 DOI: https://doi.org/10.12345/bdai.v5i8.20942

Visible-Infrared Pedestrian Target Detection Deception Based on Differential Evolution

Zhiyang Hu[1,2] Haoli Xu[2*] Haoqi Gao[2] Bin Qu[3] Mengjiang Wu[3]

1. Hefei University of Technology, Hefei, Anhui, 230009, China
2. College of Electronic Engineering, National University of Defense Technology, Hefei, Anhui, 230027, China
3. Jianghuai Advance Technology Center, Hefei, Anhui, 230032, China

Abstract

Current research mostly focuses on adversarial attacks on single-spectral target detectors, while multi-spectral detectors are more practical in real-world scenarios. In order to evaluate the security of multi-spectral detectors more effectively, this paper proposed a Unified Multispectral Adversarial Attack (UMAA), which can attack both visible and infrared detectors. The texture, position and rotation angle of the adversarial samples are optimized by differential evolutionary algorithm to generate visible light adversarial samples, which are then grayed out to generate infrared adversarial samples to attack both detectors. Experimental results show that the method has significant effectiveness and robustness.

Keywords

adversarial samples; multi-spectral target detector; differential evolutionary algorithm

基于差分进化的可见—红外行人目标检测诱骗

胡至洋[1,2] 许颢砾[2*] 高皓琪[2] 瞿斌[3] 邬梦江[3]

1. 合肥工业大学，中国·安徽 合肥 230009
2. 国防科技大学电子对抗学院，中国·安徽 合肥 230027
3. 江淮前沿技术协同创新中心，中国·安徽 合肥 230032

摘　要

当前研究多集中于单波段目标检测器的对抗攻击，而现实场景中多波段检测器更为实用。为更有效地评估多波段检测器的安全性，论文提出一种统一结构对抗样本（Unified Multispectral Adversarial Attack，UMAA），能同时攻击可见光和红外检测器。通过差分进化算法优化对抗样本的纹理、位置和旋转角度，生成可见光对抗样本，再灰度化生成红外对抗样本，从而对两种检测器发起攻击。实验结果表明，该方法具有显著的有效性和鲁棒性。

关键词

对抗样本；多波段目标检测器；差分进化算法

1 引言

深度神经网络在计算机视觉和自然语言处理方面取得了惊人的表现，被广泛地应用于各个领域中。如：移动支付[1]、自动驾驶[2]、目标检测[3]、医疗诊断[4]等，特别是目标检测领域，它不仅适用于可见光、红热外成像，还适用于多波段数据融合检测。

然而，尽管深度学习在各个领域已经取得了巨大的成功，但最近研究表明，现有的深度神经网络模型并不是绝对可靠和安全的。因此，研究对抗攻击方法评估目标检测器的安全可靠性至关重要。目前研究主要集中在单模态对抗攻击，如对可见光目标检测器的对抗攻击研究[5-8]和对红外目标检测器对抗攻击的研究[9-12]，尽管已经有研究尝试了对跨模态检测器的对抗攻击，但这些方法通常存在一些局限性。例如，Kim 等人[13]首次提出一种多波段对抗样本生成框架（MAP），但是无法对行人进行多角度攻击，随后，Kim 等人[14]又提出的多光谱隐形涂层方法利用了透明的低辐射（Low-E）膜[15]进行物理攻击，虽然在物理世界中表现出色，但其设计复杂且实施成本较高。近期的研究提出了一些新的

【基金项目】江淮前沿技术协同创新中心追梦基金课题（项目编号：2023-ZM01D006）；国家自然科学基金青年科学基金项目（项目编号：62305389）。

【作者简介】胡至洋（2000-），男，中国安徽六安人，在读硕士，从事智能对抗研究。

【通讯作者】许颢砾（1993-），男，中国江西赣州人，博士，讲师，从事智能对抗研究。

方法来克服上述挑战。Wei 等人 [16] 提出统一补丁的对抗攻击的方法，能多角度、多尺度的攻击行人检测系统，该方法成本低、制作简单方便，并且可见光和红外的对抗样本形状结构相同，易于部署至物理世界。此外，Hu 等人 [17] 提出的 Two-stage Optimized Unified Adversarial Patch（TOUAP）是一种针对跨模态检测器的物理攻击方法。它通过两阶段优化过程，分别优化红外和可见光对抗样本，最终生成统一的对抗样本。实验结果表明，该方法不仅在数字环境中表现出色，在物理世界中的跨模态攻击中也具有很强的鲁棒性和有效。

