针对目标检测的可迁移性对抗补丁生成方法

燕庆龙 [1] 向昕宇 [1] 张 浩 [1] 马佳义 [1, 2]

摘 要 随着目标检测模型在实际应用中的广泛部署, 其安全性问题日益成为研究热点. 对抗攻击技术通过精心设计对抗补丁, 能够有效诱导模型产生错误预测, 揭示深度神经网络在决策过程中存在的内在脆弱性. 为提升对抗补丁在不同检测器上的攻击迁移性, 现有方法大多依赖静态权重融合策略进行联合优化, 难以充分协调不同检测器在脆弱性分布及优化动态上的差异, 导致攻击效果无法在各模型间兼顾, 迁移性受到显著限制. 针对这一挑战, 提出一种基于多任务动态重加权机制的可迁移性对抗补丁生成框架. 该框架设计全局校正因子和局部校正因子, 分别从任务间整体优化进度及单任务细粒度收敛行为两个层面动态调整任务权重, 实现多模型联合优化过程中的协调与鲁棒性提升. 通过系统性的数字域与物理域实验验证, 所提方法显著增强了对抗补丁在不同目标检测器上的对抗攻击迁移性, 并且在真实物理域的部署中表现出优秀的攻击效果.

关键词 目标检测; 对抗攻击; 跨模型攻击; 动态重加权; 攻击迁移性

引用格式 燕庆龙, 向昕宇, 张浩, 马佳义. 针对目标检测的可迁移性对抗补丁生成方法. 自动化学报, 2026, **52**(4): 693−708

DOI 10.16383/j.aas.c250300 **CSTR** 32138.14.j.aas.c250300

A Transferable Adversarial Patch Generation Method for Object Detection

YAN Qing-Long[1] XIANG Xin-Yu[1] ZHANG Hao[1] MA Jia-Yi[1, 2]

Abstract With the widespread deployment of object detection models in real-world applications, their security issues have increasingly become a research focus. Adversarial attack techniques, by carefully designing adversarial patches, can effectively induce models to produce erroneous predictions, thereby revealing the inherent vulnerabilities of deep neural networks in the decision-making process. To enhance the transferability of adversarial patches across different detectors, most existing methods rely on static weight fusion strategies for joint optimization. However, such approaches struggle to fully reconcile the discrepancies in vulnerability distributions and optimization dynamics among detectors, leading to imbalanced attack effectiveness across models and significantly limiting the transferability. To address this challenge, this paper proposes a transferable adversarial patch generation framework based on a multi-task dynamic reweighting mechanism. The framework introduces a global correction factor and a local correction factor, which dynamically adjust task weights from two perspectives: The overall optimization progress among tasks and the fine-grained convergence behavior of individual tasks. This design enables better coordination and improved robustness during multi-model joint optimization. Extensive experiments in both the digital and physical domains demonstrate that the proposed method significantly enhances the adversarial transferability of patches across various object detectors and achieves strong attack performance in deployments under real-world physical domain.

Keywords object detection; adversarial attack; cross-model attack; dynamic reweighting; attack transferability

Citation Yan Qing-Long, Xiang Xin-Yu, Zhang Hao, Ma Jia-Yi. A transferable adversarial patch generation method for object detection. *Acta Automatica Sinica*, 2026, **52**(4): 693−708

近年来, 深度神经网络通过其卓越的非线性建模能力, 促进了各类计算机视觉任务的发展, 例如图像分类[1]、目标检测[2]、语义分割[3]. 这种以数据驱动为核心的技术演进显著提升了机器对物理世界的解析能力, 促进了智能安防、自动驾驶、人脸检测等一系列关键安全任务在实际生活中的部署应用, 使复杂场景的智能认知达到实用化水平[4−10].

然而, 研究表明通过设计特定范式的扰动或补丁生成算法, 在原始数据中添加对抗性扰动或补丁, 能够误导深度神经网络产生巨大的预测置信度偏移, 从而做出错误的决策[11−16], 该类技术被称为对抗攻击技术; 而生成扰动或补丁的过程则称为对抗样

收稿日期 2025-07-06 录用日期 2025-11-29
Manuscript received July 6, 2025; accepted November 29, 2025
国家自然科学基金 (U23B2050, 62473297) 资助
Supported by National Natural Science Foundation of China (U23B2050, 62473297)
本文责任编委 樊彬
Recommended by Associate Editor FAN Bin
1. 武汉大学电子信息学院 武汉 430072 2. 武汉大学机器人学院 武汉 430072
1. Electronic Information School, Wuhan University, Wuhan 430072 2. School of Robotics, Wuhan University, Wuhan 430072

本生成技术, 其能够构造高拟真的虚假人脸以欺骗生物识别系统[17]、误导交通标志识别模型将标识错误分类为禁止通行标志[18], 或在医学影像中注入微小扰动以导致疾病漏检等各种安全隐患[19]. 因此, 对抗攻击作为深度学习模型安全性的严峻挑战, 其攻击机制揭示了各类模型可被恶意利用或者攻击的固有脆弱性. 随着智能检测系统在各类实际场景中的广泛部署与应用, 系统性地研究针对目标检测模型的对抗样本生成方法对于提升其安全性与鲁棒性具有重要的理论意义与现实应用价值.

根据攻击域的不同, 可以将对抗样本生成技术分为针对数字域和物理域的攻击. 数字域攻击直接在数字图像上修改其像素值, 通过添加人眼难以察觉的扰动误导模型决策[20−23]. 相较于数字域攻击, 物理域对抗攻击则强调将对抗性扰动具象化为现实世界中的物理媒介, 通过在物理世界对象或环境中添加对抗补丁构成对抗样本, 然后经过传感器成像, 实现对智能决策模型的误导[24−27]. 与数字域攻击相比, 物理域攻击对社会安全保障构成更大的威胁, 引发人们的严重担忧.

在面向实际环境部署的对抗补丁生成任务中, 提升补丁在多种目标检测器间的迁移攻击能力具有重要研究意义. 现实应用中往往难以获知所采用的具体检测模型, 因而对抗补丁需具备良好的跨模型迁移能力, 能够在不同架构的检测器上均有效诱导错误预测, 从而提高其实用性与攻击鲁棒性. 如图 1(a) 所示, 传统方法通常基于单一模型独立优化攻击损失, 生成模型专属的对抗补丁. 尽管此类策略可在目标模型上取得显著的攻击效果, 但其生成的补丁往往高度依赖特定模型结构, 难以在未知或未参与训练的检测器上保持有效性, 迁移能力受限. 为提升补丁的跨模型适应性, 部分研究 (如 NAP[28]、AdvBulb[29] 等) 提出如图 1(b) 所示的联合优化策略. 该策略通过对多个检测模型的攻击损失进行加权求和, 构建统一的优化目标, 以训练一个在多个模型间具有迁移性的共享补丁. 该策略在一定程度上突破传统单模型定向优化的局限, 向更具迁移性的对抗攻击方式迈进. 然而, 不同目标检测模型在网络结构、训练范式及推理机制等方面存在一定差异, 导致其对对抗扰动的敏感性不一致. 这种差异进一步反映在联合训练过程中的收敛速度与优化动态上, 造成优化难以在各模型之间实现有效协调. 因此, 直接采用静态加权的简单融合策略, 往往使对抗补丁在各模型上的攻击效果趋于“平均化”, 即在所有模型上仅维持中等程度的攻击强度, 难以达到协同且高效的攻击效果.

受多任务学习思想的启发[30−33], 部分研究者尝试从多任务联合优化的视角审视针对多个模型的统一补丁生成过程, 并提出迁移性增强策略[34−40]. 在现有研究中, AdaEA[34] 和 SMER[35] 是两类典型的面向分类器的扰动式迁移攻击方案. 具体而言, AdaEA 属于一种测试驱动的加权方法, 其依赖于不同模型对攻击目标贡献的差异自适应地分配权重. 对于某一模型, AdaEA 通过测试由其梯度生成的对抗样本在其他模型上的攻击性能来评估其潜在的迁移性, 进而根据跨模型的攻击比率调整集成权重. 然而, 这一机制不仅需要频繁地进行跨模型测试, 带来显著的计算开销, 而且其权重更新依赖于测试评估结果, 而非直接利用优化过程中的动态信息. 与之相比, SMER 采用强化学习框架来调节多模型权重, 但其策略更新依旧未能充分建模多模型在优化过程中的演化差异, 因而难以解决收敛速度和优化动态不一致所导致的协调性问题. 如图 1(c) 所示, 本文提出一种基于多任务动态重加权 (dy-

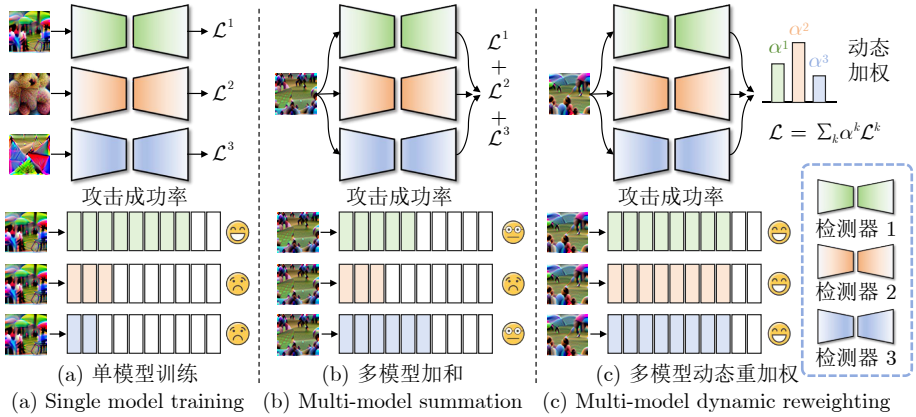


图 1 针对多个目标检测器的对抗样本优化过程

Fig. 1 An adversarial sample optimization process for multiple object detectors

namic reweighting, DR) 的迁移性对抗补丁生成方法 DRPatch, 旨在提升多模型攻击的攻击迁移性. 具体而言, 该框架在优化迭代过程中, 通过设计全局校正因子与局部校正因子, 分别从宏观和微观层面动态评估各攻击任务的收敛状态. 全局校正因子量化不同任务间整体优化进度的差异, 进而对任务训练过程进行宏观协调, 实现多模型间的平衡与协同; 局部校正因子则聚焦单一任务的内部优化表现, 确保每个模型的攻击任务均能够充分优化. 两者的协同作用为多任务联合优化生成了动态权重, 指导补丁训练过程, 避免传统静态加权策略导致的性能平均化现象, 实现多模型间攻击效果的均衡提升. 总体来说, 本文贡献如下:

1) 提出一种面向多个目标检测器的迁移性对抗补丁生成框架 DRPatch, 创新性引入多任务动态重加权机制, 有效缓解了由不同检测器在脆弱性分布与优化动态上的差异所导致的联合优化失调问题.

