激光与光电子学进展 先进成像

多模态自适应特征融合的目标检测

高小强[1]，常侃[1,2*]，凌铭阳[1]，银梦雨[1]

[1]广西大学计算机与电子信息学院，广西 南宁 530004；

[2]广西多媒体通信与网络技术重点实验室，广西 南宁 530004

摘要 随着深度学习的发展，基于卷积神经网络(CNN)的目标检测方法取得巨大成功。现有的基于CNN的目标检测模型通常采用单一模态的RGB图像进行训练和测试，但在低光照环境下，检测性能显著下降。为解决此问题，提出了一种基于YOLOv5构建的多模态目标检测网络模型，将RGB图像和热红外图像相结合，以充分利用多模态特征融合信息，从而提升目标检测精度。为了实现多模态特征信息的有效融合，提出了一种多模态自适应特征融合(MAFF)模块。该模块通过自适应地选择不同模态特征并利用各模态间的互补信息，实现多模态特征融合。实验结果表明：所提算法能有效融合不同模态的特征信息，从而显著提高检测精度。

关键词 卷积神经网络；多模态；YOLOv5；多模态目标检测；自适应特征融合

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP230856

Object Detection via Multimodal Adaptive Feature Fusion

Gao Xiaoqiang[1], Chang Kan[1,2*], Ling Mingyang[1], Yin Mengyu[1]

[1]*School of Computer and Electronic Information, Guangxi University, Nanning* 530004, *Guangxi, China*;

[2]*Guangxi Key Laboratory of Multimedia Communications and Network Technology, Nanning* 530004, *Guangxi, China*

Abstract With the advancement of deep learning, object detection methods based on convolutional neural networks (CNNs) have achieved tremendous success. Existing CNN-based object detection models typically employ single-modal RGB images for training and testing; however, their detection performance is significantly degraded in low-light conditions. To address this issue, a multimodal object detection network model built on YOLOv5 is proposed, which integrates RGB and thermal infrared imagery to fully exploit the information provided by the fusion of multi-modal features, increasing the object detection accuracy. To achieve effective fusion of multimodal feature information, a multimodal adaptive feature fusion (MAFF) module is introduced. It facilitated multimodal feature fusion by adaptively selecting diverse modal features and exploiting the complementary information between modalities. The experimental results indicate the efficacy of the proposed algorithm for seamlessly merging features from distinct modalities, which significantly increases the detection accuracy.

Key words convolution neural network; multimodality; YOLOv5; multimodal object detection; adaptive feature fusion

1 引 言

近年来，基于卷积神经网络(CNN)的目标检测方法越来越流行，其中YOLO系列算法[1-4]因其较快的检测能力而广受欢迎。目前，基于CNN的目标检测主要采用RGB图像进行训练和检测[5-11]。在现实环境中，RGB成像容易受到光照和天气变化的影响。因此，在低能见度、低光照和夜间等不利条件下，基于RGB图像的目标检测方法效果不佳。

为解决上述问题，将RGB图像与热红外图像信息进行融合[12]。热红外图像能突出显著的目标[13-15]，而RGB图像具备比热红外图像更丰富的纹理和结构信息。因此，RGB图像和热红外图像具有天然互补的优势。根据此研究思路，出现一些结合RGB图像和热红外图像的多模态目标检测方法[16-20]。

在先前的工作探索中[21-28]，根据RGB图像和热红外图像信息融合阶段可以概括为3种融合方式：早期融合、中期融合、后期融合。文献[21-23]证明了中期

收稿日期：2023-03-13；**修回日期**：2023-04-09；**录用日期**：2023-04-12；**网络首发日期**：2023-04-22

基金项目：国家自然科学基金(62171145)

通信作者：*pandack0619@163.com

融合效果最好，然而这些方法都通过级联或按元素相加的简单方式进行特征融合，未充分利用两种模态特征的内在互补性。为了更好地进行不同模态特征的融合，Zhang 等[25]使用循环中期融合方式平衡两种模态特征之间的互补性和一致性，但需要依赖语义分割信息进行监督。Zhou 等[26]从一种模态增强另一种模态解决了特征模态不平衡问题，但依赖光照对齐模块实现自适应融合。Zhang 等[27]在模态间和模态内注意模块的指导下实现 RGB 图像和热红外图像特征的完全自适应融合，但该网络需要使用复杂注意力模块进行多模态的特征融合，且需要真值掩码进行训练。Fang 等[28]将 Transformer[29]应用于多模态目标检测中，利用 Transformer 的自我注意力机制实现不同模态特征的自适应融合，但 Transformer 的计算效率并不理想。上述方法部分解决不同模态特征的融合问题，但并未充分挖掘和利用不同模态特征之间的互补特性，检测准确度仍有提升空间。

为了高效、准确地进行不同模态特征融合，本文提出了一种多尺度特征融合的多光谱目标检测框架，并提出了一种新的多模态自适应特征融合（MAFF）模块。该模块对来自不同模态、不同语义的特征进行自适应加权，从而充分利用不同模态的互补信息，有效提高目标检测精度。将 MAFF 模块嵌入到 YOLOv5 网络中得到多个特征尺度，构建 MAFF 网络（MAFFNet），在不同的数据集上与其他方法相比，所提方法在保持较低复杂度的同时，获得最高的检测精度。

2　YOLOv5 算法介绍

作为单阶段目标检测的代表性方法，YOLO 系列模型取得极大的成功。近年来，在 YOLOv4[4]的基础上，Ultralytics 公司提出 YOLOv5 模型。YOLOv5 继承之前版本网络的优点，在检测速度和检测精度上都有所提升。因此，选择 YOLOv5 模型作为基础目标检测框架。为了适应不同的应用场景，YOLOv5 分为 YOLOv5s、YOLOv5m、YOLOv5l 和 YOLOv5x 等 4 个版本。所提算法选择的目标检测模型为 YOLOv5l，后续出现的 YOLOv5 均指 YOLOv5l。YOLOv5 的主体结构如图 1 所示，主要包括 3 个部分：主干（Backbone）、颈部（Neck）、头部（Head）。

