DOI: 10.16452/j.cnki.sdkjzk.2024.01.010 文章编号:1672-3767(2024)01-0089-11

基于多尺度模态融合的 RGB-T 目标跟踪网络

程竹轩[1,2,4],范慧杰[2,4],唐延东[2,4],王 强[3]
(1. 沈阳化工大学 信息工程学院,辽宁 沈阳 110142;
2. 中国科学院沈阳自动化研究所 机器人学国家重点实验室,辽宁 沈阳 110016;
3. 沈阳大学 辽宁省装备制造综合自动化重点实验室,辽宁 沈阳 110044;
4. 中国科学院 机器人与智能制造创新研究院,辽宁 沈阳 110016)

摘 要:可见光-热红外(RGB-T)目标跟踪因受光照条件限制较小受到关注。针对不同尺度特征的分辨率与语义信息存在差异、可见光与热红外两种模态信息不一致的特点,以及现有网络在多模态融合策略上的不足,提出一种 RGB-T 目标跟踪网络。网络采用孪生结构,首先将主干特征提取网络输出的模板图像特征与搜索图像特征从单尺度拓展到多尺度,并对可见光与热红外模态在不同尺度上分别进行模态融合,然后将得到的融合特征通过注意力机制增强特征表示,最后通过区域建议网络得到预测结果。在 GTOT 与 RGBT-234 两个公开 RGB-T 数据集上的实验结果表明,该网络跟踪精度和成功率较高,可以应对复杂的跟踪场景,相比于其他网络具有更高的跟踪性能。

关键词:目标跟踪;可见光与热红外;多尺度特征;模态融合;深度学习
中图分类号:TP391 **文献标志码**:A

RGB-T object tracking network based on multi-scale modality fusion

CHENG Zhuxuan[1,2,4], FAN Huijie[2,4], TANG Yandong[2,4], WANG Qiang[3]
(1. College of Information Engineering, Shenyang University of Chemical Technology, Shenyang 110142, China;
2. State Key Laboratory of Robotics, Shenyang Institute of Automation Chinese Academy of Sciences, Shenyang 110016, China;
3. Key Laboratory of Manufacturing Industrial Integrated in Shenyang University, Shenyang 110044, China;
4. Institute of Robotics and Intelligent Manufacturing Innovation, Chinese Academy of Sciences, Shenyang 110016, China)

Abstract: RGB-T (RGB-Thermal) object tracking has received much attention in the field of object tracking because it is less restricted by lighting conditions. An RGB-T object tracking network was proposed to address the differences in resolution and semantic information of features at different scales, the inconsistency between visible and thermal infrared modal information, and the shortcomings of existing networks in multimodal fusion strategies. The network adopted Siamese structure to expand the template image features and search image features output by the backbone feature extraction network from single scale to multiple scales. The modal fusion for visible and thermal infrared modalities at different scales was performed separately. Then the obtained fused features were enhanced by the attention mechanism to enhance the feature representation. Finally, the prediction results were obtained by the region suggestion network. The experimental results on two publicly available RGB-T datasets, GTOT and RGBT-234, show that the network, with high tracking precision and success rate, can cope with complex tracking scenarios and has higher tracking performance compared with other networks.

Key words: object tracking; RGB-Thermal; multi-scale features; modality fusion; deep learning

收稿日期:2023-02-14
基金项目:国家自然科学基金项目(62273339);国家自然科学基金联合基金重点支持项目(U20A20200).
作者简介:程竹轩(1998—),男,河南鹤壁人,硕士研究生,主要从事多模态目标跟踪研究.
范慧杰(1985—),女,河南周口人,副研究员,硕士生导师,主要从事计算机视觉、图像处理研究,本文通信作者.
E-mail:fanhuijie@sia.cn

目标跟踪[1]是计算机视觉领域的一个热门研究课题,在视频监控、姿态分析、行为识别等领域被广泛应用,然而 RGB 跟踪器在遇到剧烈光照变化、低光照、雨天及大雾等跟踪场景时,输入图像的质量会受到很大影响,导致跟踪器性能严重下降,而基于可见光-热红外(RGB-Thermal,RGB-T)的多模态目标跟踪可以有效整合可见光与热红外图像信息,能克服单模态跟踪对可见光强度敏感的局限性,提高跟踪性能,因此相较于仅使用可见光模态信息进行跟踪来说,RGB-T 跟踪鲁棒性更高。然而,能否设计出高效的多模态融合策略、提取两种模态的优势特征并融合出一个包含二者优势互补信息的中间态特征,将直接影响 RGB-T 跟踪器的整体性能。

早期的模态融合策略中,一种方法是将两个模态的特征通过级联的方式进行融合,如 Zhang 等[2]提出一种基于多域卷积的 RGB-T 目标跟踪网络,先对卷积神经网络(convolutional neural network,CNN)输出的两种模态特征执行级联操作进行融合,再将得到的融合特征输入指定层进行二分类来识别目标;另一种方法是通过逐元素相加的方式进行融合,如 Zhang 等[3]提出一种基于模态感知的 RGB-T 目标跟踪网络,通过模态感知层获得一种中间模态,并将中间模态特征分别与可见光特征和热红外特征进行逐元素相加获得融合特征。这两种融合方式并未考虑到不同模态信息在不同跟踪场景下的差异,如在图 1 所示的跟踪场景中,可见光模态的质量明显高于热红外模态的质量,热红外模态无法提供有效信息,此时若将两模态特征进行级联或逐元素相加会引入无效信息,削弱可见光模态的信息,降低跟踪器性能。Zhu 等[4]提出一种新的融合方法,通过建立自适应聚合子网络,在跟踪过程中学习层权重和模态权重,网络在线跟踪速度仅为 1.3 帧/秒(frames per second,FPS),无法达到实时跟踪的效果;Zhang 等[5]将孪生跟踪网络引入 RGB-T 跟踪任务当中,利用一对孪生网络分别提取可见光与热红外特征,用级联方式融合,形成融合后的模板特征与搜索区域特征,并对这两种特征进行互相关操作得到最终的响应图,此方法跟踪速度较高,但由于缺少有效的融合策略以及未对多尺度特征信息加以利用,跟踪精度较低。