2) 动态重加权策略设计了全局校正因子与局部校正因子, 分别从任务间整体优化进度和单任务细粒度收敛两个层面精细建模并动态调整多任务权重, 推动攻击优化向更加协调与稳健的方向演进.

3) 通过数字域与物理域的系统实验, 验证了对抗补丁在多个目标检测器上的攻击迁移能力显著提升, 并在真实物理域上展现出良好的攻击性能.

1 相关工作

对抗样本生成技术作为人工智能安全领域的关键研究方向, 可依据不同维度进行多种分类. 在篡改实施场域维度上, 主要分为数字域攻击与物理域攻击. 前者直接在数字层面修改图像像素, 而后者通过现实环境中的实体干扰影响模型感知. 从攻击目标模型的功能类型考量, 可分为针对图像分类、目标检测、语义分割、目标跟踪、光流预测等不同模型的攻击技术. 这种多维分类体系为理解对抗样本生成技术的攻击机制提供了系统化的分析框架.

1.1 基于不同篡改实施场域的攻击

数字域攻击. 在数字域对抗攻击中, 研究者通常基于一个理想化的假设: 攻击者能够完全访问数字图像, 并在图像输入至智能决策模型之前, 对其像素值进行精细操控. 这种假设允许对图像进行全局扰动、局部区域篡改, 甚至达到单像素级别的修改. 尽管这种攻击模式在现实物理场景中存在较大局限, 但它能深刻揭示深度神经网络的内在脆弱性. 在经典的数字域攻击方法中, 快速梯度符号方法 FGSM[41] 和投影梯度下降方法 PGD[42] 广受关注. FGSM 通过利用神经网络的梯度信息, 直接沿着损失函数对输入最敏感的方向施加微小扰动, 使得输入数据在不被人眼察觉的情况下迅速令模型产生错误预测; 而 PGD 则在固定的扰动约束内, 采用多步迭代的方式逐步逼近损失函数的极大值, 通过反复“扰动–投影”的过程克服了单步方法容易陷入局部最优的问题, 从而生成攻击效果更强、对抗性更高的样本.

物理域攻击. 相较于数字域攻击, 物理域对抗攻击则强调将对抗性扰动以物理媒介的形式呈现在现实环境中, 通过光学成像过程将干扰信号传递给决策模型, 从而突破数字仿真场景的局限. 典型的物理媒介既包括可穿戴设备, 例如对抗纹理服饰、定制眼镜框架或经过改造的汽车牌照; 也涵盖静态环境中的隐蔽部署, 例如交通标志篡改贴纸或经过精心设计的广告海报. 例如, Sharif 等[43] 针对面部生物识别系统提出一种可打印的对抗性眼镜框生成方法, 能够使佩戴这副眼镜的攻击者逃避识别或模仿另一个人的身份, 并提出一种不可打印性评分策略约束对抗性眼镜的物理可实现优化. Xu 等[44] 提出一种对抗性 T 恤, 该方法考虑了人体姿态变化引起的服饰非刚性变形并对其进行建模, 使攻击者能够在物理域中成功逃避目标检测. Tan 等[45] 提出一种对抗性贴纸, 从颜色、边缘和纹理等视觉特征出发, 选取具有自然感知属性的卡通图像作为基础, 并引入投影函数进行优化, 旨在增强对抗补丁的视觉自然性, 实现了在人类感知与模型检测层面的双重规避. Yin 等[46] 提出一种对抗人脸识别系统的化妆生成方法, 使用混合模块在人脸眼眶区域上生成不可察觉的眼影, 在数字和物理场景下都能够产生隐蔽性攻击.

总体而言, 数字域和物理域对抗攻击分别从不同攻击域揭示现有深度学习系统的安全隐患: 前者借助对像素级修改的精确控制, 直观展示了模型对微小扰动的敏感性; 后者则通过将扰动转化为现实世界中的物理干扰, 强调了在实际应用环境中防护措施的重要性. 这两类攻击方法不仅为理论研究提供了宝贵的数据支持, 也为实际应用中模型安全防护策略的设计指明了方向. 由于本文的攻击算法面向真实世界中的迁移性部署而设计, 因此属于物理域攻击的范畴.

1.2 基于不同攻击目标模型的攻击

自 Szegedy 等[47] 首次揭示深度学习分类器对精心优化的对抗扰动存在脆弱性以来, 这一安全性

挑战迅速成为计算机安全领域的核心议题. 近年来, 对抗攻击研究已渗透至计算机视觉全任务链条, 不同视觉任务根据各自特性演化出多样化的攻击范式与技术路线.

针对图像分类的攻击. 在针对图像分类任务的对抗攻击中, Brown 等[48] 提出的物理对抗补丁具有里程碑意义. 该方法通过攻击分类损失来优化补丁, 在香蕉旁部署打印补丁后, 成功诱导分类器以 99% 的高置信度将原本属于香蕉的图像错误分类为烤面包机, 这也同时揭示了数字域对抗攻击向物理世界迁移的可行性. 而 Xue 等[49] 提出一种新颖的框架 Diff-PGD, 将扩散模型的强大先验知识融入对抗样本生成中, 确保对抗样本保持接近原始数据分布的同时仍能发挥攻击有效性.

针对语义分割的攻击. 与图像分类不同, 作为一种密集攻击任务, 针对语义分割任务对抗攻击需要实现全像素级的同步干扰, 因此其攻击难度更大. Gu 等[20] 提出的 SegPGD 是一种先进的语义分割攻击算法, 该方法通过在攻击损失中动态调整权重, 解决了错误分类的像素在更新扰动时被忽略, 进而导致经过多次攻击迭代后可能重新被正确分类的问题. 面对密集型攻击难度大的问题, Rony 等[50] 提出一种基于近邻分裂的攻击方法, 用于生成具有更小 l_∞ 范数的对抗扰动. 该方法通过增强拉格朗日方法在非凸性最小化框架内处理大量约束, 实现了有效攻击.

针对目标检测的攻击. 相较于图像分类中单一的判决目标, 目标检测任务同时涵盖分类和定位, 其决策过程更为复杂, 应用场景也更为广泛. 目标检测器的输出通常包含三个核心要素: 边界框坐标、目标存在概率及类别置信度. 因此, 针对目标检测的对抗攻击方法主要围绕这些要素设计. Thys 等[24] 针对 YOLOv2 检测模型构造双重抑制损失函数, 一方面降低目标区域的存在判定置信度, 另一方面同步压制特定类别的置信度, 从而有效扰乱检测器的判断. Hu 等[28] 提出一种视觉自然性补丁生成方法 NAP, 其利用预训练的生成对抗网络 (generative adversarial network, GAN) 来生成对抗样本, 由于 GAN 所学习的潜在空间能够较好地近似自然图像的流形, 因此生成的对抗补丁具有较强的视觉隐蔽性. Wei 等[51] 引入一种相机无关的物理对抗攻击方法 CAP, 通过构建一个可微分的相机图像信号处理代理网络, 以弥补物理域到数字域的转换差距并模拟不同相机的拍摄差异, 其生成的补丁在不同硬件成像下均实现有效的物理域攻击.

上述研究不仅从理论上揭示了深度学习模型在不同视觉任务中的脆弱性, 也为在实际应用中设计更鲁棒的视觉系统及构建有效的防御机制提供了重要启示. 鉴于目标检测在自动驾驶、视频监控、人脸识别、医疗影像分析、工业质量控制、无人机监测等多个领域的关键应用, 针对目标检测的对抗攻击研究显得尤为重要. 本文也进一步探讨了这一领域的迁移性攻击挑战与应对策略.