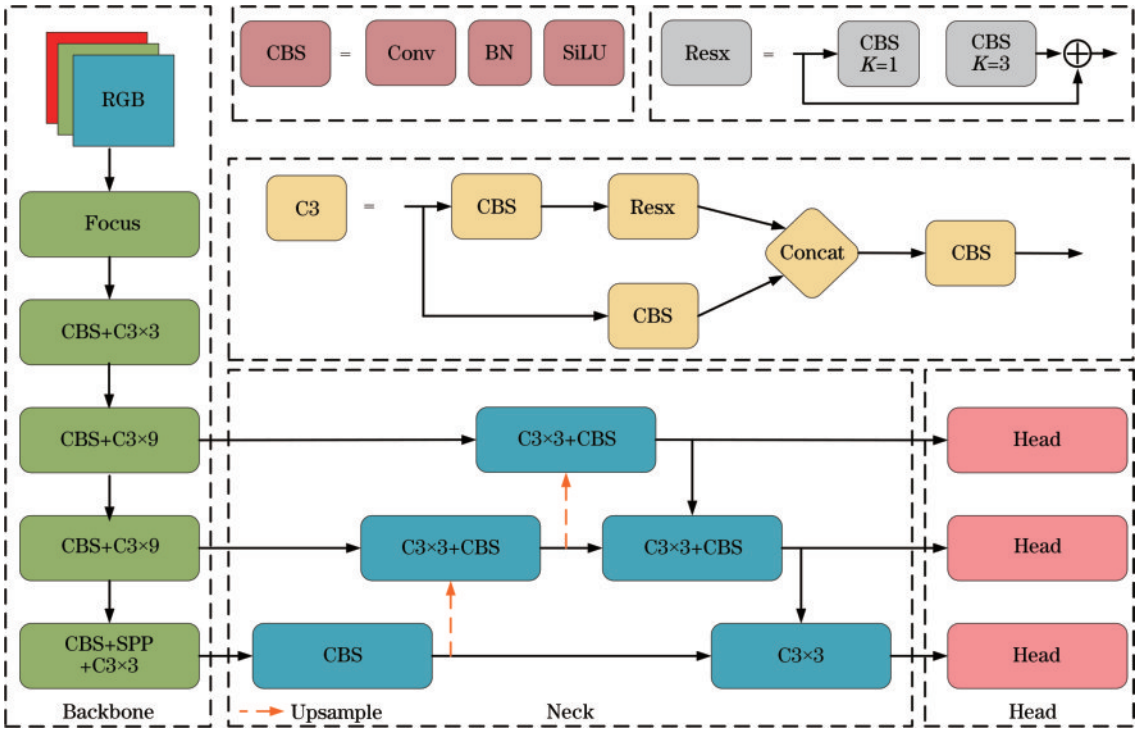

图 1　YOLOv5 结构

Fig. 1　Structure of YOLOv5

Backbone 包括 3 个模块：Focus、C3、空间金字塔池化（SPP）。Focus 模块的作用是对输入图片进行切片、拼接和卷积操作，在尽可能少丢失图片特征信息的情况下，提升特征提取的速度。C3 模块为改进的 BottleneckCPS 结构，优化了其网络结构和激活函数，具体结构如图 1 所示。SPP 模块由 3 个不同尺度的最大池化层和级联模块组成，实现融合局部和全局特征，有效提升网络的感受野。

Neck 采用路径聚合网络（PANet），融合 Backbone 中不同层次的特征，来进一步提高网络特征提取能力。PANet 基于特征金字塔网络（FPN），增加自底向上的特征金字塔结构，能更好地将底层信息传递到高层，利用高层和底层的信息互补，进而提升小目标的检测性能。

Head用于检测目标，对来自Neck的不同尺度的特征进行处理。使用非极大抑制（NMS）对多个目标预测框进行筛选，去除冗余的预测框，进而增强网络的目标检测能力。

3　多模态自适应特征融合网络

为了解决传统基于RGB图像的目标检测算法在低光照条件下检测性能较低的问题，提出了一种新的MAFFNet。MAFFNet作为端到端的目标检测模型网络，一方面能有效利用多模态特征的互补性，通过MAFF模块实现自适应融合不同模态的特征信息，获取更丰富的图像特征，进而提升目标检测性能；另一方面，其能在检测性能和检测速度之间取得良好的平衡，在较低的复杂度代价下有效提升目标检测性能。

3.1　多模态自适应特征融合网络总体框架

基于YOLOv5模型，构建MAFFNet，该网络结构如图2所示。MAFFNet包含4个主体部分：双支路主干（Backbone）、融合（Fusion）、颈部（Neck）、头部（Head）。

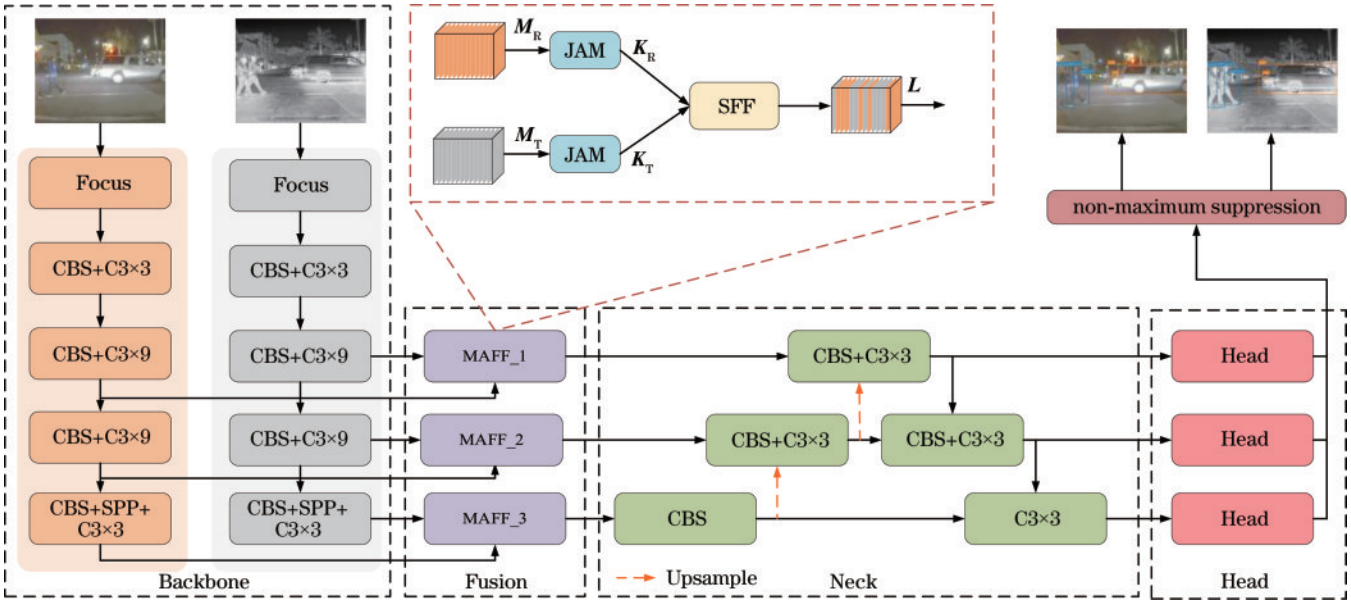

图2　MAFFNet结构

Fig. 2　Structure of MAFFNet

为充分利用不同模态特征间的互补性以获取更全面、精确的特征表示，通过MAFF模块对相同尺度的特征进行融合。为更好地捕捉不同尺寸目标特征，获取更丰富的信息和空间细节，在不同尺度的特征层分别独立地应用MAFF模块进行融合。此外，基于中期融合的方法较为有效，将融合模块嵌入在Backbone和Neck之间。MAFFNet具体实现步骤如下：首先，将RGB图像和热红外图像输入两路并行的特征提取Backbone，每路Backbone与YOLOv5的Backbone一致。其次，在Backbone的3个特征尺度嵌入MAFF模块，采用MAFF模块融合所提取的相同尺度的RGB特征和热红外特征。以输入尺寸高$H=640$、宽$W=640$、通道数量$C=3$为例，这3处特征层的尺寸分别为(80,80,256)、(40,40,512)和(20,20,1024)。这3个MAFF模块分别命名为MAFF_1、MAFF_2和MAFF_3。然后，将不同尺度的融合特征依次输入Neck中，实现多尺度特征融合。最后，输入Head部分进行目标检测，获得目标框和对应类别信息。

