计算机科学
COMPUTER SCIENCE

面向可见光与红外多模态目标检测的对抗攻防综述

郑海斌, 林秀豪, 陈靖文, 陈晋音

引用本文

郑海斌, 林秀豪, 陈靖文, 陈晋音. 面向可见光与红外多模态目标检测的对抗攻防综述[J]. 计算机科学, 2025, 52(11): 349-363.

ZHENG Haibin, LIN Xiuhao, CHEN Jingwen, CHEN Jinyin. Survey of Adversarial Attack and Defense for RBG and Infrared Multimodal Object Detection [J]. Computer Science, 2025, 52(11): 349-363.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200151

面向可见光与红外多模态目标检测的对抗攻防综述

郑海斌[1,2,3]　林秀豪[1]　陈靖文[1]　陈晋音[1,3]
1 浙江工业大学信息工程学院　杭州 310000
2 四川大学数据安全防护与智能治理教育部重点实验室　成都 610000
3 浙江工业大学计算机科学与技术学院、软件学院　杭州 310000
(haibinzheng320@gmail.com)

摘　要　目标检测作为计算机视觉中的一项基本任务被广泛应用,而基于深度学习的目标检测算法以其强大的特征提取能力,成为了当前研究的主流。然而大多数目标检测算法仅对可见光图像或红外图像进行单模态检测。通常情况下,可见光图像在天气恶劣、夜间、目标被遮挡等场景成像较差,导致检测性能下降。利用红外图像可以改善上述问题,但红外图像会缺失目标的部分细节信息。因此,基于可见光和红外图像的多模态融合检测算法逐渐兴起。然而,现有的研究集中于改善多模态目标检测算法的性能,对于其安全性的研究相对零散。基于现有的研究工作,围绕多模态目标检测对抗安全性进行综述。首先对多模态目标检测及攻防进行理论分析;然后按照不同时段的融合检测对多模态目标检测方法进行分类归纳,再对现有的目标检测对抗攻击方法与对抗防御方法进行归纳整理,梳理了现有的多模态目标检测数据集与主要评价指标;最后探讨了多模态目标检测未来潜在的研究方向,进一步推动多模态目标检测对抗安全研究发展和应用。

关键词:目标检测;深度学习;多模态目标检测;对抗攻击;防御
中图分类号　TP391.4

Survey of Adversarial Attack and Defense for RBG and Infrared Multimodal Object Detection

ZHENG Haibin[1,2,3],LIN Xiuhao[1],CHEN Jingwen[1] and CHEN Jinyin[1,3]
1 School of Information Engineering,Zhejiang University of Technology,Hangzhou 310000,China
2. Key Laboratory of Data Protection and Intelligent Management Ministry of Education,Sichuan University,Chengdu 610000,China
3 College of Computer Science and Technology,College of Software,Zhejiang University of Technology,Hangzhou 310000,China

Abstract　Object detection,as a fundamental classic task in the field of computer vision,has a wide range of applications. Deep learning based object detection algorithms have become the mainstream of current research due to their superior performance. However,most object detection algorithms only perform single-mode detection on visible or infrared images. In general,visible images have poor imaging in harsh weather,nighttime,and scenes,where targets are obstructed,leading to a decrease in detection performance. The use of infrared images can improve the above issues,but infrared images may miss some details of the target. Therefore,multimodal fusion detection algorithms based on visible light and infrared images are gradually emerging. However,existing research has focused on improving the performance of multimodal object detection algorithms,and research on their security is relatively scattered. Based on existing research work,this paper provides an overview of the security of multimodal object detection in adversarial situations. Firstly,a theoretical analysis of multimodal object detection and attack and defense is conducted. Secondly,multimodal object detection methods are classified and summarized according to fusion detection in different time periods. Then,existing methods of object detection and adversarial defense are summarized and organized,and the existing dataset and main evaluation indicators of multimodal object detection are summarized. Finally,potential research directions for multi-modal object detection in the future are discussed,further promoting the development and application of multimodal object detection

到稿日期:2024-12-19　返修日期:2025-04-23
基金项目:国家自然科学基金(62406286);浙江省自然科学基金(LDQ23F020001);四川大学数据安全防护与智能治理教育部重点实验室放课题(SCUSAKFKT202402Z);北京生命科技研究院有限公司开放基金(2024200CD0210)
This work was supported by the National Natural Science Foundation of China(62406286),Zhejiang Provincial Natural Science Foundation(LDQ23F020001),Key Laboratory of Data Protection and Intelligent Management,Ministry of Education,Sichuan University(SCUSAKFKT202402Z) and Beijing Life Science Academy(BLSA)(2024200CD0210).
通信作者:陈晋音(chenjinyin@zjut.edu.cn)

in adversarial security research.

Keywords Object detection, Deep learning, Multimodal object detection, Adversarial attacks, Defense

1 引言

目标检测是一个经典的计算机视觉问题，深度学习的快速发展极大地推动了该项技术的创新与突破[1]。它帮助人们高效率、高精度地自动识别图像或视频中的特定目标物体。

由于基于单模态图像（例如可见光）的目标检测技术在天气恶劣、夜间、目标被遮挡等极端场景下的检测性能较差，因此为了提高检测效果，可以将可见光图像丰富的目标纹理信息与红外图像显示的目标热分布相融合，获得更加完备的图像信息[2]。目前，针对多模态图像融合的目标检测技术得到了广泛研究，使得可见光传感器和热红外传感器在自动驾驶、安全监控等安全关键任务中得到了广泛的应用[3-5]。

现有大部分研究集中于改善融合检测的性能，忽略了对其安全性的研究[6-11]。已有大量研究表明，基于深度学习的目标检测模型容易受到对抗性补丁攻击[12-14]，即恶意攻击者故意制造带有对抗扰动的物体来欺骗机器学习模型并导致错误预测。目前对抗性补丁正在以各种形式发展，对现实世界的检测器构成显著的威胁。例如，犯罪分子可以穿着对抗服装实现“隐身”来躲避自动监控系统；蓄意报复社会者可以通过发射对抗激光束来入侵自动驾驶汽车的视觉系统，造成难以预想的交通事故。为了预防这种由对抗性补丁引起的灾难性后果，研究物理攻击和开发更为鲁棒的模型对于现实世界的检测模型是必要的。

在物理世界中，由于可见光传感器和红外传感器具有不同的成像机制，单模态物理攻击无法同时攻击多模态目标检测器。具体来说，红外传感器无法捕捉到可见光模态中产生的对抗扰动，而针对红外传感器产生的对抗扰动缺乏纹理色彩信息，也无法直接攻击可见光传感器。因此，目前针对单模态领域目标检测的物理攻击方法无法揭示多模态目标检测模型的对抗安全性。例如，一些研究[15-17]在可见光模态下攻击了目标检测器，一些研究[18-20]则在红外模态下对目标检测器进行了攻击，而针对多模态目标检测的对抗攻击研究较少[21-23]。本文围绕多模态目标检测安全性进行综述，总结现有的多模态目标检测算法及对抗攻防安全性研究，统计常用数据集与评价指标，并探讨了未来研究方向，以期进一步推动可见光与红外多模态目标检测对抗安全领域的研究与发展。

本文第2章概述多模态目标检测技术及其面临的对抗威胁；第3章基于不同的融合策略归纳现有的多模态目标检测方法；第4章介绍当前多模态目标检测的对抗攻击方法；第5章介绍当前多模态目标检测的对抗防御方法；第6章总结多模态目标检测安全领域的数据集及不同任务下的评估指标；最后总结未来关于多模态目标检测安全领域可能的研究方向。本文章节内容与关系图如图1所示。

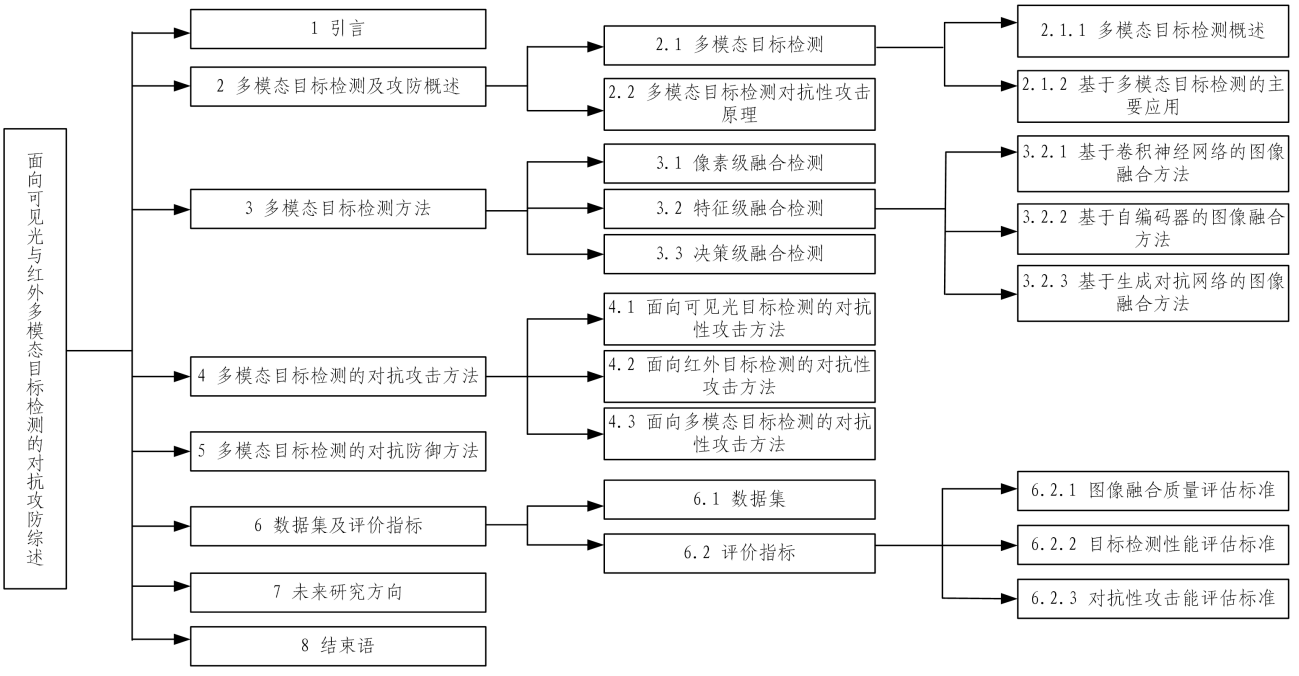

图1 面向可见光和红外多模态目标检测的对抗攻防综述

Fig. 1 Survey of adversarial attack and defense for RBG and infrared multimodal object detection