2 可迁移性对抗补丁生成方法设计

2.1 问题描述

在目标检测任务中, 给定干净的图像 $I_{clean} \in \mathbf{R}^{H\times W\times 3}$ 和对抗补丁 $\delta \in \mathbf{R}^{h\times w\times 3}$, 其中, H 和 W 分别表示干净图像的高和宽, h 和 w 分别表示对抗补丁的高和宽. 对于干净图像 I_{clean}, 通过将补丁 δ 粘贴至目标表面形成对抗图像 $I_{adver} \in \mathbf{R}^{H\times W\times 3}$, 又称为对抗样本. 设 $\mathcal{F}: I \to Y$ 和 θ 分别表示检测器及其参数, 则将干净图像输入检测器时, 预训练的目标检测器可以预测与真实标签相匹配的标签 Y_{clean}, 包括边界框位置 V_{pos}^{clean}、目标置信度分数 V_{obj}^{clean} 和类别概率 V_{cls}^{clean}, 公式表示如下:

$$Y_{clean} = [V_{pos}^{clean},\ V_{obj}^{clean},\ V_{cls}^{clean}] = \mathcal{F}(I_{clean},\ \theta) \quad (1)$$

其中, 边界框位置用于回归目标在图像中的空间位置, 目标置信度分数则衡量预测框中是否包含真实目标, 类别概率用于分类预测以判定目标所属类别. 得分最高的边界框被视为最终检测到的目标. 类似地, 对抗图像 I_{adver} 的检测结果可以表示为:

$$Y_{adver} = [V_{pos}^{adver},\ V_{obj}^{adver},\ V_{cls}^{adver}] = \mathcal{F}(I_{adver},\ \theta) \quad (2)$$

对抗性补丁 δ 的目标是成功攻击目标检测器, 使 V_{obj}^{adver} 尽可能减小, 从而让目标避开检测器的检测. 因此, 对抗过程的主要损失函数可以描述为:

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}(V_{obj}^{adver})^n \quad (3)$$

其中, N 表示目标的数量. 通过该损失函数对对抗补丁 δ 进行优化, 可以实现针对单个检测器的攻击. 而在针对多个目标检测器的集成攻击任务中, 期望通过生成一个统一的对抗补丁, 能够同时降低所有检测器对目标的检测置信度. 设共有 K 个待攻击的目标检测器, 现有方法 (如 NAP[28]、AdvBulb[29] 等) 通过对各个检测器的攻击损失进行求和以构造总体的损失函数, 从而实现对多个检测器的联合攻击, 损失表示如下:

$$\mathcal{L}_{obj}^{sum} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}^{k} \tag{4}$$

然而, 由于针对不同检测器的攻击难度存在差异, 由式 (4) 所示的优化方式生成的对抗补丁通常无法针对所有检测模型实施有效攻击. 因此, 针对这一问题, 本文设计了一种基于多任务动态重加权的对抗补丁生成方法, 通过各任务的平衡训练促进补丁的攻击迁移性.

2.2　方法实现

对于本文设计的可迁移性对抗补丁生成方法, 其整体流程如图 2 所示. 本文以单模型针对性攻击中表现最优的 AdvPatch[24] 作为基线方法, 沿用其补丁生成机制. 具体而言, 首先初始化一个灰色补丁 δ, 并将其设置为可学习的参数, 通过后续的损失函数对其进行像素值层面的优化. 考虑到在真实物理域部署中, 成像环境的复杂变化会对补丁表观产生影响, 例如多变的成像角度和复杂的天气条件, 从而造成对抗补丁的性能退化. 因此, 根据 AdvPatch 的设置, 对补丁 δ 施加 EOT (expectation over transformation) 算法, 表示如下:

$$\delta_{EOT} = \mathrm{EOT}(\delta) \tag{5}$$

其中, EOT 表示一组随机变换, 设计了关于尺度、旋转、噪声、亮度和对比度等不同的变换, 以模拟真实物理环境中的补丁表观变化, 从而保证其在物理域中的鲁棒性攻击.

随后, 将经过 EOT 扰动的对抗补丁 δ_{EOT} 粘贴于待攻击的干净图像 I_{clean}, 该过程表示如下:

$$I_{adver} = I_{clean} \odot (1 - M) + \delta'_{EOT} \odot M \tag{6}$$

其中, I_{adver} 表示粘贴补丁后的对抗图像; $\odot$ 表示哈达玛积; δ'_{EOT} 由对抗补丁 δ_{EOT} 形成, 其图像分辨率与 I_{clean} 相同; $M \in \{0, 1\}^{H\times W\times 1}$ 是一个掩码矩阵, 用于确定目标上对抗补丁所粘贴的位置. 在粘贴补丁的区域, $M_{i,j}$ 的像素值设为 1, 其余位置 (即保留干净图像的位置) 的像素值设为 0.

对于对抗图像 I_{adver}, 将其送入 YOLOv2、YOLOv3 和 Faster R-CNN 等目标检测器, 由式 (2) 得到对应置信度分数, 并根据式 (3) 计算针对各检测器的攻击损失 $\mathcal{L}^{yolov2}$、$\mathcal{L}^{yolov3}$ 和 $\mathcal{L}^{frcnn}$. 受方法 [30−33] 启发, 本文将针对各个检测器的攻击视为多个任务, 并在损失函数中引入动态重加权策略, 以应对多任务收敛差异造成攻击不强的问题. 具体而言, 整体损失函数将由式 (4) 修改为如下形式:

$$\mathcal{L}_{obj}^{reweight} = \sum_{k}\alpha^{k}\mathcal{L}^{k} \tag{7}$$

其中, $k \in \{yolov2, yolov3, frcnn\}$ 表示多个攻击任务; α^{k} 是针对第 k 个任务的重加权系数, 其随着迭代优化的过程而不断动态变化. 当通过反向传播的链式法则对补丁 δ 进行优化更新时, 可以得到:

$$\nabla_{\delta}\mathcal{L}_{obj}^{reweight} = \sum_{k}\alpha^{k}\frac{\partial\mathcal{L}^{k}}{\partial\delta} \tag{8}$$

$$\delta \leftarrow \delta - \eta\cdot\mathrm{Adam}(\nabla_{\delta}\mathcal{L}_{obj}^{reweight}) \tag{9}$$

其中, η 表示参数更新的学习率. 可以看到, 针对各个任务所引入的动态权重被直接应用于其对应的梯度项. 通过对各任务梯度的自适应加权, 可以实现针对梯度幅度和方向的协调调控, 从而有效缓解任务间的收敛差异, 提升整体训练的稳定性与性能表现. 如图 3 所示, 为了实现此目标, 本文从局部收敛状态和全局收敛状态两个角度考察各任务的优化情况, 并根据它们确定动态权重 α^{k} 的取值.

全局校正因子. 如第 2.1 节所述, 对抗攻击的目标是降低检测器输出的置信度分数, 即优化 V_{obj}^{adver},

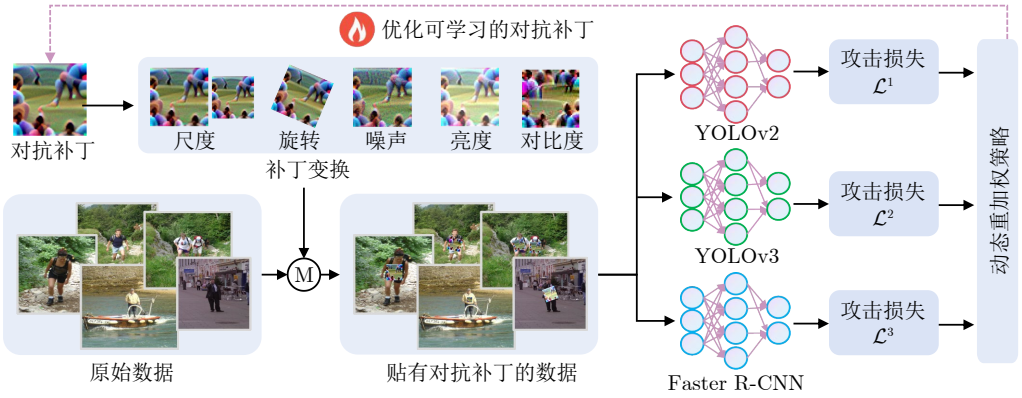


图 2　可迁移性对抗补丁生成的整体流程

Fig. 2　The overall workflow of the transferable adversarial patch generation

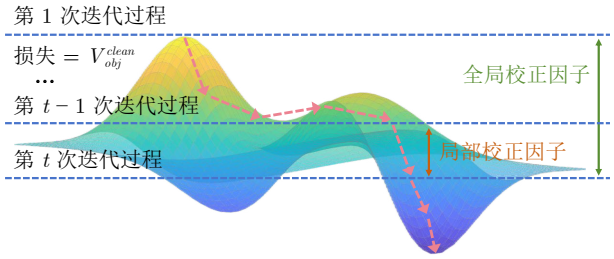


图 3 动态重加权策略

Fig. 3 The dynamic reweighting strategy

使其相对干净数据的置信度分数 V_{obj}^{clean} 逐渐降低. 针对这一宏观的优化过程, 设计了全局校正因子 g^k, 定义为当前迭代下损失的值与干净置信度分数之比, 公式如下:

$$g^k(iter) = \frac{\mathcal{L}_{iter}^k}{\frac{1}{N}\sum_{n=1}^{N}(V_{obj}^{clean})^{nk}} \tag{10}$$

其中, $iter$ 表示当前迭代次数. 由于所攻击的各个目标检测器均在 COCO 等大型基准数据集上进行了充分训练, 因此它们在干净数据上均能产生接近 1 的高置信度分数 V_{obj}^{clean}, 即各任务的初始攻击损失是几乎一致的. 而在后续优化中, 随着攻击难度的不一致, 补丁优化会更倾向于收敛较快、梯度主导的任务, 而对难以收敛的任务关注不足, 最终影响整体模型的迁移能力与任务平衡表现. 因此, g^k 能够给予攻击困难的检测器更大的权重, 以增强其在梯度更新中的影响, 实现充分优化.