3.2　多模态自适应特征融合模块

为了进一步提升多模态目标检测的性能，提出了MAFF模块，该模块利用不同模态之间的互补性进行多模态特征融合，其结构如图2所示。在MAFF模块中，主要包含了SFF单元和JAM单元。首先，采用JAM单元分别对来自RGB图像和热红外图像的特征图$\boldsymbol{M}_{\mathrm{R}}$、$\boldsymbol{M}_{\mathrm{T}}$在特征空间和通道三维空间上进行重加权，获得加权输出$\boldsymbol{K}_{\mathrm{R}}$与$\boldsymbol{K}_{\mathrm{T}}$；其次，将$\boldsymbol{K}_{\mathrm{R}}$与$\boldsymbol{K}_{\mathrm{T}}$输入到SFF单元进行自适应选择融合，获得融合特征$\boldsymbol{L}$。在采用端到端训练后，网络可以通过JAM单元和SFF单元自适应学习到两个模态数据的互补特征并进行有效融合。对这两个单元的具体结构进行描述。

3.2.1　SFF单元

在多模态目标检测任务中，将不同模态的特征进行融合至关重要。有效的融合方法能实现不同模态特征的自适应融合，充分利用不同模态之间固有的互补性，提升检测性能。最常用的融合方法包括按元素相加或级联，然而，这两种方法都过于简单，不能有效利用多模态的互补性质。影响RGB图像和热红外图像最主要的因素为光照条件，RGB图像对光照敏感，而热红外图像在夜间能更好捕捉显著特征。因此，可以

设计一种根据光照条件自适应融合RGB特征和热红外特征的融合模块。受Selective kernel networks (SKNet)[30]启发，提出了SFF单元用于融合不同模态特征，其结构如图3所示。所提的SFF单元与SKNet的区别在于：1) SFF单元融合阶段使用自适应一维卷积[31]，避免通道向量的降维操作。2) SFF单元添加最大池化层支路，丰富显著目标的特征信息。SFF单元的实现可分为以下3个步骤：

步骤1　获取不同模态融合信息。将两个携带不同模态的并行卷积特征图$\boldsymbol{K}_{\mathrm{R}}$、$\boldsymbol{K}_{\mathrm{T}}$按元素相加进行融合，$\boldsymbol{K}=\boldsymbol{K}_{\mathrm{R}}+\boldsymbol{K}_{\mathrm{T}}$。其中，$\boldsymbol{K}_{\mathrm{R}}\in\mathbb{R}^{C\times H\times W}$、$\boldsymbol{K}_{\mathrm{T}}\in\mathbb{R}^{C\times H\times W}$。

步骤2　利用融合特征生成通道特征向量。为了更好聚合空间信息，在局部信息中融入全局信息，同时采用全局平均池化(GAP)和全局最大池化(GMP)来获取更精细的通道特征向量。其中，GAP获取的特征信息更关注背景信息，GMP获取的特征信息则是更关注纹理信息。将融合特征$\boldsymbol{K}$分别采用GAP和GMP生成通道向量$\boldsymbol{s}$、$\boldsymbol{t}$，$\boldsymbol{s}$、$\boldsymbol{t}$可表示为

$$\boldsymbol{s}=F_{\mathrm{gap}}(\boldsymbol{K})=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{K}(i,j),\tag{1}$$

$$\boldsymbol{t}=F_{\mathrm{gmp}}(\boldsymbol{K})=\mathrm{Max}\left[\boldsymbol{K}(i,j)\right],\tag{2}$$

式中：$\boldsymbol{s}$、$\boldsymbol{t}\in\mathbb{R}^{1\times1\times C}$；$F_{\mathrm{gap}}(\cdot)$、$F_{\mathrm{gmp}}(\cdot)$分别为GAP、GMP。

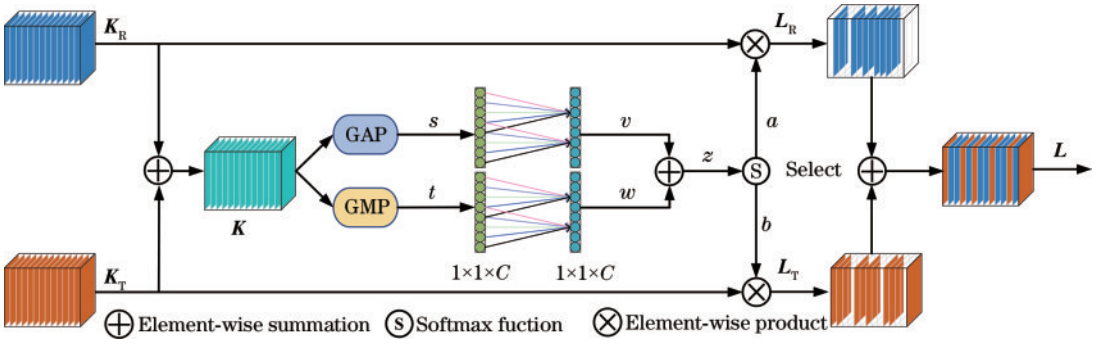

图3　SFF结构

Fig. 3　Structure of SFF

在经过GAP、GMP操作后，因为普通全连接层网络参数量较大，使用自适应一维卷积来生成维度为$1\times1\times C$的一维特征$\boldsymbol{v}$、$\boldsymbol{w}$，但该操作并不降低通道数，其能够有效避免通道维度先降再升，在保持先进性能的同时极大地降低模型的复杂性[31]。