2 多模态目标检测及攻防概述

本章首先介绍了多模态目标检测的概念、工作原理及主要应用方向，然后给出不同模态的对抗攻击方法的定义。

2.1 多模态目标检测

首先介绍目标检测的概念与发展阶段，然后由其局限性

引出多模态目标检测，最后分析目前基于多模态目标检测的主要应用。

2.1.1 多模态目标检测概述

目标检测作为计算机视觉领域的关键研究方向之一，旨在找到图像中特定关注的物体，并精准地实现定位与分类。然而，各种物体的形态各不相同且在成像时由于光照、遮挡引起的干扰，使得目标检测极具挑战。而多模态目标检测可以改善上述问题。

可见光(Visible Light)与红外(Infrared)多模态目标检测旨在融合可见光模态与红外模态图像的互补信息，以提高目标检测性能。其中可见光模态的图像包含更多的细节与纹理特征；而红外图像不受光照因素的干扰，在极端情况如雨雾、强光、遮挡等情况下能提供更多的目标特征。因此融合两者的图像信息，能提高目标检测的准确率。

2.1.2 基于多模态目标检测的主要应用

多模态目标检测技术主要应用于自动驾驶、农业监测、医疗影像诊断、灾害预测、军事侦查等领域。

1)自动驾驶领域：自动驾驶检测通过车辆上安装的摄像头获取图像，包括路面、静态物体与动态物体。通过对外界信息的采集、处理与分析，车辆可以检测到各种移动与静止的障碍物，从而定位静态与动态对象，如建筑物、车辆、行人等，识别可行驶空间。此外，路面信息包括可行驶区域、车道线、交通标志、绿化带、红绿灯等，这些信息能帮助车辆了解周边驾驶环境，准确规划出可驾驶路径。而多模态目标检测技术能帮助车辆收集更多信息，做出更安全可靠的选择。

2)农业监测领域：农业监测领域是指通过对农业生产过程中的各种数据进行收集、分析和处理，实现对农业生产过程的全面监测和预警。然而过多的太阳光、雨水或雾等情况会改变相机成像条件，利用可见光单模态的图像信息无法对农作物进行有效检测。因此，通过融合可见光与红外多源图像的特征信息可以实现对作物生长过程的全面监测和预警。

3)多模态医疗影像领域：医学影像是医疗诊断过程的关键依据之一，多模态医疗影像相对于单模态影像，能有效地帮助医生从多个维度一起判断病灶，提高诊断效果。在医学影像中，利用不同体征参数的成像能力，红外和可见光的多模态目标检测技术可以增加图像的信息内容，从而提高疾病的检出率和定位的准确性。

4)灾害预测领域：灾难预测领域是指通过对灾难发生的各种数据进行收集、分析和处理，以实现对灾难发生的全面监测和预警。其技术应用包括：气象预测、地震预测、洪水预测、火灾预测等。红外与可见光的多模态目标检测可以提供更多、更完整的信息，帮助发现并分析地面目标，有助于人们更好地了解灾难的发生规律和趋势。

5)军事监察与侦查领域：在复杂战场环境中，常有遮挡物多、光线条件差的情况，如目标周围有伪装物体、雨雾烟尘环境、目标身着伪装衣裤等，利用红外线可穿透烟雾和黑夜，对目标进行探测和监视，提高目标检测效果。

2.2 多模态目标检测的对抗攻击原理

自 Szegedy 等[24]首次提出对抗样本的概念以后，一些经典的对抗攻击方法也陆续被提出，如 Fast Gradient Sign Method[25](FGSM)，Basic Iterative Method[26](BIM)，Projected Gradient Descent[27](PGD)等。下面以 FGSM 为例，说明对抗攻击的工作原理。

FGSM 通过梯度上升最大化损失函数生成对抗样本，属于无目标攻击，即对于一个微小的扰动量 ε，沿着梯度的 l_∞ 范数方向进行一步扰动：

$$x' = x + \varepsilon \cdot sign(\nabla_x L(x, y; \theta)) \tag{1}$$

其中，$sign$ 函数在自变量大于 0 时取 1，小于 0 时取 -1；x 为模型输入样本；y 为样本标签；θ 为参数；L 为损失函数；∇ 为取偏导数。

Xie 等[28]在 2017 年首次将对抗攻击应用到目标检测中，提出了 DAG(Dense Adversary Generation)方法，其定义如下：

$$r_m = \sum_{t_n \in T_n} [\nabla_{X_m} f_{l,n}(X_m, t_n) - \nabla_{x_n} f_{ln}(X_m, t_n)] \tag{2}$$

$$X_{m+1} = X_m + \frac{\partial r_m}{\| r_m \|} \tag{3}$$

其中，X_m 代表经过 m 次迭代后的图片；f 代表目标检测函数；t_n 为模型输入样本中 T_m 个目标中的其中一个目标，l_n 为该目标的正确类别，l_n' 为错误标签；r_m 为对抗梯度；∂ 为学习率。

在多模态目标检测领域中，其对抗攻击原理与上述类似，旨在生成同时攻击可见光与红外模态的对抗扰动，在数字空间以像素扰动呈现[29]，而物理域中则通常采用可见光扰动与红外扰动层叠的形式实现[23]。

3 多模态目标检测方法

根据融合时期的不同，基于可见光和红外图像的多模态融合目标检测算法可以分为像素级融合检测、特征级融合检测和决策级融合检测。

3.1 像素级融合检测

像素级融合检测方法首先基于传统的图像处理技术和一些数学模型来融合可见光与红外图像，再将融合图像输入目标检测模型实现融合检测。这些方法主要包括均值相加、小波变换、拉普拉斯金字塔等像素级操作，通常基于人工设计的特征和规则，需要人工选择和优化参数。主流方法有多尺度变换[6,30-32]、表示学习[7,33-37]以及显著性方法[38]。

基于多尺度变换的方法首先通过数学变换将图像分解成一系列多尺度表示，然后基于融合规则融合多尺度表示，最后通过相应的逆变换来重建融合图像。多尺度变换侧重于从源图像中提取不同尺度的特征，适用性强，且与人类视觉系统一致[39]，获得的融合图像具有良好的视觉效果，能提高目标检测的性能。

基于表示学习的方法从机器学习的角度分析图像，试图将原始数据转换为新的表示形式，这些新的表示形式可以更好地反映数据的本质特征，有利于数据理解和处理。在表示学习领域[40]，最常见的方法是稀疏表示(Sparse Representation，SR)和低秩表示(Low-Rank Representation，LRR)。在图像融合任务中，稀疏表示和低秩表示的主要作用是压缩图像，从而更好地把握和融合其主要特征。

基于显著性的方法主要关注图像的显著性，即图像中的某些区域或物体在视觉上更具有吸引力和重要性，更容易被

人眼所关注和记忆。它可以保持重要对象区域的完整性和像素强度，并提高融合图像的视觉质量[41]。

在像素级融合检测方法中，不同技术各有优劣。多尺度变换方法通过数学变换将图像分解为多尺度表示并融合，能有效提取不同尺度的特征，生成视觉效果良好的图像，且与人类视觉系统一致，适用性强，但仍需人工选择和优化参数。表示学习方法通过稀疏表示和低秩表示等技术，将图像转化为新的表示形式，以便更好地捕捉主要特征并提升融合效果，然而其计算复杂度较高，对特定数据的依赖性较强。显著性方法则侧重于增强图像中视觉上更为重要的区域，能有效提升目标区域的视觉质量，但可能忽视某些不显眼但同样重要的细节信息。

3.2 特征级融合检测

特征级融合检测通过提取每个模态图像的高维特征再进行融合，得到最终包含可见光纹理细节与红外热信息的融合图像，以提高目标检测的准确度。

3.2.1 基于卷积神经网络的图像融合方法

卷积神经网络(Convolutional Neural Network，CNN)是目前人工智能领域在图像识别与处理相关应用中的关键技术之一[8]。其具有强大的特征提取和表达能力，经过训练的模型可以自适应地获取不同图像的特征，因此可以使用 CNN 来提取可见光与红外图像的特征信息，实现图像融合。具体框架如图 2 所示。

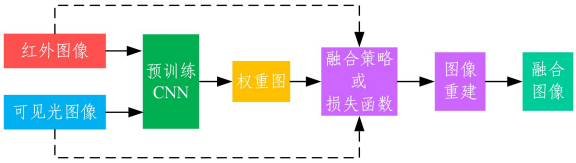

图 2 基于预训练 CNN 的红外和可见光图像融合算法框架

Fig. 2 Framework of infrared and visible light image fusion algorithm based on pre-trained CNN

近年来，相关研究[9,42]尝试用 CNN 构造各种独特的网络结构，或者采用特殊设计的损失函数约束，以针对复杂场景下的数据集开发出适应性更强的深度融合方法。一方面，卷积网络在自适应性上的提升有利于缓解对大量数据集的训练需求；另一方面，在开发新网络的过程中越来越多的数据集也被设计和发布出来，有利于支持图像融合领域的网络训练和综合评估。

Ma 等[42]使用显著目标掩膜及其取反的值分别标注红外图像的显著区域和可见光图像的背景区域，将其送入一个特殊设计的损失函数，指导网络学习并生成更加符合期望的融合图像，实现端到端的图像融合。

Tang 等[9]提出了一种基于光照感知的渐进式图像融合网络 PIAFusion。他们设计了一个光照感知子网络来评估图像的光照条件，并采用一个包含跨模态差分感知融合(Cross-Modality Differential Aware Fusion，CMDAF)模块的渐进式特征提取器，充分提取和融合多模态图像中的互补信息，随后融合可见光图像和红外图像的互补特征和共同特征，最后使用图像重建器将融合后的特征变换成融合图像。

3.2.2 基于自编码器的图像融合方法

基于自编码器(Autoencoder，AE)的图像融合方法首先在大规模自然图像数据集上训练一个自编码器，随后利用编码器进行图像的特征提取，再对以上提取的特征进行融合，最后用解码器对融合特征进行解码，重建融合图像。其框架如图 3 所示。

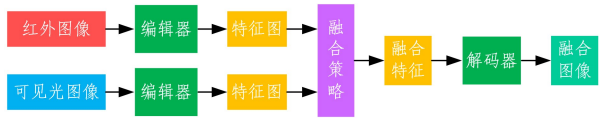

图 3 基于自编码器的多模态图像融合框架

Fig. 3 Infrared and visible light image fusion framework based on autoencoder

Li 等[43]提出的 DenseFuse 由编码器、融合层和解码器 3 部分构成。首先利用编码器提取样本特征；然后通过解码器得到融合图像；最后，通过融合策略对融合后的图像进行重构。