局部校正因子. 为提升在局部优化中的协调性, 引入局部校正因子 l^k, 根据每个任务的局部收敛行为动态调整其优化速度. 具体地, l^k 为每个攻击任务计算了当前迭代损失与前一次迭代损失之比:

$$l^k(iter) = \frac{\mathcal{L}_{iter}^k}{\mathcal{L}_{iter-1}^k} \tag{11}$$

当某个任务的损失相比前一次迭代出现上升趋势时, 表明其偏离了理想的收敛路径, 此时对应的 l^k 值会随之增加并超过 1, 从而增强该任务梯度的权重, 引导优化过程优先对其进行调整以修正次优收敛表现. 鉴于在小批量训练中样本质量波动可能导致损失波动, 进而造成系数不稳定, 因此引入了指数移动平均 (exponential moving average, EMA) 策略, 对历史损失进行平滑处理:

$$\tilde{\mathcal{L}}_{iter-1}^k = \beta\tilde{\mathcal{L}}_{iter-2}^k + (1-\beta)\mathcal{L}_{iter-1}^k \tag{12}$$

其中, $\beta \in [0, 1]$ 是一个控制平滑程度的超参数. β 的引入赋予历史损失一定的权重, 从而保证局部校正因子 l^k 更加平滑, 有效减缓局部波动和异常值带来的干扰. 随后, 式 (11) 可以修改为:

$$l^k(iter) = \frac{\mathcal{L}_{iter}^k}{\tilde{\mathcal{L}}_{iter-1}^k} \tag{13}$$

动态重加权. 全局校正因子 g^k 和局部校正因子 l^k 分别从全局和局部收敛状态衡量了各任务的优化情况. 因此, 动态重加权权重 α^k 由两者组合而成, 以促进不同任务的平衡优化, 公式表示如下:

$$\alpha^k = \frac{\exp\left(g^k \cdot \frac{l^k}{\tau}\right)}{\sum_{j=1}^{K}\exp\left(g^j \cdot \frac{l^j}{\tau}\right)} \tag{14}$$

其中, τ 表示温度系数. 通过动态重加权权重 α^k, 可以更有效地优先处理更需要关注的任务, 提升多个攻击任务学习过程的整体性能, 从而促进对抗补丁 δ 针对不同检测器的可迁移攻击.

最后, 在优化中还引入了总变分损失以鼓励补丁颜色的平滑过渡:

$$\mathcal{L}_{tv} = \sum_{i,j}\sqrt{(\delta_{i+1,j}-\delta_{i,j})^2 + (\delta_{i,j+1}-\delta_{i,j})^2} \tag{15}$$

$\mathcal{L}_{tv}$ 可以防止对抗补丁 δ 出现杂乱的噪声模式. 因此, 完整的损失函数包含了针对多检测器攻击的重加权攻击损失以及针对补丁表观约束的总变分损失:

$$\mathcal{L}_{total} = \mathcal{L}_{obj}^{reweight} + \mathcal{L}_{tv} \tag{16}$$

3 实验

3.1 实验设置

实验细节. 本文使用深度学习框架 PyTorch 实现模型的搭建, 整个训练过程在单个 NVIDIA TITAN RTX GPU 上进行. 优化过程中采用 Adam 优化器, 初始学习率设为 0.01, 动量因子设置为 $\beta_1 = 0.500$、$\beta_2 = 0.999$, 训练过程所采用的批量大小为 4, 对抗补丁的分辨率设为 300 × 300 × 3, 像素值全部初始化为 128. 对于待攻击的目标检测器, 本文使用了 YOLOv2[52]、YOLOv3[53]、YOLOv3-tiny、YOLOv4[54]、YOLOv4-tiny 和 Faster R-CNN[55]. 上述方法基于卷积神经网络架构实现, 为验证所生成补丁对于更先进的基于 Transformer 架构检测器的迁移性, 本文还选择 DETR[56] 进行验证实验. 而对于攻击方法, 选取 UPC[57]、NAP[28] 和 DAP[58] 作为对比方法, 同时引入白色补丁、灰色补丁以及随机噪声构成的补丁作为参照. 对于各个补丁, 本文将其按照目标真实边界框的中心位置进行 0.2 倍的缩放和粘贴. 在该设置下, 不同方法之间实现了公平比较.

数据集. 在数字域实验中, 本文采用 InriaPerson 数据集[59] 的训练集进行训练, 并在该数据集的测试集上进行验证, 随后直接在更大规模的数据集 COCO-Person 测试集和 CCTV-Person 测试集上进行泛化性测试, 这两个数据集分别包括 2 693 张和 559 张图像. 如图 4 所示, InriaPerson 数据集中每个人物边界框的面积占整张图像的比例多变 (从 0.002 至 0.448), 充分体现了该数据集的多样性与复杂性. 因此, 在 InriaPerson 训练集上进行优化, 有助于学习到更具普适性的攻击特征, 这为评估对抗补丁在不同尺寸目标上的攻击鲁棒性提供了良好的实验基础. 在实验中, 所有用于训练和测试的图像均被统一缩放至 416 × 416 的分辨率, 然后送入目标检测器. 为了进一步验证对抗补丁在真实物理环境下的攻击有效性, 本文还设计了物理域实验. 具体而言, 将生成的对抗补丁打印成 45 cm × 45 cm 的实物贴纸, 并粘贴于 T 恤表面. 随后使用摄像机对穿戴者进行拍摄, 采集得到多段行人场景视频. 对于采集的视频, 每间隔 5 帧对视频进行采样, 生成包含 355 张图像的数据集, 对检测结果进行评估. 视频内容涵盖室内与室外多种环境, 包括教室、走廊、操场等典型场景, 并在不同距离与光照条件下采集, 以保证实验的丰富性与代表性. 此外, 由于攻击策略针对预测框中目标存在性的置信度分数进行干扰, 因此在原理上具有跨类别的普适性. 为说明其适用性, 本文还在车辆目标上进行了验证实验.

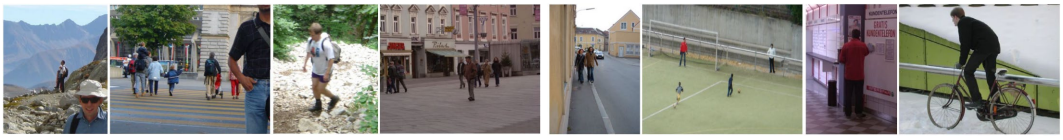
评估指标. 本文采用平均精度 (average precision, AP) 作为评估对抗补丁攻击性能的主要定量指标. AP 综合考虑了检测模型在不同置信度阈值下的精确率与召回率, 是目标检测任务中广泛使用且具有代表性的性能度量方式. 在本研究中, 将目标检测器在干净图像上输出的边界框检测结果视为真实框, 即没有粘贴对抗补丁时, 检测器的 AP 将达到 100%. 而 AP 值越低, 表示模型越难以正确检测到目标, 从而说明对抗补丁具备更强的干扰能力.

3.2　数字域对比实验

定量比较. 为评估在多目标检测器场景下的对抗攻击迁移能力, 本节在表 1 中展示了不同攻击策略下的检测精度对比结果, 相应的对抗补丁如图 5 所示. 在表 1 的第一列中, “Ours-Ensemble-sum”表示在原始静态加权策略下针对多个检测器的联合攻击结果, “Ours-Ensemble-reweight”表示在本文重加权策略下的联合攻击结果. 类似地, “攻击方法–检测器”表示所对比的各个攻击方法针对该检测器训练后的测试结果. 越低的 AP 表示越优秀的攻击效果.

表 1 首先比较了针对单一检测器训练的对抗补丁 ($P_1 \sim P_6$). 从结果来看, 此类方法在各自目标检测器上表现出显著的攻击效果. 例如, 针对 YOLOv2 的攻击将其检测精度降至 2.68%, 针对 YOLOv3-tiny 的攻击降至 6.79%, 而 YOLOv4 的攻击则使其精度下降至 4.37%. 这表明, 当攻击优化专注于单一模型时, 能够实现较强的破坏性. 然而, 这类方法在其他目标检测器上的攻击迁移能力普遍较弱. 例如, 针对 YOLOv3-tiny 优化的对抗补丁虽然在其本身上效果显著, 但在 YOLOv2、YOLOv3、YOLOv4 和 Faster R-CNN 等模型上仍保持较高的检测精度, 均在 40% 以上, 难以满足实际多模型对抗的需求.

实验还引入了基于静态权重的集成攻击策略作为对比, 通过对多个目标检测模型的损失函数加权求和进行联合优化. 实验结果表明, 相较于针对单一模型的针对性优化方式, 联合优化生成的对抗补丁在未参与训练的检测器上表现出一定程度的迁移能力提升, 但在训练中所涉及的检测器上攻击强度有较大的下降. 具体而言, 联合补丁在 YOLOv2、YOLOv3 及 Faster R-CNN 上训练后, 针对 YOLOv4 模型的攻击, 其效果 (16.49%) 优于针对 YOLOv3 (44.39%)、YOLOv3-tiny (58.74%) 甚至是 YOLOv4-tiny (51.15%) 单独训练的补丁. 这表明联合优化在一定程度上提升了对未知模型的适应能力. 然而, 整体来看, 该策略在多模型攻击中的表现呈现出“平均化”特征, 即在各个目标检测模型上的攻击效果均有所下降, 难以实现针对单一模型的最优破坏效果. 这表明简单的损失求和方法难以充分协调多模型间的优化目标, 限制了多模型联合攻击的整体性能. 相比上述静态权重融合策略, 本文



(a) 训练集中的图像
(a) Images in the training set

(b) 测试集中的图像
(b) Images in the test set

图 4　InriaPerson 数据集中的行人图像

Fig. 4　Pedestrian images on the InriaPerson dataset

表 1 在 InriaPerson 数据集上攻击性能的定量比较 (%)

Table 1 Quantitative comparison of attack performance on the InriaPerson dataset (%)