然而，尽管这些新方法在特定场景下表现优异，跨模态对抗攻击仍面临诸多挑战，如何在保持物理可行性的同时，最大化对抗攻击效果，以及如何设计更为灵活和通用的对抗样本，以应对不同类型的跨模态检测器，并且实施场景主要集中在自动驾驶领域，没有针对航拍场景进行对抗攻击的研究。

因此，笔者提出了 UMAA 对抗攻击方法，通过差分进化算法优化对抗样本的纹理结构、位置和旋转角度，以生成对可见光检测器具有较强攻击性的对抗样本。具体而言，差分进化算法是一种基于种群的全局优化算法，它通过不断调整样本的各项参数，寻找最优的对抗样本，以最大化对目标检测器的误导效果。在此基础上，生成的可见光对抗样本通过灰度化处理，转化为适用于红外光检测器的对抗样本。由于可见光和红外光在物理特性上的差异，通过统一的纹理结构来生成对抗样本，使其能够同时对两种检测器进行有效攻击。

2 方法

2.1 问题定义

给定多模态行人数据集 *Dv*，其中 *Ip*、*Lp* 分别代表行人的图像以及对应的标签，*Ip* ∈ *Dp*，对 Yolov5 检测器进行微调，分别训练出红外和可见光的模型 *fMulti*( )，*vis*, *ir* ∈ *Multi*，将 *Dp* 内的数据经过训练好的 *fMulti*( ) 多模态检测器中，*fMulti*(*Ip*) → *Lpred*，得到预测结果 *Lpred*，其中 *Lpred*=(*Vpos*, *Vobj*, *Vcls*)。论文的目的是让行人躲避目标检测，因此，论文重点关注 *Vobj* 信息，定义如下：

$$\arg\min_{\min}\left(V_{obj} \leftarrow f(I_p)\right) \tag{1}$$

2.2 UMAA 对抗攻击方法

为了保证行人能够多角度、多尺度地躲避多波段目标检测系统，论文将采用多补丁联合攻，为了找到更好的对抗样本，论文拟采用差分进化算法 [18]，优化上文提到的对抗样本的优化参数。差分进化算法是一种有效的进化算法，用于全局优化，具有较好的全局搜索能力，它对参数设置不敏感，即使在参数选择不是最优的情况下也能有良好的表现，并且控制参数较少，通常只有种群大小、交叉概率和变异三个主要参数，能够适应不同类型和规模的优化问题。