(a) 可见光模态 (b) 热红外模态

图 1 模态质量差距较大的跟踪场景

Fig. 1 Tracking scenarios with large modal quality gap

针对上述 RGB-T 目标跟踪的特点以及现有网络的不足,本研究提出一种基于多尺度模态融合的 RGB-T 目标跟踪网络,首先通过主干特征提取网络获得可见光与热红外各自的模板和搜索区域的特征,然后分别进行特征尺度拓展以引入不同语义级别的特征,并在 3 种尺度上分别进行模态信息融合,再将融合特征通过增强模块增强特征表示,最后通过区域建议网络(region proposal networks,RPN)获得预测结果。

1 相关工作

1.1 基于区域建议的孪生网络

Li 等[8]提出的基于区域建议的孪生网络(SiamRPN)由一对主干特征提取网络和区域建议网络组成,网络分为模板分支和搜索分支。主干特征提取网络用于获得模板和搜索区域的初始特征;区域建议网络由分

类分支和回归分支组成,分类分支用于区分跟踪目标与背景,回归分支用于调整候选框的大小与位置,为图像中的每个样本被预测为跟踪目标或是背景的概率提供一个置信分数,将主干特征提取网络输出的初始特征进行互相关操作。具体来说,该操作会将模板特征变成批大小×通道数个卷积核,每个卷积核的大小为模板特征的宽×高,再将搜索区域特征分为批大小×通道数个组后进行卷积操作,得到分类分支和回归分支对应的响应图,最后利用非极大抑制对候选框进行筛选得到跟踪结果。

1.2 基于 SiamRPN 的 RGB-T 目标跟踪网络

FSRPN(fuison SiamRPN tracker)[9]是一种基于 SiamRPN 的 RGB-T 目标跟踪网络,将 SiamRPN 拓展到可见光与热红外两种模态,该网络将特征叠加的融合策略应用于孪生网络框架,通过主干特征提取网络 ResNet-50[10]获得深层特征,并利用通道注意力对模板特征与搜索区域特征进行增强,然后将增强后的特征直接相加获得融合特征,再利用这些融合的深层特征与区域建议网络跟踪目标。该方法在一定程度上融合了两种模态的信息,但该网络仅使用深层特征,未考虑不同尺度特征对后续融合以及互相关操作的影响,忽视了两种模态特征间的差异性,难以获得鲁棒的融合特征,限制了网络的跟踪性能。

2 多尺度模态融合网络

2.1 网络整体结构

本研究提出的网络模型采用孪生网络结构,整体结构如图 2 所示。网络相较于基线网络 FSRPN,在多尺度特征的利用、模态融合、特征增强方面进行了改进。不同于基线网络 FSRPN 仅使用单尺度特征以及直接相加的模态融合方式,本研究设计了一个特定的模块,在将特征拓展到多个尺度的同时,在不同尺度上分别对可见光、热红外两种模态信息进行自适应融合,并根据模板图像和搜索区域图像的特征分布差异,使用带有残差的通道自注意力以及卷积块注意力模块(convolutional block attention module,CBAM)[11]改进基线网络中对融合特征增强的过程。

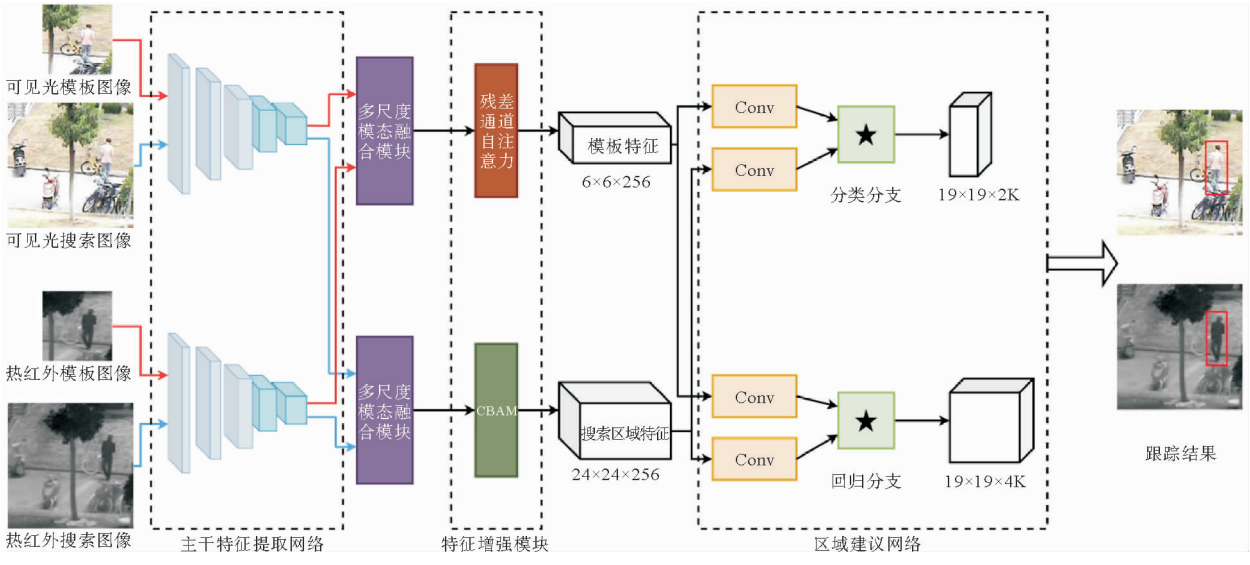


图 2 多尺度模态融合网络结构图

Fig. 2 Structure map of multi-scale modality fusion network

如图 2 所示,本研究提出的网络模型由一对 AlexNet[12]组成的主干特征提取网络、多尺度模态融合模块、特征增强模块和区域建议网络组成。考虑模型跟踪速度,采用一对 AlexNet 作为主干特征提取网络,用于提取可见光与热红外对应的模板图像和搜索区域图像的初始特征。多尺度模态融合模块主要实现对可见光与热红外对应的初始特征在 3 种尺度上的拓展,实现将两种模态信息在 3 种尺度上分别进行模态融合。特征增强模块由残差通道自注意力和 CBAM 并联组成,作用是对输入的特征图进行加权操作,增强目标区

域的特征表示,抑制背景信息的特征表示,提高目标区域特征在区域建议网络中的贡献,提升跟踪效果。区域建议网络用来获得分类分支和回归分支对应的响应图,最终输出跟踪结果。