方法	YOLOv2	YOLOv3	YOLOv3-tiny	YOLOv4	YOLOv4-tiny	Faster R-CNN	DETR
(P_1) Ours-YOLOv2	**2.68**	22.51	8.74	12.89	**4.74**	39.41	22.05
(P_2) Ours-YOLOv3	36.47	15.07	66.20	44.39	39.85	57.38	38.84
(P_3) Ours-YOLOv3-tiny	57.61	67.81	*6.79*	58.74	44.45	64.52	46.97
(P_4) Ours-YOLOv4	28.33	31.71	44.35	*4.37*	30.74	44.11	33.14
(P_5) Ours-YOLOv4-tiny	58.46	63.11	44.23	51.15	9.88	70.18	46.86
(P_6) Ours-Faster R-CNN	7.55	*7.09*	10.27	10.89	10.18	*18.47*	*18.19*
(P_7) Ours-Ensemble-sum	12.27	14.98	43.70	16.49	27.10	41.35	38.89
(P_8) Ours-Ensemble-reweight	*5.90*	**1.86**	**3.01**	**3.94**	*6.81*	**15.38**	**8.89**
(P_9) Gray Patch	72.66	74.17	67.52	66.52	60.93	61.54	52.86
(P_{10}) White Patch	69.63	74.93	66.45	72.48	65.13	65.40	46.52
(P_{11}) Random Noise	75.03	73.75	78.91	76.71	76.66	73.00	51.18
(P_{12}) UPC-YOLOv2	48.62	54.40	63.82	64.21	63.03	61.87	47.58
(P_{13}) DAP-YOLOv3	35.66	30.48	47.36	37.11	39.33	62.30	40.74
(P_{13}) DAP-YOLOv3-tiny	59.19	58.73	6.99	41.62	30.42	70.43	48.41
(P_{14}) DAP-YOLOv4	25.40	45.22	46.73	24.27	51.33	55.05	39.59
(P_{15}) DAP-YOLOv4-tiny	23.90	47.54	23.08	50.62	12.99	60.00	37.08
(P_{16}) NAP-YOLOv2	12.06	43.05	32.12	50.56	39.47	52.54	41.78
(P_{18}) NAP-YOLOv3	56.67	34.93	41.46	56.29	53.59	61.78	43.17
(P_{19}) NAP-YOLOv3-tiny	31.61	28.81	10.02	65.13	31.62	55.08	35.17
(P_{20}) NAP-YOLOv4	44.27	56.59	56.61	22.63	58.23	59.42	44.76
(P_{21}) NAP-YOLOv4-tiny	34.68	37.79	21.69	46.80	23.70	59.97	37.78
(P_{22}) NAP-Faster R-CNN	28.26	39.05	37.06	51.46	28.68	42.47	45.09

注: 在下文的表格和图中, YOLOv3-tiny 和 YOLOv4-tiny 分别简写为 YOLOv3t 和 YOLOv4t, Faster R-CNN 简写为 FRCNN, Ensemble-reweight 和 Ensemble-sum 分别简写为 Ense.-rewe. 和 Ense.-sum.

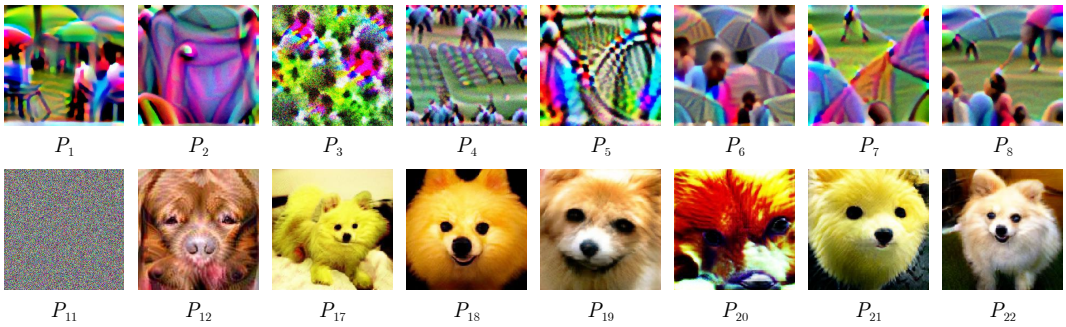


图 5 InriaPerson 数据集上各方法生成的对抗补丁

Fig. 5 Adversarial patches generated by various methods on the InriaPerson dataset

设计的动态重加权策略显著改善了多模型联合攻击的整体表现. 正如表 1 所示, 本文方法不仅相较于静态权重集成策略具有更强的攻击鲁棒性, 且多项攻击效果优于单模型攻击. 这表明, 本文所提出的动态重加权机制能够实现训练过程中的自适应任务平衡与动态协同, 从而在多目标检测器联合攻击中取得更强的攻击能力和更高的跨模型迁移性.

上述结果展示了在卷积神经网络架构下的迁移性能. 为进一步评估其迁移攻击能力, 实验扩展至基于 Transformer 的检测器 DETR. 如表 1 所示, 各设置下生成的对抗补丁被直接应用于 DETR 进行评估. 可以看到, 不同检测器上针对性训练的补丁均取得了一定的攻击效果: Ours-YOLOv2、Ours-YOLOv3、Ours-YOLOv3-tiny、Ours-YOLOv4 与 Ours-YOLOv4-tiny 的攻击分别将 AP 降至 22.05%、38.84%、46.97%、33.14% 和 46.86%. 相比之下, 静态加权方式下生成的补丁使得 AP 降低至 38.89%, 而本文方法 DRPatch 提出的动态重加权策略则表现最为突出, 使得 AP 进一步下降至 8.89%. 值得强调的是, 即便在没有针对 DETR 进行特定攻击训练

的情况下, DRPatch 依然在该先进检测器上取得了最佳的攻击效果, 从而验证了所提策略在更复杂检测模型上的攻击迁移性.

定性比较. 为进一步直观验证所提方法在多目标检测器上的攻击效果, 本文在定性分析中选取了多个具有代表性的补丁, 在 InriaPerson 数据集上对不同检测模型的攻击表现进行了对比分析. 所选补丁包括: 在定量评估中表现突出的补丁 P_1 和 P_6, 由静态权重融合策略生成的补丁 P_7, 基于本文所提出动态重加权机制生成的补丁 P_8 以及作为参考的随机噪声补丁 P_{11}, 以对比不同策略下补丁对多个检测器攻击能力的表现差异.

如图 6 所示, 图中以红框标注各检测器对图像中行人目标的检测结果. 实验结果表明, 在无补丁扰动的原始图像中, YOLOv2、YOLOv3、YOLOv3-tiny、YOLOv4、YOLOv4-tiny 以及 Faster R-CNN 等主流目标检测的模型均能准确检测出图像中的目标, 呈现出稳定且高置信度的检测输出. 而在随机噪声补丁 P_{11} 的干预下, 各检测模型的输出几乎未发生显著变化, 目标检测框的位置和置信度与未攻击时基本一致. 这一现象表明, 深度神经网络虽然存在潜在的脆弱性, 但这类漏洞并非可由随机扰动轻易激活, 而是需要通过有针对性的对抗性优化过程加以精确挖掘. 相较而言, 尽管对抗补丁 P_1、P_6 和 P_7 能够诱导检测性能下降, 但仍然存在较多图像中目标被成功识别的情况, 攻击效果在不同模型之间表现不均, 反映出这些方法在模型迁移性与攻击稳定性方面的不足. 特别是静态加权策略生成的补丁 P_7, 由于未能动态适应不同任务间的优化需求, 其攻击效果表现出明显的“平均化”倾向, 即在多个模型中均产生有限的干扰, 但难以在所有模型上达到显著的攻击效果. 相比之下, 基于本文提出的多任务动态重加权机制生成的对抗补丁 P_8 展现出显著的优势. 该补丁在所有检测模型中均能诱发严重的感知偏差, 使得检测置信度显著下降. 这一结果充分地证明了动态重加权机制在提升多模型联合攻击中的作用, 该机制能够有效缓解不同检测器在结构特性与收敛动态方面的差异对训练过程带来的干扰, 从而生成具有更强迁移能力的对抗补丁. 图 7 展示了在检测器 DETR 上的两组定性结

图 6　在 InriaPerson 数据集上攻击性能的定性比较

Fig. 6　Qualitative comparison of attack performance on the InriaPerson dataset

(a) 在 DETR 检测器上的定性比较 (场景 1)
(a) Qualitative comparison on the DETR detector (Scenario 1)

(b) 在 DETR 检测器上的定性比较 (场景 2)
(b) Qualitative comparison on the DETR detector (Scenario 2)

图 7 在 DETR 检测器上的定性比较

Fig. 7 Qualitative comparison on the DETR detector

果. 可以观察到, 相较于其他方法, 本文方法 DRPatch 的动态重加权策略能够最大程度地削弱检测模型的预测能力, 展示了在跨架构黑盒攻击下的有效迁移潜力.

3.3 数字域泛化实验

本节在 COCO-Person 与 CCTV-Person 测试集上进行了跨数据集泛化测试, 并报告了定量与定性结果. 在 COCO-Person 测试集的定量结果如表 2 所示. 可以看到, 对单一检测器进行针对性训练和测试的补丁在大多数检测器上显著优于静态加权策略. 例如, 在 YOLOv2、YOLOv4 与 YOLOv4-tiny 上, 其 AP 分别降至 7.41%、15.22% 与 17.40%, 均低于静态加权对应结果 (18.80%、23.82% 与 27.87%). 而本文提出的多任务动态重加权机制在跨模型攻击迁移性上展现出更优性能. 具体而言, 该方法在多个检测模型上均取得了最低的 AP, 在 YOLOv3、YOLOv4 与 YOLOv4-tiny 上分别降至 6.09%、11.01% 与 14.79%, 显著优于单模型与静态加权策略. 在 CCTV-Person 测试集上的结果如表 3 所示, 同样验证了这一结论. 本文方法 DRPatch 在整体上获得了最优的攻击迁移性能. 例如, 在 YOLOv3 上, 相比单模型训练的 AP (8.14%) 和静态加权策略的 AP (7.42%), 动态重加权机制能够将 AP 降至 0.50%, 显著提升了攻击效果.