采用差分进化算法优化上述提到的对抗攻击方法描述如图 1 所示。

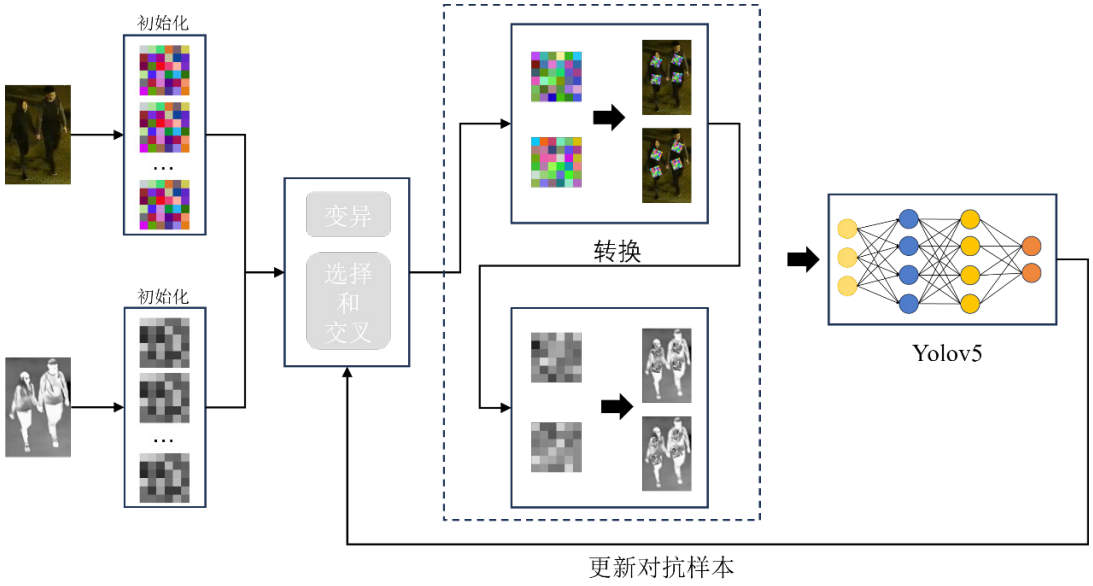

图 1 UMAA 对抗样本生成方法概述

变异：对于种群中的每个个体 *i*，执行以下步骤：

首先，随机选择三个不同的索引 r_0、r_1、r_2，直到它们互不相同且与 *i* 不同：$r_0, r_1, r_2 \sim U(0, p\text{-}1)$，$r_0 \neq r_1 \neq r_2$，其中 p 为种群大小，$U(0, p\text{-}1)$，表示区间 [0, *p*-1] 内均匀分布的随机数。

随后，计算变异后的个体，描述如下：

$$p_{tmp} = p[r_0] + (p[r_1] + p[r]) \times p_m$$
$$s.t.\ p_m = rand[0,1] \tag{2}$$

其中，*pm* 为变异概率；*ptmp* 为变异后的个体，对于 *ptmp* 的个体，如果超过了定义的范围 [min, max]，则将其设置该范围内的随机值，定义如下：

$$p_{tmp}[t]=\begin{cases}rand[0,1] & if\ p_{tmp}[t]>\max\ or\ p_{tmp}[t]<\min\\ p_{tmp}[t] & otherwise\end{cases} \tag{3}$$

最后，将变异后的结果添加到变异群体中。

选择和交叉：对于变异后种群的每个个体，执行步骤如下：

首先，选择目标个体 *ptarget=p*[*i*] 和变异个体 *pdonor=mutate*[*i*]，初始化一个空的试验个体 *ptrial*。

随后，随机选择一个交叉点 *divide_point* ~ *U* (0, *len*(*p*) - 1)，对于 *ptrial* 的每个维度 *j* 执行以下操作：

$$trial[j]=\begin{cases}p_{donor}[j] & if\ rand[0,1]<p_c\ or\ j=divide\_point\\ p_{target}[j] & otherwise\end{cases} \tag{4}$$

计算试验个体的适应度 *trial_fitness* 和目标个体的适应度 *target_fitness*。

如果试验个体的适应度优于目标个体，则用试验个体替换目标个体，并更新适应度值：

$$\begin{cases}p[i]=p_{trial} & if\ trial\_fitness<target\_fitness\\ fitness\_values[i] & if\ trial\_fitness>target\_fitness\end{cases} \tag{5}$$

通过采用上述方法优化对抗样本的纹理结构、位置和旋转角度，以此来获得最佳的对抗样本。

考虑到论文的目标是保证生成的对抗样本在可见光和红外条件下的纹理结构相同，如图 2 所示，论文定义了一个权重数组 *weights*=[0.299, 0.587, 0.114]，用于将 *vis_patch* 转换为 *ir_patch*，表述如下：

