中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-19

论文引用格式: Wang Jun, Miao Zhuang, Hou Yifan, Bi Xianghe, Wang Jiabao, Li Yang. Dynamic objective prioritization-Driven optimization method for infrared adversarial patches[J/OL]. Journal of Image and Graphics, XXXX:1-19. DOI: 10.11834/jig.250464.(王俊, 苗壮, 侯壹凡, 毕翔鹤, 王家宝, 李阳. 动态目标优先级驱动的红外对抗贴片优化方法[J/OL]. 中国图象图形学报,XXXX:1-19. DOI: 10.11834/jig.250464.)
[DOI:10.11834/jig.250464]

动态目标优先级驱动的红外对抗贴片优化方法

王俊,苗壮,侯壹凡,毕翔鹤,王家宝,李阳

陆军工程大学指挥控制工程学院,南京　210007

摘　要: **目的**　针对红外目标检测系统的白盒物理对抗攻击方法中,存在的攻击性能与物理可实现性的多目标优化失衡、物理约束不足及工艺复杂等问题,本文提出一种高效、鲁棒且易于部署的红外物理对抗攻击方法,以实现对红外行人检测器的强干扰攻击。**方法**　本文提出一种动态目标优先级驱动的对抗贴片多目标优化方法。该方法以聚合正则化和稀疏二值正则化为约束基础,同时构建动态目标优先级策略,通过实时量化各优化目标的难度自适应调整损失权重,解决传统固定权重导致的优化失衡问题;在此框架下,设计双重投影优化算法,结合投影梯度下降(projected gradient descent,PGD)与加权动量累积策略,提升对抗掩码在物理约束下的全局优化能力与收敛稳定性。**结果**　实验主要在FLIR ADAS v1_3数据集上进行,并在FLIR ADAS v2.0与LLVIP数据集上进一步验证泛化性,采用YOLOv3作为目标检测器,并在数字域和物理域场景下验证攻击效果。在数字域实验中,平均攻击成功率达75.08%,YOLOv3平均精度下降至36.18%,相较于对比实验中的最优基准方法,两项指标上分别高出4.85%和2.91%;且该方法在不同数据集上的攻击成功率与平均精度均为最优,充分证明其泛化能力;在物理域实验中,在不同角度、不同距离,室外复杂情况以及动态场景下,攻击成功率始终保持在79%以上,均超越基准方法。**结论**　本文提出的红外物理对抗攻击方法有效解决了攻击性效果与物理可实现性之间的冲突,实现了高攻击效能与低部署成本的平衡,在多种复杂真实场景中均表现出卓越的攻击效能,为红外目标检测系统的安全性评估与防御机制设计提供了重要参考。

关键词: 物理对抗攻击;目标检测;红外成像;动态目标优先级;双重投影优化

Dynamic objective prioritization-Driven optimization method for infrared adversarial patches

Wang Jun, Miao Zhuang, Hou Yifan, Bi Xianghe, Wang Jiabao, Li Yang

Command and Control Engineering College, Army Engineering University of PLA, Nanjing 210007, China

Abstract: **Objective** With the in-depth penetration of deep learning technology in the field of infrared object detectors, such systems have become the core technical support for key scenarios including night driving early warning in autonomous driving, night security deployment in surveillance systems, and long-distance body temperature screening in epidemic prevention and control. Their detection accuracy and anti-interference ability are directly related to the operational safety and reliability of society and people. The technology of infrared object detectors is widely used due to its advantages of being not restricted by lighting conditions and capable of penetrating partial occlusions. Nevertheless, like other deep learning-based vision systems, it is not immune to security vulnerabilities. However, the emergency of adversarial examples has posed

收稿日期:2025-09-24;**修回日期**:2026-03-04

基金项目:国家自然科学基金(72471240)

Supported by: National Natural Science Foundation of China (72471240)

severe challenges to the robustness and security of infrared object detection models. As a crucial tool for evaluating and enhancing model security, the research on adversarial examples holds significant importance. Existing research on infrared adversarial examples is mainly divided into two categories: digital adversarial attacks and physical adversarial attacks. Although the methods of digital adversarial attacks can achieve high attack success rates, the perturbations only exist in the digital domain and cannot be directly applied to physical domain. Currently, research on physical adversarial attacks mainly focuses on the visible light domain; for infrared images, due to their unique imaging mechanism, they cannot form high-level semantic features such as texture and color, which increases the difficulty of optimizing physical adversarial examples in the infrared domain. Despite extensive active explorations in existing studies, white-box methods which are regarded as the most effective adversarial attack means at present due to their ability to utilize information such as the internal parameters and structure of models. Although existing methods have explored rapid deployment and low production costs, they still suffer from two major drawbacks. The first drawback is imbalance in multi-objective optimization: most methods adopt fixed weights to balance the ability of attack effectiveness and physical realizability and when there is a significant difference when the optimization complexity of achieving strong adversarial effects significantly outweighs that of ensuring physical realizability, extreme situations such as insufficient attack capability or failure to ensure real-world applicability are likely to occur. The second drawback is inadequate design of physical constraints: existing methods mostly only consider basic constraints such as the size and shape of patches, while neglecting pixel-level consistency in the digital domain. This oversight often results in patches that fail under real-world infrared imaging conditions. **Method** We propose a dynamic objective prioritization-Driven optimization method for adversarial patches. First, we introduce two aggregation regularization and sparsity-binarization regularization. Aggregation regularization adopts a graph-theoretic Local Clustering Coefficient (LCC) approximation to force the mask pixels into a compact and continuous geometry, while sparsity-binarization regularization pushes the mask values toward a bi-modal distribution with a sharp boundary, enabling straightforward material cutting. Next, to avoid the optimization imbalance caused by fixed loss weights, we devise a dynamic objective prioritization strategy. Key Performance Indicators (KPI) are defined to quantify the current optimization difficulty of the attack loss, aggregation loss and binarization loss. The Focal Loss(FL) is applied to reallocates computational resources: difficult objectives receive larger gradients, whereas easy objectives are down-weighted, ensuring that the solver always focuses on the most challenging constraints. Finally, we design a dual projection optimization algorithm that synergizes PGD (Projected Gradient Descent) tailored for adversarial tasks with a momentum-carrying weighted update. The first projection confines the updated mask to the feasible set immediately after each gradient step, preventing gradient waste; the second projection recalibrates the mask after momentum fusion, smoothing local gradient noise and guaranteeing global convergence under stringent physical constraints. **Result** Experiments were primarily conducted on the FLIR ADAS v1. 3 dataset, with further validation of generalization performance on the FLIR ADAS v2. 0 and LLVIP datasets. YOLOv3 was adopted as the target detector, and the attack effectiveness was verified in both digital-domain and physical-domain scenarios. In digital-domain experiments, the average Attack Success Rate (ASR) reached 75. 08%, and the average of Average Precision (AP) across different datasets of YOLOv3 dropped to 36. 18%, which is significantly superior to state-of-the-art methods. Moreover, the proposed method achieved the optimal attack success rate and mean average precision across different datasets, fully demonstrating its strong generalization capability. To verify cross-model robustness, further tests were carried out on advanced detection models with different architectures, including single-stage detection models (YOLOv5, YOLOv8), two-stage detection model (Faster R-CNN), and Transformer-based DETR. Except for the DETR detection model, the ASR against all other models exceeded 74%, among which the ASR against Faster R-CNN reached 78. 90%. These results fully confirm the universal interference capability and strong robustness of this method across different detection architectures. In the physical domain experiments, heat-insulating gel was used as the thermal insulation material. The position and shape of the adversarial mask were optimized through digital domain algorithms, and the thermal insulation material was cut into adversarial patches for testing in both indoor controlled environments and outdoor scenarios. In the indoor scenario, ASR reached 98. 5% at frontal view and remained above 84. 5% at ±30° angles under different conditions such as multi-angle, distance variation, and posture changes. Particularly, with a single patch, a high attack success rate was still maintained even at a left/right viewing angle of 30°. In the complex outdoor environment, the

ASR still reached 90.5%, indicating that the proposed method can maintain high attack efficiency under the diverse interferences of physical scenarios with only a single patch. **Conclusion** By integrating the dynamic objective prioritization strategy and the dual-projection optimizer, the proposed infrared physical adversarial patch resolves the fundamental conflict between aggressiveness and physical realizability. The dynamic strategy balances multiple objectives, the dual projection optimization ensures strict constraint satisfaction and global convergence, and the dual-regularization formulation minimizes fabrication effort. The method delivers high digital attack efficacy and cross-model robustness, while maintaining stable performance in complex physical scenarios, achieving an effective trade-off between attack strength and deployment cost. It provides a reliable red-team tool for security evaluation of infrared detection systems and offers crucial insights for countermeasure design, especially for safeguarding infrared perception modules in autonomous driving and security applications.

Key words: physical adversarial attack; object detection; thermal infrared imaging; dynamic objective optimization; dual projection optimization

0 引　言

随着深度学习技术的飞速发展，基于深度学习的红外目标检测系统已广泛应用于自动驾驶(Dai等，2021)、安防监控(Wu等，2024)和疫情防控(Yan等，2012)等关键领域。与可见光成像相比，红外成像具有全天候工作(Hou等，2022)、被动成像(Rippa等，2023)等独特优势，在复杂环境下表现出更强的鲁棒性。然而，近期研究表明，深度神经网络极易受到对抗攻击威胁，人为设计的微小扰动即可导致模型输出错误结果(Kurakin等，2018)。这种攻击行为被定义为对抗攻击，加入对抗扰动的图像被称为对抗样本。因此，研究高效的红外对抗攻击方法，不仅是评估红外目标检测系统鲁棒性的关键手段，也是推动防御机制设计的核心前提。

对抗攻击通常可分为数字对抗攻击(叶乙轩等，2024；刘复昌 等，2022)和物理对抗攻击(Shen等，2019；Guesmi 等，2024)。数字对抗攻击是指在数字图像上对图像像素进行修改，达到欺骗深度神经网络的目的；而物理对抗攻击则是指在真实世界的物体上部署物理扰动(如贴片或饰物等)以误导深度学习模型。为实现物理对抗攻击，通常需要先对物理对抗攻击目标实施数字对抗攻击。如图1所示，物理对抗攻击首先对目标进行数字对抗攻击，优化对抗扰动后即可得到对抗贴片的形状位置信息。然后，在物理域目标上部署对抗贴片并向检测模型实施攻击。相对于数字对抗攻击，物理对抗攻击因具备现实世界威胁性(彭振邦 等，2025)，已成为对抗攻击领域的研究热点。

当前，物理对抗攻击研究主要聚焦于可见光域(Shen 等，2019；Hua 等，2021)，而红外域由于其成像机制的特殊性，面临纹理、像素等高级语义特征稀缺的困境，致使对抗样本的攻击效果与物理可实现性难以兼顾。红外物理对抗攻击可划分为黑盒攻击与白盒攻击两类。黑盒攻击是在未知目标网络详细信息的情况下，仅通过输入数据与输出反馈的映射关系进行攻击(Hu 等，2024；Jia 等，2025)；白盒攻击是利用模型架构、参数及训练数据进行攻击，能更精准地针对神经网络特性设计攻击策略。因此，白盒攻击被认为是当前效能最优的对抗攻击手段(Wang等，2022)。