步骤3　实现自适应选择融合。首先，使用Softmax函数应用于特征向量$\boldsymbol{z}$，利用跨通道的权重系数来自适应选择不同的信息空间尺度，产生通道特征向量$\boldsymbol{a}_i$和$\boldsymbol{b}_i$，并对二者进行归一化。$\boldsymbol{a}_i$、$\boldsymbol{b}_i$可表示为

$$\boldsymbol{a}_i=\frac{\exp(X_i\boldsymbol{z})}{\exp(X_i\boldsymbol{z})+\exp(Y_i\boldsymbol{z})},\tag{3}$$

$$\boldsymbol{b}_i=1-\boldsymbol{a}_i,\tag{4}$$

$$\boldsymbol{z}=\boldsymbol{v}\oplus\boldsymbol{w},\tag{5}$$

式中：$\boldsymbol{X}$、$\boldsymbol{Y}\in\mathbb{R}^{C}$；$\boldsymbol{a}$、$\boldsymbol{b}$分别为$\boldsymbol{K}_{\mathrm{R}}$、$\boldsymbol{K}_{\mathrm{T}}$的权重系数向量；$X_i$、$Y_i$分别为$\boldsymbol{X}$、$\boldsymbol{Y}$的第$i$行；$\boldsymbol{a}_i$、$\boldsymbol{b}_i$分别为$\boldsymbol{a}$、$\boldsymbol{b}$的第$i$个元素；$\oplus$为元素相加符号。

然后，通过各自模态通道权重系数获得特征图$\boldsymbol{L}_{\mathrm{R}}$、$\boldsymbol{L}_{\mathrm{T}}$，$\boldsymbol{L}_{\mathrm{R}}$、$\boldsymbol{L}_{\mathrm{T}}$可表示为

$$\begin{cases}\boldsymbol{L}_{\mathrm{R}}=\boldsymbol{a}\otimes\boldsymbol{K}_{\mathrm{R}}\\ \boldsymbol{L}_{\mathrm{T}}=\boldsymbol{b}\otimes\boldsymbol{K}_{\mathrm{T}}\end{cases},\tag{6}$$

式中：$\boldsymbol{L}_{\mathrm{R}}$、$\boldsymbol{L}_{\mathrm{T}}\in\mathbb{R}^{H\times W\times C}$；$\otimes$为元素相乘符号。

最后，得到融合特征$\boldsymbol{L}$，$\boldsymbol{L}$可表示为

$$\boldsymbol{L}=\boldsymbol{L}_{\mathrm{R}}+\boldsymbol{L}_{\mathrm{T}}。\tag{7}$$

3.2.2　JAM单元

虽然SFF单元在双模态分支之间融合信息，但还需要一种机制来共享特征张量中的信息，包括通道维度和空间维度。受Wang等[31]和Woo等[32]的启发，提出联合使用增强高效通道注意力(EECA)机制和空间注意力(SA)机制的JAM单元，其结构如图4所示。其中，EECA为改进的高效通道注意力(ECA)机制[31]。

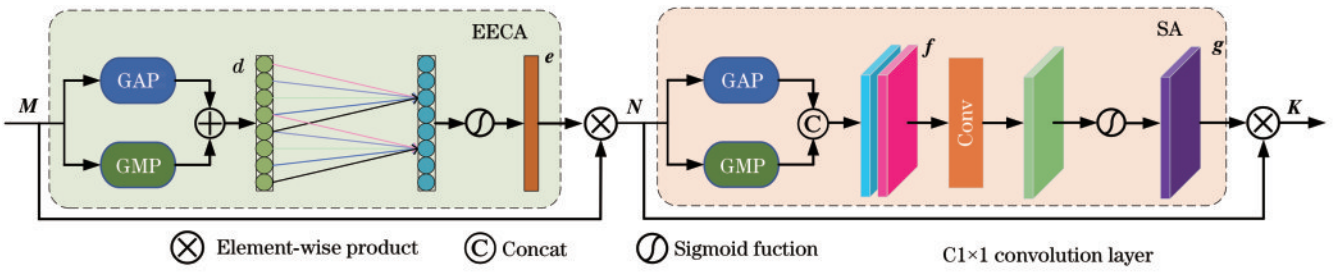

图4　JAM结构

Fig. 4　Structure of JAM

与ECA相比，EECA添加了GMP层支路。GMP层能获取更具有纹理信息的特征信息，联合GAP和GMP能获取更加有效的通道注意信息。

JAM单元能抑制几乎无用的特征，只允许包含更多有效信息的特征来进一步传递。JAM单元的具体实现公式为

$$\boldsymbol{N}=F_{\mathrm{EECA}}(\boldsymbol{M})\otimes\boldsymbol{M}, \tag{8}$$

$$\boldsymbol{K}=F_{\mathrm{SA}}(\boldsymbol{N})\otimes\boldsymbol{N}, \tag{9}$$

式中：$\boldsymbol{M}$为校准前的特征；$\boldsymbol{K}$为校准后的特征；$F_{\mathrm{EECA}}(\cdot)$为EECA操作；$F_{\mathrm{SA}}(\cdot)$为SA操作。

EECA支路的具体操作：给定一个特征图$\boldsymbol{M}\in\mathbb{R}^{H\times W\times C}$，首先在空间维度上联合应用GAP和GMP获得全局上下文信息，将获得的两个特征向量进行相加，得到通道特征向量$\boldsymbol{d}\in\mathbb{R}^{1\times1\times C}$。再通过自适应一维卷积捕获不同通道之间的信息。最后，使用Sigmoid激活函数产生通道特征向量$\boldsymbol{e}\in\mathbb{R}^{1\times1\times C}$。

SA支路的具体操作：首先，给定一个特征图$\boldsymbol{N}\in\mathbb{R}^{H\times W\times C}$，沿通道维度分别独立对输入特征应用GAP和GMP操作，并将输出结果级联起来以形成特征$\boldsymbol{f}\in\mathbb{R}^{H\times W\times2}$。然后，对特征$\boldsymbol{f}$使用1×1卷积进行降维操作。最后，使用Sigmoid激活函数获得特征$\boldsymbol{g}\in\mathbb{R}^{H\times W\times1}$。

4 实验结果与分析

4.1 实验设置

为了更好地证明所提算法的性能以及保证实验的公平性，所有实验均使用相同的学习策略。使用PyTorch深度学习框架，在单个NVIDIA GeForce RTX 3090 GPU上进行训练和测试。总训练轮次设置为100，输入块大小为640×640，每个批次为16，初始学习率为0.01，动量为0.937。使用了Mosaic数据增强方法，将4张随机图片拼接成一张图片。此外，在实验中，使用了YOLOv5检测器基于MS-COCO数据集[33]的训练模型作为预训练模型，分别对双支路进行预训练模型加载。预训练模型是基于RGB图像训练获得，因此在热红外图像输入支路对预训练模型进行微调。