$$ir\_patch=\sum_{c=1}^{3}vis\_patch[c,:,:] \tag{6}$$

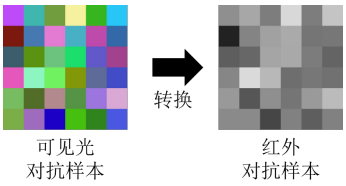

图 2 可见光对抗样本转换为红外对抗样本效果图

2.3 损失函数优化

论文的目标是保证行人能同时躲避红外和可见光目标检测器，并且生成的对抗样本要在纹理和结构上保持一致，因此，论文设定的损失函数定义如下：

$$loss_{obj}=\begin{cases}loss_{ir} & if\ f_{multi}(I_{ir\_adv})\geq f_{multi}(I_{vis\_adv})\\ loss_{vis} & otherwise\end{cases} \tag{7}$$

$$s.t.\ loss_{ir}=f_{multi}(I_{ir}),\ loss_{vis}=f_{multi}(I_{vis})$$

$$loss_{tv}=\sqrt{(I_{i,j}-I_{i+1,j})^2+(I_{i,j}-I_{i,j+1})^2} \tag{8}$$

$$loss=loss_{obj}+loss_{tv} \tag{9}$$

lossir 和 *lossvis* 分别代表带有对抗样本的可见光和红外图像分别经过多波段目标检测系统检测后得到的目标识别分数。损失函数总体包含两个部分，*lossobj* 是保证行人可以同时躲避可见光和红外目标检测系统，而 *losstv* 目的是保证生成的对抗样本更加的平滑。

3 实验

3.1 实验设置

3.1.1 数据集

论文的目的是使行人躲避多波段目标检测系统，因此，论文选择 LLVIP 数据集 [19] 作为实验数据集，并从中选取了 1500 张图像，其中 1000 张作为训练集，500 张作为验证集，将其命名为 LLVIP-mini。

3.1.2 目标检测器

论文选择使用 Yolov5 检测器 [20] 作为攻击目标，并对 Yolov5 检测器进行微调，通过训练 LLVIP-mini，通过训练 LLVIP-mini 可见光部分的 AP（Average Precision）值为 96.4%，训练 LLVIP-mini 红外部分的 AP 值为 99.3%。

3.1.3 评估标准

论文采用 AP 和 ASR（Attack Success Rate），干净样本的数量为 N，贴有对抗样本经过检测器检测后的数量为 M，ASR 定义为 (N-M)/N，ASR 越高，说明对抗样本的攻击效果越好，AP 为横坐标为召回率，纵坐标为精度的曲线，AP 值越低，攻击效果越好。

3.1.4 其他细节

论文采用差分进化算法优化对抗样本的纹理结构、位置和旋转角度，论文将种群的规模设为 100，迭代次数设为 3，交叉率为 0.5，变异率为 0.1，所有的代码都在 pytorch 中实现，并在 RTXA100 上运行。

3.2 实验结果分析

论文采用多补丁联合攻击方法，采用差分进化优化算法，在目标图像中精心设计多个对抗样本，保证行人能够同时躲避可见光和红外目标检测器。为了保证这些对抗样本在视觉上不易被察觉，同时又能有效地欺骗目标检测器，论文选择将对抗样本的数量（n）从 1~4 变化，间隔为 1；对于占用目标检测框的比例（s），论文选择了 0.2~0.3 的范围，每次间隔 0.02。论文设计了几组消融实验，系统地变化 n 和 s 的值，以评估它们对攻击效果的影响。实验结果表明，对抗样本的数量和占用目标框的比例都对攻击成功率有显著影响。具体来说，通过表 1 可知，随着 n 的增加，攻击成功率提高，但当 n 超过 3 时，边际效益开始减少。对于 s，我们发现在 0.26 左右时，攻击效果最佳。在考虑将对抗样本应用于物理世界时，我们特别关注了对抗样本的数量为 2 时，不同 s 值对攻击效果的影响。通过表 2 发现对抗样本所占目标框比例的增加，可以提高攻击的成功率，这对于后续的物理迁移研究具有重要意义。攻击效果如图 3 所示。

表 1 对抗样本的数量在可见光和红外目标检测器下的攻击效果

模态	评估准则	n			
		1	2	3	4
可见光	AP（%）	85.9	52.6	53.0	59.7
	ASR（%）	22.3	50.1	50.8	56.1
红外	AP（%）	93.6	79.6	75.2	66.5
	ASR（%）	11.2	33.3	39.5	52.3