Li 等[44]随后又提出了 NestFuse，从多尺度的角度保存大量的输入数据信息。首先，将样本输入编码器，提取多尺度的图像特征；然后融合每个尺度的特征；最后通过嵌套链接的解码器对融合图像进行重构。虽然自编码器能够充分提取图像特征，但在融合策略的选择上往往不够灵活，如 DenseFuse 只采用了两种固定的融合策略，存在局限性。Li 等[10]设计了一种基于残差结构的残差融合网络(Residual Fusion Network，RFN)实现可学习的融合策略。

3.2.3 基于生成对抗网络的图像融合方法

作为一种生成模型，生成对抗网络(Generate Adversarial Network，GAN)[45]及其衍生模型常用于为深度学习中的数据增强、数据预处理方法生成样本，并在图像处理、生物医学、网络与信息安全等领域有着广泛的应用场景[46]。

Ma 等[47]提出了一种使用生成对抗网络的图像融合方法 FusionGAN。其中，生成器的功能为生成具有红外强度和可见光梯度的融合图像，判别器使融合图像中保留更多的可见光图像纹理细节。

针对 FusionGAN 在训练时无法充分平衡红外和可见光图像的模态权重的问题，Ma 等[48]提出了双判别器条件生成对抗网络(DDcGAN)，能够融合不同分辨率的红外和可见光图像。此外，Ma 等[49]在 DDcGAN 双判别器的基础上又提出了一种新的具有多分类约束的生成对抗网络(GAN-Mcc)，使用多分类器作为判别器，同时估计可见光和红外图像的分布。生成器在多分类约束下提升融合图像被判定为既是红外图像又是可见光图像的概率，判别器降低融合图像与两种图像相似的概率，从而提高了融合图像的对比度和纹理细节。

3.3 决策级融合检测

目前很少有融合架构通过探测器集成来探索单模态探测的晚期(决策级)融合，大多数研究只是简单地采用启发式(加权)平均置信度得分[11,50-51]。

Guan 等[50]设计了一种新的光照感知加权机制，使两个卷积网络分别学习白天和夜间的多光谱特征，最终检测结果采用启发式加权平均置信度得分。Li 等[11]设计了一个统一网络，通过语义分割与行人检测进行联合优化，对不同模态的

输出以及两个阶段的输出进行积分，得到最终的检测结果。Zhang 等[51]设计了一种新的多模态融合方法，将提取的特征进行重加权操作来选择更有效的特征并抑制无用的特征。

Chen 等[52]针对平均分[53-54]或最大投票[55]等后期融合方法在处理缺失模态时效果较差的问题，首次提出检测器概率集成方法 ProbEn 作为多模态检测的融合方法。该方法中，给定真实标签的单模态信号彼此条件独立，而最优融合策略由贝叶斯规则给出[56]。

4 多模态目标检测的对抗攻击方法

现有针对多模态目标检测的对抗攻击方法[23,57]大多仍是分开设计、作用的，即分别对不同模态的检测器进行对抗攻击。因此，面向单模态目标检测的对抗攻击方法有一定的参考意义。本章将分别介绍面向可见光目标检测的对抗攻击方法、面向红外目标检测的对抗攻击方法与面向多模态目标检测的对抗攻击方法，具体情况如表 1 所列。

表 1 面向多模态目标检测的对抗攻击方法分类

Table 1 Classification of adversarial attack methods for multimodal object detection

类别	方法	攻击目标	扰动类型	数据集	目标模型	评价指标	物理域可复现	攻击形式	物理复现材料
面向可见光模态目标检测的对抗攻击方法	DAG	人、自行车等	全局扰动	Pascal VOC	Faster R-CNN,R-FCN	ASR	×	白盒	—
	ShapeShifter	人、自行车等	全局扰动	MS COCO	Faster R-CNN	ASR	×	白盒	—
	RAP	人、自行车等	全局扰动	MS COCO	Faster R-CNN	mAP	×	黑盒	—
	CAP	人、自行车等	全局扰动	Pascal VOC,MS COCO	Faster R-CNN	mAP	×	白盒	—
	Category-wise	人、自行车等	全局扰动	Pascal VOC,MS COCO	Faster R-CNN,SSD 等	ASR,ATR,mAP	×	白盒	—
	Daedalus	人、自行车等	全局扰动	COCO	YOLO v3,SSD 等	FP,mAP	×	黑盒	—
	UEA	人、自行车等	全局扰动	Pascal VOC	Faster R-CNN,SSD	mAP	×	白盒	—
	TOG	人、自行车等	全局扰动	Pascal VOC,MS COCO 等	Faster R-CNN,SSD 等	mAP	×	白盒	—
	G-UAP	人、自行车等	全局扰动	Pascal VOC,KITTI	Faster R-CNN,R-FCN 等	mAP,PSNR	×	黑盒	—
	Thys 等	人	局部扰动	Inria	YOLO v2	AP	√	白盒	贴纸
	Adversarial t-shirt	人	局部扰动	video frames,COCO	YOLO v2,Faster-RCN	ASR	√	白盒	纺织物
	UPC	人	局部扰动	Pascal VOC	Faster R-CNN	precision	√	白盒	纺织物、贴纸
	Sharma 等	人	全局扰动	Glioma,M_tumour 等	YOLO v5	mAP	×	白盒	—
	Li 等	车	局部扰动	COCO	Faster R-CNN	mAP	×	白盒	—
	AdvART	人	局部扰动	INRIA,MPII	YOLO v4	mAP	√	白盒	纺织物
	Cui 等	车	局部扰动	VisDrone,UAV	YOLO v3,Faster R-CNN	ASR	√	白盒	投影仪
面向红外模态目标检测的对抗攻击方法	Infrared Invisible Clothing	人	局部扰动	FLIR	YOLO v3	AP	√	白盒	气凝胶
	adversarial infrared patches	人	局部扰动	FLIR	YOLO v3	ASR,AP	√	白盒	气凝胶
	HOTCOLD Block	人	局部扰动	FLIR	YOLO v5	ASR,AP	√	黑盒	加热膏、冷却膏
	AdvIB	人	局部扰动	FLIR	YOLO v3	ASR,AP	√	黑盒	加热膏、冷却膏
	Wei 等	人	局部扰动	FLIR	YOLO v3	ASR,AP	√	黑盒	气凝胶
	Zhu 等	车	局部扰动	FLIR	Faster R-CNN	ASR	√	黑盒	铝纸
	Wang 等	人	局部扰动	FLIR	YOLO v3,Faster R-CNN	ASR,AP	√	黑盒	气凝胶
	AdviC	人	局部扰动	FLIR	YOLOv3	ASR	√	黑盒	冷却膏
面向多模态目标检测的对抗攻击方法	Yu 等	人、车等	全局扰动	多光谱语义分割数据集[77]	MFNet	mAcc,mIoU	×	白盒	—
	MAP	人	局部扰动	自建数据集[21]	UMPDF[107]	AP	√	白盒	铝、钢、砂纸
	UAP	人	局部扰动	LLVIP	YOLO v3,Faster R-CNN	ASR,AP	√	黑盒	气凝胶
	Kim 等	人	局部扰动	FLIR	UMPDF[107]	AP	√	白盒	Low-e 薄膜
	Wang 等	人	局部扰动	MSRS,RoadScene,VIFB	YOLO v8	mAP	×	白盒	—
	Tarchoun 等	人	局部扰动	Wildtrack,MVDet	YOLO 系列	mAP	×	白盒	—
	Wei 等	人	局部扰动	LVIIP,KAIST 等	YOLO 系列、Faster R-CNN 等	ASR,AP	√	黑盒	气凝胶

4.1 面向可见光目标检测的对抗攻击方法

面向可见光目标检测的对抗攻击方法根据攻击的目标检测模型类别可以分为针对两阶段目标检测模型的对抗攻击、针对一阶段目标检测模型的对抗攻击以及针对这两类目标检测模型的对抗攻击[53]。

1)针对两阶段目标检测模型的攻击。Xie 等[28]提出密集对抗样本生成(Dense Adversary Generation,DAG)方法，首次将对抗攻击应用到目标检测中。此外，Chen 等[58]设计了一种 ShapeShifter 的对抗攻击方法，该方法基于 C&W[59]和 EOT[60]，首次实现针对 Faster R-CNN 模型的有目标对抗攻击。Li 等[15]通过降低 RPN 网络目标建议框的置信度，提出了一种鲁棒对抗性干扰(Robust Adversarial Perturbation，RAP)方法，使得目标检测模型将样本中的目标分类为背景，达到对抗效果。Zhang 等[16]利用了上下文区域概念，在生成对抗样本时提取目标的强特征，提出上下文对抗扰动(Contextual Adversarial Attack，CAP)方法，提升了对抗攻击效果。

2)针对一阶段目标检测模型的攻击。Liao 等[17]利用样本中重要区域的语义信息对模型进行对抗攻击，提出了 CA(Category-wise Attack)算法。Wang 等[61]发现 YOLO 模型中的非极大抑制(Non-Maximum Suppression，NMS)操作存在漏洞，设计了 Daedalus 方法，通过破坏其筛选机制，让候选框变小，实现对抗攻击。

3)针对两类目标检测模型的攻击。Wei 等[62]使用生成

对抗网络生成对抗样本，并添加了多尺度注意力特征损失，在加快生成效率的同时，提高了对抗样本的迁移性。Chow等[63]从目标检测多任务的角度进行算法设计，同时攻击两阶段目标检测模型和一阶段目标检测模型。Wu等[64]提出一种黑盒对抗攻击方法G-UAP，通过诱导RPN网络将目标物体错误识别为背景。Sharma等[65]提出了一种通过将对抗扰动嵌入图像，采用图像内图像隐写术的技术，使扰动对人眼几乎不可见，但却足以显著降低置信度。Li等[66]提出了一种方法，旨在通过引入可控制的对抗扰动来克服这些局限性，使其可以在不同的目标检测系统中应用。

然而针对数字域的对抗攻击大多都强调扰动的不可察觉性，通过将微小的扰动添加到图像全局范围上来影响模型的预测结果，这种扰动一般无法进行物理复现，不具有太大的实际应用价值。物理对抗攻击在目标物体的局部区域内添加一个对抗补丁[12](Patch)，从而对目标检测模型实现对抗攻击。

