:一是模型泛化能力,通过在不同模型上进行测试并与现有的多种对抗补丁算法进行对比实验;二是数据集可迁移性,通过采用不同的数据集进行测试验证,检验算法在不同数据集下的攻击表现。实验结果分别如表 2 和表 3 所示。

在模型泛化能力实验中,如表 2 所示,采用 YOLOv2 作为白盒模型进行训练,并在 5 种其他模型上进行实际攻击测试,将实验结果分别与灰色、随机、白色 3 种基本对照补丁以及 NPAP、AdvPatch、AdvCloak 和 T-SEA 4 种现有对抗补丁生成算法进行对比分析。从结果上不难看出,MSBR 在 5 种黑盒模型上的攻击效果要明显优于上述 4 种现有常用的对抗补丁生成方法。此外,与基本对照组的实验结果对比也能证明 MSBR 的有效性并非随机实验的偶发性结果,可见 MSBR 具备较强的模型泛化能力。

在数据集可迁移性实验中,如表 3 所示,采用 YOLOv5 作为白盒模型进行训练,训练数据集依旧采用与 T-SEA 相同的 INRIA Person 数据集,分别在 COCOPerson 和 CCTVPerson 两个测试数据集上进行可迁移性验证。实验结果表明,MSBR 相比于基线算法 AdvPatch 以及现有的先进算法 T-SEA 在数据集可迁移性上存在明显优势。

表 2　黑盒模型可迁移性对比分析结果

Table 2　Comparison of black-box model transferability results

模型	White Box↓	Black Box↓					Black Box
	YOLO v2	YOLO v3	YOLO v3tiny	YOLO v4	YOLO v4tiny	YOLO v5	Avg↓
Grey	67.75	76.22	80.69	75.22	76.89	81.86	—
Random Noise	70.67	75.80	82.44	75.10	78.74	81.79	—
White	68.52	74.89	80.20	74.73	76.09	80.09	—
NPAP	38.03	56.85	58.04	67.74	67.43	66.85	**61.45**
AdvCloak	33.74	54.77	53.42	67.57	56.12	68.05	**59.42**
AdvPatch	5.66	40.26	18.07	48.49	24.44	43.38	**36.46**
T-SEA	3.98	13.81	10.82	23.07	16.40	6.41	**16.26**
MSBR(ours)	**3.36**	**12.71**	**8.54**	**16.75**	18.99	**4.84**	**10.87**

表 3　MSBR 在 YOLOv5 上的跨数据集迁移性结果对比

Table 3　Comparison of cross-dataset transferability results of MSBR on YOLOv5

数据集	Method	White Box↓	Black Box↓
COCOPerson	AdvPatch	45.83	52.54
	T-SEA	37.28	38.87
	MSBR	**34.41**	**30.27**
CCTVPerson	AdvPatch	38.07	34.08
	T-SEA	38.71	19.91
	MSBR	**32.54**	**15.69**

3)多尺度攻击特性

为进一步探讨 MSBR 在不同尺度目标上的抑制效果，依据 COCO 协议将目标划分为小(s)、中(m)、大(l)三类，并分别计算 mAP。表 4 给出了 T-SEA 与 MSBR 在不同尺度下的 mAP 指标对比结果。从实验结果可以明显看出，MSBR 相比于 T-SEA 算法在不同模型不同尺度的预测结果上，mAp 值均有明显下降，这直接体现了 MSBR 算法的多尺度抑制的有效性。

图 3 给出了 T-SEA 与 MSBR 在小、中、大 3 个尺度上对抗攻击结果的可视化对比，可以更为直观的看出 MSBR 有效的减少了多尺度抑制不均衡问题导致的检测框残留现象。

为了更加显著的展示多尺度抑制的有效性和理论一致性，从单一模型角度和多模型角度分别进行分析。通过对表 4 中数据进行后处理得到图 4 和图 5，图 4 所示为基于 YOLOv3 模型训练的多尺度 mAP 对比结果，从折线图可以明显看出 MSBR 在中、大目标上相较于基线算法 mAP 指标明显下降，整体折线图比 T-SEA 方法更加平整，表明其在弥补 T-SEA 多尺度攻击不均衡方面发挥了作用，这也进一步说明了 MSBR 在不同尺度预测头上的压制力更均衡。类似现象也出现在 YOLOv4-tiny 上。