在白盒攻击方法中，Zhu 等人(2021)首次提出灯泡阵列作为红外扰动源，通过调控灯泡发热在红外图像中形成高亮区域，使行人检测器失效。但该方法需持续手持电路板，隐蔽性与实施性均存在不足。随后，Zhu 等人(2022)进一步设计了以气凝胶材料为核心的“二维码”(Quick Response Code，QR)红外隐身服装，并通过对服装表面热

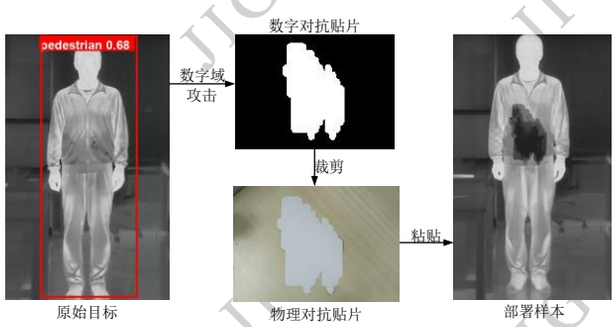


图1　物理对抗攻击过程示意图

Fig. 1　Schematic diagram of the physical adversarial attack process

隔离图案优化实现多角度攻击。但是该方法面临实用性瓶颈,其图案制作流程要求人工逐一裁剪气凝胶块,复杂性较高。Zhu等人(2023)提出了一种基于柔性碳纤维加热片的对抗服装,通过优化加热片的布局模式,进一步解决了阻热材料的阻热不彻底问题。Wei等人(2023)通过构建"攻击性-物理可实现性"多目标优化损失函数,获得连续聚合的数字对抗扰动。在物理部署时仅需裁剪气凝胶热绝缘材料即可获得对抗贴片,有效简化了物理对抗攻击的实现工艺。虽然前期的研究工作已经在提升对抗攻击的成功率同时简化了物理对抗攻击制作工艺,但现有方法仍存在两个不足:一是现有方法容易产生优化失衡问题,导致物理对抗攻击成功率较低。在对抗扰动优化过程中,多目标损失存在攻击效果和物理可实现性多目标优化矛盾,进而导致某一单一目标损失项主导整个优化进程。二是优化过程中未充分考虑红外对抗攻击中物理约束性问题。采用传统无约束梯度下降类算法,可能导致生成的对抗贴片在实际物理环境中失效。

为解决以上问题,本文提出了一种动态目标优先级驱动的红外对抗贴片优化方法。具体而言,在引入聚合正则化与稀疏二值正则化构建多目标优化损失基础上,提出了一种更适合物理对抗样本优化任务的动态目标优先级策略。该策略从人类任务管理和认知负荷规律出发(Bellotti 等,2004;Kember等,2004),根据不同损失优化难度,通过动态调整损失权重实现攻击效果与物理可实现性多目标优化平衡。同时,本文提出了双重投影优化算法,可以有效解决对抗样本优化中的约束性和非凸困境问题。

综上所述,本文的主要贡献总结如下:

1)提出了一种动态目标优先级驱动的对抗贴片多目标优化方法。该方法创新地融合了动态目标优先级策略与双正则化约束机制,提升了攻击效能。

2)提出了动态目标优先级策略。通过关键绩效指标与焦点损失自适应调整多目标损失权重,解决了对抗贴片优化过程中的多目标优化失衡问题。

3)构建了双重投影优化算法。融合了适合红外物理对抗攻击的PGD投影机制与加权动量累积机制,有效解决了物理对抗攻击中物理约束问题和非凸困境,提升了全局优化能力与收敛稳定性。

4)通过大量实验验证了本文方法的有效性。在数字域实验中,在多个数据集上本文方法的平均攻击成功率达75.08%,显著优于主流方法,验证了其有效性和泛化能力。在物理域实验中,生成的对抗贴片在多变视角、不同距离、姿态变化、复杂室外环境以及动态场景下,攻击成功率均超过79%,优于基准方法。

1 相关工作

1.1 可见光域行人目标物理对抗攻击方法

可见光域的物理对抗攻击,其本质是利用纹理,颜色等高级语义特征,在物理实体上嵌入能够欺骗深度神经网络的扰动。实施这种攻击的关键在于优化双重目标:既要显著提升攻击的有效性,又要有视觉自然性。

Thys等人(2019)提出了一种针对行人目标的对抗贴片方法,并将生成的对抗贴片打印在硬纸板上,成功攻击了YOLOv2检测模型。Huang等人(2020)提出的通用物理伪装(universal physical camouflage,UPC)通过联合欺骗区域提议网络(region proposal network,RPN),加强了对抗贴片攻击性能优化,可以有效攻击属于同一对象类别的所有实例。

生成具有视觉自然性的对抗贴片是可见光域对抗攻击的研究重点。Hu等人(2021)为了平衡攻击效果和视觉自然性,利用对抗生成网络(generative adversarial network,GAN)在真实世界拍摄的图像进行预训练,生成更自然的对抗贴片,并将其打印在衣物上欺骗检测器。此方法为对抗贴片的视觉自然性和攻击性平衡提供了新思路。Guesmi等人(2023)提出一种生成对抗艺术图案的方法,无需计算密集GAN即可生成视觉自然的对抗贴片。之后,该团队又提出针对动态场景的对抗贴片方法(dynamic adversarial patch,DAP)(Guesmi 等,2024),该方法直接修改贴片中的像素值,显著降低计算量,同时通过期望增强框架(expectation over transformation,EOT),模拟物理迁移时可能出现的光照,扭曲等变化,增强对抗样本的鲁棒性。

相比于以往方法生成不规则色块,部分研究则从生成生活常见物体图像形式优化对抗贴片。Tan等人(2021)从衣物装饰性图片角度,提出生成与卡通图案视觉相似的贴片方法(legitimate adversarial patches,LAP),并从颜色特征、边缘特征和纹理特征三个角度评估其合理性,在数字域与物理域实验中

均表现出显著攻击效果。Hu等人(2022)则提出了一种可以任意扩展的对抗纹理，即任意裁剪局部区域均可欺骗检测器。解决了因拍摄不完整导致攻击失效的问题。

Lapid等人(2023)提出了一种黑盒无梯度方法，该方法使用预先训练的GAN的学习图像流形生成自然的物理对抗贴片。这是首个直接对目标检测模型实施黑盒物理攻击的方法，具备黑盒无关性。

1.2 红外域行人目标物理对抗攻击方法

目前，大部分研究集中于可见光域，红外域物理对抗攻击研究仍处于起步阶段。与可见光图像相比，红外图像中缺少颜色、纹理等高级语义特征，这不仅导致可见光域方法难以直接迁移，还因红外图像特性，显著增加了攻击效果设计与物理可实现性的难度。从技术路径划分，现有红外物理对抗攻击可归纳为白盒攻击与黑盒攻击两大类。

图2展示了现有方法部署效果图。其中，黑盒攻击主要利用启发式算法优化对抗样本。Wei等人(2023)提出一种用冷热贴片作为物理干扰的对抗攻击方法(hotcold block，HCB)。通过粒子群优化算法优化贴片的形状和位置，并将贴片附着于在衣物内部，增强了隐蔽性。Hu等人(2024)进一步开发出可实现多角度攻击的方法(adversarial infrared blocks，AdvIB)，通过建模冷热离散贴片，使用差分进化算法优化贴片角度与位置，弥补了以往方法无法实现多角度攻击的不足。Tiliwalidi等人(2025)对多角度对抗攻击也进行了探索，提出一种针对红外行人检测器的多视角攻击的对抗服饰方法(adversarial infrared grid，AdvGrid)。首先对红外网格形状进行建模，使用遗传算法获得最优网格位置。在物理域实验中实现了多视角的物理对抗攻击。Jia等人(2025)提出一种基于Catmull-Rom样条曲线的攻击方法(adversarial infrared Catmull-Rom spline，AdvICRS)。通过用启发式算法优化控制点生成具有攻击性的闭合曲线形状。

这些方法虽然对物理对抗攻击的实用性进行了探索，但仍不能完全脱离模型信息，需要对模型输入数据及输出进行大量查询反馈获取优化方向，导致计算开销居高不下；同时由图2中黑盒方法示例可见，目前这类方法受限于阻热材料的物理属性如形态可塑性等，贴片形状的建模空间被显著压缩，难以实现复杂拓扑结构的对抗扰动。

而相比于黑盒方法，白盒方法则基于神经网络内部决策逻辑，对对抗样本形状位置进行精准优化，可高效地生成最佳的样本，为快速部署贴片提供了技术可能，也为神经网络鲁棒性提升提供精准依据，具体示例图2所示。

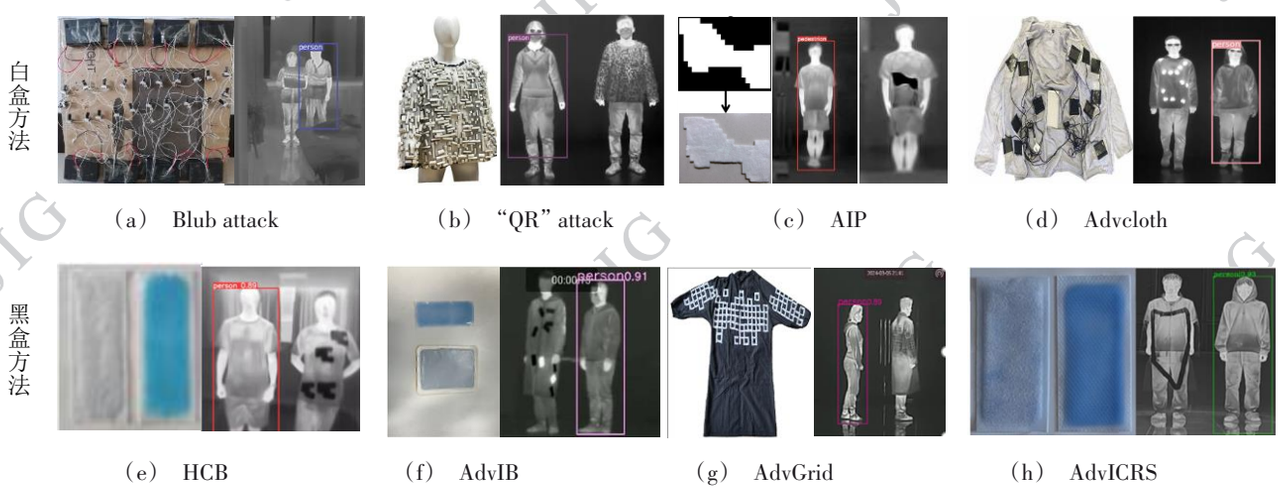


图2 红外物理对抗攻击不同方法示例对比

Fig. 2 Examples of different methods

Zhu等人(2021)首次对红外域物理攻击展开研究，提出了一种附着小灯泡硬纸板作为对抗扰动的白盒攻击方法。其原理是用高斯函数模拟小灯泡的红外成像像素特点，通过梯度优化获得最佳灯泡位置，而后部署欺骗红外行人检测器。Zhu等人(2022)提出一种对抗服饰方法。该方法利用阻热凝