Thys等[67]针对YOLO模型提出一种物理世界的对抗补丁生成方法Adversarial-yolo，如图4所示，使目标检测模型无法检测到行人，达到“隐身”效果。Xu等[13]设计了一种对抗T恤以实现物理对抗攻击，并设计一种TPS技术模拟正常生活中T恤的形变情况，保持对抗样本的攻击性能。Huang等[68]提出UPC(Universal Physical Cam-ouflage)方法，实现有目标的通用物理对抗攻击。Guesmi等[69]设计了一种新的框架AdvART，它通过引入语义约束来优化生成的伪装图案，使其不仅自然且具有艺术性，同时也能够最大化对抗攻击的效果，解决了因受限于潜在空间的范围而难以生成既自然又具有欺骗性的对抗扰动的问题。Cui等[70]提出了一种新的对抗攻击生成方法，通过投影转换和自适应的对抗补丁生成，解决了传统方法在生成补丁时存在的形态学问题。

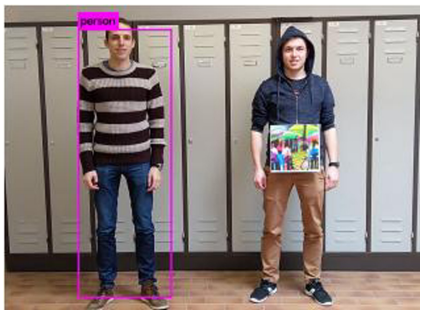

图4 物理世界实现的对抗补丁[67]

Fig. 4 Adversarial patches implemented in the physical world[67]

4.2 面向红外目标检测的对抗攻击方法

尽管目前红外行人检测系统被广泛使用，但是很少有人去关注该系统的安全隐患。受到将布料制成衣服过程的启发，Zhu等[14]首先在数字域设计了一个类二维码的对抗扰动，如图5所示，他们在扰动优化环节引入Gumble-softmax技术和EOT变换，实现梯度优化离散值的同时，提高了物理对抗扰动的鲁棒性。在数字域完成对抗纹理的生成之后，在物理域将对抗纹理进行平铺，制成带有对抗纹理的衣服，并在不同距离范围和角度范围对生成的物理对抗纹理进行评估。在类二维码的对抗纹理的基础上，Wei等[18]利用梯度方法优化对抗扰动的形状和位置，同时在优化过程中加入聚类的正则项，确保具有相似像素值的像素点聚在一起，并且具有对抗性，便于后续物理域对抗纹理在具有对抗性的同时易于迁移。类二维码对抗扰动过于复杂的纹理给物理域迁移和迁移之后对抗纹理的逼真性带来了巨大的挑战。为了实现简单的对抗纹理，Wei等[19]提出将数字域对抗扰动生成定义为一个黑盒优化问题，通过PSO算法优化对抗扰动的纹理、尺寸和位置。受Hotcold Block的启发，Hu等[20]提出了一种新的黑盒物理攻击方法，名为对抗红外块(AdvIB)，用于针对红外行人检测器。通过优化红外块的物理参数(如位置、角度、长度和颜色等)，设计了一种新的物理攻击方法，可以对红外行人检测器进行多视角攻击，并确保在不同的距离和视角下都能有效执行攻击。

图5 物理红外攻击演示[14]

Fig. 5 Demonstration of physical infrared attack[14]

与可见光对抗攻击相同，数字域的对抗扰动在实际应用中难以完全模拟物理世界中的情况，因为在物理环境中，图像中的扰动往往难以准确捕捉且具有显著的物理限制。为了真正评估和提升红外物体检测系统的安全性，红外对抗攻击必须从数字域成功迁移到物理域。近年来，物理域红外对抗攻击方法不断被提出。

Wei等[71]提出了一种简单、可实施的物理攻击方法，通过将热绝缘材料附加到目标物体上，来操控其热辐射，从而制造对抗扰动。Zhu等[72]提出了一种基于3D建模的物理攻击方法，旨在使红外检测器无法检测到现实世界中的汽车。对抗补丁采用了3D控制点平滑算法，能够有效地增强对抗补丁的效果并方便实施。Wang等[73]通过设计红外检测器的敏感区域(Sensitive Region，SR)来进行攻击。具体而言，通过在物体表面附加热绝缘材料，调整其形状和位置，改变物体的热辐射分布，从而使得目标在红外检测器中的可见性减少或消失。Hu等[74]提出了一种名为AdviC(Adversarial Infrared Curves)的新型物理对抗攻击方法，使用贝塞尔曲线生成红外对抗补丁，在物理环境中制造热辐射扰动。

4.3 面向多模态目标检测的对抗攻击方法

目前针对多模态目标检测的对抗性攻击的研究很少，且大部分集中于激光雷达点云图像与可见光图像的多模态场景。

Yu等[29]为了填补多模态数据融合和对抗鲁棒性之间的研究空白，首次使用FGSM和PGD等经典对抗性攻击方法对多模态目标检测模型开展攻击，分析并强调了多模态数据融合的对抗鲁棒性，探索了多模态深度学习模型中可能存在

的对抗性漏洞。

Wang 等[75]进而也提出了一种添加全局扰动的对抗扰动生成方法，旨在通过注入微小的不可察觉扰动来干扰可见光和红外图像融合模型的表现，进而影响后续的目标检测等高层任务。

Tarchoun 等[76]提出了两种新的对抗攻击方法。第一种方法通过投影梯度攻击，针对每个视角分别计算局部梯度，并通过将这些梯度汇总来生成对抗补丁。通过这种方式，攻击能够有效影响多视角检测系统，但这种方法无法适应包括变换器的多视角系统，因此提出了第二种方法。针对含有变换器的多视角检测系统，通过重新设计攻击补丁生成流程，在标准的多视角系统框架内，成功地使补丁能够适应不同视角下的特征。

Kim 等[21]观察到现有的多模态鲁棒性研究只分析了数字空间中的对抗性扰动，而不是物理世界中存在的脆弱性。因此他们生成了一个多光谱对抗贴片（MAP），可以通过欺骗物理世界中的多光谱探测器来评估多模态人体检测模型在物理世界中的鲁棒性，在其早期工作中提出的可见光与红外多模态目标检测模型[77-79]上实现了有效攻击。在提出的生成框架中，他们利用材料的发射率作为替代解决方案来表示热模式。根据斯蒂芬-玻尔兹曼定律，热像的强度不仅与温度有关，还与材料的发射率有关。因此，他们预先定义了具有不同发射率的材料，使其适合制作各种红外和可见光图案。利用预定义的材料发射率，又提出了一种交叉光谱映射（Cross-Spectral Mapping，CSM）方法，以便在生成贴片时考虑贴片在两个传感器上的外观，其物理实验效果图如图 6 所示。

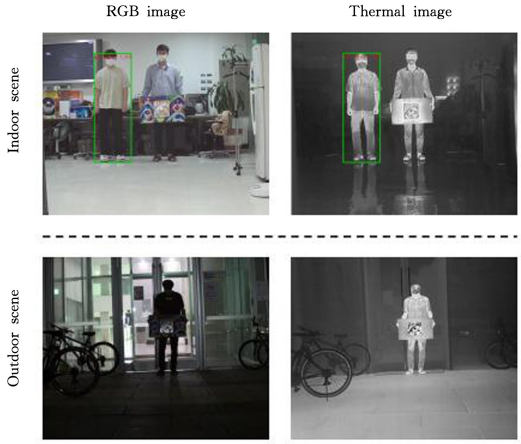

图 6 MAP 多模对抗攻击的物理域检测结果[21]

Fig. 6 Physical domain detection results of MAP multimodal adversarial attack[21]

Wei 等[22]为了解决现有的物理攻击只能针对单模态场景的问题，设计了一种统一的物理世界跨模态攻击，即通过单个补丁同时欺骗可见光和红外目标探测器。他们设计的这种新的边界限制形状优化方法，通过实现紧凑和光滑的形状，从而使其易于在物理世界中实现。此外，由于形状灵活，因此可以提供巨大的搜索空间，以找到最优形状，从而实现成功的跨模态攻击。此外，为了在优化过程中平衡可见光探测器和红外探测器之间的愚弄程度，还提出了一种分数感知的迭代评估方法。该方法可以引导对抗补丁迭代降低多模态传感器的预测分数。当应用于物理工具时，只需要在数字世界中打印模拟结果，并用绝缘材料进行补丁裁剪即可。

Kim 等[23]使用多光谱隐形涂层来研究多光谱探测器在物理攻击中的脆弱性，将透明 Low-e 薄膜附着在可见光攻击的 patch 上，实现针对多模态目标检测模型的物理对抗攻击。此外，还应用该物理方法制造了一种多光谱隐形服，可以在多光谱探测器的多个视角下隐藏人，具体效果如图 7 所示。

图 7 层叠可见光-红外多模态物理攻击[23]

Fig. 7 Laminated visible-infrared multimodal physical attacks[23]

Wei 等[80]设计了一种统一的对抗补丁生成框架。该方法的核心创新在于，使用一种边界限制的形状优化技术来生成紧凑且适应性强的对抗补丁，既能在物理世界中实施，又能在数字环境中达到良好的攻击效果。利用红外和可见光图像的互补性，提出了一种基于仿射变换的增强策略，优化了跨模态对抗样本的生成，如图 8 所示。

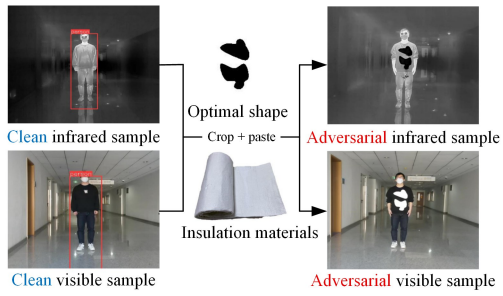

图 8 多模态基于形状优化技术的物理攻击检测结果[80]

Fig. 8 Multimodal physical attack detection results based on shape optimization techniques[80]

上述多模态目标检测的对抗攻击方法可以从攻击类型、攻击目标以及攻击方法这几个维度考虑。

1）攻击类型可以根据攻击的方式和产生的效果分类，可分为对抗性扰动和对抗性贴片，其中对抗性贴片包含在局部扰动中。对抗性扰动这种攻击方式通过对输入数据施加微小的扰动，使得目标检测模型产生错误的预测。扰动通常是不可见的，但却能显著影响模型的输出，例如 Yu 等[29]通过全局扰动进行攻击。对抗性贴片这种攻击方式通过向图像中添加或修改特定区域的图像，生成对抗样本来干扰多个模态的输入，从而导致模型在多模态数据融合中的判断错误，例如 MAP[21]通过添加特定的对抗补丁进行攻击。