4.2 数据集和评价指标

4.2.1 数据集

FLIR数据集[34]由FLIR公司于2018年发布，其包含约1万张手动注释的热红外图像及其相应的参考RGB图像，这些图像都在白天和夜间采集。FLIR数据集是图像未对齐的多光谱多物体检测数据集，由RGB图像（由分辨率为1280 pixel×1024 pixel的FLIR BlackFly RGB相机拍摄）和热红外图像（由分辨率为640 pixel×512 pixel的FLIR Tau2热相机拍摄）组成。为了便于与其他算法进行比较，采用对齐版本的FLIR数据集[25]，为了便于图像融合研究，将所有图像的大小调整为640 pixel×512 pixel，手动删除了严重未对齐的多光谱图像对，保留了5142个对齐良好的多光谱图像对，其中4129对用于训练模型，1013对用于测试模型。为了方便描述，后续出现的FLIR数据集均指代对齐版本的FLIR数据集。

LLVIP数据集由Jia等[35]提出，该数据集使用多光谱摄像机以俯视的监控视角进行采集，包含大量行人和骑行者的街景图像。其中，大部分图像都是在非常黑暗的场景中拍摄的。LLVIP数据集是一个单目标多光谱配对的数据集，其中行人和骑行者都被标注为“人”，且在时间和空间上都严格对齐。该数据集包含15488对图像，其中12025对图像用于训练模型，3463对图像用于测试模型。

4.2.2 评价指标

将MAFFNet算法与其他方法进行比较来测量其性能。实验模型均采用MS-COCO数据集[33]提出的目标检测指标平均精度均值（mAP）来进行评估。其中，mAP50、mAP75和mAP分别表示交并比（IoU）等于0.50、0.75和0.50∶0.95时所有类别的平均精度（AP）的平均值。此外，还将网络模型在模型尺寸、计算复杂度、检测速度等方面进行比较。模型尺寸采用参数衡量，计算复杂度采用千兆浮点运算（GFLOPs）衡量，检测速度采用每秒帧数（FPS）衡量。

4.3 消融实验

为了更好地深入探究MAFF模块对目标检测性能的影响，分别对MAFF模块嵌入网络的数量和MAFF模块单个组成结构所起作用进行消融实验。在FLIR数据集上使用网络参数、GFLOPs、FPS、mAP50等指标进行综合评估。

4.3.1 MAFF模块数量对目标检测结果的影响

探索了融合模块MAFF嵌入的数量和嵌入的位置对目标检测的性能的影响如表1所示。此外，在未嵌入MAFF模块的特征层，使用级联的方式进行不同模态特征的融合。级联融合具体操作如下：首先，进行两个特征的级联；然后，使用1×1卷积进行降通道输出。由表1可知，嵌入3个MAFF模块获得的检测性能最好，其mAP50为0.849，没有嵌入MAFF模块的检测性能最差，其mAP50为0.809。且嵌入3个MAFF模块相比于没有嵌入MAFF模块的FPS仅下降2帧，而网络参数下降2.75×10^{6}，检测性能提升4个百分点。因此，3个不同层次的特征融合对目标检测性能的提升是最高的。由表1还可知，在逐个嵌入MAFF模块时，检测性能也在逐步提升。

4.3.2 MAFF模块的单个组件对目标检测结果的影响

MAFF模块的单个组件对目标检测性能的影响如表2所示。其中，仅包含JAM单元的变体模型使用级联融合方式取代SFF单元实现两种模态的融合。与MAFF模块（JAM+SFF）的mAP50相比，仅包含

表 1 MAFF 模块数量对目标检测结果的影响

Table 1 Influence of the number of MAFF modules on object detection results

MAFF module	Parameter quantity /10^6	GFLOPs	FPS /(frame·s^{-1})	mAP50			
				Person	Car	Bicycle	All
Without MAFF module	76.47	195.8	25	0.864	0.912	0.652	0.809
MAFF_1	76.34	194.1	24	0.868	0.914	0.676	0.819
MAFF_2	75.95	194.1	24	0.872	0.910	0.681	0.821
MAFF_3	74.37	197.1	24	0.867	0.910	0.668	0.815
MAFF_1+MAFF_2	75.82	192.4	23	0.882	0.918	0.701	0.834
MAFF_1+MAFF_3	74.24	192.4	24	0.879	0.913	0.690	0.827
MAFF_2+MAFF_3	73.85	192.4	23	0.879	0.914	0.736	0.843
MAFF_1+MAFF_2 +MAFF_3	73.72	190.8	23	0.886	0.922	0.740	0.849

表 2 融合模块单个组件对目标检测结果的影响

Table 2 Influence of single component of fusion module on object detection results

Module	Parameter quantity /10^6	GFLOPs	FPS /(frame·s^{-1})	mAP50			
				Person	Car	Bicycle	All
JAM	76.47	195.8	23	0.857	0.905	0.685	0.816
SFF	73.72	190.7	24	0.883	0.917	0.721	0.840
JAM+SFF	73.72	190.8	23	0.886	0.922	0.740	0.849

JAM 单元的变体模型的检测性能下降最严重(降低 3.3 个百分点),这表明 SFF 单元在 MAFF 模块中的重要性。在 MAFF 模块中,SFF 单元的作用为融合不同模态的特征,进而提升目标检测性能。而仅 JAM 单元不能有效利用不同模态之间的互补性,导致网络检测性能变差。MAFF 模块(JAM+SFF)与仅包含 SFF 单元的变体模型相比 mAP50 提升 0.9 个百分点,而参数量和 GFLOPs 几乎没有增加,FPS 也仅下降 1 帧。这表明 JAM 单元对融合前的特征处理,有助于提升多模态目标检测性能。因此,组成 MAFF 模块的 JAM 单元和 SFF 单元都有助于提升多模态目标的检测性能。

此外,为进一步探索 SFF 单元的有效性,将仅保留 GAP 支路的 SFF 单元变体模型称为 SFF-GAP;将 SFF 单元的自适应一维卷积替换成的降通道尺度卷积的变体模型,称为变通道选择特征融合(DSFF)。如表 3 所示,MAFF 模块(JAM+SFF)比仅使用 GAP 支路的 JAM+SFF-GAP 模块的 mAP50 高 0.6 个百分点。JAM+DSFF 的 mAP50(0.844)比 MAFF 模块(JAM+SFF)低 0.5 个百分点,且在网络参数量上也增加 0.51×10^6。因此,在 SFF 单元中,GMP 支路和自适应一维卷积都能增强不同模态特征的融合,进而提高检测精度。

为了探索 EECA 模块的有效性,将 JAM 单元中的 EECA 模块替换成 ECA 模块的变体模型,称为 JAM-ECA。由表 3 可知,相对于 MAFF 模块(JAM+SFF),JAM-ECA+SFF 模块的性能下降 0.04 个百分点,这证明了 EECA 模块的有效性。

表 3 不同组件对目标检测结果的影响

Table 3 Influence of different component of fusion module on object detection results