2.2 主干特征提取网络

对于两种模态下图像的特征提取,采用一对非共享权值的 AlexNet 网络作为主干特征提取网络,其网络参数如表 1 所示。其中,Conv 表示卷积核大小不同的卷积,MaxPooling 表示全局最大池化。在对数据集中每段视频序列进行跟踪时,将首帧的目标中心作为该序列的模板图像,每次跟踪只对模板图像进行一次特征提取,后续跟踪过程不对模板特征进行在线更新,这样可以有效降低运算量,提高网络跟踪速度,同时跟踪目标也不受背景信息影响,在长时间跟踪中可以有效提高跟踪性能,避免遇到遮挡时在线更新学习到背景信息。

表 1 主干特征提取网络参数

Table 1 Backbone feature extraction network parameters

网络层	卷积核/池化核	数量	步长	模板图像	搜索区域图像
输入	—	—	—	127×127×3	255×255×3
Conv1	11×11	96	2	59×59×96	131×131×96
MaxPooling1	3×3	—	2	29×29×96	65×65×96
Conv2	5×5	256	1	25×25×256	61×61×256
MaxPooling2	3×3	—	2	12×12×256	30×30×256
Conv3	3×3	384	1	10×10×384	28×28×384
Conv4	3×3	384	1	8×8×384	26×26×384
Conv5	3×3	256	1	6×6×256	24×24×256

可见光与热红外对应的模板图像与搜索区域图像在进入主干特征提取网络之前预先裁剪成尺寸 127×127 和 256×256 大小,最终输出的特征尺度为 6×6×256 和 24×24×256,此过程可由式(1)表示:

$$A_{\mathrm{v}}^{t}=\varphi(z^{\mathrm{v}})\ ,\ A_{\mathrm{v}}^{s}=\varphi(x^{\mathrm{v}})\ 。\tag{1}$$

式中:z^{v} 与 x^{v} 分别表示可见光对应的模板图像与搜索区域图像,$\varphi()$表示特征提取操作。热红外分支处理过程同理。

使用 AlexNet 作为主干特征提取网络可以有效保证跟踪速度,并且 SiamDW[13] 通过实验表明,对于孪生网络,使用更深的主干网络并不能有效提升跟踪效果,因为更深层的主干网络在提取深层特征时会提高网络的感受野,而孪生网络的最佳感受野为整个输入图像的 60%~80%。此外,深层的主干网络还会降低特征间的区分度,导致跟踪性能降低。

2.3 多尺度模态融合模块

深层特征含有更加丰富的语义信息,但缺点是特征图的分辨率很低,无法充分表达对应的空间信息;浅层特征可以很好地表达特征中所包含的空间信息,但语义信息表达能力较弱。因此,如何将深层特征与浅层特征相结合,使不同尺度的特征起到互补效果是多尺度操作的关键。本研究提出一种多尺度模态融合模块,如图 3 所示。首先将初始特征经过卷积核大小为 1×1 的卷积调整特征的通道数,然后经过一个由若干残差卷积组成的瓶颈层(Bottleneck)改变特征的语义级别,不同级别的融合可以得到更加稳定的语义信息,利用稳定的语义信息可以使跟踪过程不再受到目标物体外观变化的影响,通过残差结构可以避免梯度消失所导致的退化问题,并且残差卷积可以通过构建恒等映射层以实现卷积层的自适应组合[14],从而构建出更加高效的卷积结构。此过程以可见光搜索区域图像为例,可由式(2)表示:

$$A_{\mathrm{v}}^{6\times6\times1024}=\mathrm{B}_1[\mathrm{L}(\mathrm{Conv}(A_{\mathrm{v}}^{s}))]\ ,\ A_{\mathrm{v}}^{12\times12\times512}=\mathrm{B}_2[\mathrm{L}(\mathrm{Conv}(A_{\mathrm{v}}^{s}))]\ 。\tag{2}$$

式中:B_1、B_2 表示瓶颈层,L 表示 LeakyReLU 激活函数,热红外分支对应的操作同理。通过此模块,模板图像分支与搜索区域图像分支将会各得到两种模态特征尺寸为 1 024×6×6、512×12×12、256×24×24 的共计 6 种特征。

在得到每个模态的不同尺度特征后,通过自适应融合方式在 3 种尺度上分别进行模态融合。不同模态信息各有优点:可见光图像可以提供丰富的背景信息,更好地区分目标与背景;热红外图像可根据热成像原理,全天候提供准确的目标轮廓信息。为了利用两种模态的互补信息,本研究通过生成模态权重的融合结构,对原特征进行自适应加权的方式融合两种模态信息,如图 4 所示。图 4 中,GAP 表示全局平均池化,FC 表示公共全连接层,FC_1 与 FC_2 为两个非共享权重的全连接层,w_v 与 w_t 表示可见光与热红外的模态权重,Cat 表示级联操作。

该结构中,首先将可见光与热红外特征叠加,然后利用全局平均池化以及全连接层和 Softmax 函数生成每个模态对应的自适应权重向量并加权。以 6×6×1 024 尺度的特征为例,该过程可表示为:

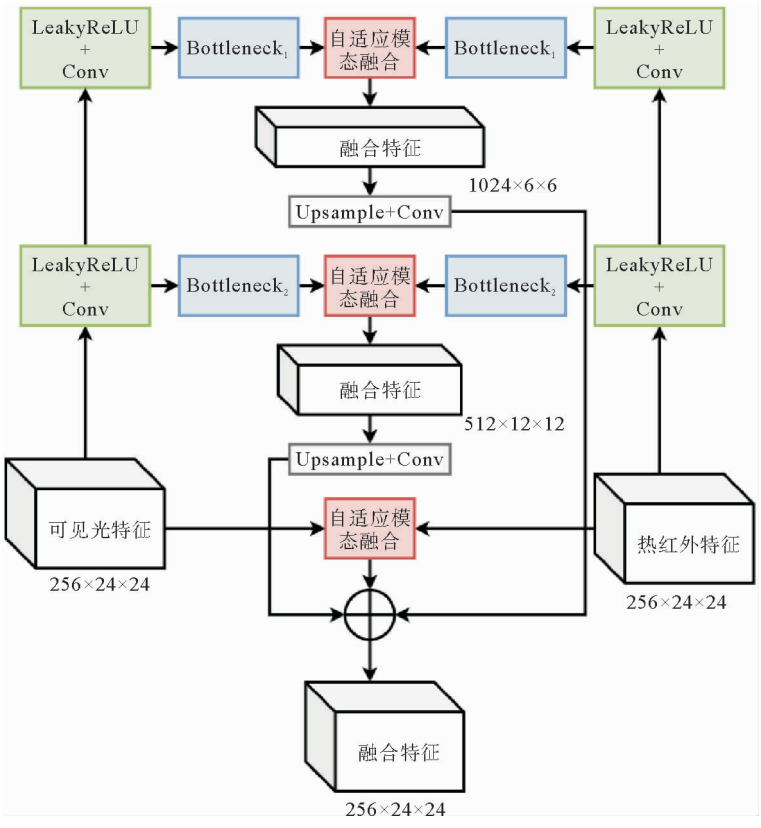


图 3　多尺度模态融合模块

Fig. 3　Multi-scale modality fusion module

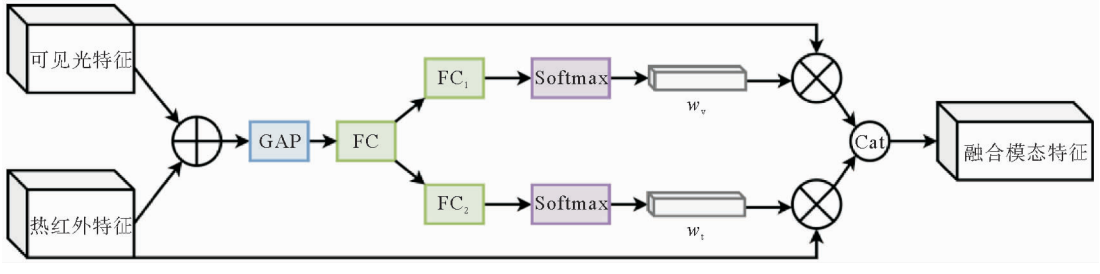


图 4　自适应模态融合

Fig. 4　Adaptive modality fusion

$$w_g = \mathrm{FC}(\mathrm{GAP}(A_v^{6\times6\times1\,024} + A_t^{6\times6\times1\,024})), \tag{3}$$

$$w_v = \frac{e^{FC_1(w_g)}}{e^{FC_1(w_g)} + e^{FC_2(w_g)}}, \quad w_t = \frac{e^{FC_2(w_g)}}{e^{FC_1(w_g)} + e^{FC_2(w_g)}}, \tag{4}$$

$$A_f = \mathrm{Cat}(w_v \times A_v^{6\times6\times1\,024} + w_t \times A_t^{6\times6\times1\,024})。 \tag{5}$$

式中:GAP 表示池化核大小为 1 的全局平均池化,FC 表示公共全连接层,w_g 表示公共权重,FC_1 与 FC_2 表示可见光与热红外各自的全连接层,w_v 与 w_t 分别表示各自生成的权重,Cat 表示级联操作,A_f 表示模块输出的每个尺度的融合特征。最后通过上采样与卷积操作使 3 种尺度特征归一化并叠加得到最终的融合特征,该融合策略通过自适应的方式避免了对有效模态信息的削弱,融合出的特征相较于级联和逐元素相加的方法有更强的鲁棒性。

2.4 特征增强模块

特征增强模块由残差通道自注意力和 CBAM 并联组成,通过特征增强模块从特征中学习权重分布,利用学到的权重分布,改变原特征的特征分布,从而达到增强目标特征并抑制背景特征的目的。Hu 等[15]提出一种通道自注意力结构,通过建模通道之间的关系自适应地改变通道特征分布。由于模板特征是目标最显著的特征,包含背景信息较少,使用深层的注意力机制会破坏模板图像的特征分布。本研究在通道自注意力的基础上,设计一种包含捷径连接的残差通道自注意力,在增强融合后的模板分支特征的同时,最大限度地保留其特征分布,结构如图 5 所示。图 5 中,Conv 表示卷积核为 1 的卷积。

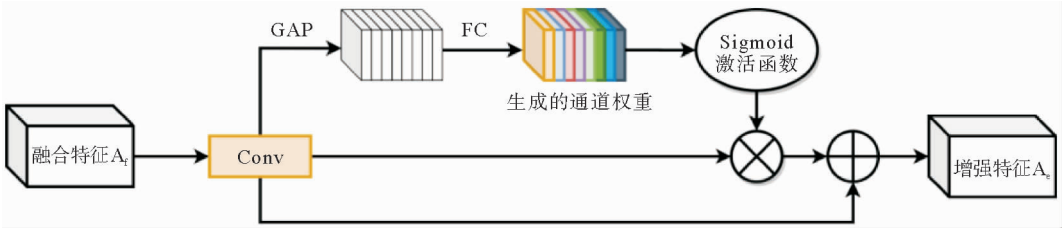


图 5 残差通道自注意力

Fig. 5 Residual channel self-attention

首先通过一个 1×1 的卷积调整输入特征的通道数,然后利用全局平均池化操作将特征的空间维度压缩成一个点,得到一个通道数维度的特征向量,之后通过全连接层与 Sigmoid 激活函数生成通道权重,并将权值向量对输入特征加权得到增强特征,最后与捷径连接相加得到最终输出,该过程可表示为:

$$A_e = \mathrm{Sigmoid}[\mathrm{FC}(\mathrm{GAP}(\mathrm{Conv}(A_f)))] \otimes \mathrm{Conv}(A_f) + \mathrm{Conv}(A_f) \text{ 。} \tag{6}$$

对于搜索区域图像,采用 CBAM 进行特征增强。CBAM 比通道自注意力的结构更加复杂,可以对特征在通道和空间位置两个方面进行增强,因在不同网络结构和不同任务中具有适用性强的特点,可在任何卷积神经网络架构中灵活使用,在计算量较小的同时增强特征的表达,其结构如图 6 所示。

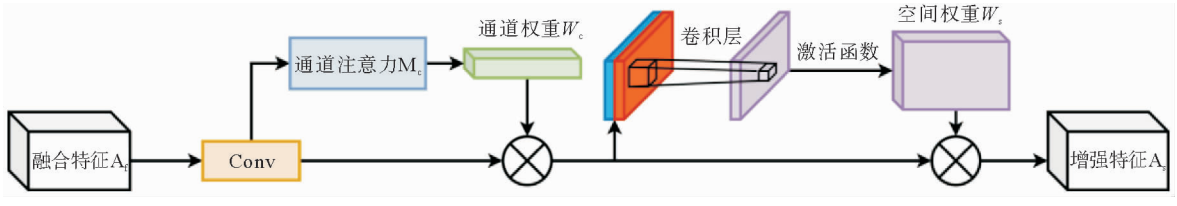


图 6 CBAM 结构图

Fig. 6 Framework map of CBAM

CBAM 会依次通过通道自注意力 M_c 和空间自注意力 M_s 求出对应的通道权重与空间权重,并对输入特征进行加权操作,得到增强后的特征,该过程可由式(7)表示:

$$A_c = M_c(A_f) \otimes A_f \text{ ，} A_s = M_s(A_c) \otimes A_c \text{ 。} \tag{7}$$