2）攻击目标是对抗攻击所针对的特定部分，这是多模态攻击方法所特有的，可以分为多模态数据攻击、融合层攻击。针对多个模态的联合攻击通过干扰不同模态的数据，导致模型的多模态融合过程失败。融合层攻击通过攻击多模态目标检测模型中的融合层，使得不同模态的特征融合失效，从而导

致整体性能下降。

3)攻击方法可分为数字域攻击和物理域攻击,而数字域攻击又可以分为白盒攻击与黑盒攻击。数字域的白盒攻击,攻击者可以完全访问目标模型的结构和参数,可以直接在模型内部进行攻击。这类攻击通常能够生成更强的对抗扰动,因为攻击者可以利用模型的完整信息。数字域的黑盒攻击,攻击者不访问模型的内部结构或参数,仅通过输入和输出进行攻击,例如使用优化算法生成对抗样本,通过模拟输入输出的关系来推测模型的敏感性。物理域攻击,攻击者通过改变物理世界中的环境来欺骗传感器或目标检测系统,例如在MAP[21]中制作带有对抗图案的攻击补丁,并用不同反射率的材料制作红外纹理,对检测器进行攻击。

5 多模态目标检测的对抗防御方法

现有针对目标检测的对抗防御研究集中于可见光领域,对于红外领域与多模态领域的研究较少,而同目标检测任务下各种模态对抗防御都有一定的参考意义,因此本章将一起介绍各领域目标检测的对抗防御方法,具体如表2所列。

表2 面向多模态目标检测的对抗防御方法分类

Table 2 Classification of adversarial defense methods for multimodal object detection

类别	方法	扰动类型	数据集	模型	评价指标	防御对象	防御类型
面向可见光模态目标检测的对抗防御方法	MTD	全局扰动	PASCAL VOC,MS-COCO	SSD	mAP	DAG,RAP	基于训练
	Chen 等	全局扰动	PASCAL VOC,MS-COCO	SSD	mAP	CWA,DAG	基于训练
	RobustDet	全局扰动	PASCAL VOC,MS-COCO	SSD	mAP	CWA,DAG	基于训练
	Chiang 等	全局扰动	MS-COCO	YOLOv3	AP,certified AP	DAG	基于训练
	ROSA	全局扰动	MSRA-B 数据集、HKU-IS 数据集等	RFCN,DSS 等	MAE,precision,recall	—	基于模型
	Zhou 等	局部扰动	Inria 数据集、PASCAL VOC	Darknet,Tiny YOLO	mAP	Thys 等[67]	基于模型
	APM	局部扰动	COCO 数据集	YOLOv3,RetinaNet	mAP	ROC	基于模型
	ADT	全局扰动	Cityscapes,Foggy Cityscapes,BDD100K	Faster R-CNN	mAP	—	基于训练
面向红外模态目标检测的对抗防御方法	POD	局部扰动	FLIR	YOLO v5	AP	AdvPatch,HCB	基于数据
	Yu 等	全局扰动	FLIR	Faster R-CNN	mIOU,Precision,Recall	PGD,FGSM,ZOO,CW	基于数据
	Chen 等	全局扰动	GTSRB,TT100K	Yolov4	mAP	—	基于模型
面向多模态目标检测的对抗防御方法	Lee 等	全局扰动	RGB-thermalMFNet	RTFNet	mAcc,mIoU	FGSM,PGD	基于训练
	MSCL	全局扰动	RGB-thermalMFNet	RTFNet	IoU,IoUR	FGSM,PGD	基于数据
	Li 等	全局扰动	MFNet,M3FD	A^2RNet	mAP,mIoU	PGD	基于模型
	Suttapak 等	全局扰动	GOT-10k,VOT2018 等	MUNet	precision,recall,acc	CSA,DFA	基于模型
	Yuan 等	全局扰动	DroneVehicle,KAIST Pedestrian	CAGTDet	mAP,IoU	—	基于模型
	Zhao 等	全局扰动	VEDAI,M3FD,LLVIP,FLIR	MFMGF-Net	Precision,Recall,mAP	—	基于数据
	DARD	全局扰动	SYSU-MM01,RegDB	Faster R-CNN	mAP	—	基于数据

早期关于对抗训练的研究主要集中于图像分类领域。Zhang 等[57]将对抗训练方法扩展到可见光场景的目标检测领域中,从多任务学习角度揭示了目标检测中对抗攻击的共同底层机制,开发了一种对抗训练方法以实现目标检测任务的对抗防御。Chen 等[81]提出了一种新的类感知鲁棒对抗训练方法,用于目标检测任务。该方法将总损失分解为类的损失,并根据类的对象数量对每个类的损失进行归一化,不仅可以平衡每个类的影响,而且可以有效、均匀地提高训练模型对所有目标类的对抗鲁棒性。

Dong 等[82]对以上对抗训练方法进行分析研究指出,干净样本和对抗样本之间特征不同,导致模型精度急剧下降,其鲁棒性是有限的。对此,他们提出了一个鲁棒检测模型(RobustDet),学习不同组的卷积核,并基于对抗图像鉴别器(Adversarial Image Discriminator,AID)自适应地为它们分配权重,此外还采用了一致性特征重构(CFR)策略,通过应用重构约束,使模型提取的特征能够重构为尽可能干净的图像,从而驱动模型对干净图像和对抗图像提取更鲁棒的特征。

Chiang 等[83]提出了第一个针对目标检测器的对抗性攻击的认证防御,他们将复杂的检测网络视为黑箱,提出基于中值平滑认证目标检测模型鲁棒性的方法,将目标检测认证鲁棒性问题构建为回归问题,利用预测框的 IoU 计算将鲁棒性认证从图像分类任务扩展到目标检测任务。

Li 等[84]为了提高现有密集标记方法的鲁棒性和保持效率,提出了鲁棒的显著性目标检测(Robust Salient Object Detection,ROSA)方法来抵御对抗攻击,通过引入一些新的通用噪声来破坏对抗性扰动,然后学习自适应地预测针对新引入噪声的显著性映射。

Zhou 等[85]提出了一种通用的防御方法来防御目标检测上的物理攻击。首先,他们设计了一种基于熵的提议分量来获得提议的对抗区域。其次,基于对抗补丁中包含的不光滑和噪声信息,提出了基于梯度的滤波分量,对所提出的区域进行进一步滤波,得到更精确的对抗区域。由于该方法可以看作对对抗性图像的预处理,因此对所有数据集与模型都通用。这种方法对高频对抗样本防御效果较好,并且在干净样本上对精度影响较小,但是无法抵御低频对抗攻击。

Chiang 等[86]考虑到对抗防御的实用性,提出对抗像素掩模(Adversarial Pixel Masking,APM)方法去除高频和低频噪声的影响,用于针对预训练对象检测器的物理攻击。APM 通过将数据预处理网络添加到给定的目标检测器上来改变对抗训练,并让预处理网络学习屏蔽输入图像中存在的对抗物理扰动的能力。

Wang 等[87]提出了一种教师-学生框架(Adversarial Defense Teacher,ADT),用于改进对象检测模型在面对对抗性攻击时的鲁棒性。通过引入微小的扰动来增强模型的泛化能力,从而提升其在复杂场景下的表现。通过在图像输入过程中对目标图像进行放大和缩小操作,确保模型能够从下采样

物体中提取更精细的特征，有效应对了在低可见度环境中小物体检测的挑战。

目前针对红外场景目标检测的对抗防御研究十分有限。Strack 等[88]首次研究了针对行人红外探测的对抗性补丁攻击的防御策略，设计了一种计算效率高且有效的防御方法，称为基于补丁的闭塞感知检测（Patch-based Occlusion-aware Detection，POD），该方法可以有效地用随机补丁增强训练样本并对其进行检测。同时，POD 具有很强的迁移性，对未知的对抗攻击都有较强的防御效果。此外，与典型的对抗训练不同，POD 不会降低模型的主任务性能，反而能提高其在干净数据集检测中的检测精度。

Yu 等[89]利用生成器和判别器共同作用来进行红外小目标的对抗防御，通过向干净样本中加入噪声来生成对抗样本。他们提出的防御方法能够提升训练样本的多维特征扭曲，从而提高模型的整体性能。

Zhang 等[90]为了应对来自实际环境中红外模态的对抗样本，提出了一种基于形状优化的对抗攻击方法。其设计了一个全新的框架，这一框架的核心思想是通过共同训练“检测”和“恢复”模块来有效消除对抗扰动，确保在多个不同条件下，红外检测器仍然能够保持高精度的目标检测性能。

在可见光与红外多模态领域，Lee 等[91]提出了一种不容易受到对抗性攻击的鲁棒多传感器数据融合方法，通过保留随机化的融合特征，抵御对抗性攻击，首次对多传感器数据融合模型的对抗防御开展研究。此外，Yu 等[92]为了实现针对对抗性攻击的鲁棒多传感器数据融合网络，提出了一种新的鲁棒训练方案，称为多传感器累积学习（Multi-Sensor Cumulative Learning，MSCL）。MSCL 允许多传感器融合网络从单个传感器学习鲁棒特征，然后从多个传感器学习复杂的联合特征，其分步框架使网络能够将预训练的鲁棒性知识与来自多个传感器的新联合信息相结合。

Li 等[93]提出了一种新的对抗攻击防御网络 A^2RNet，它通过在训练过程中引入反攻击损失函数，帮助模型应对对抗攻击。A^2RNet 利用 U-Net 架构，该架构在图像上采样和下采样的过程中，能够有效抵抗攻击带来的扰动。此外，还提出了一种基于变换器的防御模块，称为防御性重细化模块（Defensive Refinement Module，DRM），旨在加强特征学习和减小噪声扰动。

Suttapak 等[94]提出了一种基于 MUNet 新型架构的模型，结合了像素级去噪和特征级防御技术，并且适用于目标追踪，是一种专门为抗击对抗性攻击而设计的目标跟踪防御模型。

Yuan 等[95]提出的防御方法主要包括两个模块：用于特征对齐的平移-尺度-旋转对齐（TSRA）模块和用于捕捉互补特征的互补融合变换器（CFT）。该方法通过使用级联对齐引导变换器（CAGT）来提升 RGB-IR 目标检测的效果。

Zhao 等[96]提出了 MFMG-Net，它是一种用于多光谱无人机地面检测的新架构。通过采用基于掩模的数据增强方法，研究团队旨在解决不同光谱数据之间可能存在的特征偏差问题，从而改善多模态特征的提取效果。通过增强特征提取能力，促进光谱网络之间的跨光谱信息交换。接着，利用注意力机制融合通道和空间维度上的特征，进一步提升了多光谱无人机地面检测的性能。