胶阻断人体热源的方式对红外行人检测器实施物理攻击，首先在数字域中生成可周期性扩展的对抗性“二维码（quick response code，QR）”图案，其每个局部都具有对抗性特征，所以可以任意裁剪仍具有对抗效果。然后裁剪阻热凝胶将对抗图案制作为红外对抗服饰（“QR” attack）。在物理实验中实现了不同角度实攻击。Zhu等人（2023）提出一种对抗服饰方法（adversarial clothes，AdvCloth），采用柔性碳纤维加热器作为干扰源附着于衣服内部，制作对抗服饰，并提出一种新损失函数减少加热器的重合部分，使之更符合实际部署情况。以上方法虽然做出了积极探索，但由于对抗服饰制作工艺比较复杂，实现成本较高，无法实现快速部署。Wei等人（2023a）提出一种新型白盒攻击方法（adversarial infrared patches，AIP）。该方法通过梯度信息同时优化对抗贴片的位置和形状，开发形状位置等低级语义特征的攻击潜力。物理部署时仅需通过裁剪热绝缘材料制作成不规则的贴片来实施物理对抗攻击，极大降低制作成本。并通过大量数字域和物理域实验验证了方法的有效性。

2 本文方法

为了生成能够高效欺骗红外检测模型的对抗样本，本文提出了一种动态目标优先级驱动的红外对抗贴片优化方法，整体方法框架如图3所示。该方法基于白盒优化范式，通过梯度精准优化对抗贴片的形状与位置。本方法攻击过程分为数字域优化与物理域迁移两阶段。图3上侧为数字域流程框架，数字域流程以“攻击性效果-物理可实现性均衡优化”为核心目标。首先，构建融合双正则化的多目标优化损失，引导随机初始化的掩码生成紧凑、连续的几何形状，为后续物理域裁剪部署奠定基础。为了规避传统固定权重导致的单目标主导优化问题，流程中嵌入动态目标优先级策略，依据关键绩效指标自适应调整损失权重，实现多目标资源的动态均衡分配。最后采用双重投影优化算法作为核心优化器，高效优化对抗掩码，并与干净样本结合得到数字域对抗样本。图3下侧展示了将数字域对抗样本向物理域的迁移。根据数字域生成的对抗掩码形状裁剪热绝缘材料，将贴片部署于行人衣物表面后，即可对检测模型形成干扰。

2.1 多目标优化损失

2.1.1 攻击性损失定义

给定干净的红外图像$\boldsymbol{x}$，$f(\cdot)$表示参数为θ的红外目标检测模型。相应地，$f(\boldsymbol{x};\theta)\rightarrow y$为模型的预测输出，$y$包含目标物体的输出边界框$\{b_i|i=1,\ldots,n\}$和置信度分数$\{s_i|i=1,\ldots,n\}$。模型将置信度分数最高的边界框作为目标边界框。本文旨在通过构建对抗样本$\boldsymbol{x}_{\text{adv}}$，使红外目标检测器不能检测到目标，所以通过最小化所有输出边界框的最高置信度得分，直到它低于检测阈值，使检测器失效。于是攻击性损失可以被定义为：

$$L_{\text{k}}\left(f\left(\boldsymbol{x}_{\text{adv}};\theta\right)\right)=\max_{i\in\{1,\ldots,n\}}\left(s_i\right) \tag{1}$$

$$\boldsymbol{x}_{\text{adv}}=(1-\boldsymbol{m})\odot\boldsymbol{x}+\boldsymbol{m}\odot\delta \tag{2}$$

式中，$\boldsymbol{x}_{\text{adv}}$是红外对抗样本，$\odot$表示逐元素相乘，$x\in\mathbf{R}^{h\times w}$表示干净目标（其检测模型预测边界框长宽为$h\times w$），$\boldsymbol{m}\in\{0,1\}^{h\times w}$表示约束对抗贴片的形状和位置的二进制掩码，$\delta$为红外对抗贴片在数字域对应像素值。红外对抗贴片可定义为$\boldsymbol{m}_{ij}=1$所覆盖区域，所以掩码$\boldsymbol{m}$决定贴片的位置和形状，δ决定贴片数字域对应像素值。在现实世界中，一旦用于制造对抗贴片的材料选定，贴片对应像素值δ就可以确定，所以现在只需要优化掩码$\boldsymbol{m}$确定对抗贴片

的位置和形状，以最小化输出边界的最高置信度得分作为优化目标。

2.1.2 聚合正则化与二值正则化

为精准定位贴片的形状与位置，优化后的掩码$\boldsymbol{m}$需满足两大核心要求：一是值为1的像素应高度聚合形成连续几何体，避免呈现离散碎片化分布；二是掩码像素值需严格趋近于0或1的二值化分布状态。然而直接求解公式（1）会导致掩码$\boldsymbol{m}$中的值在0到1之间连续变化。为此，本文引入聚合正则化与稀疏二值正则化双重约束。聚合正则化通过空间连续性惩罚项强制高值像素形成紧凑区域；稀疏二值正则化采用阈值惩罚促使像素值向极值0或1收敛。最终，两者共同构建兼具空间连续性与数值离散性的优化掩码。

为引导掩码$\boldsymbol{m}$中值为1像素形成连续的几何形状，本文引入了聚合正则化，强制对抗样本形成紧凑的簇状结构。其核心实现方式是根据图论中的局部聚类系数（local clustering coefficient，LCC），通过计算所有元素聚合度的负平均值作为正则化迫使高值

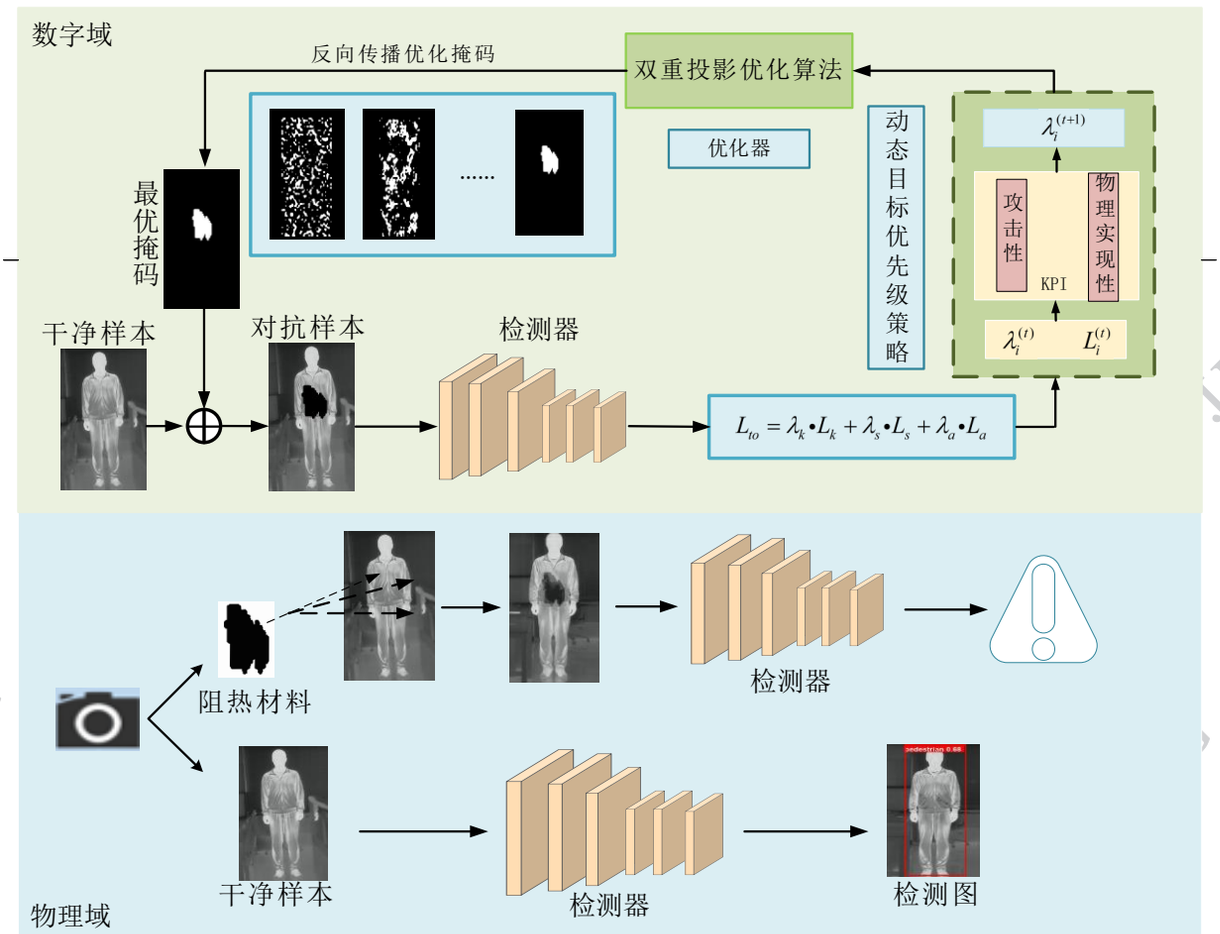


图3　本文方法框架

Fig. 3　Framework of our method

像素形成紧凑区域。

为了简化计算，使用卷积操作来近似计算局部聚类系数。定义一个核矩阵$\boldsymbol{K}$，其中每个元素对应掩码$\boldsymbol{m}$中每个位置$\boldsymbol{m}_{ij}$的邻居数量。则$\boldsymbol{m}_{ij}$的聚合度C_{ij}表示为：

$$C_{ij}=\frac{(\boldsymbol{K}*\boldsymbol{A})_{ij}}{n(n-1)} \tag{3}$$

式中，$\boldsymbol{A}$是衰减因子矩阵，用于考虑顶点值及其邻居值的影响。n为固定值8，表示每个位置共有8个邻居，$*$表示卷积操作。衰减因子矩阵$\boldsymbol{A}$的每个元素$\boldsymbol{A}_{ij}$定义为：

$$\boldsymbol{A}_{ij}=V_i\times\frac{1}{|L_i|}\sum_{v_j\in L_i}V_j \tag{4}$$

式中，V_i和V_j分别是顶点v_i和v_j的值。L_i表示与顶点v_i相连的顶点的集合，$|L_i|$是顶点v_i的相连顶点的数量，$v_j\in L_i$表示集合L_i中每个顶点v_j。

最终，聚合正则化损失$L_{\mathrm{a}}(m)$定义为所有元素聚合度的负平均值：

$$L_{\mathrm{a}}(\boldsymbol{m})=-\frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w}C_{ij}\cdot\boldsymbol{m}_{ij} \tag{5}$$