通过两种注意力结构的并联使用,可以有效地增强融合后的特征表示。

2.5 区域建议网络

区域建议网络首先由 Faster R-CNN[16]提出,可以根据输入的特征图在原图像上生成候选框,结构包含分类分支与回归分支,分别用于区分前景和背景以及对候选框位置进行回归。对特征增强模块输出的增强特征进行互相关操作后得到分类和回归响应图:

$$A^{\mathrm{cls}}_{w\times h\times 2k} = [A_s]_{\mathrm{cls}} \bigstar [A_e]_{\mathrm{cls}} \text{ ，} A^{\mathrm{reg}}_{w\times h\times 4k} = [A_s]_{\mathrm{reg}} \bigstar [A_e]_{\mathrm{reg}} \text{ 。} \tag{8}$$

式中:★表示互相关操作,分类响应图上的每个点都是一个通道数为 2 000 的向量,代表原图像上锚点属于正样本或是负样本,即目标或是背景信息。而回归响应图上的每个点都是一个通道数为 4 000 的向量,代表锚点在原图像上的位置信息。本研究设置锚点数量为 5,其宽高比分别为 3、2、1、1/2、1/3,网络得到的响应图尺寸为

19×19,则原图像上的锚点数为 1 805,之后通过非极大抑制进行筛选,计算所有锚点对应锚框与目标框真值的交并比,其中大于 0.6 为正样本,小于 0.3 为负样本,最终选择出 16 个正样本和 48 个负样本供网络学习。

2.6 损失函数

采用交叉熵函数作为分类分支的损失函数,采用 L_1 平滑损失作为回归分支的损失函数,定义候选框与目标框真值之间的标准距离为:

$$\lambda_0=\frac{g_x-d_x}{d_w},\lambda_1=\frac{g_y-d_y}{d_h},\lambda_2=\ln\frac{g_w}{d_w},\lambda_3=\ln\frac{g_h}{d_h}。\tag{9}$$

式中:g_x、g_y、g_w、g_h 为目标框真值的坐标,d_x、d_y、d_w、d_h 为锚点相较于目标框真值的偏移量。L_1 平滑损失为:

$$L_1(\lambda)=\begin{cases}0.5\lambda^2,|\lambda|<1;\\|\lambda|-0.5,|\lambda|\geqslant 1。\end{cases}\tag{10}$$

因此,回归分支的损失可以表示为:

$$L_{\text{reg}}=\sum_{i=0}^{3}L_1(\lambda_i),\tag{11}$$

分类分支的损失可以表示为:

$$L_{\text{cls}}=-\frac{1}{N}\sum_{i}^{N}y_i\log_2 p_i+(1-y_i)\log_2(1-p_i)。\tag{12}$$

式中:N 为样本数量;p_i 是网络对样本的预测值;y_i 是样本的真实标签,若为正样本则 y_i 为 1,若为负样本则 y_i 为 0。

网络总的损失函数 L 可以表示为:

$$L=L_{\text{cls}}+\mu L_{\text{reg}}。\tag{13}$$

式中,μ 为控制两种损失函数平衡的超参数,用来确保二者在数值上处于同一数量级,设为 1。

3 实验结果与分析

3.1 实验数据集

将提出的目标跟踪网络在 GTOT、RGBT-234 两个公开的 RGB-T 目标跟踪数据集上进行实验。其中,GTOT 包括 50 个不同场景下的可见光与热红外视频序列,共 7 500 个帧对,每帧图片都由人工进行真实边界框的标注,并且视频中包含了遮挡、尺寸变化、快速移动、低光照、热红外交叉、小目标、形变等 7 种挑战属性;RGBT-234 是一个规模更大更复杂的数据集,包括 234 个不同场景下的可见光与热红外视频序列,共 117 000 个帧对,并且视频中包含了 12 种挑战属性。LasHeR[17]是一个大型数据集,由 1 224 个可见光和热红外视频序列和 730 000 个帧对组成,其目标类别达到 32 个。