Wei 等[97]提出了一种名为双重对抗表示解耦（DARD）的模型，用于学习身份区分特征，同时有效地解耦模态特定信息。通过图像级别对抗学习和特征级别的对抗学习来消除不同模态图像之间的差异，进而提高模态共享特征的提取效果。

根据前文所述的方法内容，可以将多模态目标检测中的防御方法分为基于训练的防御方法、基于模型的防御方法以及基于数据增强的防御方法。以下为具体分析。

1）基于训练的防御方法通过改变训练过程或训练策略，增强多模态目标检测模型对对抗攻击的鲁棒性。Zhang 等[57]将此方法扩展到可见光场景的目标检测，显著提升了模型鲁棒性。该方法同样适用于多模态目标检测，通过引入不同模态的对抗样本，提升跨模态的鲁棒性。

2）基于模型的防御方法通过设计更加鲁棒的模型架构或增强现有模型的特性来抵抗对抗攻击。Dong 等[82]提出的鲁棒检测模型（RobustDet）通过对抗图像鉴别器（AID）自适应调整卷积核权重，提高了对抗攻击下的鲁棒性。其利用一致性特征重构（CFR）策略，确保提取特征接近干净样本，减少对抗样本的影响。

3）基于数据的防御方法主要通过对输入数据进行处理或修改，增强其对抗鲁棒性。这些方法通常包括数据预处理、对抗样本生成与增强等技术，以确保模型能够在不同的数据扰动下依然保持较高的鲁棒性。在多模态目标检测中，数据增强尤为重要，它能够帮助模型适应多模态输入中的变化和扰动。

6 数据集及评价指标

已有研究针对多模态目标检测性能及对抗攻击效果，采用相对固定的数据集与评价指标，方便不同方法之间的性能比较，因此本章总结了常用的数据集和评价指标。其中，数据集的基本情况如表 3 所列。

表 3 红外与可见光多模态目标检测常用数据集

Table 3 Datasets for infrared and visible multimodal object detection

名称	来源	类别	数量	大小	是否有类标
M3FD	Liu 等	已配准且打好标签的红外与可见光数据对，主要场景为校园和道路	4500 对	1024×768	有
RoadScene	Xu 等	车辆和行人	221 对	506×270 等	无
RGB-T234	Li 等	汽车、行人、风筝等	234000 对	630×460 等	有
VIFB	Zhang 等	行人、汽车等	21 对	630×460 等	无
TNO	Toet 等	已配准的军事目标等	261 对	590×426 等	无
KAIST	Hwang 等	已配准的红外与可见光数据对，主要目标为行人	95000 对	640×480	有
LLVIP	Jia 等	已配准的暗光条件下的行人目标	16836 对	1280×1024	有
FLIR	Teledyne FLIR 公司	未配准的可见光-红外数据，目标包括自行车、汽车、狗、人等	14000 对	1600×1800	有

6.1 数据集

1)M3FD[98]数据集是一个多场景多模态基准数据集,包含高分辨率的红外和可见光图像,涵盖各种场景下的不同对象类型。由特制的红外成像设备拍摄,能够自动获得配准的红外和可见光图像。M3FD 总共有 4 200 对图像(用于融合、检测和基于融合的检测),其中 300 对独立场景用于融合。图像格式分别为 24 位灰度位图的红外图像和 24 位彩色位图的可见光图像,图像大小为 1 024×768 像素。本实验使用了其中用于融合的 300 对独立场景图像。

2)RoadScene[99]数据集包含 221 对预配准的红外和可见光图像,涵盖道路、车辆和行人等丰富场景。由 FLIR 红外成像仪拍摄获得,具有较高的分辨率,并对原始红外图像中的背景热噪声进行了预处理,精确对齐可见光和红外图像对,再剪切出精确的配准区域以形成该数据集。

3)RGB-T234[100]数据集出自安徽大学李成龙课题组。数据集包括 234 个 RGB-Themal 视频序列对及其对应的目标检测框真值(Ground Truth)。视频序列标注中有 12 个目标属性,包含汽车、行人、风筝等。总的帧数为 234 000,最长的视频序列有 8 000 帧。

4)VIFB 是由 Zhang 等[101]提出的红外和可见光图像融合基准,其提供的测试数据集包含 21 对可见光和红外图像。这些图像对涵盖了室内、室外、低照明和过度曝光等场景,能够测试融合算法的泛化能力。

5)TNO[102]数据集是一组多光谱图像融合数据集,包含了强化视觉、近红外和长波红外 3 种波段的夜间图像,涵盖了不同的军事和监控场景。该数据集包含了 4 个子集,分别是 Athena,DHV,FEL 和 TRICLOBS,每个子集中都有多对可见光和红外图像,以及一些视频序列。

6)KAIST[103]数据集是由 Hwang 等采集设计,包含了多种场景下的行人图像,并进行了严格配准,图像大小为 640×480,总共包括 95 328 张图片、103 128 个密集注释。

7)LLVIP[104]数据集是由北京邮电大学实验团队提出,包含 16 836 对进行了严格配准的图像,其中大多数图像是在黑夜环境下拍摄的。

8)FLIR 数据集由 Teledyne FLIR 公司[105]发行,总共有 26 442 个完全注释的帧,包含 15 种不同的对象类。

6.2 评价指标

评价指标主要分为 3 个部分,分别是图像融合质量评估指标、目标检测性能评估指标与对抗攻击性能评估指标。

6.2.1 图像融合质量评估指标

1)信息熵(Entropy,EN):用于衡量图像的复杂性和信息量,高信息熵意味着图像包含更丰富的信息和细节。在红外和可见光融合图像评估中,通过测量信息熵,可以评估图像中细节的丰富程度。计算式如下:

$$EN=-\sum_{i=1}^{N}p(i)\log_2(p(i)) \tag{4}$$

2)空间频率(Spatial Frequency,SF):反映图像在空间域内的总体活跃程度,即图像灰度的变化率。主要用于衡量图像中的细节清晰度和边缘信息。空间频率越高,表示图像具有更多的纹理和边缘信息。计算式如下:

$$\begin{cases} SF=\sqrt{RF^2+CF^2} \\ RF=\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j+1)-F(i,j))^2} \\ CF=\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j+1)-F(i,j))^2} \end{cases} \tag{5}$$

3)互信息(Mutual Information,MI):用于度量两个随机变量之间的相关性。在融合图像评估中,互信息可用于评估红外和可见光图像之间的一致性和相关性。较高的互信息表示两个图像之间的信息重叠度高,融合效果较好。计算式如下:

$$MI=MI_{V,F}+MI_{I,F} \tag{6}$$

其中,$MI_{V,F}$和$MI_{I,F}$分别表示可见光图像和红外图像传递到融合图像的互信息,计算式如下:

$$MI_{X,F}=\sum_{x,f}p_{X,F}(x,f)\log\frac{p_{X,F}(x,f)}{p_X(x)p_F(f)} \tag{7}$$

其中,x 表示红外或者可见光图像,$p_{X,F}(x,f)$表示源图像 x 和融合图像 f 的联合直方图,$p_X(x)$和$p_F(f)$分别表示源图像和融合图像的边缘直方图。

4)结构相似性(Structural Similarity,SSIM):度量两幅图像相似度的指标,表示融合图像和源图像在亮度、对比度和结构上有多少共同特征。结构相似性越高,两幅图像越接近,计算式如下:

$$SSIM_{V,F}=\frac{2\mu_V\mu_F+C_1}{\mu_V^2+\mu_F^2+C_1}\cdot\frac{2\sigma_{VF}+C_2}{\sigma_V^2+\sigma_F^2+C_2}\cdot\frac{\sigma_{VF}+C_3}{\sigma_V\sigma_F+C_3} \tag{8}$$

其中,V 表示可见光图像;F 表示融合图像;μ_V 和μ_F 分别表示 V 和 F 的平均值;σ_V 和σ_F 分别表示 V 和 F 的标准差;σ_{VF} 表示 V 和 F 之间的协方差;C_1,C_2 和C_3 是常数,用于稳定计算,一般取较小的正值。SSIM 值为 0～1,越接近 1 表示两个图像越相似,即融合图像的视觉效果越偏向真实。

5)均方根误差(Root Mean Square Error,RMSE):表示融合图像和源图像之间的像素差异。RMSE 值越小,表示融合图像与原始图像之间的像素差异越小,两者越相似。计算式如下:

$$RMSE=\frac{RMSE_{V,F}+RMSE_{I,F}}{2} \tag{9}$$

其中,$RMSE_{V,F}$和$RMSE_{I,F}$分别表示可见光图像和红外图像对于融合图像的均方根误差,计算式如下:

$$RMSE_{X,F}=\sqrt{\frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}(X(m,n)-F(m,n))^2} \tag{10}$$

$X=V$ or I

6.2.2 目标检测性能评估指标

1)精确率(Precision):表示正确检测出的目标数量与总检测出的目标数量之比。

$$Precision=\frac{TP}{TP+FP} \tag{11}$$

其中,TP(True Positive)表示被分为正样本,并且分类正确;FP(False Positive)表示被分为正样本但是分类错误。

2)召回率(Recall):表示正确检测出的目标数量与实际存在的目标数量之比。

$$Recall=\frac{TP}{TP+FN} \tag{12}$$

其中,FN (False Negative)表示被分为负样本但是分类错误。

3）mAP（mean Average Precision）：对每个类别的精度-召回率曲线下的面积进行平均。通常在不同的 IoU（Intersection over Union）阈值下计算，然后取平均值。mAP 用于衡量训练出来的模型在感兴趣的类别上的检测能力的好坏。

6.2.3 对抗攻击性能评估指标

1）攻击成功率（ASR）：通过计算目标检测模型在对抗攻击前后的 mAP 值下降的程度来衡量攻击性能。ASR 越大表明攻击效果越好，目标检测模型出错越多。ASR 的定义如下：

$$ASR=\frac{mAP_{\text{clean}}-mAP_{\text{adv}}}{mAP_{\text{clean}}} \tag{13}$$

2）生成时间：利用对抗攻击算法在整个数据集上生成对抗样本的平均时间。

3）扰动程度：是指生成的对抗样本与原样本之间的相似度，一般使用 PSNR（Peak Signal to Noise Ratio）来进行评估：

$$PSNR=10\log_{10}\frac{(2^n-1)^2}{MSE} \tag{14}$$

MSE 的计算式如下：

$$MSE=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{M}(x(i,j)-\hat{x}(i,j))^2 \tag{15}$$