通过最小化$L_{\mathrm{a}}(\boldsymbol{m})$，可以使掩码$\boldsymbol{m}$中的高值像素聚集形成有效的贴片形状。

由于掩码$\boldsymbol{m}$应该是二值的（即元素值为0或1），本文引入了稀疏二值正则化。使用均方误差（mean square error，MSE）来构造正则化项，首先定义了一个类似ReLU函数的阈值函数H：

$$\mathrm{H}(\boldsymbol{m}_{ij})=\begin{cases}1 & \boldsymbol{m}_{ij}<V_{\mathrm{thre}}\\ \boldsymbol{m}_{ij} & \boldsymbol{m}_{ij}\geqslant V_{\mathrm{thre}}\end{cases} \tag{6}$$

式中，V_{thre}是一个阈值为0.1。为了确保H的可微性，在阈值点处设置了导数为零。均方误差项表示为：

$$L_{\mathrm{MSE}}(\mathrm{H}(\boldsymbol{m}),\boldsymbol{I})=\frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w}\left(\mathrm{H}(\boldsymbol{m}_{ij})-1\right)^2 \tag{7}$$

式中，$\boldsymbol{I}$是每个元素均为1的矩阵。通过均方误差项

表示 $\boldsymbol{m}$ 中大于阈值 V_{thre} 的像素值与所需像素1之间的差值。

最终的稀疏二值正则化损失为：

$$L_{\text{s}}(\boldsymbol{m}) = \|\boldsymbol{m}\|_1 + \gamma \cdot L_{\text{MSE}}(\text{H}(\boldsymbol{m}), \boldsymbol{I}) \tag{8}$$

式中，$\|\boldsymbol{m}\|_1$ 表示掩码 $\boldsymbol{m}$ 的 L_1 范数，即掩码中所有元素的绝对值之和。在优化过程中，L_1 范数有助于促进掩码的稀疏性，使得更多的元素值为0。γ 是一个权重因子，用于平衡 L_1 范数项和均方误差项的影响。

现在，将上述攻击性损失，聚合正则化，稀疏二值正则化合并为总损失，最终的目标函数由以下公式表示：

$$L_{\text{to}} = \lambda_{\text{k}} \cdot L_{\text{k}}\left(f\left(\boldsymbol{x}_{\text{adv}}; \theta\right)\right) + \lambda_{\text{s}} \cdot L_{\text{s}}(\boldsymbol{m}) + \lambda_{\text{a}} \cdot L_{\text{a}}(\boldsymbol{m}) \tag{9}$$

其中 λ_{k}，λ_{s}，λ_{a} 是为对应损失项的可动态调节权重参数。

2.2 动态目标优先级策略

公式(9)将攻击性损失、聚合正则化与稀疏二值正则化整合为总损失，意在通过多目标协同优化，实现对抗样本的高效攻击性与物理可实现性。但在实际优化过程中，传统方法通常用经验值固定权重的方式，这种静态分配机制易导致优化过程中出现显著的目标主导现象，具体表现为某一目标的优化进程过度消耗计算资源，而其他目标的性能指标出现系统性退化，这种资源错配问题严重影响了对抗样本生成的攻击效果与物理可实现性之间的平衡。

为突破这一限制，本文受多任务学习(multi-task learning，ML)中动态任务优先级策略(dynamic task prioritization，DTP)(Guo 等，2018)启发，提出更适配物理对抗攻击的动态目标优先级策略(dynamic objective prioritization，DOP)。与 DTP 相比，DOP 策略针对对抗攻击任务中物理约束缺乏显式评价指标的特性，设计了基于损失占比的隐式优先级度量，通过损失项权重的自适应调整实现资源合理分配。

2.2.1 目标优先级机制设计

不同于多任务学习中传统策略如课程学习采用优先分配资源给简单任务，再逐步引入复杂任务。这类策略在对抗样本优化场景中存在显著缺陷，若聚合正则化快速收敛，会导致其权重继续增大，最终导致对抗样本攻击性损失优化不足。这并不符合均衡优化目标。所以，本文提出的 DOP 策略采用与 DTP 相同的难优先资源分配原则。具体而言，本文引入关键绩效指标(key performance indicator，KPI)(Guo 等，2018)量化任务难度，通过实时评估各目标的优化难度，动态增加困难目标的损失权重，同时降低简单目标的资源占比。确保在整个训练过程中各目标的均衡推进。

然而，DTP 策略依赖于明确的任务性能指标 KPI 如目标检测任务中的平均精度 AP 来量化难度。反观物理对抗攻击场景，其物理可实现性约束(例如掩码的空间连续性、二值化特性)通常缺乏此类可直接用于评估优化进度的显式指标，这导致无法直接移植基于性能指标的 KPI 定义。因此，需要设计一个适用于多目标损失优化的全新优先级度量标准。

为此，本文提出了一种基于归一化损失占比和焦点损失的目标优先级机制。该机制通过计算各损失项在总损失中的归一化占比作为关键绩效指标 KPI，并根据其优化困难程度动态调整损失权重，当某一损失项占比过低时，表明该目标优化难度大，需提升其权重以加速收敛；反之则降低权重，确保优化过程中各项约束得到均衡处理。

2.2.2 关键性能指标 KPI

在对抗样本生成过程中，本文构建了一个包括多项损失函数的总损失函数如公式(9)所示，具体包括攻击损失函数、稀疏二值化损失函数、聚合损失函数。其中 λ_{k}，λ_{s}，λ_{a} 是对应损失项的可调节权重参数。

每次迭代记录当前各损失项的数值大小，并进行归一化处理，计算其占总损失的比例作为 KPI，第 i 个损失项在第 t 次迭代时的初始 KPI 表示为：

$$\text{KPI}_i^{(t)} = \frac{L_i^{(t)}}{\sum_i L_i^{(t)}} \tag{10}$$

式中，$L_i^{(t)}$ 为第 i 项损失在第 t 次迭代时的数值。

为了抑制因单次迭代中瞬时波动导致权重频繁变化的问题，本文引入滑动平均机制对 KPI 进行平滑处理。更新后的 KPI 结合了前一次的历史值与当前计算值，其更新公式如下：

$$\text{KPI}_i^{(t)} = \beta \cdot \text{KPI}_i^{(t-1)} + (1-\beta) \cdot \text{KPI}_i^{(t)} \tag{11}$$

式中，β 为平滑系数(取 $\beta = 0.9$)，用于平衡历史信息与当前状态的影响，从而提高权重调整的稳定性与鲁棒性。

2.2.3　基于KPI损失的参数自适应机制

为增强对困难目标的关注度，本文引入焦点损失(focal loss，FL)(Lin 等，2017)的思想，作为一种通过动态调整损失权重，抑制易优化目标贡献，增加模型资源向困难目标倾斜的方法。动态权重$\lambda_i^{(t)}$根据平滑后的KPI自适应调整：

$$\lambda_i^{(t)} = \lambda_i^{(t-1)} \cdot \left(1 \pm \mu \cdot \mathrm{FL}\left(\mathrm{KPI}_i^{(t)}\right)\right) \tag{12}$$

$$\mathrm{FL}(p) = -(1-p)^{\alpha}\log(p) \tag{13}$$

式中，$\mu = 0.008$为固定缩放因子，$\alpha = 1.5$为聚焦参数，p为经公式(10)(11)所得关键绩效指标值KPI，代表某个损失项在当前迭代中占总损失的比例。同时考虑到各损失存在量化差异，以及焦点损失对于占比微小波动十分敏感，本文设计了安全区间机制，当超出区间时，触发调整机制：

$$\lambda_i^{(t)} = \mathrm{clip}\left(\lambda_{\min}, \lambda_{\max}\right) \tag{14}$$

式中，clip(·)为强制截断，将各损失项权重参数$\lambda_i^{(t)}$截断至可行域内。该机制在保留焦点损失对不同目标资源分配影响的同时，规避因量化差异导致的权重过度震荡，提升优化稳定性。最终通过动态目标优先级策略实现公式(9)中各目标损失权重动态自适应。

2.3　双重投影优化算法

在本文的对抗样本优化任务中，本质上是一类具有显式约束的优化任务，其核心约束体现为：所生成的对抗扰动必须严格限定在预先设定的可行域内。通过公式(9)求解掩码$\boldsymbol{m}$，传统求解框架普遍沿用随机梯度下降算法(stochastic gradient descent，SGD)及其变体如带动量的随机梯度下降(stochastic gradient descent with momentum，Momentum SGD)(Zhu 等，2023；Wei 等，2023a)，但这些算法的设计初衷是面向无约束优化场景，即默认解可在全空间自由移动。然而，无约束范式在约束场景中暴露出根本性缺陷：一是难以高效处理边界约束，迭代过程中极易越出可行域而违反约束；一旦发生越界，往往需通过截断等手段强行拉回，导致梯度方向携带的丰富信息未被充分利用而削弱优化效率，还可能破坏搜索方向与参数空间的连续性。二是掩码参数优化面临多重挑战，尤为突出的是对抗损失函数的强非凸特性，导致优化轨迹极易陷入局部最优，严重限制了最终解的全局质量。

为解决以上问题，本文针对本任务中对抗样本优化问题创新性地提出双重投影优化算法。该算法引入与对抗样本优化需求高度适配的投影机制，以此解决对抗掩码优化中的约束满足难题。投影机制的核心功能在于：当更新后的参数偏离可行域，将其映射回可行域内距该点最近的合法位置，从而确保迭代过程中每一个中间解均严格满足约束条件。

具体而言，该方法融合了梯度投影下降算法PGD(Madry 等，2017)与加权动量累积策略，通过双阶段投影与加权动量融合实现对约束优化过程的精细化控制。其中，加权动量累积机制借助历史梯度信息动态校正迭代方向。相较于传统动量累积中固定比例的历史信息复用模式，双重投影优化算法显著提升了历史梯度的累积权重占比，使迭代方向更贴合长期优化趋势，从而有效平滑局部梯度的剧烈波动，避免因瞬时梯度突变导致的掩码结构畸变；投影机制于每一步迭代中对掩码像素施加针对约束空间的严格投影操作。与无约束优化中“先更新后截断”的策略相比，显著抑制局部像素突变导致的对抗特征失真，进而提升攻击效果的稳定性。上述机制协同作用，在保障迭代稳定性的同时，大幅提升了算法的收敛效率。

在动量加速的投影梯度下降框架中，首先执行临时投影操作：将当前掩码$\boldsymbol{m}_t$沿着损失函数的梯度方向$\nabla_{\boldsymbol{m}} L_{\mathrm{to}}$移动步长$\eta$，经投影操作得到临时变量$\boldsymbol{Z}_t$：

$$\boldsymbol{Z}_t = \prod_{[0,1]}\left(\boldsymbol{m}_t + \eta \nabla_{\boldsymbol{m}} L_{\mathrm{to}}\right) \tag{15}$$