3.2 实验配置

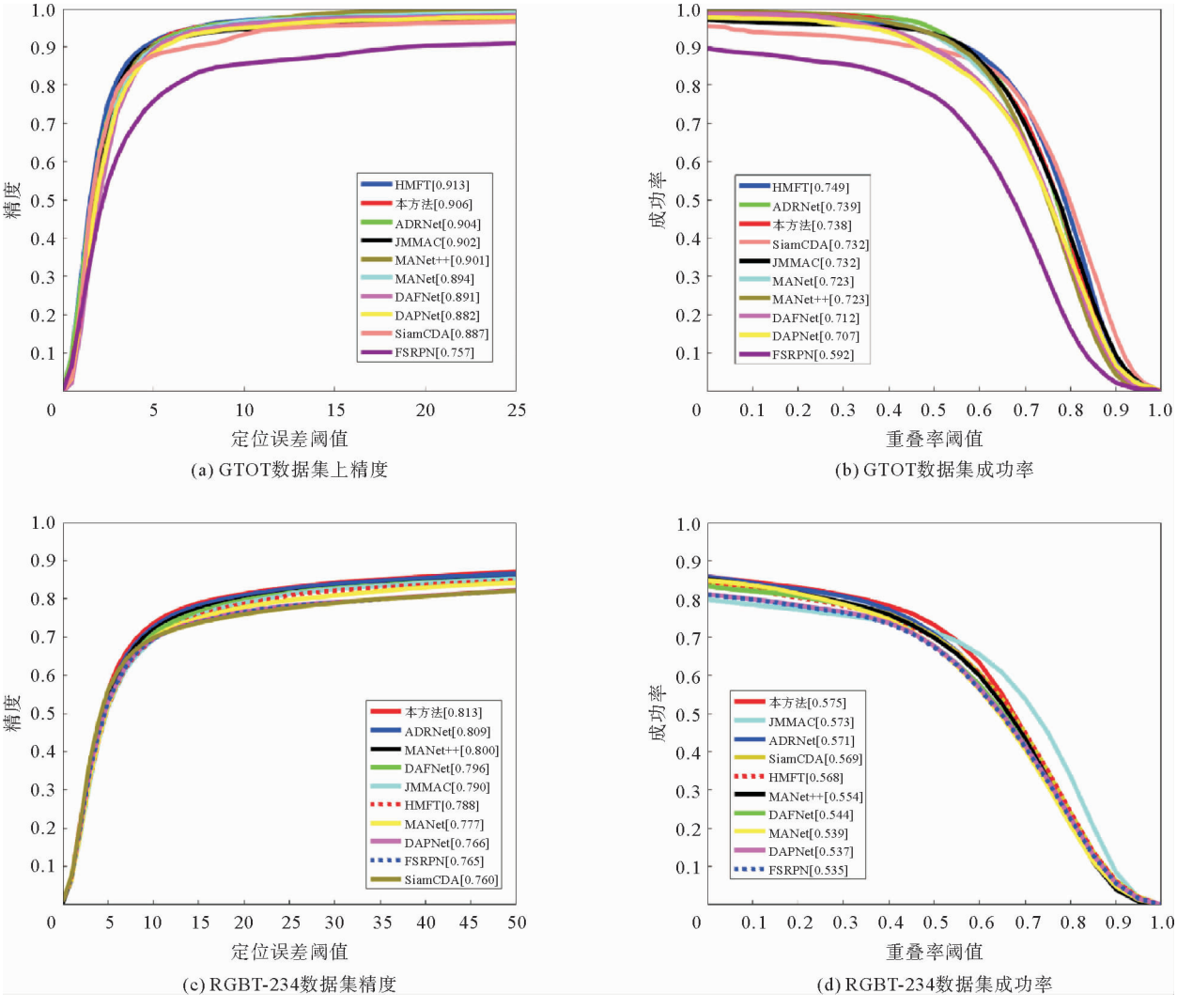
所提出的网络基于深度学习框架 Pytorch 实现,实验使用的软件环境为 Windows 10,CUDA 11.0.197,Python 3.7,硬件配置为 NVIDIA TITAN XP;网络初始学习率为 0.01,随着训练的进行衰减至 0.000 01,整个网络采用端到端的方式训练迭代 50 次,批大小设置为 32;使用 AlexNet 的预训练参数对 Conv1、Conv2、Conv3 的参数进行初始化,并在前 10 次迭代中冻结这 3 层参数以防止权值被破坏,使用随机梯度下降法优化损失函数;使用 RGBT-234 和 LasHeR 作为数据集训练一个网络,并在 GTOT 数据集上测试,然后使用 GTOT 和 LasHeR 数据集训练另一个网络,并在 RGBT-234 上测试。

3.3 评价指标

本研究采用 RGB-T 跟踪中最常用的精度(precision rate,PR)和成功率(success rate,SR)两种指标来评估所提出的 RGB-T 目标跟踪网络的性能。精度是在给定的距离阈值内预测边界框的中心与目标真实边界框中心距离小于阈值的图像帧数占所有帧数的比值,成功率是预测边界框与目标真实边界框之间的交并比大于阈值的图像帧数占所有帧数的比值。两种指标数值越高,表示网络的跟踪性能越好。

3.4 实验分析

在 GTOT 数据集和 RGBT-234 数据集上对网络进行测试,并将实验结果与已有的先进方法(HMFT[18]、ADRNet[19]、JMMAC[20]、FSRPN、MANet++[21]、DAFNet[22]、DAPNet[23]、SiamCDA[24])进行对比,对比结果如图 7 所示,图注中每种方法后的数值表示该方法在不同阈值下的平均精度或平均成功率。可以看出,在 GTOT 和 RGBT-234 数据集中,本研究所提网络的精度和成功率分别比基线网络 FSRPN 高 14.9%、14.6%和 4.8%、4%,证明了本研究所提网络结构的有效性,并且在测试过程中的平均跟踪帧率为 37 FPS,可以达到实时跟踪的效果。



(a) GTOT数据集上精度 (b) GTOT数据集成功率

(c) RGBT-234数据集精度 (d) RGBT-234数据集成功率

图 7 不同网络在两个数据集上的对比结果

Fig. 7 Comparison results of different networks on two datasets

RGBT-234 数据集包含 12 种挑战属性,分别为背景、相机移动、形变、快速移动、严重遮挡、低光照、低分辨率、运动模糊、无遮挡、部分遮挡、尺度变化、热红外交叉,与其他网络的对比结果如表 2 所示。表 2 中每种挑战表现最优结果以黄色表示,次优结果以蓝色表示。

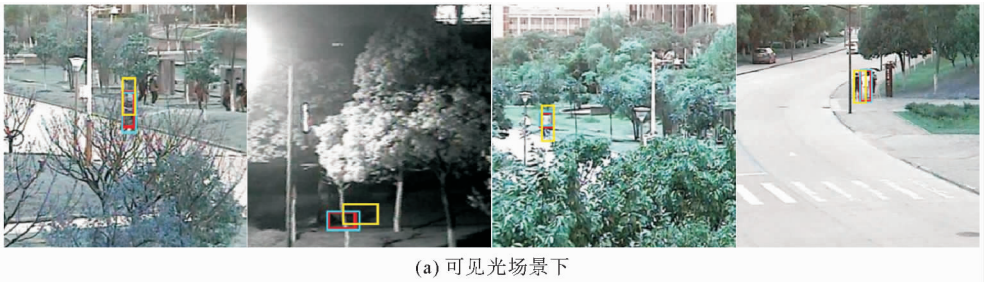
从表 2 可以看出,所提出的网络在绝大多数挑战属性中的表现优于基线网络 FSRPN 及其他网络,背景、形变、严重遮挡、热红外交叉等 4 种属性优于其他所有网络,表明通过多尺度模态融合以及对融合后模态特征的增强为网络提供了目标更加丰富的语义信息和细节特征,可以有效解决目标形变、快速移动、严重遮挡等导致的目标跟踪性能不佳的问题。