其中，$x(i,j)$为原样本的像素值，$\hat{x}(i,j)$为生成的对抗样本的像素值。

7 未来研究方向

现有针对多模态目标检测的研究集中于改善特征融合效果以提高检测性能，其安全性尚未得到充分探索，未来关于多模态目标检测的研究可从以下几个方面着手。

1）设计更好的特征融合策略，以应对极端成像情况。例如，可见光场景下亮度过高（低），红外场景下温差太小，图像非对齐等情况。现有特征级融合检测工作集中于改善可见光与红外多模态融合的性能[42-44]，对于一些极端成像情景仍未考虑。Tang 等[9]提出的 PIAFusion 考虑了可见光场景下光照不足导致纹理细节信息缺失的问题，但仍未解决图像未对齐以及红外场景下对比度过小导致的融合图像重影、红外信息丢失等问题。未来的研究可着眼于解决图像融合过程中图像未对齐、亮度太低以及对比度太小等问题，设计出更好的特征融合策略，提高特征级多模态融合目标检测的鲁棒性。尤其是在自动驾驶、安防监控等高风险场景下，极端成像条件频繁出现，鲁棒融合策略的研究显得尤为迫切。

2）设计更成熟的通用物理对抗攻击方法。现有针对可见光与红外多模态目标检测的对抗攻击方法只能攻击某个融合策略的多模态目标检测。MAP[21]方法通过交叉谱映射设计了针对特征级融合多模态目标检测的对抗攻击方法，而 UAP[22]方法通过差分进化算法优化扰动形状设计了针对决策级融合多模态目标检测的对抗攻击方法。尽管它们都取得了不错的效果，但是仍未考虑通用性。未来的研究可考虑同时实现针对像素级、特征级以及决策级融合多模态目标检测的通用物理对抗攻击方法。

3）多模态对抗攻击的语义性。在真实物理环境中，具有语义性的自然对抗噪声显得尤为重要，能够避免引起注意，增加隐蔽性。当前研究[21,23]生成的物理对抗样本形状怪异，较为艳丽，十分惹眼，隐蔽性不足，未来关于多模态物理对抗样本生成技术应更多地考虑语义性，生成自然、隐蔽的物理对抗样本。

4）实现变尺度的全向多模态物理对抗攻击。真实物理世界检测器的成像过程包含了成像尺度、角度的因素，因此，若要实现有意义的多模态物理对抗攻击，那么在生成多模态对抗样本时就要考虑尺度、俯仰角、旋转角等因素。现有研究[22-23]仅考虑了成像过程的旋转角的变换，对于尺度以及俯仰角等因素仍未探索。未来的研究可以考虑实现变尺度的全向多模态物理对抗攻击，增加实际应用价值。

5）提高多模态物理对抗攻击的可迁移性。现有关于多模态物理对抗攻击的研究还未考虑其物理对抗样本的可迁移性[21-23]，这意味着针对多个目标检测器实现对抗攻击时，要分别训练对抗样本，这会增加开销，在实际中是不可取的。因此，提高物理对抗攻击的可迁移性，实现多模态目标检测模型层面的通用攻击或许是未来的一大研究方向。

6）发掘更优的对抗样本物理赋形材料。由于可见光与红外之间的频谱差异，对于对抗样本物理赋形材料的需求也不相同，而层叠可见光扰动与红外扰动会对各自模态的对抗样本的攻击性能产生影响。尽管已有研究[23]考虑使用 Low-e 薄膜来减轻影响，但其仍会在一定程度上影响对抗样本的攻击性能。若未来要实现全向的多模态对抗攻击，可见光与红外对抗样本的层叠就必不可少。因此，发掘更优的物理赋形材料对于多模态物理对抗攻击的发展也是非常重要的。

7）物理世界中鲁棒的多模态目标检测方法。目前关于对抗防御的研究主要集中于可见光领域[81-83]，且大多防御手段采取对抗训练[57,88]，然而，对抗训练的实施需要为特定的模型和攻击策略生成对抗样本，并进行模型的重训练，这不仅需要大量的额外计算资源，还可能对目标检测模型的任务准确率产生负面影响。此外，目前针对多模态目标检测对抗攻击的防御方法[91-92]只在数字域实现，针对的是 FGSM，PGD 对抗攻击，无法对物理世界的对抗补丁实现有效防御。因此，如何在物理世界设计出更鲁棒、高效的针对可见光与红外多模态目标检测防御策略是一个亟待解决的问题。

结束语 多模态目标检测能融合不同模态图像互补特征，提高目标检测效能，在日常生活中的作用越来越大。本文基于现有的研究工作，围绕可见光与红外多模态目标检测的对抗安全性进行综述。首先对多模态目标检测及攻防进行概述；其次按照不同时段的融合检测对多模态目标检测方法进行分类归纳；然后对现有的可见光与红外多模态目标检测对抗攻击与对抗防御方法进行归纳整理，并梳理了常用的多模态目标检测数据集与主要评价指标；最后探讨了多模态目标检测未来潜在的研究方向，以期进一步推动多模态目标检测对抗安全研究发展和应用。

参 考 文 献

[1] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：Towards real-time object detection with region proposal networks [C]// Advances in Neural Information Processing Systems. 2015.

[2] MA Y, WU Z Y, JIANG X. Object Detection based on Feature Fusion of Infrared and Visible Images [J]. Missiles and Space Vehicles, 2022, 389(5): 83-87.

[3] KIEU M, BAGDANOV A D, BERTINI M, et al. Task-conditioned domain adaptation for pedestrian detection in thermal imagery [C]// European Conference on Computer Vision. Cham: Springer, 2020: 546-562.

[4] DEVAGUPTAPU C, AKOLEKAR N, SHARMA M, et al. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[5] ZHANG L, ZHU X, CHEN X, et al. Weakly aligned cross-modal learning for multispectral pedestrian detection [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5127-5137.

[6] PANCHOTIYA B, ISRANI D, PATEL R. An efficient image fusion of visible and infrared band images using integration of anisotropic diffusion and discrete wavelet transform [J]. Journal of Communications Software and Systems, 2020, 16(1): 30-36.

[7] CAO Y Q, YANG S C. Image fusion method based on convolutional sparse representation [J]. Navigation and Control, 2020, 19(2): 97-105.

[8] LIN D, PAN L, YI P. Research progress on robustness of convolutional neural networks for image recognition [J]. Chinese Journal of Network and Information Security, 2022, 8(3): 111-122.

[9] TANG L, YUAN J, ZHANG H, et al. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware [J]. Information Fusion, 2022, 83: 79-92.

[10] LI H, WU X, KITTLER J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images [J]. Information Fusion, 2021, 73: 72-86.

[11] LI C, SONG D, TONG R, et al. Multispectral pedestrian detection via simultaneous detection and segmentation [J]. arXiv: 1808.04818, 2018.

[12] LIU X, YANG H, LIU Z, et al. Dpatch: An adversarial patch attack on object detectors [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 2849-2858.

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

[14] ZHU X, HU Z, HUANG S, et al. Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13317-13326.

[15] LI Y, TIAN D, CHANG M C, et al. Robust adversarial perturbation on deep proposal-based models [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 8082-8091.

[16] ZHANG H, ZHOU W, LI H. Contextual adversarial attacks for object detection [C]// IEEE International Conference on Multimedia and Expo. 2020: 1-6.

[17] LIAO Q, WANG X, KONG B, et al. Category-wise attack: Transferable adversarial examples for anchor free object detection [C]// IEEE International Conference on Multimedia and Expo. 2021: 1-6.

[18] WEI X, YU J, HUANG Y. Physically adversarial infrared patches with learnable shapes and locations [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 12334-12342.

[19] WEI H, WANG Z, JIA X, et al. Hotcold block: Fooling thermal infrared detectors with a novel wearable design [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023: 15233-15241.

[20] HU C, SHI W, JIANG T, et al. Adversarial infrared blocks: A multi-view black-box attack to thermal infrared detectors in physical world [J]. Neural Networks, 2024, 175: 106310.

[21] KIM T, LEE H J, RO Y M. MAP: Multispectral adversarial patch to attack person detection [C]// ICASSP 2022 — 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 4853-4857.

[22] WEI X, HUANG Y, SUN Y, et al. Unified adversarial patch for cross-modal attacks in the physical world [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4445-4454.

[23] KIM T, YU Y, RO Y M. Multispectral invisible coating: Laminated visible-thermal physical attack against multispectral object detectors using transparent Low-e films [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023: 1151-1159.

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

[25] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [J]. arXiv: 1412.6572, 2014.

[26] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world [J]. arXiv: 1607.02533, 2016.

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

[28] XIE C, WANG J, ZHANG Z, et al. Adversarial examples for semantic segmentation and object detection [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 1369-1378.

[29] YU Y, LEE H J, KIM B C, et al. Investigating vulnerability to adversarial examples on multimodal data fusion in deep learning [J]. arXiv: 2005.10987, 2020.

[30] SHEN Y, XIANG K, CHEN X, et al. A noisy infrared and visible light image fusion algorithm [J]. The Journal of Information Processing Systems, 2021, 17(5): 1004-1019.

[31] HU D, SHI H. Infrared and visible image fusion based on empirical curvelet transform and phase congruency [J]. Ukrainian Journal of Physics, 2021, 22: 128-137.

[32] XING X, LIU C, LUO C, et al. Infrared and visible image fusion based on nonlinear enhancement and NSST decomposition [J]. EURASIP Journal on Wireless Communications and Networking, 2020, 2020(1): 162.

[33] ZHU Y,LU Y,GAO Q,et al. Infrared and visible image fusion based on convolutional sparse representation and guided filtering [J]. Journal of Electronic Imaging,2021,30(4):043003.

[34] LIU F,CHEN L,LU L,et al. Infrared and visible image fusion via rolling guidance filter and convolutional sparse representation [J]. Journal of Intelligent and Fuzzy Systems,2021,40(6):10603-10616.

[35] LIU G,YAN S. Latent low-rank representation for subspace segmentation and feature extraction [C]// IEEE International Conference on Computer Vision. Barcelona:IEEE Computer Society,2011:1615-1622.

[36] LI H,WU X,KITTLER J. MDLatLRR:A novel decomposition method for infrared and visible image fusion [J]. IEEE Transactions on Image Processing,2020,29:4733-4746.

[37] SUN B,ZHUGE W W,GAO Y X,et al. Infrared and Visible Image Fusion Based on Latent Low-Rank Representation [J]. Infrared Technology,2022,44(8):853-862.

[38] YANG Y,ZHANG Y,HUANG S,et al. Infrared and visible image fusion using visual saliency sparse representation and detail injection model [J]. IEEE Transactions on Instrumentation and Measurement,2021,70:1-15.

[39] MA J,MA Y,LI C. Infrared and visible image fusion methods and applications:a survey [J]. Information Fusion,2019,45:153-178.