Method	Parameter quantity /10^6	GFLOPs	FPS /(frame·s^{-1})	mAP50			
				Person	Car	Bicycle	All
JAM+SFF-GAP	73.72	190.8	23	0.888	0.919	0.722	0.843
JAM+DSFF	74.23	191.2	21	0.881	0.912	0.734	0.844
JAM-ECA+SFF	73.72	190.8	23	0.891	0.920	0.725	0.845
JAM+SFF	73.72	190.8	23	0.886	0.922	0.740	0.849

4.4 不同算法的性能比较

在 FLIR 数据集上将 MAFFNet 算法与现有算法进行检测性能对比。现有算法包括 CFR[25]、GAFF[27]、CFT[28]、ProbEn[35]。为了使对比实验更具合理性,增加基于单模态 RGB 图像的 YOLOv5、基于单模态热红外图像的 YOLOv5 和 YOLOBase 算法来进行对比。其中,YOLOBase 算法是基准模型网络,由 MAFFNet 将 3 个 MAFF 模块全部替换成级联融合模块得到。级联融合模块包含两部分,级联操作和 1×1 卷积降维度操作。所有算法在 FLIR 数据集上测试的 mAP50 如表 4 所示。其中,使用基于 MS-COCO 数据集预训练模型的算法,对表 4 中的数据进行加粗,加以区分。

表 4　不同算法在 FLIR 数据集的性能比较

Table 4　Performance comparison of different algorithms in FLIR dataset

Method	Backbone	Data	mAP50			
			Person	Car	Bicycle	All
YOLOv5	CSPDarkNet	RGB	0.581	0.781	0.407	0.590
YOLOv5	CSPDarkNet	Thermal	0.791	0.887	0.538	0.739
CFR[25]	VGG16	RGB+T	0.745	0.849	0.578	0.724
GAFF[27]	VGG16	RGB+T				0.727
GAFF[27]	ResNet18	RGB+T				0.729
CFT[28]	**CSPDarkNet**	**RGB+T**	**0.822**	**0.890**	**0.640**	**0.784**
ProbEn[36]	**ResNet101**	**RGB+T**	**0.877**	**0.901**	**0.735**	**0.838**
YOLOBase(ours)	**CSPDarkNet**	**RGB+T**	**0.864**	**0.912**	**0.652**	**0.809**
MAFFNet(ours)	**CSPDarkNet**	**RGB+T**	**0.886**	**0.922**	**0.740**	**0.849**

所提算法获得最高的 mAP50(0.849)。由表 4 可知，在 3 个类别中，相对于单模态红外图像的 YOLOv5 算法，所提算法在自行车类别的 mAP50 提升最高。这证明了所提算法有效提高了对自行车的检测精度。CFT 和 ProbEn 算法在自行车类别的检测性能提升也较高。这种情况是由于自行车不会发出热量，其在热红外图像中不明显，而在 RGB 图像中表现显著，融合 RGB 图像和热红外图像能极大地提高目标检测网络对自行车的检测精度。

4.5　复杂度比较

为了进一步探索 MAFFNet 算法的性能，对 CFT、ProbEn、YOLOBase、MAFFNet 算法的复杂度和检测性能进行比较如表 5 所示。由表 5 可知，MAFFNet 算法在参数量、GFLOPs、FPS 和 mAP50 均达到最优。这表明 MAFFNet 能平衡检测性能和网络复杂度。CTF 算法的参数量最大，GFLOPs 最高，这验证了 Transformer 存在参数大、消耗算力高等缺点。ProbEn 算法使用双阶段的目标检测器进行检测，与 MAFFNet 算法相比，其 FPS 较低。与 YOLOBase 算法相比，MAFFNet 算法使用模型参数量少的 MAFF 模块，而 YOLOBase 算法使用了级联和 1×1 卷积降维度操作，因此 MAFFNet 算法的参数量更少，GPLOPs 更低。MAFFNet 算法能在 FPS 仅下降 2 帧的情况下，显著提高目标检测性能。

表 5　各模型的复杂度比较

Table 5　Comparison of complexity of various models

Data	Method	Detector	Parameter quantity /10^6	GFLOPs	FPS / (frame·s^{-1})	mAP50
RGB	YOLOv5	YOLOv5	46.64	114.6	38	0.590
Thermal	YOLOv5	YOLOv5	46.64	114.6	38	0.739
RGB+T	CFT[28]	YOLOv5	206.26	13732.5	14	0.784
RGB+T	ProbEn[36]	Faster R-CNN	107.18	339.3	17	0.838
RGB+T	YOLOBase(ours)	YOLOv5	76.47	195.8	25	0.809
RGB+T	MAFFNet(ours)	YOLOv5	73.72	190.8	23	0.849

此外，由表 5 还可知，与单模态 RGB 图像和单模态热红外图像的检测结果相比，所提的 MAFFNet 算法的 mAP50 最高(0.849)。与单模态 YOLOv5 算法相比，MAFFNet 在参数量和 GFLOPs 上分别增加 27.08×10^6 和 76.2，FPS 下降 15 帧，但其检测性能显著提高。与单模态 RGB 图像相比，MAFFNet 的 mAP50 提高 0.259；与单模态热红外图像相比，MAFFNet 的 mAP50 提高 0.110。与 mAP50 的显著提升相比，参数量、GFLOPs、FPS 可作为次要性能指标。

多模态目标检测算法在目标检测性能方面具有优势，能够充分利用不同模态之间的互补性，从而提升目标检测性能。尽管多模态目标检测算法可能导致运行时间增加，但对许多实际应用场景而言，提升目标检测性能是最关键的。此外，得益于并行计算技术的进步，采用多图形处理器(GPU)策略可以有效提高检测速度并降低运行时间。因此，在实际应用场景中，可以通过优化硬件来解决运行时间较长的问题。

4.6　可视化比较

真值图像、CFT、ProbEn、MAFFNet 算法的目标检测可视化的结果对比，如图 5 所示。其中，第一行为白天的图像，第二、第三行为晚上的图像。由图 5 可知，综合人、汽车、自行车 3 个类别的检测，无论白天或晚上，所提算法均获得最好的检测结果。此外，由于