表 2 不同网络在 RGBT-234 不同挑战属性下的 PR/SR 结果对比

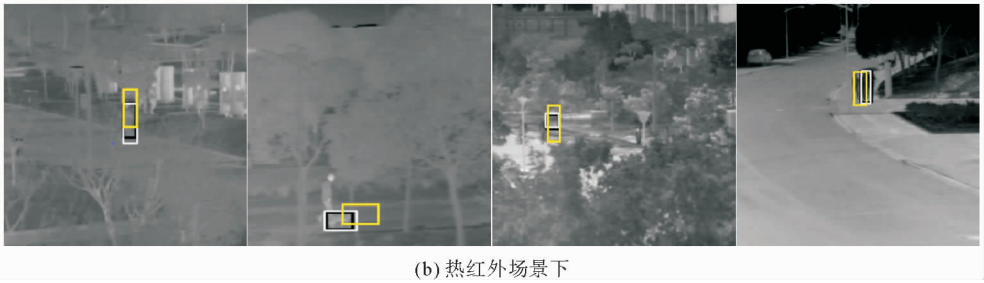
Table 2 Comparison of PR/SR results of different networks under different challenge attributes of RGBT-234

挑战属性	FSRPN		SianCDA		HMFT		JMMAC		MANet++		DAFNet		ADRNet		本研究方法	
	PR	SR	PR	SR	PR	SR	PR	SR	PR	SR	PR	SR	PR	SR	PR	SR
背景	0.716	0.479	0.739	0.527	0.730	0.485	0.687	0.485	0.752	0.472	0.791	0.493	0.789	0.527	0.811	0.539
相机移动	0.666	0.473	0.733	0.548	0.769	0.549	0.762	0.556	0.729	0.503	0.723	0.506	0.757	0.535	0.757	0.545
形变	0.717	0.515	0.749	0.574	0.763	0.567	0.706	0.529	0.777	0.543	0.741	0.516	0.743	0.529	0.778	0.559
快速移动	0.664	0.440	0.614	0.454	0.653	0.458	0.610	0.417	0.678	0.438	0.740	0.465	0.776	0.503	0.774	0.517
严重遮挡	0.658	0.456	0.675	0.495	0.656	0.457	0.677	0.483	0.703	0.466	0.686	0.459	0.708	0.491	0.730	0.506
低光照	0.775	0.528	0.822	0.603	0.815	0.572	0.840	0.588	0.772	0.511	0.812	0.542	0.802	0.551	0.819	0.565
低分辨率	0.750	0.506	0.709	0.499	0.756	0.502	0.771	0.517	0.778	0.505	0.818	0.538	0.831	0.556	0.819	0.562
运动模糊	0.653	0.467	0.635	0.474	0.696	0.498	0.751	0.549	0.720	0.501	0.708	0.500	0.727	0.530	0.710	0.513
无遮挡	0.900	0.640	0.889	0.668	0.901	0.663	0.932	0.694	0.884	0.643	0.900	0.636	0.917	0.658	0.922	0.672
部分遮挡	0.821	0.572	0.788	0.600	0.845	0.607	0.841	0.611	0.814	0.568	0.859	0.588	0.863	0.612	0.846	0.593
尺寸变化	0.780	0.540	0.743	0.568	0.793	0.582	0.837	0.616	0.781	0.555	0.791	0.544	0.790	0.562	0.814	0.571
热红外交叉	0.769	0.543	0.680	0.479	0.714	0.494	0.749	0.526	0.761	0.548	0.811	0.583	0.789	0.589	0.867	0.621
全部	0.765	0.535	0.760	0.569	0.788	0.568	0.790	0.573	0.781	0.540	0.796	0.544	0.809	0.571	0.813	0.575

图 8 展示了本研究提出的网络在 4 个复杂跟踪场景下的跟踪效果,其中蓝色框与白色框为预测边界框,红色框与黑色框为目标真实边界框,黄色框为 FSRPN 的预测边界框。



(a) 可见光场景下



(b) 热红外场景下

图 8 网络在 4 个复杂跟踪场景下的跟踪效果

Fig. 8 Tracking effect of network in four complex tracking scenarios

3.5 消融实验

为了验证网络中各个模块的有效性,本研究在 RGBT-234 数据集上进行消融实验,实验设计如下。

1) Our-ATO。仅使用特征增强模块,移除网络中的多尺度模态融合模块;

2) Our-MSO。仅使用多尺度模态融合模块,移除网络中特征增强模块;

3) Our-MSO-A。移除多尺度模态融合模块中的尺度拓展操作;

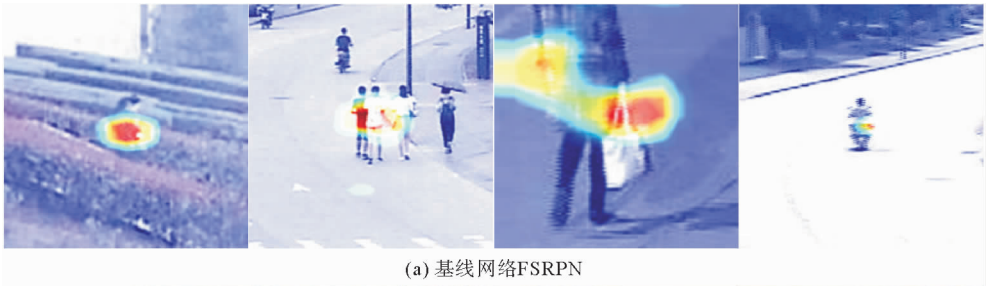
4) Our-MSO-B。移除多尺度模态融合模块中的自适应模态融合,并以特征级联替代。

表 3 为 RGBT-234 数据集上消融实验结果。由表 3 可见,Our-ATO 和 Our-MSO 的 PR 和 SR 指标均高于基线网络 FSRPN,表明两个模块的有效性,Our-MSO-A 和 Our-MSO-B 的结果均低于 Our-MSO,表明多尺度模态融合模块中的尺度拓展以及模态融合的有效性。为了更直观地展示二者对于网络性能的提升效果,图 9 以响应热力图的方式展示了 4 个跟踪场景下网络输出响应的比较,可以看出,在多尺度模态融合模块与特征增强模块的作用下,响应位置更趋近于目标中心区域,表明两个模块可以为网络提供有效的多尺度融合模态信息以获得更准确的响应,从而提高跟踪精度。