式中，$\eta = 0.1$为步长，$\prod_{[0,1]}(\cdot)$为投影操作，将结果投影至约束空间$[0,1]$内。为缓解非凸优化中的梯度震荡，双重投影优化算法引入加权动量累积机制，提高更新稳定性。在完成首次投影操作后，算法进入动量融合阶段：将历史累计梯度更新量$\boldsymbol{D}_{t-1}$与当前梯度更新量$(\boldsymbol{Z}_t - \boldsymbol{m}_t)$按照动量系数$\sigma$进行加权融合，生成新的梯度更新量$\boldsymbol{D}_t$：

$$\boldsymbol{D}_t = \sigma \boldsymbol{D}_{t-1} + (1-\sigma)\left(\boldsymbol{Z}_t - \boldsymbol{m}_t\right) \tag{16}$$

式中，σ为权重系数取0.75，调整历史梯度与当前梯度的贡献权重。

在获得融合后的梯度更新量$\boldsymbol{D}_t$后，需要将其叠加到当前掩码$\boldsymbol{m}_t$上，随后将叠加结果通过投影操作再次映射回可行域$[0,1]$内，从而生成更新后的掩码$\boldsymbol{m}_{t+1}$：

$$\boldsymbol{m}_{t+1} = \prod_{[0,1]}\left(\boldsymbol{m}_t + \boldsymbol{D}_t\right) \tag{17}$$

本文提出的整个攻击流程可归纳为如下算法：

基于动态目标优先级驱动的红外对抗贴片优化方法：

输入：干净图像$\boldsymbol{x}$，热绝缘材料数字域对应像素δ，最大迭代次数T，对抗掩码$\boldsymbol{m}_t$，损失函数中第i个损失项初始参数为$\lambda_i^{(t)}$梯度更新量$\boldsymbol{D}_t$。

输出：优化后对抗掩码$\boldsymbol{m}_T$。

1）初始化：随机生成初始掩码$\boldsymbol{m}_0$，损失函数

中第i个损失项在第1次迭代参数$\lambda_i^{(0)}$，梯度 历史更新量$\boldsymbol{D}_0=0$；

2）构造多目标优化损失函数L_{to}；

3）双重投影优化算法求解掩码$\boldsymbol{m}_{t+1}$：

计算临时投影点$\boldsymbol{Z}_t=\Pi_{[0,1]}\left(\boldsymbol{m}_t+\eta\nabla_m L_{\text{to}}\right)$；

根据公式（16）计算梯度更新量$\boldsymbol{D}_t$；

更新后掩码$\boldsymbol{m}_{t+1}=\prod_{[0,1]}\left(\boldsymbol{m}_t+\boldsymbol{D}_t\right)$；

4）动态目标优先级策略更新各损失权$\lambda_i^{(t)}$：

根据公式（10）（11）计算$\text{KPI}_i^{(t)}$为

$\text{KPI}_i^{(t)}=\beta\cdot\text{KPI}_i^{(t-1)}+\left(1-\beta\right)\cdot\text{KPI}_i^{(t)}$；

实现动态更新权重$\lambda_i^{(t)}$为

$\lambda_i^{(t)}=\lambda_i^{(t-1)}\cdot\left(1\pm\gamma\cdot\text{FL}\left(\text{KPI}_i^{(t)}\right)\right)$；

5）迭代执行步骤1）—4）T次，输出最终对抗掩码$\boldsymbol{m}_T$。

3 实　验

3.1 数据集介绍

本文使用了FLIR提供的FLIR ADAS v1_3为主要数据集（如无特殊说明，数字域实验均以此数据集为测试目标）。该数据集共收录10228张由人工精细标注的长红波外影像，涵盖行人、自行车、汽车和狗四类目标。图像采集自短视频与连续视频片段，使用FLIR Tau2热像仪（13mmf/1.0，水平视场45°，垂直视场37°，640×512像素，NETD<60mK）。针对行人检测任务，本文筛选出含行人标签的9900张有效样本，并划分训练集和测试集，其中训练集包含7873张图像，测试集包含2027张图像，并把测试集中能够被检测到目标的图像直接作为对抗攻击的目标样本。

为验证本文方法的泛化能力，进一步在FLIR ADAS v2.0和LLVIP两个数据集上开展测试。其中，FLIR ADAS v2.0数据集经筛选后，得到含行人标签的有效样本1255张，划分为训练集878张、测试集377张；LLVIP数据集经筛选后，获得含行人标签的有效样本5004张，相应划分为训练集3503张、测试集1501张。

3.2 实验设置

目标检测器：选用单阶段检测模型YOLOv3（Redmon等，2018）。实验中，按照YOLOv3的实验要求，其输入图像分辨率统一缩放至416×416。使用YOLO官方提供的预训练权重，并在本文筛选的数据集上进行了重新训练。所提检测模型在训练集与测试集上的平均精度表现如下：在FLIR ADAS v1_3、FLIR ADAS v2.0和LLVIP训练集上，平均精度分别达到95.28%、92.34%和97.21%；在对应测试集上，平均精度则为92.13%、90.44%和96.56%。

对比方法：为了评估本文方法的攻击效果，本文在数字域中与已开源的主流方法进行了对比实验，其中包括白盒攻击方法AIP（Wei等，2023a）以及黑盒攻击方法HCB（Wei等，2023），AdvIB（Hu等，2024），AdvICRS（Jia等，2025）。对比实验均在本文所述数据集和实验设备上进行了统一实验。物理域实验中，由于与黑盒方法使用热绝缘材料存在不同，对比并不公平。所以本文模拟使用同种热绝缘材料的白盒方法AIP（Wei等，2023a）实验条件，并引用其论文数据进行对比。本文方法属于白盒攻击范畴，具有对比意义和评估价值。

实验设置：本文实验均在单张NVIDIA GeForce RTX 3090上完成。对于每张输入图像，算法连续执行5轮独立优化，每轮迭代100步。

如图4（a）所示展示了物理域实验所用设备，包括一个双光谱筒型摄像机，一个三脚架以及热绝缘材料。本文使用阻热凝胶作为热绝缘材料（物理域实验中，默认材料的热绝缘性能处于稳定性状态），图4（b）展示了用阻热凝胶制作贴片部署于左侧行人目标，其在红外相机下为黑色色块，且行人目标无法被检测模型识别。数字域实验中把热绝缘材料像素值δ设置为0，对应热绝缘材料在摄像机中的最低辐射响应。

评估指标：为了量化对抗贴片对行人检测的干扰程度，本文采用两项互补指标：平均精度（average precision，AP）与攻击成功率（attack success rate，ASR）。AP通过计算精度-召回曲线下的面积反映检测器在对抗样本上的整体性能，数值越低表明干扰

(a)实验设备 (b)热绝缘材料阻热效果图

((a) experimental equipment; (b) influence of aerogel material)

图4 物理域实验设备及热绝缘材料阻热效果

Fig. 4 Experimental devices and heat insulation effect of thermal insulation materials

越强。ASR则被定义为在干净样本中可被正确检测的行人实例总数N中，经过扰动后置信度低于阈值而“消失”的实例所占比例。ASR计算公式如下：

$$\mathrm{ASR} = 1 - \frac{1}{N}\sum_{n=1}^{N} F\left(y_{obj}^{n}\right) \tag{18}$$

$$F\left(y_{obj}^{n}\right) = \begin{cases} 0 & y_{obj}^{n} < 0.5 \\ 1 & 其它 \end{cases} \tag{19}$$

式中y_{obj}^{n}表示可被正确检测的行人实例中第n个目标的检测置信度。数字域实验统一将检测阈值设为0.5。

3.3 实验结果与分析

3.3.1 与主流方法的对比

为验证本文方法的攻击效果和泛化能力，本小节从ASR和AP两个指标，将本文方法与主流物理对抗攻击方法进行比较。在多个红外数据集上进行测试包括FLIR ADAS v1_3、FLIR ADAS v2.0、LLVIP，实验结果如表1所示。

实验结果表明，本文所提方法在多个红外数据集上均展现出更卓越的攻击性能，并验证了其优异的泛化能力。平均攻击成功率达到75.08%，攻击后检测模型检测精度平均值为36.18%，均显著优于主流方法。这一结果充分说明了本文方法的跨数据集鲁棒性。

从具体数据集实验结果来看：与主流方法AIP（Wei 等，2023a）、HCB（Wei 等，2023）、AdvIB（Hu 等，2024）以及AdvICRS（Jia 等，2025）相比，本文方法在所有测试数据集上ASR和AP两个指标均为最优。表明本文方法对目标检测器的干扰最为有效。即使在成像质量较低、背景复杂度较高的LLVIP数据集上，本文方法仍能维持高达73.58%的攻击成功率，进一步证明了其在复杂真实场景中的鲁棒性。

为了更直观展示本文所提方法的实际效果，图5展示了干净样本和对抗样本的检测对比图。可以看出，本文所提方法在数字域中生成的对抗

扰动为紧凑的几何形状，意味着良好的物理可实现性，扰动加入干净目标后目标无法被检测。

3.3.2 优化器性能消融实验

为验证本文所提双重投影优化算法对于本文有约束优化问题的实际作用，本小节将本文提出的双重投影优化方法和传统无约束优化方法SGD（Robbins 等，1951）、Momentum SGD（Polyak 等，1964）、内彻斯特罗夫加速梯度下降（Nesterov accelerated gradient，NAG）（Nesterov 等，1983）、自适应矩估计（adaptive moment estimation，Adam）（Kingma 等，2014）以及有约束优化方法PGD（Madry 等，2017）在固定损失参数基础

表1 与主流方法的结果对比

Table1 Comparisons with SOTA infrared attacks in digital tests.