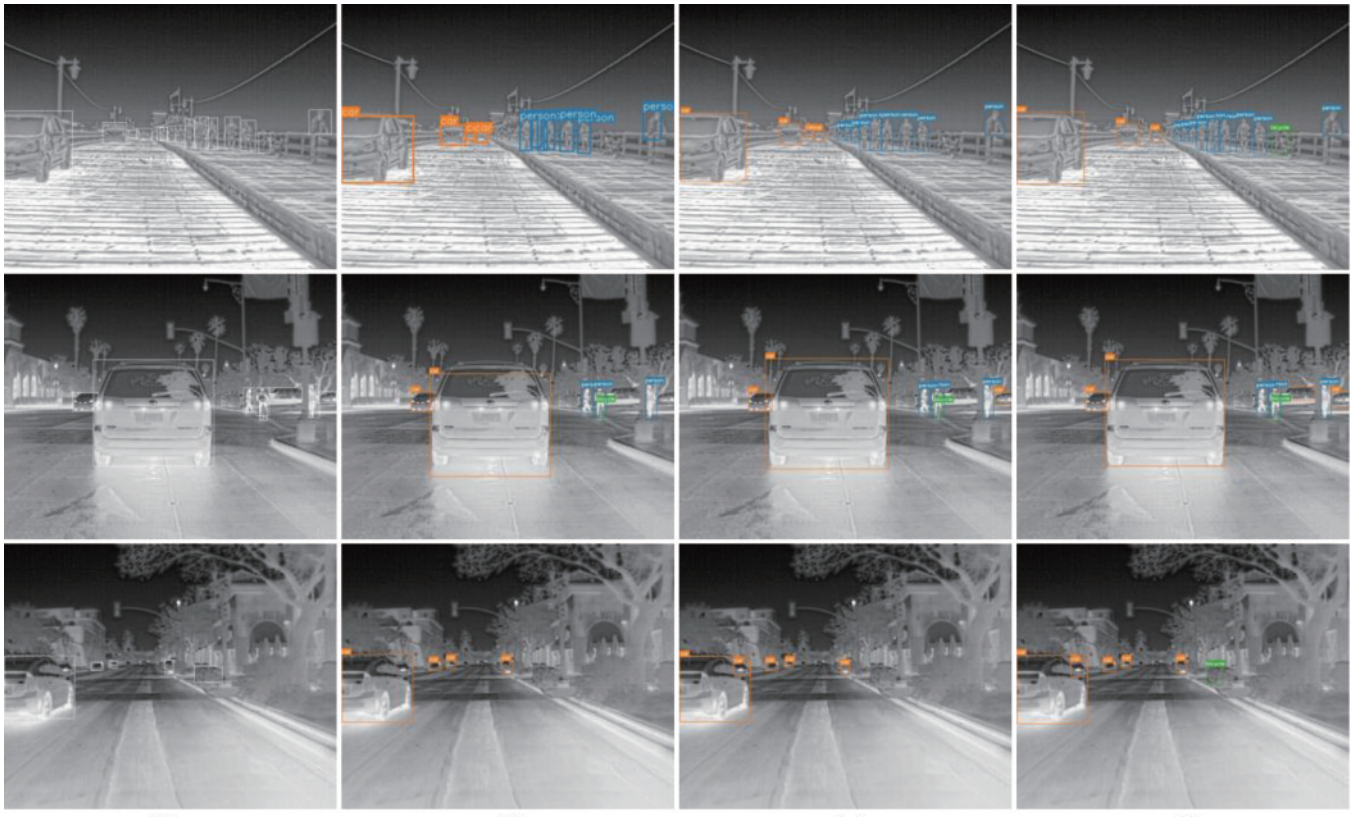

图 5 与其他算法在 FLIR 数据集的定性比较。(a)真值图像;(b)CFT;(c)ProbEn;(d)MAFFNet

Fig. 5 Qualitative comparison with other algorithms in FLIR dataset. (a) Truth image; (b) CFT; (c) ProbEn; (d) MAFFNet

自行车不发热,红外图像很难识别,只有所提算法检测到第一行和第三行的自行车,这证明了 MAFFNet 能充分利用不同模态间的互补性来提高目标检测的精度。

4.7 在 LLVIP 数据集上评估

所提算法 MAFFNet 与其他算法的在 LLVIP 数据集上的检测结果如表 6 所示。对比算法有:基于单模态 RGB 图像的 YOLOv3、基于单模态热红外图像的 YOLOv3、基于单模态 RGB 图像的 YOLOv5、基于单模态热红外图像的 YOLOv5、CFT[28]、CCIFNet[37]等 6 种算法。由表 6 可知,MAFFNet 算法的 mAP50、mAP75、mAP 评估指标都是最高的,mAP50、mAP50、mAP 分别为 0.977、0.783、0.671。总体而言,所提的 MAFFNet 在所有的 IoU 阈值下都能获得较好的检测结果,这表明所提方法可以很好地推广到不同类型的图像。

表 6 不同算法在 LLVIP 数据集的性能比较

Table 6 Performance comparison of different algorithms in LLVIP dataset

Method	Backbone	Data	mAP50	mAP75	mAP
YOLOv3[3]	DarkNet	RGB	0.859	0.379	0.433
YOLOv3[3]	DarkNet	Thermal	0.897	0.534	0.528
YOLOv5	CSPDarkNet	RGB	0.908	0.519	0.505
YOLOv5	CSPDarkNet	Thermal	0.946	0.722	0.619
CFT[28]	CSPDarkNet	RGB+T	0.975	0.729	0.636
CCIFNet[37]	ResNet50	RGB+T	0.976	0.726	0.641
MAFFNet	CSPDarkNet	RGB+T	0.977	0.783	0.671

5 结 论

为了解决基于 RGB 图像的目标检测在低光照条件下性能较低的问题,提出了一种多模态目标检测网络 MAFFNet。MAFFNet 的双支路主干能够实现不同模态的图像并行输入,通过设计 MAFF 模块充分利用不同模态之间的互补性,从而为目标检测提供更丰富的信息。此外,MAFF 模块引入通道注意力和空间注意力进一步聚焦重要特征,为特征融合提供更加有效的信息。实验结果证明,所提的 MAFFNet 算法能够充分利用多模态信息,在仅略微增加网络参数量和计算复杂度的情况下,显著提升了模型的性能。与已有算法相比,MAFFNet 算法得到的 mAP50 最高,且能在检测速度和检测性能之间达到平衡。然而,所提算法并未考虑未对齐图像对的情况,未来本团队将致力于提升网络模型的鲁棒性,以适应包含未对齐多模态图像的复杂应用场景。

参 考 文 献

[1] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and

Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[2] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6517-6525.

[3] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2023-03-07]. https://arxiv.org/abs/1804.02767.

[4] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2023-03-07]. https://arxiv.org/abs/2004.10934.

[5] 张寅, 朱桂熠, 施天俊, 等. 基于特征融合与注意力的遥感图像小目标检测[J]. 光学学报, 2022, 42(24): 2415001.
Zhang Y, Zhu G Y, Shi T J, et al. Small object detection in remote sensing images based on feature fusion and attention[J]. Acta Optica Sinica, 2022, 42(24): 2415001.

[6] 徐志京, 柏雪. 基于双重特征增强的遥感舰船小目标检测[J]. 光学学报, 2022, 42(18): 1828002.
Xu Z J, Bai X. Small ship target detection method for remote sensing images based on dual feature enhancement [J]. Acta Optica Sinica, 2022, 42(18): 1828002.