[40] LI H. Research and Application of Image Fusion Algorithms Based on Representation Learning [D]. Wuxi:Jiangnan University,2021.

[41] BAVIRISETTI D,DHULI R. Two-scale image fusion of visible and infrared images using saliency detection [J]. Infrared Physics & Technology,2016,76:52-64.

[42] MA J,TANG L,XU M,et al. STDFusionNet:An infrared and visible image fusion network based on salient target detection [J]. IEEE Transactions on Instrumentation and Measurement,2021,70:1-13.

[43] LI H,WU X. DenseFuse:A fusion approach to infrared and visible images [J]. IEEE Transactions on Image Processing,2019,28(5):2614-2623.

[44] LI H,WU X,DURRANI T. NestFuse:An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models [J]. IEEE Transactions on Instrumentation and Measurement,2020,69(12):9645-9656.

[45] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial networks [J]. Communications of the ACM,2020,63(11):139-144.

[46] WANG Z L,ZHANG B W. Review of Generative Adversarial Networks [J]. Chinese Journal of Network and Information Security,2021,7(4):68-85.

[47] MA J,YU W,LIANG P,et al. FusionGAN:A generative adversarial network for infrared and visible image fusion [J]. Information Fusion,2019,48:11-26.

[48] MA J,XU H,JIANG J,et al. DDcGAN:A dual-discriminator conditional generative adversarial network for multi-resolution image fusion [J]. IEEE Transactions on Image Processing,2020,29:4980-4995.

[49] MA J,ZHANG H,SHAO Z,et al. GANMcC:A generative adversarial network with multiclassification constraints for infrared and visible image fusion [J]. IEEE Transactions on Instrumentation and Measurement,2021,70:1-14.

[50] GUAN D,CAO Y,YANG J,et al. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection [J]. Information Fusion,2019,50:148-157.

[51] ZHANG L,ZHU X,CHEN X,et al. Weakly aligned cross-modal learning for multispectral pedestrian detection [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:5127-5137.

[52] CHEN Y T,SHI J,YE Z,et al. Multimodal object detection via probabilistic ensembling [C]// European Conference on Computer Vision. Cham:Springer,2022:139-158.

[53] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. Imagenet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems. 2012:1097-1105.

[54] DOLLAR P,WOJEK C,SCHIELE B,et al. Pedestrian detection:A benchmark [C]// CVPR. 2009.

[55] XU P,DAVOINE F,DENOEUX T. Evidential combination of pedestrian detectors [C]// British Machine Vision Conference. 2014:1-14.

[56] PEARL J. Probabilistic reasoning in intelligent systems:Networks of plausible inference[M]. Elsevier,2014.

[57] ZHANG H C,WANG J Y. Towards adversarially robust object detection [C]// Proceedings of IEEE/CVF International Conference on Computer Vision(ICCV). Piscataway,NJ:IEEE,2020:421-430.

[58] CHEN S T,CORNELIUS C,MARTIN J,et al. Shapeshifter:Robust physical adversarial attack on Faster R-CNN object detector [C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham:Springer,2019:52-68.

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

[60] ATHALYE A,ENGSTROM L,ILYAS A,et al. Synthesizing robust adversarial examples [C]// International Conference on Machine Learning. 2018:284-293.

[61] WANG D,LI C,WEN S,et al. Daedalus:Breaking nonmaximum suppression in object detection via adversarial examples [J]. IEEE Transactions on Cybernetics,2021,52(8):7427-7440.

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

[63] CHOW K H,LIU L,LOPER M,et al. Adversarial objectness gradient attacks in real-time object detection systems [C]// IEEE International Conference on Trust,Privacy and Security in Intelligent Systems and Applications. 2020:263-272.

[64] WU X,HUANG L,GAO C,et al. G-UAP:Generic universal adversarial perturbation that fools RPN-based detectors [C]// Asian Conference on Machine Learning. 2019:1204-1217.

[65] SHARMA G, GARG U. Unveiling vulnerabilities: evading YOLOv5 object detection through adversarial perturbations and steganography [J]. Multimedia Tools and Applications, 2024, 83: 74281-74300.

[66] LI G, XU Y, DING J, et al. Toward generic and controllable attacks against object detection [J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-12.

[67] THYS S, VAN RANST W, GOEDEME T. Fooling automated surveillance cameras: Adversarial patches to attack person detection [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[68] HUANG L, GAO C, ZHOU Y, et al. Universal physical camouflage attacks on object detectors [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 720-729.

[69] GUESMI A, BILASCO I M, SHAFIQUE M, et al. AdvART: Adversarial art for camouflaged object detection attacks [C] // 2024 IEEE International Conference on Image Processing (ICIP). IEEE, 2024: 666-672.

[70] CUI J, GUO W, HUANG H, et al. Adversarial examples for vehicle detection with projection transformation [J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-18.

[71] 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: 1928-1944.

[72] ZHU X, LIU Y, HU Z, et al. Physical adversarial attacks for infrared object detection [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024: 24284-24293.

[73] WANG X, LI W. Physical adversarial attacks for infrared object detection [C] // 2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE). IEEE, 2024: 64-69.

[74] HU C, SHI W, YAO W, et al. Adversarial infrared curves: An attack on infrared pedestrian detectors in the physical world [J]. Neural Networks, 2024, 178: 106459.

[75] WANG Y, LI X, YANG L, et al. Adaptive oriented adversarial attacks on visible and infrared image fusion models [C] // 2024 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2024: 1-6.

[76] TARCHOUN B, ALAM Q M, ABU-GHAZALEH N, et al. Fool the Hydra: Adversarial attacks against multi-view object detection systems [J]. arXiv: 2312. 00173, 2023.

[77] HA Q, WATANABE K, KARASAWA T, et al. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes [C] // 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017: 5108-5115.

[78] WOHLBERG B. Efficient algorithms for convolutional sparse representations [J]. IEEE Transactions on Image Processing, 2015, 25(1): 301-315.

[79] YUAN L, LI X M, PAN Z X, et al. A review of adversarial samples for object detection [J]. Chinese Journal of Image and Graphics, 2022, 27(10): 2873-2896.

[80] WEI X, HUANG Y, SUN Y, et al. Unified adversarial patch for visible-infrared cross-modal attacks in the physical world [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(4): 2348-2363.

[81] CHEN P C, KUNG B H, CHEN J C. Class-aware robust adversarial training for object detection [C] // Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2021: 10415-10424.

[82] DONG Z Y, WEI P X, LIN L. Adversarially-aware robust object detector [C] // European Conference on Computer Vision. Berlin: Springer, 2022: 297-313.

[83] CHIANG P Y, CURRY M J, ABDELKADER A, et al. Detection as regression: Certified object detection by Median smoothing [J]. arXiv: 2007. 03730, 2020.

[84] LI H F, LI G B, YU Y Z. ROSA: Robust salient object detection against adversarial attacks [J]. IEEE Transactions on Cybernetics, 2020, 50(11): 4835-4847.

[85] ZHOU G Z, GAO H C, CHEN P, et al. Information distribution based defense against physical attacks on object detection [C] // Proceedings of IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Piscataway, NJ: IEEE, 2020: 1-6.

[86] CHIANG P H, CHAN C S, WU S H. Adversarial pixel masking: A defense against physical attacks for pre-trained object detectors [C] // Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 1856-1865.

[87] WANG K, SHEN Y, LAUER M. Adversarial defense teacher for cross-domain object detection under poor visibility conditions [J]. arXiv: 2403. 15786, 2024.

[88] STRACK L, WASEDA F, NGUYEN H H, et al. Defending against physical adversarial patch attacks on infrared human detection [J]. arXiv: 2309. 15519, 2023.

[89] YU T, XUE Y, HE Y, et al. Adversarial defense technology for small infrared targets [J]. Computers, Materials & Continua, 2024, 81(1): 1235.

[90] ZHANG Y, ZHAO S, WEI X, et al. Defending adversarial patches via joint region localizing and inpainting [C] // Pattern Recognition and Computer Vision: 7th Chinese Conference, PRCV 2024. Springer, 2024: 236-250.

[91] LEE H J, RO Y M. Adversarially robust multi-sensor fusion model training via random feature fusion for semantic segmentation [C] // 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021: 339-343.

[92] YU Y, LEE H J, KIM B C, et al. Towards robust training of multi-sensor data fusion network against adversarial examples in semantic segmentation [C] // ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 4710-4714.

[93] LI J, YU H, CHEN J, et al. A^2RNet: Adversarial Attack Resilient Network for Robust Infrared and Visible Image Fusion [C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2025: 4770-4778.

[94] SUTTAPAK W, ZHANG J, ZHAO H, et al. Multi-Model U-Net: An Adversarial Defense Mechanism for Robust Visual Tracking [J]. Neural Process Letters, 2024, 56: 132.

[95] YUAN M,SHI X,WANG N,et al. Improving RGB-infrared object detection with cascade alignment-guided transformer [J]. Information Fusion,2024,105:102246.

[96] ZHAO F,LOU W,FENG H,et al. MFMG-Net:Multispectral Feature Mutual Guidance Network for Visible-Infrared Object Detection [J]. Drones,2024,8:112.

[97] WEI Z,YANG X,WANG N,et al. Dual-Adversarial Representation Disentanglement for Visible Infrared Person Re-Identification [J]. IEEE Transactions on Information Forensics and Security,2024,19:2186-2200.

[98] LIU J,FAN X,HUANG Z,et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:5802-5811.

[99] XU H,MA J,LE Z,et al. FusionDN:A unified densely connected network for image fusion [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:12484-12491.

[100]LI C,LIANG X,LU Y,et al. RGB-T object tracking:Benchmark and baseline [J]. Pattern Recognition,2019,96:106977.

[101]ZHANG X,YE P,XIAO G. VIFB:A visible and infrared image fusion benchmark [J]. arXiv:2002. 03322,2020.

[102]TOET A,HOGERVORST M A. Progress in color night vision [J]. Optical Engineering,2012,51(1):010901.

[103]HWANG S,PARK J,KIM N,et al. Multispectral pedestrian detection:Benchmark dataset and baseline [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:1037-1045.

[104]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.

[105]FLIR. Free FLIR thermal dataset for algorithm training [EB/OL]. https://www.flir.com/oem/adas/adas-dataset-form/.

ZHENG Haibin, born in 1995, Ph.D, lecturer. His main research interests include deep learning and artificial intelligence security.

CHEN Jinyin, born in 1982, Ph.D, professor. Her main research interests include artificial intelligence security, graph data mining and evolutionary computing.

(责任编辑:喻藜)