方法	技术路径	部署方式	FLIR ADAS v1_3		FLIR ADAS v2.0		LLVIP		Average	
			AP	ASR	AP	ASR	AP	ASR	AP	ASR
HCB（Wei 等，2023）	黑盒	对抗贴片	46.40	63.53	49.66	60.61	51.35	59.44	49.14	61.19
AdvIB（Hu 等，2024）	黑盒	对抗贴片	43.28	68.69	46.87	64.37	42.93	66.64	43.11	66.57
AdvICRS（Jia 等，2025）	黑盒	对抗贴片	37.16	71.43	<u>38.44</u>	<u>70.56</u>	<u>41.66</u>	<u>68.71</u>	<u>39.09</u>	<u>70.23</u>
AIP（Wei 等，2023a）	白盒	对抗贴片	<u>36.94</u>	<u>73.56</u>	41.67	68.27	42.25	67.43	39.31	69.75
本文方法	白盒	对抗贴片	**33.43**	**77.51**	**36.86**	**74.14**	**38.25**	**73.58**	**36.18**	**75.08**

注：ASR、AP单位均为%，加粗数值为各列最优结果，加下划线数值为次优结果

图5　本文方法数字域对抗样本示例

Fig. 5　Examples of adversarial samples in the digital tests

上进行的消融实验对比。实验结果如表2所示。

从表2实验数据可以看出，本文提出的双重投影优化算法ASR为76.60%，均优于其他优化器。对抗样本平均优化时间为14.46s。表明双重投影优化算法对于对抗样本生成强约束优化任务中可以有效规避无约束方法在高维空间中的低效

探索，以及单一投影PGD方法对有效梯度分量的过度压制，通过动量项保持梯度更新的方向一致性，加速跨越损失函数的局部洼地，但是计算开销也有所增大。

表2　与传统优化器方法的消融结果对比

Table2　Comparison results with traditional optimizer methods

方法	ASR (%)	时间(s)
SGD（Robbins 等，1951）	73.56	10.74
Momentum SGD（Polyak 等，1964）	73.56	16.42
NAG（Nesterov 等，1983）	72.64	15.74
Adam（Kingma 等，2014）	75.38	**10.33**
PGD（Madry 等，2017）	74.77	10.80
双重投影优化算法	**76.60**	14.46

注：加粗数值为各列最优结果。

3.3.3　模块消融实验

为了验证本文提出的DOP策略与双重投影优化算法的各模块有效性，本小节对这两个算法进行了消融实验。采用带有传统动量机制的PGD优化方法替代本文提出双重投影优化算法以及固定权重方式代替DOP算法，评价指标为ASR和对抗样本平均优化时间。实验结果如表3所示。

表3实验数据中，与带有传统动量机制的PGD和固定权重的基线组相比，仅启用DOP可将ASR从

表3　消融实验结果

Table3　Ablation experimental results

带动量的PGD	固定权重	DOP	双重投影优化算法	ASR (%)	时间(s)
√	√			75.08	10.80
√		√		76.29	9.13
	√		√	76.60	14.46
		√	√	77.51	10.38

注："√"指使用对应方法策略。

75.08%提升至76.29%，同时显著缩短优化时间至9.13秒，表明DOP能通过自适应调整损失权重来智能分配资源、加速收敛。

而仅启用双重投影优化算法时，尽管计算开销增大导致优化时间增至14.46秒，但ASR提升至76.60%，表明双重投影算法能有效探索解空间。当DOP与双重投影优化算法协同工作时，取得了最高的ASR（77.51%）和优化时间（10.38秒），说明DOP可有效弥补双重投影优化算法的计算效率短板，形成协同增效效应，最终达成精度与速度的同步优化目标。

3.3.4　攻击鲁棒性实验

为评估本文提出的动态目标优先级驱动的红外对抗贴片优化方法在各种主流检测器上的攻击鲁棒性，本小节将本文方法对不同架构检测器包括单阶段检测模型YOLOv3（Redmon 等，2018）、YOLOv5、YOLOv8（Varghese 等，2024）、两阶段检测模型Faster R-CNN（Ren 等，2015）以及DETR（Carion 等，2020）进行攻击，评估指标为AP和ASR。实验结果如表4所示，验证了本文方法对于大部分主流检测器的攻击鲁棒性。

表4　本文提出物理对抗攻击方法在不同架构检测器上的实验结果

Table4　The results of the method in this paper on various detectors

检测器	w/o 攻击		w/ 攻击	
	AP (%)	ASR (%)	AP (%)	ASR (%)
YOLOv3	92.13	–	33.43	77.51
YOLOv5	91.80	–	32.83	76.16
YOLOv8	94.57	–	37.46	74.45
Faster R-CNN	90.47	–	28.96	78.90
DETR	92.33	–	73.94	38.18

注：w/o 攻击指无攻击，w/ 攻击指有攻击，“–”指无攻击场景下，无此指标。

从表4实验数据中可以看出，在未进行攻击时，所有检测器在测试集上的AP值均超过90%，表明这些模型本身具有很强的检测能力。在本文方法进行攻击后，绝大多数模型的AP值均出现断崖式下降，同时具有高攻击成功率，表明本文提出的攻击方法具有极强的泛化能力。具体来看，对于YOLO系列和Faster R-CNN等主流检测器，攻击后的ASR均超过了74%，其中Faster R-CNN甚至达到了78.90%。这说明本文方法生成的对抗扰动能够有效欺骗多种基于深度学习的检测模型。然而，DETR模型表现出相对较强的抗噪性，其ASR仅为38.18%，AP仅从92.33%下降到73.94%，相比于其他模型被攻击后有更高的AP和更低的ASR。这种差异可能源于YOLO系列和Faster R-CNN检测模型依赖卷积操作进行局部特征的层级聚合，感受野扩张具有局部性与层级性。

本文方法生成的对抗贴片可通过破坏局部关键特征实现有效干扰。而DETR模型特征提取网络是基于Transformer设计的，能够捕获全局范围的特征依赖关系，形成融合全局上下文的特征表达，这种全局特征融合能力可稀释局部对抗扰动的影响，因此具有更强抗噪能力。从特征鲁棒性设计上来看，基于卷积神经网络检测模型的卷积核参数固定，导致对局部扰动的敏感性更高。而Transformer的自注意力权重可自适应聚焦目标关键语义区域，对本文方法生成的局部非语义对抗扰动具有更强的抑制能力，进一步提升了其抗扰性能，使DETR模型对本文所生成的特定类型对抗扰动不那么敏感。总体而言，表4的实验结果充分验证了本文方法在跨模型攻击上的强大鲁棒性，证实了其作为一种通用红外物理对抗攻击手段的潜力。同时，DETR模型的低ASR也揭示了未来研究中提升攻击普适性的研究方向。

3.3.5　超参数分析

为验证双重投影优化算法中更新步长η与动量系数σ对抗贴片性能的影响，本节在固定损失参数基础上，对两项参数进行网格扫描，并比较ASR和AP指标，实验结果如表5所示。

从表5的实验结果可以看出，参数η与σ的取值对攻击性能存在显著影响。$\eta = 0.1$时，不同η配置下的ASR和AP均为最优值；$\eta < 0.1$时，步长不足，导致损失函数难以在有限迭代次数内收敛到最优解；$\eta > 0.1$时，步长过大，由于算法中存在两次投影可行域约束，过大的步长会被强制截断，导致梯度信息丢失，最终劣化指标。

在最优步长$\eta = 0.1$时，随着σ从0.5增至0.75，ASR从57.75%持续上升至76.60%（最优），

AP值从53.56%下降至34.62%(最优)。表明此时充分累积了历史梯度,抑制了多目标损失强非凸特性带来的局部噪声,同时保持对损失曲面变化的灵敏响应,最终收敛至全局更优解。

为研究公式(12)和公式(13)中,参数α和μ对对抗贴片性能的影响,本小节在固定参数$\eta=0.1$,$\sigma=0.75$,$\beta=0.9$基础上,通过比较ASR与AP指标,以确定最优参数组合。具体实验结果如表6所示。

表5　超参数η、σ对对抗贴片性能的影响

Table5　Influence of hyperparameters ηandσ

η	$\sigma=0.5$		$\sigma=0.6$		$\sigma=0.7$		$\sigma=0.75$		$\sigma=0.8$		$\sigma=0.85$		$\sigma=0.9$	
	ASR	AP	ASR	AP	ASR	AP	ASR	AP	ASR	AP	ASR	AP	ASR	AP
0.05	38.60	69.37	45.29	64.77	54.71	56.91	57.75	53.76	56.23	54.97	54.94	55.53	48.63	61.36
0.10	57.75	53.36	64.74	43.26	74.77	33.12	**76.60**	**34.62**	75.08	33.56	72.34	37.73	67.47	41.64
0.15	52.28	55.33	59.27	51.70	65.65	46.27	69.30	38.28	69.00	38.77	66.26	43.38	63.53	46.47
0.20	41.64	69.94	49.24	58.65	55.02	54.93	58.66	53.38	57.14	54.83	54.10	56.55	48.63	60.99
0.25	30.09	77.89	36.17	72.71	41.95	69.35	44.68	66.21	43.16	68.81	40.73	70.29	37.99	72.23
0.30	20.97	91.04	25.23	84.79	30.40	77.43	32.22	76.63	31.30	77.24	28.88	82.48	24.32	84.31

注:ASR、AP单位均为%,加粗为各组合指标最优值,下滑线为各列指标最优值。

表6　参数α、μ对对抗贴片性能的影响

Table6　Influence of hyperparameters αandμ

α	$\mu=0.004$		$\mu=0.006$		$\mu=0.008$		$\mu=0.010$		$\mu=0.012$		$\mu=0.014$		$\mu=0.016$	
	ASR	AP	ASR	AP	ASR	AP	ASR	AP	ASR	AP	ASR	AP	ASR	AP
0.50	74.47	35.28	75.15	35.1	75.91	34.89	76.21	34.81	75.38	35.05	74.23	35.35	73.56	35.52
1.00	76.21	34.79	76.82	34.64	77.13	34.56	**77.51**	33.91	76.57	34.72	75.99	34.86	73.77	35.43
1.50	76.52	34.72	77.06	34.58	**77.51**	**33.43**	77.37	34.62	75.45	35.04	75.68	34.98	73.38	35.54
2.00	75.84	34.91	76.90	34.61	77.20	34.56	76.60	34.71	75.76	34.94	74.32	35.33	72.80	35.74
2.50	74.70	35.22	75.45	35.04	76.75	34.71	75.60	34.98	75.32	35.06	72.95	35.70	71.43	36.12
3.00	73.86	35.45	73.97	35.42	75.08	35.08	74.77	35.25	74.16	35.41	72.64	35.82	70.82	36.19

注:ASR、AP单位均为%,加粗为各组合指标最优值,下滑线为各列指标最优值。

从表6中实验结果可看出,随着α从0.50增至3.00,ASR和AP整体呈现先提升后退化的趋势,变化拐点集中在[1,1.5]区间。当$\alpha<1$时,增大参数取值,则提升权重调整幅度与困难目标聚焦能力,从而优化攻击效果;当$\alpha>1.5$时,参数过大,易导致权重更新震荡,忽视其他目标的优化需求,削弱攻击效果。参数μ的变化拐点集中在[0.008,0.010]。$\mu<0.008$时,权重响应不足,DOP策略无法发挥效用;而$\mu>0.010$时,导致权重更新步长过大,破坏优化过程的稳定性。

当$\alpha=1$,$\mu=0.01$和$\alpha=1.5$,$\mu=0.008$时,ASR均取得最优结果77.51%,但从对检测器的干扰能力指标AP来看,当$\alpha=1.5$,$\mu=0.008$时,AP值取得更优结果,低至33.43%,所以这里取$\alpha=1.5$,$\mu=0.008$为最优参数组合。

为研究公式(11)中,参数β对贴片性能的影响,本小节在参数$\eta=0.1$、$\sigma=0.75$、$\alpha=1.5$、$\mu=0.008$基础上,将β在0.1至0.9内取值,其中0.1起始段采用等差取值,0.9附近则进行细化取值以提升关键区间的实验精度,并比较ASR与AP指标,以确定最优参数。实验结果图6所示。