进而, 在 COCO-Person 与 CCTV-Person测试集上的定性结果如图 8 所示. 从图中可以直观地观察到, 与单模型训练方法和静态加权策略生成的对抗补丁相比, 本文提出的多任务动态重加权机制所生成的补丁在不同检测模型上均能够持续诱发更加显著的感知偏差. 具体而言, 该补丁能够有效降低检测置信度, 从而大幅削弱检测器对目标的识别能力. 即便在复杂场景、背景干扰和目标尺度显著变化等挑战条件下, 本文方法依然能够始终保持鲁棒的攻击效果.

表 2 在 COCO-Person 数据集上的定量比较 (%)

Table 2 Quantitative comparison on the COCO-Person dataset (%)

方法	YOLOv2	YOLOv3	YOLOv3t	YOLOv4	YOLOv4t	FRCNN
针对性训练	**7.41**	30.92	12.82	15.22	17.40	27.13
Ours-Ense.-sum	18.80	25.13	37.43	23.82	27.87	40.87
Ours-Ense.-rewe.	9.84	**6.09**	**11.02**	**11.01**	**14.79**	**24.17**

表 3 在 CCTV-Person 数据集上的定量比较 (%)

Table 3 Quantitative comparison on the CCTV-Person dataset (%)

方法	YOLOv2	YOLOv3	YOLOv3t	YOLOv4	YOLOv4t	FRCNN
针对性训练	**0.82**	8.14	4.43	0.43	5.95	17.16
Ours-Ense.-sum	9.41	7.42	39.20	2.60	21.76	37.89
Ours-Ense.-rewe.	2.89	**0.50**	**2.88**	**0.26**	**4.00**	**12.20**

3.4 数字域消融实验

为验证所提出动态重加权机制中各组成部分的有效性, 本文设计并开展了系统性的消融实验. 具体地, 分别构建了两种变体模型: 一是仅引入全局校正因子, 该因子主要用于捕捉各子任务在整体优化过程中的相对进展; 二是仅引入局部校正因子, 该因子从微观层面建模单一子任务的收敛趋势. 实验结果如表 4 所示, 与完整引入全局校正因子 g^k 与局部校正因子 l^k 的联合模型相比, 这两种单一因子模型在对抗攻击迁移性方面仍存在一定差距. 这说明两者的共同参与能够实现对多任务优化过程更细致建模与精确调控, 显著提升了对抗补丁针对不同检测器的攻击有效性.

(a) 在 COCO-Person 数据集上的定性比较
(a) Qualitative comparison on the COCO-Person dataset

(b) 在 CCTV-Person 数据集上的定性比较
(b) Qualitative comparison on the CCTV-Person dataset

图 8　在 COCO-Person 和 CCTV-Person 数据集上的定性比较

Fig. 8　Qualitative comparison on the COCO-Person and CCTV-Person datasets

表 4　消融实验的定量比较 (%)

Table 4　Quantitative comparison of ablation experiments (%)

方法	YOLOv2	YOLOv3	YOLOv3t	YOLOv4	YOLOv4t	FRCNN
Only with g^k	19.36	18.05	34.99	24.87	27.69	38.37
Only with l^k	10.98	19.46	41.58	21.98	16.47	42.37
Ours-Ense.-rewe.	**5.90**	**1.86**	**3.01**	**3.94**	**6.81**	**15.38**

3.5　数字域对抗防御实验

为验证对抗补丁针对防御模型的攻击效果, 本节选取了通用型防御方法 Oddefense[60]、专用于对抗补丁设计的防御方法 PBCAT[61] 和 FNS[62]. 如表 5 所示, 选用 Faster R-CNN、YOLOv3 和 YOLOv3-tiny 作为防御实验的验证对象, 并选择针对这三类检测器的相关补丁进行比较. 值得注意的是, 本文方法 DRPatch 在对抗补丁优化过程中能够有效习得多类型检测模型的目标特征感知漏洞, 因此 DRPatch 在面向通用型防御方法 Oddefense 时依旧保持较高的攻击性能. 针对对抗补丁设计的防御方法 PBCAT 通过结合小区域梯度引导的对抗性补丁和覆盖整个图像、难以察觉的全局对抗性扰动来优化模型, 有效防御了物理世界下面向检测器的攻击. 如表 5 所示, 在 PBCAT 的防御下, DRPatch 的攻击效果有所减弱, 但是依旧能对检测器实现一定程度上的攻击. 同样, 在引入 FNS 防御机制后, DRPatch 仍能保持较好的攻击效果. 在 YOLOv3 检测器上, 对其进行针对性训练的对抗补丁可将 AP 降至 15.07%, 而当面对 FNS 防御时, AP 升至 30.13% (干净检测的 AP 为 100%). 在静态加权策略下, AP 从攻击时的 14.98% 升至防御时的 29.98%. 而本文的动态重加权策略下, AP 在攻击时为 1.86%, 而面向防御时则升至 13.16%, 这甚至与其他两种攻

表 5　典型防御模型下的攻击性能定量评估 (%)

Table 5　Quantitative evaluation of attack performance under typical defense models (%)

方法	Ours-FRCNN	Ours-Ense.-sum	Ours-Ense.-rewe.
No defense	18.47	41.35	15.38
Oddefense	30.41	44.66	18.16
PBCAT	37.94	55.57	24.73
方法	Ours-YOLOv3	Ours-Ense.-sum	Ours-Ense.-rewe.
No defense	15.07	14.98	1.86
FNS	30.13	29.98	13.16
方法	Ours-YOLOv3t	Ours-Ense.-sum	Ours-Ense.-rewe.
No defense	6.79	43.70	3.01
FNS	6.96	47.45	4.59

击在未经防御情况下的攻击能力相当. 因此, 本文提出的基于动态重加权策略的迁移性方法不仅在未防御情况下获得最优攻击性能, 在面对防御时也能最好地保持攻击效果.

3.6 与迁移性攻击方法的比较

本节选择迁移性攻击方法 DAS[63] 和 AdaEA[34] 进行比较. 对于方法 DAS, 将实验范围扩展至另一代表性目标, 即针对车辆的检测, 这可以同时验证本文方法在多种目标下的适用性. 数字域下的定量结果如表 6 所示, DAS 的攻击使 YOLOv2、YOLOv3、YOLOv4 与 Faster R-CNN 的 AP 分别下降至 37.32%、35.34%、32.92% 和 39.09%, 而本文方法 DRPatch 则进一步将 AP 降低至 33.05%、30.97%、25.20% 和 32.95%, 表现出更强的攻击效果. 在轻量化的 YOLOv3-tiny 和 YOLOv4-tiny 上, DAS 的攻击结果分别为 2.51% 和 1.16%, 而本文方法则进一步下降至 1.32% 和 0.68%. 可以看到, 在六个主流检测器上, DRPatch 均取得更低的检测精度, 体现出更强的攻击效果和跨模型迁移性. 进一步地, 物理域下针对车辆场景的定性检测结果如图 9 所示. 与方法 DAS 相比, DRPatch 在真实环境中能够更有效地干扰多个检测器的识别, 使目标车辆难以被正确检测. 而对于对比方法 AdaEA, 在 InriaPerson 数据集上进行迁移性比较. 可以看到, 本文方法 DRPatch 在大多数检测器上获得了更优秀的攻击性能. 例如, 在 YOLOv2 上, AdaEA 使得 AP 降低至 12.81%, 而 DRPatch 使得 AP 降低至 5.90%. 在 Faster R-CNN 上, AdaEA 使得 AP 降低至 34.01%, 而 DRPatch 使得 AP 降低至 15.38%.

表 6 与迁移性攻击方法 DAS 和 AdaEA 的定量比较 (%)
Table 6 Quantitative comparison with transferable attack methods DAS and AdaEA (%)

方法	YOLOv2	YOLOv3	YOLOv3t	YOLOv4	YOLOv4t	FRCNN
DAS	37.32	35.34	2.51	32.92	1.16	39.09
Ours	**33.05**	**30.97**	**1.32**	**25.20**	**0.68**	**32.95**
Δ	4.27 ↓	4.37 ↓	1.19 ↓	7.72 ↓	0.48 ↓	6.14 ↓
AdaEA	12.81	2.68	5.57	4.09	**5.28**	34.01
Ours	**5.90**	**1.86**	**3.01**	**3.94**	6.81	**15.38**
Δ	6.91 ↓	0.82 ↓	2.56 ↓	0.15 ↓	−1.53 ↓	18.63 ↓

同时, 本节还对训练时间进行了定量分析. 具体而言, AdaEA 训练 1 个 epoch 的平均时间为 1 342 s, 而 DRPatch 仅为 434 s, 节省了 908 s. 在相同资源条件下 (单个 NVIDIA TITAN RTX GPU), DRPatch 的批大小可设置为 4, 而 AdaEA 仅能使用批大小为 1, 相同的批量需增加大量计算资源. 因此, 无论在攻击效果还是训练效率上, 本文方法 DRPatch 均表现出更优秀的性能.