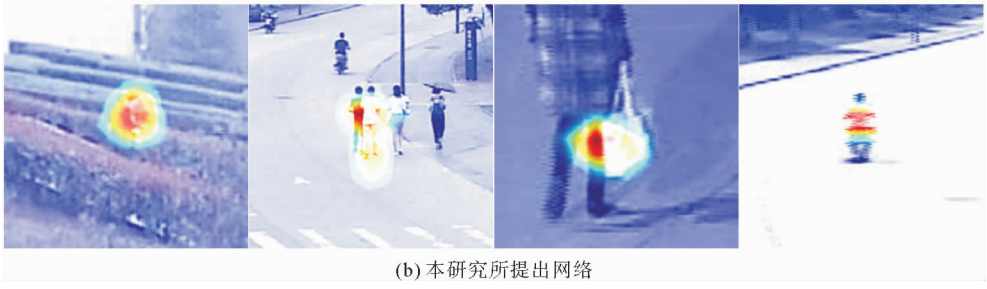
表 3 RGBT-234 数据集上消融实验结果

Table 3 Ablation experimental results on RGBT-234

方法	RGBT-234 数据集	
	PR	SR
FSRPN	0.765	0.535
Our-ATO	0.775	0.549
Our-MSO	0.784	0.552
Our-MSO-A	0.769	0.542
Our-MSO-B	0.771	0.550
本研究方法	0.813	0.575



(a) 基线网络FSRPN



(b) 本研究所提出网络

图 9 网络输出的响应热力图比较

Fig. 9 Comparison of network output response thermograms

4 结论

本研究提出的 RGB-T 目标跟踪网络可以在不同尺度融合两种模态信息获得更加鲁棒的模态互补特征,并且通过特征增强模块进一步增强特征表示,可以有效应对目标快速移动、目标遮挡、热红外交叉等复杂跟踪场景。在两个 RGB-T 跟踪数据集上的实验结果表明,本网络与其他网络相比具有更高的跟踪性能,可以通过双模态信息互补的方式获得更加准确的目标响应,提高不同场景、不同挑战下的跟踪效果。

未来考虑通过改进网络结构提升运动模糊以及像机移动跟踪场景下对目标特征的捕捉能力。

参考文献:

[1] 姚云翔,陈莹. 注意力机制下双模态交互融合的目标跟踪网络[J]. 系统工程与电子技术,2022,44(2):410-419.
YAO Yunxiang,CHEN Ying. Object tracking network based on dual-modal interactive fusion under attention mechanism [J]. Journal of Systems Engineering and Electronics,2022,44(2):410-419.

[2] ZHANG X M,ZHANG X H,DU X D,et al. Learning multi-domain convolutional network for RGB-T visual tracking[C]// 2018 11th International Congress on Image and Signal Processing. Beijing,Oct. 13-15,2018:1-6.

[3] ZHANG H,ZHANG L,ZHUO L,et al. Object tracking in RGB-T videos using modal-aware attention network and com-

petitive learning[J/OL]. Sensors,2020,20(2). DOI:10.3390/s20020393.
[4] ZHU Y B,LI C L,TANG J,et al. Quality-aware feature aggregation network for robust RGBT tracking[J]. IEEE Transactions on Intelligent Vehicles,2021,6(1):121-130.
[5] ZHANG X C,YE P,PENG S Y,et al. SiamFT:An RGB-infrared fusion tracking method via fully convolutional Siamese Networks[J]. IEEE Access,2019,7:122122-122133.
[6] LI C L,CHENG H,HU S Y,et al. Learning collaborative sparse representation for grayscale-thermal tracking[J]. IEEE Transactions on Image Processing,2016,25(99):5743-5756.
[7] LI C L,LIANG X Y,LU Y J,et al. RGB-T object tracking:Benchmark and baseline[J/OL]. Pattern Recognition,2019,96. DOI:10.1016/j.patcog.2019.106977.
[8] LI B,YAN J J,WU W,et al. High performance visual tracking with Siamese region proposal network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City,2018:8971-8980.
[9] LI H,WU X J,KITTLER J,et al. The seventh visual object tracking VOT2019 challenge results[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul,2019:2206-2241.
[10] HE K M,ZHANG X Y,REN S Q,et al. Deep residual learning for image recognition[C]//2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas,2016:770-778.
[11] WOO S H,PARK J C,LEE J Y,et al. CBAM:Convolutional block attention module[C]//European Conference on Computer Vision. Cham,2018:3-19.
[12] KRIZHEVSKYA,SUTSKEVER I,HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Association for Computing Machinery. New York,2012:1097-1105.
[13] ZHANG Z P,PENG H W. Deeper and wider Siamese networks for real-time visual tracking[C]//2019 IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach,2019:4586-4595.
[14] 徐岩,李晓振,吴作宏,等. 基于残差注意力网络的马铃薯叶部病害识别[J]. 山东科技大学学报(自然科学版),2021,40(2):76-83.
XU Yan,LI Xiaozhen,WU Zuohong,et al. Potato leaf disease recognition via residual attention network[J]. Journal of Shandong University of Science and Technology(Natural Science),2021,40(2):76-83.
[15] HU J,SHEN L,SUN G,et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,42(8):2011-2023.
[16] REN S Q,HE K M,GIRSHICK R B,et al. Faster R-CNN:Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.
[17] LI C L,XUE W L,JIA Y Q,et al. LasHeR:A large-scale high-diversity benchmark for RGBT tracking[J]. IEEE Transactions on Image Processing,2021,31:392-404.
[18] ZHANG P Y,ZHAO J,WANG D,et al. Visible-thermal UAV tracking:A large-scale benchmark and new baseline[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans,2022:8876-8885.
[19] ZHANG P Y,WANG D,YANG X. Learning adaptive attribute-driven representation for real-time RGB-T tracking[J]. International Journal of Computer Vision,2021,129(9):2714-2719.
[20] ZHANG P Y,ZHAO J,WANG D,et al. Jointly modeling motion and appearance cues for robust RGB-T tracking[J]. IEEE Transactions on Image Processing,2021,30:3335-3347.
[21] LU A D,LI C L,YAN Q Y,et al. RGBT Tracking via multi-adapter network with hierarchical divergence loss[J]. IEEE Transactions on Image Processing,2021,30:5613-5625.
[22] GAO Y,LI C L,ZHU Y B,et al. Deep adaptive fusion network for high performance RGBT tracking[C]// 2019 IEEE/CVF International Conference on Computer Vision Workshop(ICCVW). Seoul,2019:91-99.
[23] ZHU Y B,LI C L,LUO B,et al. Dense feature aggregation and pruning for RGBT tracking[C]// Proceedings of the 27th ACM International Conference on Multimedia. Nice,2019:21-25.
[24] ZHANG T,LIU X,ZHANG Q,et al. SiamCDA:Complementarity-and distractor-aware RGB-T tracking based on Siamese network[J]. IEEE Transactions on Circuits and Systems for Video Technology,2022,32(3):1403-1417.

(责任编辑:傅　游)