由图6可直观看出,当β从0.1逐步增大至0.9时,ASR提升至峰值77.51%,AP则持续下降至最优

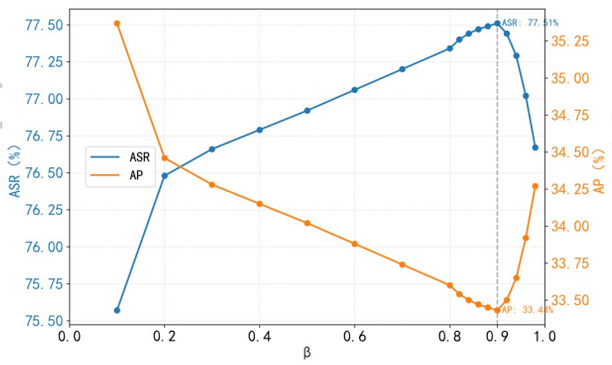


图6　参数β对对抗贴片性能的影响

Fig. 6　Influence of hyperparameters

值33.43%。这一趋势表明，随着β增大，历史KPI信息的权重占比提升，有效抑制了瞬时波动对权重调整的干扰。当β超过0.9时，两项指标均呈现退化趋势。原因在于此时，历史KPI信息主导作用，削弱了对权重调整的反馈灵敏度，多目标优化均衡性被打破，攻击效能退化。

综上，公式(11)中β的最优取值为0.9，此取值下KPI平滑效果与动态反馈灵敏度达到最佳平衡，可最大化DOP策略对多目标优化失衡问题的解决效能，使贴片取得最优的攻击性能。

3.3.6　物理域实验

为评估红外对抗贴片在实际条件下的有效性，本文模拟AIP(Wei 等，2023a)论文中物理实验条件进行实验。实验在室内低温环境(约18摄氏度)中，红外相机正对行人目标固定拍摄，距离为4米，采集其正面视角视频数据。在保持相同距离4米，使志愿者变化角度±10°、±20°、±30°、±40°和±50°进行拍摄，以考察视角偏移对检测性能的影响(由于AIP方法(Wei 等，2023a)论文中物理域实验最大视角偏移为±30°，所以对比实验时仅取视角最大偏移量±30°)。为进一步分析距离因素的作用，将拍摄距离从4米增加至6米。在姿态变化方面，将行人的站立状态调整为坐姿，以验证不同身体姿态下的攻击效果。此外，还将测试场景拓展至室外高温环境(约32摄氏度)，以检验方法在下环境温度变化及复杂背景下的鲁棒性。各类实验场景的直观效果如图7所示。针对每种情况，视频以10帧/秒的速率录制20秒，共获得200帧图像。采用YOLOv3作为检测模型，设定置信度阈值为0.5，并据此计算ASR。图8展示了对抗贴片部署后的实际效果。上侧图展示了干净样本可以直接

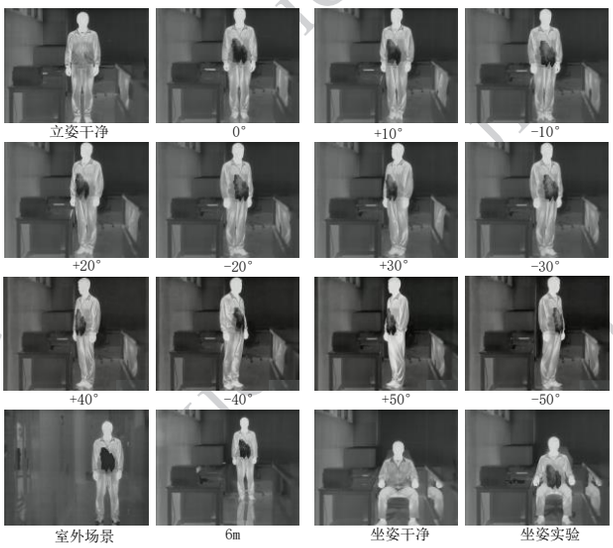


图7　不同场景下实施物理攻击视觉示例

Fig. 7　Visual examples of physical attacks with infrared patches under various conditions.

图8　实施物理攻击后可视化结果

Fig. 8　Visualization results of physical attacks with adversarial infrared patches.

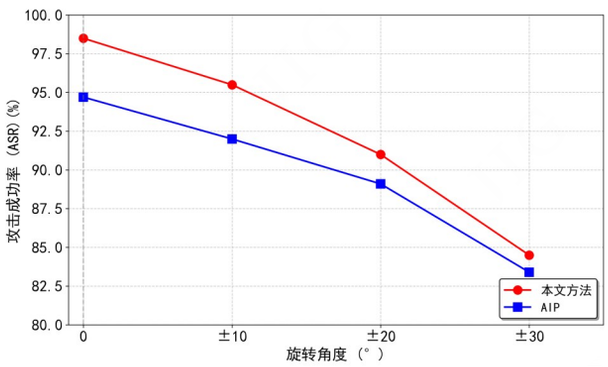
被检测出；下侧图展示了在各类情况下目标在部署贴片无法被检测的可视化结果。

表7实验数据全面验证了本文提出的对抗贴片在多样化物理场景下的卓越攻击效果与强鲁棒性。在正面视角(0°)下，ASR高达98.50%，即使观测角度偏移至±30°，ASR仍能保持在84.50%，表明对抗贴片对视角变化±30°内具有良好的容忍度，但是当视角偏移进一步扩大至±40°和±50°时，攻击成功率出现了断崖式下降，仅为13.25%和2.50%。结合图示分析可知，大角度视角偏移(超过±30°)，单张贴片在摄像机下无法完全成像，导致攻击性能急剧下降。当拍摄距离从4米增加至6米时，ASR下降至

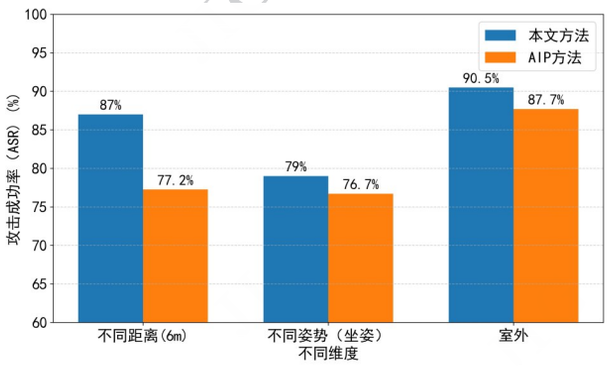
87.00%，说明其具备一定的远距离攻击能力，但同时也说明在距离较远时对抗样本的轮廓细节容易模糊，攻击能力有所减弱；在行人姿态由站立变为坐姿的情况下，ASR下降至79.00%，相较于角度和距离影响变化最大，可能是姿势的变化导致对抗贴片产生非刚性形变的结果；在复杂的室外高温环境（约32摄氏度）中，ASR依然达到90.50%，说明本文方法生成的对抗贴片对于温度以及复杂环境的耐受性。

图9(a)和图9(b)展示了不同维度下本文方法和AIP(Wei 等，2023a)的ASR对比图。通过对比，可以发现本文方法生成的对抗贴片在面对视角，距离，姿势及环境等多种现实挑战均比AIP方法表现更优异。在视角影响实验中如图8(a)所示，0°、±10°、±20°和±30°场景下比AIP方法高出3.83%，3.46%，1.95%和1.12%；同时针对距离，姿势，以及室外场景分别高出9.75%，2.3%，2.8%。充分证明了本文方法在物理域中的高适应性和鲁棒性。

为了进一步验证对抗贴片在动态场景性能，采用相同YOLOV3检测器，阈值设为0.6。并让志愿者附着对抗贴片面对红外摄像机，沿着相同路径从7m缓慢行走到3m，如图9展示了实验中不同距离动态场景中物理攻击的示例图，我们录制了45s视屏，按照每秒10帧，共获得450帧图像，并使用相同指标ASR。



（a） 不同角度ASR对比



（b） 距离，姿势，室外不同场景ASR对比

((a) different angles; (b) other conditions))

图9 与AIP方法多维度场景下ASR对比

Fig. 9 Comparisons with AIP in physical tests.

根据表7结果，在动态场景下，依然展现出优

表7 本文方法在不同实验场景下的实验结果

Table7 Experimental results under different experimental scenarios

	0°	±10°	±20°	±30°	±40°	±50°	距离6m	姿势	室外	动态场景
ASR	98.50%	95.50%	91.00%	84.50%	13.25%	2.50%	87.00%	79.00%	90.50%	80.89%

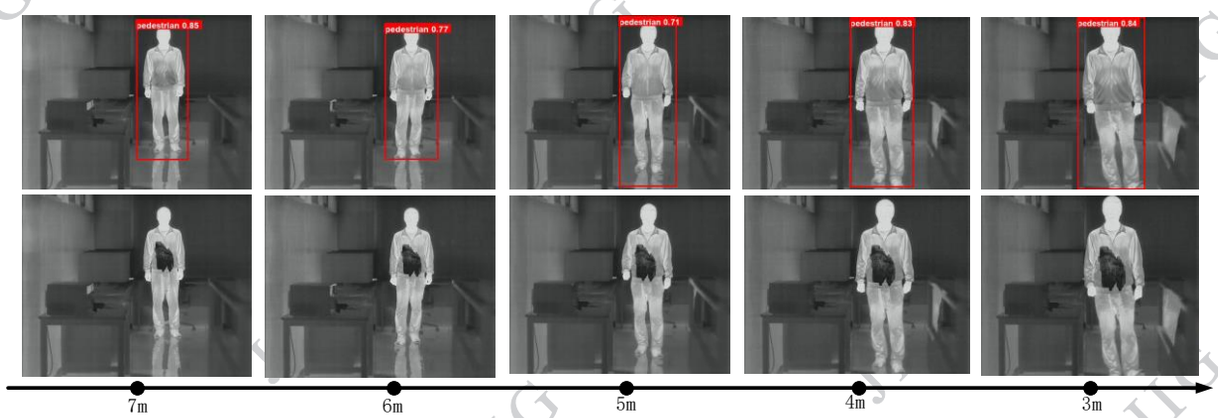


图10 动态场景中物理攻击的示例图

Fig. 10 Visualization results of physical attacks with adversarial infrared patches.