另一方面，从多模型角度做进一步验证。如图 5 所示，其为综合评估各检测器在不同尺度上的 mAP 结果，图中显示的数值为各尺度在 6 个检测器上的 mAP 的平均值。可以直观的从图中看出两种不同算法对于不同尺度的攻击效果，其中 MSBR 的抑制效果更好，在 3 个尺度上均优于 T-SEA 方法。这进一步验证了 MSBR 在不同尺度预测头上的压制力更均衡并非某一单一模型的偶发性结果。

表 4　基于 INRIA Person 数据集不同检测器攻击效果对比

Table 4　Comparison of attack performance on different detectors based on the INRIA Person dataset

检测器	Bucket	T-SEA	MSBR(ours)	Δ mAP
YOLOv2	s	0.00	0.00	0.00
YOLOv2	m	24.0	**23.1**	−0.9
YOLOv2	l	**56.4**	57.4	+0.1
YOLOv3	s	4.6	**3.0**	−1.6
YOLOv3	m	20.7	**17.3**	−3.4
YOLOv3	l	47.0	**36.7**	−10.3
YOLOv3-tiny	s	10.0	**5.0**	−5.0
YOLOv3-tiny	m	29.1	**21.7**	−7.4
YOLOv3-tiny	l	61.7	**61.1**	−0.6
YOLOv4	s	4.0	**2.9**	−1.1
YOLOv4	m	26.0	**23.1**	−2.9
YOLOv4	l	60.0	**59.4**	−0.6
YOLOv4-tiny	s	3.3	**1.4**	−1.9
YOLOv4-tiny	m	31.7	**29.5**	−2.2
YOLOv4-tiny	l	59.7	**57.2**	−2.5
YOLOv5	s	**4.0**	4.3	+0.3
YOLOv5	m	35.2	**32.9**	−2.3
YOLOv5	l	55.5	**54.9**	−0.6

此外，还发现 MSBR 对于不同尺度的抑制性也有差别，其主要针对中目标和大目标的抑制效果更强，而对于小目标抑制效果并不明显。在部分检测器上，如 YOLOv3-tiny 与 YOLOv4-tiny，小目标的 mAP 有明显下

图 3 T-SEA 与 MSBR 的在不同尺度上的可视化对比

Fig. 3 Visualization comparison between T-SEA and MSBR across different scales

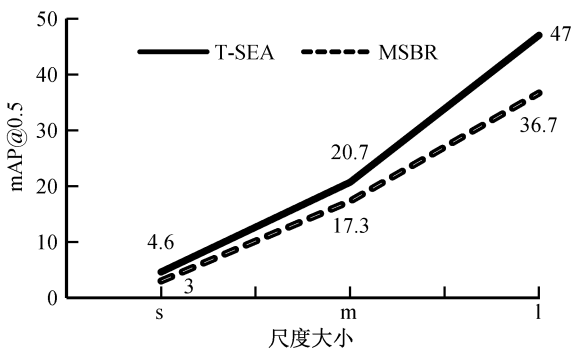

图 4 基于 YOLOv3 模型训练的多尺度 mAP 对比

Fig. 4 Scale-wise mAP comparison trained on YOLOv3 model

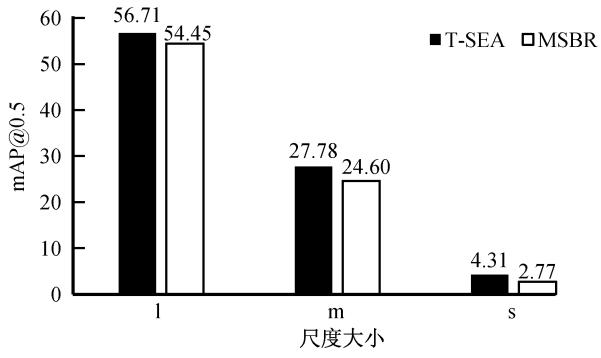

图 5 不同尺度在多检测器上的 mAP 对比

Fig. 5 Scale-wise mAP comparison across different detectors

降,然而在 YOLOv5 上,小目标的 mAP 却略有上升(+0.003),这一变化幅度极小,基本可以认为小目标抑制保持不变,但也体现出 MSBR 并未削弱对小目标的攻击效果,只是保持不低于基线算法的攻击能力。这说明 MSBR 的主要贡献并不在进一步压制小目标,而在于均衡 3 种尺度之间的下降幅度,这也符合 MSBR 的均衡正则理论。由于小目标攻击效果本就接近于消失攻击,而对于中、大目标的置信度仍有较多的攻击空间,因此对于置信度更高的中、大目标抑制效果会更明显。可见 MSBR 通过在训练中最小化不同尺度检测头置信度的方差,实现了更均衡的压制,显著提升了中、大目标上的攻击强度,并增强了跨模型的一致性和鲁棒性。