表 2 对抗样本占目标比例的大小对可见光和红外目标检测器的攻击效果

模态	评估准则	size（%）					
		0.20	0.22	0.24	0.26	0.28	0.30
可见光	AP（%）	88.4	72.9	73.0	52.6	54.4	60.2
	ASR（%）	22.8	43.2	44.5	50.1	54.0	56.2
红外	AP（%）	85.0	88.0	73.0	79.6	85.8	80.6
	ASR（%）	20.1	20.7	23.3	33.3	25.6	29.2

图 3 对抗样本在可见光和红外条件下的攻击效果图

4 结论

论文提出了一种新的多波段对抗样本生成方法，因为生成的对抗样本在可见光和红外条件下纹理和结构相同，可保证行人同时躲避多波段目标检测器，采用差分进化算法优化对抗样本的纹理结构、位置和旋转角度，以此找到最具有攻击效果的对抗样本。未来的研究可以继续深化对跨模态检测器的对抗攻击研究，特别是针对更复杂的多传感器系统（如可见光—红外—雷达系统）的攻击。此外，随着对抗攻击方法的不断发展，探索更加有效的防御策略，如增强的对抗训练和动态伪装技术，将是应对这些威胁的重要方向。跨模态对抗攻击是一个充满挑战和机遇的研究领域，其发展将对未来的自动驾驶和智能监控系统的安全性评估产生深远影响。

参考文献

[1] Meiyan D. Mobile payment recognition technology based on face detection algorithm: Mobile Payment Recognition Technology[J]. Concurrency and Computation Practice and Experience, 2018, 30:e4655.

[2] Badue C, Guidolini R, Carneiro R V, et al. Self-driving cars: A survey[J]. Pergamon, 2021.

[3] Redmon J, Farhadi A. 2017. YOLO9000: Better,Faster, Stronger. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 6517–6525. IEEE Computer Society.

[4] A S B, A P K R M, B Q V P, et al. Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey[J].Sustainable Cities and Society, 2020.

[5] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[J].Computer Science, 2014.

[6] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015:27-36.

[7] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP),2017: 39-57.

[8] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations (ICLR), 2018.

[9] Zhu X, Hu Z, Huang S, et al. Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[10] Zhu X, Li X, Li J, et al. Fooling thermal infrared pedestrian detectors in real world using small bulbs. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021.

[11] Zhu X, Liu Y, Hu Z, et al. Infrared Adversarial Car Stickers[J]. IEEE, 2024.

[12] Wei X, Yu J, Huang Y. Infrared Adversarial Patches with Learnable Shapes and Locations in the Physical World[J].International Journal of Computer Vision, 2024, 132(6):1928-1944.

[13] Hwang S, Park J, Kim N, et al. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015:1037-1045.

[14] Kim T, Yu Y, Ro Y M. Multispectral Invisible Coating: Laminated Visible-Thermal Physical Attack against Multispectral Object Detectors Using Transparent Low-E Films[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023,37(1):1151-1159.

[15] Winckler L. Low-Emissivity, Energy-Control, Retroft Window Film. Technical report, Cpflms Incorporated, Fieldale, VA (United States),2012.

[16] Wei X, Huang Y, Sun Y, et al. Unified Adversarial Patch for Cross-modal Attacks in the Physical World[J]. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023:4422-4431.

[17] Hu C, Shi W. (2023). Two-stage optimized unified adversarial patch for attacking visible-infrared cross-modal detectors in the physical world. *ArXiv, abs/2312.01789*.

[18] Price K V. Differential evolution[M]. Handbook of Optimization. Springer, Berlin, Heidelberg, 2013.

[19] Jia X, Zhu C, Li M, et al. LLVIP: A visible-infrared paired dataset for low-light vision[C]//Proceedings of the IEEE/CVF international conference on computer vision, 2021:3496-3504.

[20] Jocher G, Chaurasia A, Stoken A, et al. ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations[J]. Zenodo, 2022.