异的攻击效能，成功率达到了80.89%。这一结果充分表明，本文所提对抗贴片不仅在静态场景中具备稳定的攻击性能，验证了其在动态红外监控场景中的可行性。

4 结 论

本文提出的红外物理对抗攻击方法通过动态目标优先级策略与双重投影优化算法的协同设计，有效弥合了攻击性与物理可实现性之间的优化矛盾。在数字域实验中展现出卓越的攻击性能、优化效率；在物理域面对角度、姿势、场景多维

度环境具备鲁棒性，相比于主流方法具有显著的优势。

但需客观指出的是，本文方法虽然对于主流大部分红外目标检测器均具有鲁棒性，但针对DETR及其衍生变体模型的适配效果仍存在优化空间。此外，物理域实验结果可以看出，当前方法生成的对抗贴片尚未实现全视角覆盖的攻击效果，同时在隐蔽性层面仍可被人眼识别。如何实现低成本、快速部署的全视角对抗攻击样本，以

及如何在纹理、像素等高级语义特征缺失的情况下，提高视觉的自然性，并保持攻击性效果，仍然是该领域研究的难点和方向。同时热绝缘材料在长期复杂环境中的性能稳定性仍需进一步验证。

即便存在上述待改进方向，本文方法生成的对抗贴片依然在复杂真实场景中具备高攻击成功率、低成本与快速部署的核心优势，为红外目标检测系统的安全性评估与防御机制研究提供了可靠的技术支撑与实践参考。

参考文献(References)

Bellotti V, Dalal B, Good N, Flynn P, Bobrow D G and Ducheneaut N. 2004. What a to-do: studies of task management towards the design of a personal task list manager// Proceedings of Conference on Human factors in computing systems. Vienna, Austria: ACM: 735-742 [DOI: 10.1145/985692.985785]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers // Proceedings of European Conference on Computer Vision (ECCV 2020). Cham: Springer International Publishing: 213 - 229 [DOI: 10.1007/978-3-030-58452-8_13]

Dai X R, Yuan X and Wei X Y. 2021. TIRNet: Object detection in thermal infrared images for autonomous driving. Applied Intelligence, 51(3): 1244 - 1261 [DOI: 10.1007/s10489-020-01882-2]

Guesmi A, Hanif M A, Ouni B, Abid M and Ben Ayed H. 2023. Physical adversarial attacks for camera-based smart systems: current trends, categorization, applications, research challenges, and future outlook. IEEE Access, 11: 109617 - 109668 [DOI: 10.1109/ACCESS.2023.3321118]

Guesmi A, Ding R, Hanif M A, Alouani I and Shafique M. 2024. DAP: A dynamic adversarial patch for evading person detectors // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24595 - 24604 [DOI: 10.1109/cvpr52733.2024.02322]

Guesmi A, Bilasco I M, Shafique M and Alouani I. 2023. AdvART: Adversarial art for camouflaged object detection attacks [EB/OL]. [2025-02-13]. https://arxiv.org/pdf/2303.01734.pdf

Guo M, Haque A, Huang D A, Yeung S and Li F. 2018. Dynamic task prioritization for multitask learning // Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 270 - 287 [DOI: 10.1007/978-3-030-01270-0_17]

Hou F J, Zhang Y, Zhou Y, Zhang M, Lv B and Wu J Q. 2022. Review on infrared imaging technology. Sustainability, 14(18): 11161 [DOI: 10.3390/su141811161]

Hu C Y, Shi W W, Jiang T S, Yao W, Tian L, Chen X Q and Li W. 2024. Adversarial infrared blocks: a multi-view black-box attack to thermal infrared detectors in physical world. Neural Networks, 175: 106310 - 106328 [DOI: 10.1016/j.neunet.2024.106310]

Hu Y C T, Kung B H, Tan D S, Chen J C, Hua K L and Cheng W H. 2021. Naturalistic physical adversarial patch for object detectors // Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7848 - 7857 [DOI: 10.1109/iccv48922.2021.00775]

Hu Z H, Huang S Y, Zhu X P, Sun F C, Zhang B and Hu X L. 2022. Adversarial texture for fooling person detectors in the physical world // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13307 - 13316 [DOI: 10.1109/cvpr52688.2022.01295]

Hua L, Zhuang Y, Gu F and Chen W. 2021. AdVLP: Unsupervised visible light positioning by adversarial deep learning. Measurement Science and Technology, 32(6): 064003 [DOI: 10.1088/1361-6501/abd2de]

Huang L F, Gao C Y, Zhou Y, Xie C H, Yuille A, Zou C Q and Liu N. 2020. Universal physical camouflage attacks on object detectors // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 720 - 729 [DOI: 10.1109/cvpr42600.2020.00080]

Jia Z J, Hu C Y, Zhang J R, Lu G, Tiliwalidi K, Tian L, Li X and Kang X. 2025. Adversarial infrared Catmull-Rom spline: a black-

box attack on infrared pedestrian detectors in physical world. Information Sciences, 122263 [DOI: 10.1016/j.ins.2025.122263]

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization [EB/OL]. [2025-03-10]. https://arxiv.org/pdf/1412.6980.pdf

Kurakin A, Goodfellow I J and Bengio S. 2018. Adversarial examples in the physical world // Fiedler B, ed. Artificial Intelligence Safety and Security. Boca Raton: Chapman and Hall/CRC: 99 - 112 [DOI: 10.1201/9781351251389-8]

Lapid R, Mizrahi E and Sipper M. 2023. Patch of invisibility: naturalistic black-box adversarial attacks on object detectors [EB/OL]. [2025-04-14]. https://arxiv.org/pdf/2303.04238.pdf

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017. Focal loss for dense object detection // Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2980 - 2988 [DOI: 10.1109/iccv.2017.324]

Liu F C, Nan B and Miao Y W. 2022. Point cloud replacement adversarial Attack based on saliency map. Journal of Image and Graphics, 27(2): 500 - 510 (刘复昌, 南博, 缪永伟. 2022. 基于显著性图的点云替换对抗攻击. 中国图象图形学报, 27(2): 500 - 510) [DOI: 10.11834/jig.210546]

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2017. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2025-04-18]. https://arxiv.org/pdf/1706.06083.pdf

Nesterov Y. 1983. A method for solving the convex programming problem with convergence rate O($1/k^2$). Dokl. Akad. Nauk SSSR, 269 (3): 543 [DOI: 10.1007/bf01069607]

Peng Z B, Zhang Y, Dang Y, Chen J Q, Shi Z W and Zou Z X. 2025. Review of physical adversarial attacks against visual deep learning models. Journal of Image and Graphics, 30(6): 2082 - 2119 (彭振邦, 张瑜, 党一, 陈剑奇, 史振威, 邹征夏. 2025. 针对视觉深度学习模型的物理对抗攻击研究综述. 中国图象图形学报, 30(6): 2082 - 2119) [DOI: 10.11834/jig.240442]

Polyak B T. 1964. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 4(5): 1 - 17 [DOI: 10.1016/0041-5553(64)90137-5]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2024-12-10]. https://arxiv.org/pdf/1804.02767.pdf

Ren S Q, He K M, Girshick R and Sun J. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (6): 1137 - 1149 [DOI: 10.1109/tpami.2016.2577031]

Rippa M, Pagliarulo V, Napolitano F, Valente T and Russo M. 2023. Infrared imaging analysis of green composite materials during inline quasi-static flexural test: monitoring by passive and active approaches. Materials, 16 (8): 3081 [DOI: 10.3390/ma16083081]

Robbins H and Monro S. 1951. A stochastic approximation method. The Annals of Mathematical Statistics, 22(3): 400 - 407 [DOI: 10.1214/aoms/1177729]

Shen M, Liao Z L, Zhu L H, Xu K and Du X J. 2019. VLA: a practical visible light-based attack on face recognition systems in physical world // Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. London, UK: ACM: 1 - 19 [DOI: 10.1145/3351261]

Tan J, Ji N, Xie H and Xiang X S. 2021. Legitimate adversarial patches: evading human eyes and detection models in the physical world // Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM: 5307 - 5315 [DOI: 10.1145/3474085.3475653]

Thys S, Van Ranst W V and Goedemé T. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE: 0-0 [DOI: 10.1109/cvprw.2019.00012]

Tiliwalidi K, Hu C Y, Lu G X, Jia M and Shi W W. 2025. Advgrid: a multi-view black-box attack on infrared pedestrian detectors in the physical world. Applied Soft Computing, 112981 [DOI: 10.1016/j.asoc.2025.112981]

Varghese R and Sambath M. 2024. YOLOv8: a novel object detection algorithm with enhanced performance and robustness // Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems. Chennai, India: IEEE: 1 - 6 [DOI: 10.1109/adics58448.2024.10533619]

Wang D, Yao W, Jiang T, Tang G and Chen X. 2022. A survey on physical adversarial attack in computer vision [EB/OL]. [2024-10-11]. https://arxiv.org/pdf/2209.14262.pdf

Wei H, Wang Z X, Jia X M, Zheng Y Q, Tang H, Satoh S and Wang Z. 2023. Hotcold block: fooling thermal infrared detectors with a novel wearable design // Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 15233 - 15241 [DOI: 10.1609/aaai.v37i12.26777]

Wei X, Yu J and Huang Y. 2023a. Physically adversarial infrared patches with learnable shapes and locations // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12334 - 12342 [DOI: 10.1109/cvpr52729.2023.01187]

Wu J, He Y and Zhao J L. 2024. An infrared target images recognition and processing method based on the fuzzy comprehensive evaluation. IEEE Access, 12: 12126 - 12137 [DOI: 10.1109/access.2024.3355157]

Yan W R, Nie S F, Xu B, Dong H J, Palm L and Diwan V K. 2012. Establishing a web-based integrated surveillance system for early detection of infectious disease epidemic in rural China: a field

experimental study. BMC Medical Informatics and Decision Making, 12(1): 4 [DOI: 10.1186/1472-6947-12-4]

Ye Y X, Du X, Chen S, Zhu S Z and Yan Y. 2024. Sparse adversarial patch attack based on QR code mask. Journal of Image and Graphics, 29(7): 1889 - 1901 (叶乙轩, 杜侠, 陈思, 朱顺痣, 严严. 2024. 二维码掩膜下的稀疏对抗补丁攻击. 中国图象图形学报, 29(7): 1889 - 1901) [DOI: 10.11834/jig.230453]

Zhu X P, Hu Z H, Huang S Y, Li J M and Hu X L. 2022. Infrared invisible clothing: hiding from infrared detectors at multiple angles in real world // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13317 - 13326 [DOI: 10.1109/cvpr52688.2022.01296]

Zhu X P, Hu Z H, Huang S Y, Li J M, Hu X L and Wang Z Y. 2023. Hiding from infrared detectors in real world with adversarial clothes. Applied Intelligence, 53(23): 29537 - 29555 [DOI: 10.1007/s10489-023-05102-5]

Zhu X P, Li X, Li J M, Wang Z Y and Hu X L. 2021. Fooling thermal infrared pedestrian detectors in real world using small bulbs // Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Conference: AAAI Press: 3616 - 3624 [DOI: 10.1609/aaai.v35i4.16477]

作者简介

王俊，1997年生，男，硕士研究生，研究方向为物理对抗攻击。E-mail：3253305122@qq.com

李阳，通信作者，男，副教授，主要研究方向为计算机视觉与图像检索。 E-mail：solarleeon@outlook.com

苗壮，男，教授，主要研究方向为图像视频处理。E-mail：emiao_beyond@163.com

侯壹凡，男，博士研究生，主要研究方向为深度伪造。E-mail：201921002000@smail.xtu.edu.cn

毕翔鹤，男，硕士研究生，主要研究方向为图像融合，深度学习。E-mail：614307276@qq.com

王家宝，男，副教授，主要研究方向为计算机视觉与机器学习。E-mail：jiabao_1108@163.com