3.7 物理域行人检测验证实验

对于物理域下拍摄的行人视频帧数据集, 本节在六个检测器上提供了物理域实验的定量评估结果, 如表 7 所示. 结果表明, 相较于静态加权策略, 本文提出的动态重加权策略在所有检测器上均

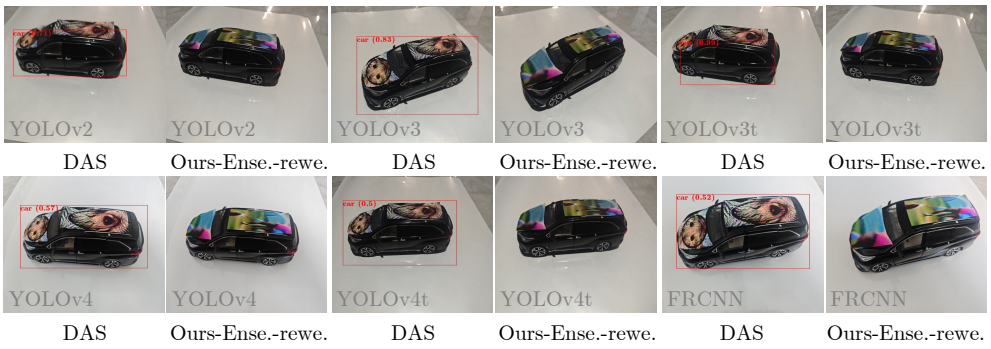


图 9 在车辆目标上与 DAS 方法的定性比较
Fig. 9 Qualitative comparison with DAS method on car targets

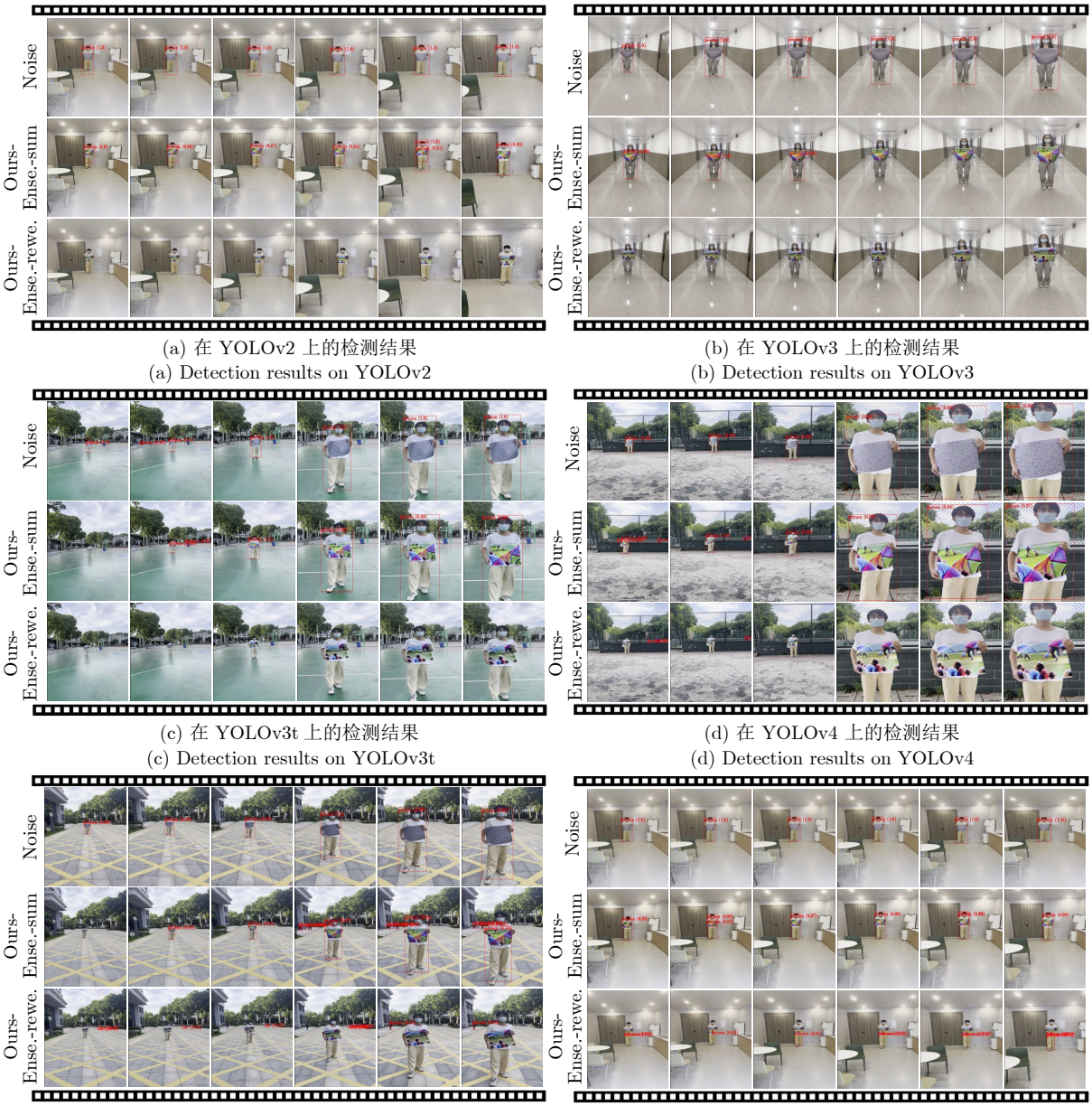
表 7 在物理域视频上攻击性能的定量比较 (%)
Table 7 Quantitative comparison of attack performance on physical-domain videos (%)

方法	YOLOv2	YOLOv3	YOLOv3t	YOLOv4	YOLOv4t	FRCNN
Ours-Ense.-sum	49.62	33.56	47.32	34.00	48.16	33.60
Ours-Ense.-rewe.	**32.94**	**17.70**	**19.55**	**22.60**	**35.10**	**16.78**
Δ	16.68 ↓	15.86 ↓	27.77 ↓	11.40 ↓	13.06 ↓	16.82 ↓

取得了显著更低的 AP 指标, 最低改善幅度达 11.40%. 例如, 在 YOLOv3-tiny 上, AP 由 47.32% 降至 19.55%, 在 Faster R-CNN 上由 33.60% 降至 16.78%. 这些结果充分说明, 所提出的方法能够在物理域中实现更强的跨模型攻击迁移性, 进一步验证了其在真实场景中的实用性. 六个检测器上的定性比较结果如图 10 所示, 其中检测到的目标以红色边框标注. 对比方法包括: 随机噪声补丁、静态加权策略生成的补丁以及本文提出方法生成的补丁. 实验结果显示, 随机噪声补丁几乎无法对检测器产生干扰; 静态加权策略生成的补丁虽然在一定程度上降低了检测置信度, 但攻击效果有限; 而本文方法生成的补丁能够使目标检测置信度大幅降低, 显著削弱检测器对目标的识别能力, 充分验证了本方法在现实场景下跨检测器攻击中的迁移有效性.

4　结束语

针对对抗补丁在不同检测器下跨模型攻击困难的问题, 本文提出一种基于多任务动态重加权机制



(a) 在 YOLOv2 上的检测结果
(a) Detection results on YOLOv2

(b) 在 YOLOv3 上的检测结果
(b) Detection results on YOLOv3

(c) 在 YOLOv3t 上的检测结果
(c) Detection results on YOLOv3t

(d) 在 YOLOv4 上的检测结果
(d) Detection results on YOLOv4

(e) 在 YOLOv4t 上的检测结果
(e) Detection results on YOLOv4t

(f) 在 FRCNN 上的检测结果
(f) Detection results on FRCNN

图 10　真实物理域中检测结果的定性比较

Fig. 10　Qualitative comparison of detection results in the real physical-domain

的可迁移对抗补丁生成框架. 该方法从任务间全局优化进度与单任务局部收敛行为两个层面出发, 设计全局与局部校正因子以动态调整模型任务权重, 从而有效协调不同检测模型在脆弱性分布和优化动态上的差异, 显著提升了联合攻击优化的稳定性与协同性. 数字域实验证明所提方法在多个主流目标检测模型中均表现出优越的对抗攻击迁移性, 而物理域实验展示了对抗补丁在复杂真实物理环境中的部署可行性, 对提升目标检测器的安全性评估具有重要理论意义和应用价值.

参考文献

1 Liu C, Dong Y P, Xiang W Z, Yang X, Su H, Zhu J, et al. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *International Journal of Computer Vision*, 2025, **133**(2): 567−589

2 Xiao C, An W, Zhang Y F, Su Z, Li M, Sheng W D, et al. Highly efficient and unsupervised framework for moving object detection in satellite videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 11532−11539

3 Zhou T F, Wang W G. Prototype-based semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(10): 6858−6872

4 Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(7): 3523−3542

5 Yi X P, Tang L F, Zhang H, Xu H, Ma J Y. Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 2024, **110**: Article No. 102450

6 Masana M, Liu X L, Twardowski B, Menta M, Bagdanov A D, van de Weijer J. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(5): 5513−5533

7 Yi X P, Ma Y, Li Y S, Xu H, Ma J Y. Artificial intelligence facilitates information fusion for perception in complex environments. *The Innovation*, 2025, **6**(4): Article No. 100814

8 Bär A, Houlsby N, Dehghani M, Kumar M. Frozen feature augmentation for few-shot image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). Seattle, USA: IEEE, 2024. 16046−16057

9 Yu Y, Da F P. On boundary discontinuity in angle regression based arbitrary oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(10): 6494−6508

10 Sung C, Kim W, An J, Lee W, Lim H, Myung H. Contextrast: Contextual contrastive learning for semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). Seattle, USA: IEEE, 2024. 3732−3742

11 Chen Jin-Yin, Shen Shi-Jing, Su Meng-Meng, Zheng Hai-Bin, Xiong Hui. Black-box adversarial attack on license plate recognition system. *Acta Automatica Sinica*, 2021, **47**(1): 121−135
(陈晋音, 沈诗婧, 苏蒙蒙, 郑海斌, 熊晖. 车牌识别系统的黑盒对抗攻击. 自动化学报, 2021, **47**(1): 121−135)

12 Wang Lu-Yao, Cao Yuan, Liu Bo-Han, Zeng En, Liu Kun, Xia Yuan-Qing. Ensemble adversarial training defense for time series classification models. *Acta Automatica Sinica*, 2025, **51**(1): 144−160
(王璐瑶, 曹渊, 刘博涵, 曾恩, 刘坤, 夏元清. 时间序列分类模型的集成对抗训练防御方法. 自动化学报, 2025, **51**(1): 144−160)