[7] 琚长瑞, 秦晓燕, 袁广林, 等. 尺度敏感损失与特征融合的快速小目标检测方法[J]. 电子学报, 2022, 50(9) 2119-2126
Ju C R, Qin X Y, Yuan G L, et al. Fast small object detection method with scale-sensitivity loss and feature fusion[J]. Acta Electronica Sinica, 2022, 50(9)2119-2126

[8] 李翔, 何淼, 罗海波. 一种面向遮挡行人检测的改进YOLOv3算法[J]. 光学学报, 2022, 42(14): 1415003.
Li X, He M, Luo H B. Occluded pedestrian detection algorithm based on improved YOLOv3[J]. Acta Optica Sinica, 2022, 42(14): 1415003.

[9] 王友伟, 郭颖, 邵香迎. 基于改进级联算法的遥感图像目标检测[J]. 光学学报, 2022, 42(24): 2428004.
Wang Y W, Guo Y, Shao X Y. Remote sensing image target detection based on improved cascade algorithm[J]. Acta Optica Sinica, 2022, 42(24): 2428004.

[10] 薛俊达, 朱家佳, 张静, 等. 基于FFC-SSD模型的光学遥感图像目标检测[J]. 光学学报, 2022, 42(12): 1210002.
Xue J D, Zhu J J, Zhang J, et al. Object detection in optical remote sensing images based on FFC-SSD model [J]. Acta Optica Sinica, 2022, 42(12): 1210002.

[11] Lin J P, Haberstroh F, Karsch S, et al. Applications of object detection networks in high-power laser systems and experiments[J]. High Power Laser Science and Engineering, 2023, 11(1): e7.

[12] Osornio-Rios R A, Antonino-Daviu J A, de Jesus Romero-Troncoso R. Recent industrial applications of infrared thermography: a review[J]. IEEE Transactions on Industrial Informatics, 2019, 15(2): 615-625.

[13] Dai X R, Yuan X, Wei X Y. TIRNet: object detection in thermal infrared images for autonomous driving[J]. Applied Intelligence, 2021, 51(3): 1244-1261.

[14] 何自芬, 陈光晨, 陈俊松, 等. 多尺度特征融合轻量化夜间红外行人实时检测[J]. 中国激光, 2022, 49(17): 1717002.
He Z F, Chen G C, Chen J S, et al. Multi-scale feature fusion lightweight real-time infrared pedestrian detection at night[J]. Chinese Journal of Lasers, 2022, 49(17): 1717002.

[15] 宋子壮, 杨嘉伟, 张东方, 等. 基于无监督域适应的低空海面红外目标检测[J]. 光学学报, 2022, 42(4): 0415001.
Song Z Z, Yang J W, Zhang D F, et al. Low-altitude Sea surface infrared object detection based on unsupervised domain adaptation[J]. Acta Optica Sinica, 2022, 42(4): 0415001.

[16] Xu D, Ouyang W, Ricci E, et al. Learning cross-modal deep representations for robust pedestrian detection[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5363-5371.

[17] Li C Y, Song D, Tong R F, et al. Multispectral pedestrian detection via simultaneous detection and segmentation[EB/OL]. (2018-08-14)[2023-03-07]. https://arxiv.org/abs/1808.04818.

[18] Devaguptapu C, Akolekar N, Sharma M M, et al. Borrow from anywhere: pseudo multi-modal object detection in thermal imagery[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 16-17, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 1029-1038.

[19] Zhang L, Zhu X Y, Chen X Y, et al. Weakly aligned cross-modal learning for multispectral pedestrian detection [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 5126-5136.

[20] 刘通, 高思洁, 聂为之. 基于多模态信息融合的多目标检测算法[J]. 激光与光电子学进展, 2022, 59(8): 0815002.
Liu T, Gao S J, Nie W Z. Multitarget detection algorithm based on multimodal information fusion[J]. Laser & Optoelectronics Progress, 2022, 59(8): 0815002.

[21] Wagner J, Fischer V, Herman M, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks[C]//Proceedings of 2016 European Symposium on Artificial Neural Networks (ESANN), April 27-29, 2016, Bruges, Belgium. [S.l.: s.n.], 2016: 509-514.

[22] Liu J J, Zhang S T, Wang S, et al. Multispectral deep neural networks for pedestrian detection[EB/OL]. (2016-11-08)[2023-03-07]. https://arxiv.org/abs/1611.02644.

[23] Konig D, Adam M, Jarvers C, et al. Fully convolutional region proposal networks for multispectral person detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), July 21-26, 2017, Honolulu, USA. New

York: IEEE Press, 2017: 243-250.

[24] Kieu M, Bagdanov A D, Bertini M, et al. Task-conditioned domain adaptation for pedestrian detection in thermal imagery[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12367: 546-562.

[25] Zhang H, Fromont E, Lefevre S, et al. Multispectral fusion for object detection with cyclic fuse-and-refine blocks[C]//Proceedings of 2020 IEEE International Conference on Image Processing (ICIP), September 25-28, 2020, Abu Dhabi, United Arab Emirates. New York: IEEE Press, 2020: 276-280.

[26] Zhou K L, Chen L S, Cao X. Improving multispectral pedestrian detection by addressing modality imbalance problems[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12363: 787-803.

[27] Zhang H, Fromont E, Lefevre S, et al. Guided attentive feature fusion for multispectral pedestrian detection[C]// 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), January 3-8, 2021, Waikoloa, HI, USA. New York: IEEE Press, 2021: 72-80.

[28] Fang Q Y, Han D P, Wang Z K. Cross-modality fusion transformer for multispectral object detection[EB/OL]. (2021-10-30)[2023-03-07]. https://arxiv.org/abs/2111.00273.

[29] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of 2017 Advances in Neural Information Processing Systems, December 4-9, 2017, Long Beach, CA, USA. Cambridge: MIT Press, 2017: 5998-6008.

[30] Li X, Wang W H, Hu X L, et al. Selective kernel networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 510-519.

[31] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11534-11542.

[32] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[33] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[34] FLIR: Flir thermal dataset for algorithm training[EB/OL]. (2018-07-10) [2023-03-07]. https://www.flir.in/oem/adas/adas-dataset-form/.

[35] Jia X Y, Zhu C, Li M Z, et al. LLVIP: a visible-infrared paired dataset for low-light vision[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), October 11-17, 2021, Montreal, BC, Canada. New York: IEEE Press, 2021: 3489-3497.

[36] Chen Y T, Shi J, Ye Z, et al. Multimodal object detection via probabilistic ensembling[M]//Avidan S, Brostow G, Cissé M, et al. Computer vision-ECCV 2022. Lecture notes in computer science. Cham: Springer, 2022, 13669: 139-158.

[37] Yan C Q, Zhang H, Li X L, et al. Cross-modality complementary information fusion for multispectral pedestrian detection[J]. Neural Computing and Applications, 2023: 1-26.