4 结 论

本文提出了一种基于多尺度均衡正则的对抗补丁生成方法,该方法在 T-SEA 框架下引入跨尺度置信度约束,使不同尺度检测头在训练过程中保持均衡下降,从而提升了攻击的全面性与稳定性。多尺度均衡的策略能够有效缓解 FPN 检测器在不同预测头之间抑制不均衡的问题,尤其是在中、大目标上显著增强了补丁的攻击效果。该方法无需引入额外教师模型或多模型集成,具备较高的训练效率与可扩展性,为通用物理攻击与多场景安全评估提供了新的思路。未来工作可进一步探索 MSBR 与动态特征选择、跨模态迁移及物理实现的结合,以提升其在复杂环境下的可用性与稳定性。同时,当前研究主要针对 YOLO 系列检测器进行验证,后续可在 Transformer 架构及实时检测场景中进一步评估其通用性与实际应用潜力。

参考文献

[1] BAR A, WANG X, KANTOROV V, et al. DETReg: Unsupervised pretraining with region priors for object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022:14605-14615.

[2] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. ArXiv preprint arXiv:1312.6199, 2013.

[3] WEI X X, LIANG S Y, CHEN N, et al. Transferable adversarial attacks for image and video object detection[C]. Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019:954-960.

[4] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. ArXiv preprint arXiv:1706.06083, 2017.

[5] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]. 2017 IEEE Symposium on Security and Privacy. IEEE, 2017: 39-57.

[6] 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. An overview of adversarial attack and defense for target detection[J]. Journal of Communications, 2023,44(11): 260-277.

[7] 武阳,刘靖.面向图像分析领域的黑盒对抗攻击技术综述[J].计算机学报,2024,47(5):1138-1178.
WU Y, LIU J. An overview of black-box adversarial attack techniques for image analysis [J]. Acta Computer Sinica, 2024,47(5): 1138-1178.

[8] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[C]. 2016 Acm Sigsac Conference on Computer and Communications Security, 2016: 1528-1540.

[9] BROWN T B, MANÉ D, ROY A, et al. Adversarial patch[J]. ArXiv preprint arXiv:1712.09665, 2017.

[10] LIU X, YANG H, LIU Z, et al. Dpatch: An adversarial patch attack on object detectors[J]. ArXiv preprint arXiv:1806.02299, 2018.

[11] ZHAO Y, ZHU H, LIANG R, et al. Seeing isn't believing: Towards more robust adversarial attack against real world object detector[C]. 2019 ACM SIGSAC Conference on Computer and Communications Security, 2019: 1989-2004.

[12] THYS S, RANST W V, GOEDEME T. Fooling automated surveillance cameras: adversarial patches to attack person detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[13] XU K D, ZHANG G Y, LIU S J, et al. Adversarial T-shirt! Evading person detectors in a physical world[C]. 16th European Conference on Computer Vision, 2020: 665-681.

[14] WANG J K, LIU AI SH, BAI X, et al. Universal adversarial patch attack for automatic checkout using perceptual and attentional bias[J]. IEEE Transactions on Image Processing, 2021, 31:598-611.

[15] HUANG H, CHEN Z Y, CHEN H R, et al. T-SEA: Transfer-based self-ensemble attack on object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 20514-20523.

[16] JOSEPH K J, KHAN S, KHAN F S, et al. Towards open world object detection [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021:5826-5836.

[17] 李海龙,黄孙港,饶兴昌.跨尺度特征融合的自适应水下目标检测算法[J].电子测量技术,2025,47(13):129-138.
LI H L, HUANG S G, RAO X CH. Adaptive underwater target detection algorithm based on cross-scale feature fusion [J]. Electronic Measurement Technology, 2025,47(13): 129-138.

[18] LUO Y H, CAO X, ZHANG J T, et al. Dynamic multi-scale loss optimization for object detection[J]. Multimedia Tools and Applications, 2023, 82(2): 2349-2367.

作者简介

谢家乐,硕士研究生,主要研究方向为计算机视觉、对抗攻击。

E-mail:xiejiale@stu.xmu.edu.cn

赵宇熙,硕士研究生,工程师,主要研究方向神经网络鲁棒性、对抗攻击。

E-mail:zhaoyx@cae.avic

曾念寅,博士,教授,博士生导师,主要研究方向空天信息智能。

E-mail:zny@xmu.edu.cn

王若(通信作者),博士,高级工程师,硕士生导师,主要研究方向人工智能算法测试、群体智能、反智能。

E-mail:wang@cae.avic