13 Xu Chang-Kai, Feng Wei-Dong, Zhang Chun-Jie, Zheng Xiao-Long, Zhang Hui, Wang Fei-Yue. Research on black-box attack algorithm by targeting ID card text recognition. *Acta Automatica Sinica*, 2024, **50**(1): 103−120
(徐昌凯, 冯卫栋, 张淳杰, 郑晓龙, 张辉, 王飞跃. 针对身份证文本识别的黑盒攻击算法研究. 自动化学报, 2024, **50**(1): 103−120)

14 Xiang X Y, Yan Q L, Zhang H, Ding J F, Xu H, Wang Z Y, et al. Cross-modal stealth: A coarse-to-fine attack framework for RGB-T tracker. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI, 2025. 8620−8627

15 Wei H, Tang H, Jia X M, Wang Z X, Yu H X, Li Z B, et al. Physical adversarial attack meets computer vision: A decade survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 9797−9817

16 Li H, Dang K L, Gong M G, Qin A K, Zhou Y, Wu Y, et al. Sparse unmixing guided adversarial attack for hyperspectral image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026, **36**(2): 2318−2331

17 Jia S, Yin B J, Yao T P, Ding S H, Shen C H, Yang X K, et al. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: Curran Associates Inc., 2022. Article No. 2474

18 Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C W, et al. Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1625−1634

19 Finlayson S G, Bowers J D, Ito J, Zittrain J L, Beam A L, Kohane I S. Adversarial attacks on medical machine learning. *Science*, 2019, **363**(6433): 1287−1289

20 Gu J D, Zhao H S, Tresp V, Torr P H S. SegPGD: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In: Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022. 308−325

21 Su J W, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019, **23**(5): 828−841

22 Wei X X, Yan H Q, Li B. Sparse black-box video attack with reinforcement learning. *International Journal of Computer Vision*, 2022, **130**(6): 1459−1473

23 Wei X X, Zhu J, Yuan S, Su H. Sparse adversarial perturbations for videos. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI, 2019. 8973−8980

24 Thys S, van Ranst W, Goedemé T. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, USA: IEEE, 2019. 49−55

25 Wei X X, Guo Y, Yu J. Adversarial sticker: A stealthy attack method in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(3): 2711−2725

26 Xiang X Y, Yan Q L, Zhang H, Ma J Y. ACAttack: Adaptive cross attacking RGB-T tracker via multi-modal response decoupling. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). Nashville, USA: IEEE, 2025. 22099−22108

27 Zhu X P, Hu Z H, Huang S Y, Li J M, Hu X L. Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 13307−13316

28 Hu Y C T, Chen J C, Kung B H, Hua K L, Tan D S. Naturalistic physical adversarial patch for object detectors. In: Proceedings of the IEEE International Conference on Computer Vision

(ICCV). Montreal, Canada: IEEE, 2021. 7828−7837

29 Zhu X P, Li X, Li J M, Wang Z Y, Hu X L. Fooling thermal infrared pedestrian detectors in real world using small bulbs. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI, 2021. 3616−3624

30 Yu T H, Kumar S, Gupta A, Levine S, Hausman K, Finn C. Gradient surgery for multi-task learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: Curran Associates Inc., 2020. Article No. 489

31 Guo M, Haque A, Huang D A, Yeung S, Li F F. Dynamic task prioritization for multitask learning. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 282−299

32 Liu S K, Johns E, Davison A J. End-to-end multi-task learning with attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 1871−1880

33 Wu G, Jiang J J, Jiang K, Liu X M. Harmony in diversity: Improving all-in-one image restoration via multi-task collaboration. In: Proceedings of the 32nd ACM International Conference on Multimedia (MM). Melbourne, Australia: ACM, 2024. 6015−6023

34 Chen B, Yin J L, Chen S K, Chen B H, Liu X M. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 4466−4475

35 Tang B W, Wang Z, Bin Y, Dou Q, Yang Y, Shen H T. Ensemble diversity facilitates adversarial transferability. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 24377−24386

36 Wang Z B, Guo H C, Zhang Z F, Liu W X, Qin Z, Ren K. Feature importance-aware transferable adversarial attacks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 7619−7628

37 Chen J Q, Chen H, Chen K Y, Zhang Y L, Zou Z X, Shi Z W. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(2): 961−977

38 Li Z X, Yin B J, Yao T P, Guo J F, Ding S H, Chen S M, et al. Sibling-attack: Rethinking transferable adversarial attacks against face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 24626−24637

39 Guo Y, Liu W Q, Xu Q S, Zheng S J, Huang S J, Zang Y, et al. Boosting adversarial transferability through augmentation in hypothesis space. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). Nashville, USA: IEEE, 2025. 19175−19185

40 Li Y B, Hu C, Wu X J. Transferable stealthy adversarial example generation via dual-latent adaptive diffusion for facial privacy protection. *IEEE Transactions on Information Forensics and Security*, 2025, **20**: 9427−9440

41 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR). San Diego, USA: ICLR, 2015.

42 Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proceedings of the 6th International Conference on Learning Representations (ICLR). Vancouver, Canada: ICLR, 2018. 1−23

43 Sharif M, Bhagavatula S, Bauer L, Reiter M K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS). Vienna, Austria: ACM, 2016. 1528−1540

44 Xu K D, Zhang G Y, Liu S J, Fan Q F, Sun M S, Chen H G, et al. Adversarial T-shirt! Evading person detectors in a physical world. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 665−681

45 Tan J, Ji N, Xie H D, Xiang X S. Legitimate adversarial patches: Evading human eyes and detection models in the physical world. In: Proceedings of the 29th ACM International Conference on Multimedia (MM). Chengdu, China: ACM, 2021. 5307−5315

46 Yin B J, Wang W X, Yao T P, Guo J F, Kong Z L, Ding S H, et al. Adv-Makeup: A new imperceptible and transferable attack on face recognition. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI). Montreal, Canada: IJCAI.org, 2021. 1252−1258

47 Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I J, et al. Intriguing properties of neural networks. In: Proceedings of the 2nd International Conference on Learning Representations (ICLR). Banff, Canada: ICLR, 2014.

48 Brown T B, Mané D, Roy A, Abadi M, Gilmer J. Adversarial patch. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). Long Beach, USA: Curran Associates Inc., 2017. 1−5

49 Xue H T, Araujo A, Hu B, Chen Y X. Diffusion-based adversarial sample generation for improved stealthiness and controllability. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS). New Orleans, USA: Curran Associates Inc., 2023. Article No. 129

50 Rony J, Pesquet J C, Ben Ayed I. Proximal splitting adversarial attack for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 20524−20533

51 Wei H, Wang Z X, Zhang K W, Hou J Q, Liu Y W, Tang H, et al. Revisiting adversarial patches for designing camera-agnostic attacks against person detection. In: Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: Curran Associates Inc., 2024. 8047−8064

52 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 6517−6525

53 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv preprint arXiv: 1804.02767, 2018.

54 Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv: 2004.10934, 2020.

55 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems (NeurIPS). Montreal, Canada: MIT Press, 2015. 91−99

56 Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020. 213−229

57 Huang L F, Gao C Y, Zhou Y Y, Xie C H, Yuille A L, Zou C Q, et al. Universal physical camouflage attacks on object detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 717−726

58 Guesmi A, Ding R T, Hanif M A, Alouani I, Shafique M. DAP: A dynamic adversarial patch for evading person detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 24595−24604

59 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition (CVPR). San Diego, USA: IEEE, 2005. 886−893

60 Li X, Chen H, Hu X L. On the importance of backbone to the adversarial robustness of object detectors. *IEEE Transactions on Information Forensics and Security*, 2025, **20**: 2387−2398

61 Li X, Zhu Y M, Huang Y F, Zhang W, He Y Z, Shi J, et al. PB-CAT: Patch-based composite adversarial training against physically realizable attacks on object detection. arXiv preprint arXiv: 2506.23581, 2025.

62 Yu C, Chen J S, Wang Y, Xue Y Z, Ma H M. Improving adversarial robustness against universal patch attacks through feature norm suppressing. *IEEE Transactions on Neural Networks and Learning Systems*, 2025, **36**(1): 1410−1424

63 Wang J K, Liu A S, Yin Z X, Liu S C, Tang S Y, Liu X L. Dual attention suppression attack: Generate adversarial camouflage in physical world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 8561−8570

燕庆龙 武汉大学电子信息学院博士研究生. 主要研究方向为计算机视觉, 对抗攻击.
E-mail: qinglong_yan@whu.edu.cn
(**YAN Qing-Long** Ph.D. candidate at the Electronic Information School, Wuhan University. His research interests include computer vision and adversarial attack.)

向昕宇 武汉大学电子信息学院博士研究生. 主要研究方向为计算机视觉, 对抗攻击.
E-mail: xiangxinyu@whu.edu.cn
(**XIANG Xin-Yu** Ph.D. candidate at the Electronic Information School, Wuhan University. His research interests include computer vision and adversarial attack.)

张 浩 武汉大学电子信息学院博士后. 主要研究方向为计算机视觉, 信息融合, 对抗攻击.
E-mail: zhpersonalbox@gmail.com
(**ZHANG Hao** Postdoctor at the Electronic Information School, Wuhan University. His research interests include computer vision, information fusion, and adversarial attack.)

马佳义 武汉大学电子信息学院教授. 主要研究方向为计算机视觉, 机器学习, 模式识别. 本文通信作者.
E-mail: jyma2010@gmail.com
(**MA Jia-Yi** Professor at the Electronic Information School, Wuhan University. His research interests include computer vision, machine learning, and pattern recognition. Corresponding author of this paper.)