文章编号:1674-2974(2025)04-0068-11 DOI:10.16339/j.cnki.hdxbzkb.2025267

基于孪生网络的特征融合位移RGB-T目标跟踪

李海燕[1],曹永辉[1],郎恂[1†],李海江[2]
(1. 云南大学 信息学院,云南 昆明,650000;
2. 云南交通投资建设集团有限公司,云南 昆明 650000)

摘 要:为解决现有目标跟踪算法深层次特征提取困难、不能充分利用跨模态信息以及目标特征表示较弱等问题,提出了基于孪生网络的特征融合位移RGB-T目标跟踪算法.首先,基于可见光模态SiameseRPN++的目标跟踪框架,扩展设计红外模态分支,以获得多模态目标跟踪框架,设计了改进步长的ResNet50作为特征提取网络,有效挖掘目标的深层次特征.随后,设计特征交互学习模块,利用一种模态的判别信息引导另一种模态的目标外观特征学习,挖掘特征空间和通道中的跨模态信息,增强网络对前景信息的关注.然后,设计多模特征融合模块计算输入的可见光图像和红外图像的特征融合度,对不同模态的重要特征进行空间融合以去除冗余信息,并采用级联融合策略重建多模态图像,增强目标特征表示.最后,设计特征空间位移模块,分割红外模态分支的特征图并向四个不同方向移位,增强热源目标特征的边缘表示.在两个RGB-T数据集上的实验验证了提出算法的有效性,消融实验证明了设计的单个模块的优越性.

关键词:RGB-T跟踪;多模特征融合模块;特征空间位移模块;特征交互学习模块
中图分类号:TP391.4 **文献标志码**:A

FSSiamNet:Feature Fusion Shift Siamese Network for RGB-T Target Tracking

LI Haiyan[1], CAO Yonghui[1], LANG Xun[1†], LI Haijiang[2]
(1. School of Information Science and Engineering, Yunnan University, Kunming 650000, China;
2. Yunnan Communications Investment and Construction Group Co., Ltd., Kunming 650000, China)

Abstract: To solve the problems of the existing target tracking algorithms, such as inability to extract deep-level features, failure to fully exploit cross-modal information, and weak representation of target features, a feature fusion shift Siamese network for RGB-T target tracking is proposed. First, a target tracking framework based on the visible modal SiameseRPN++ is designed to extend the infrared modal branch, in order to obtain a multimodal target tracking framework. Moreover, the improved ResNet50 network with adjusted stride as a feature extraction network

* **收稿日期**:2024-01-23
基金项目:国家自然科学基金资助项目(62166048),National Natural Science Foundation of China(62166048);云南省万人计划"云岭教学名师","Yunling Famous Teacher" of Yunnan 10 000 Persons Program;云南省高校重点实验室建设计划资助项目(202101AS070031),Key Laboratory Construction Plan of Universities in Yunnan Province(202101AS070031);云南大学第十四届研究生科研创新项目(KC-22221737),The 14th Graduate Student Research and Innovation Project at Yunnan University(KC-22221737)
作者简介:李海燕(1976—),女,云南红河人,云南大学教授,博士
† 通信联系人,E-mail: langxun@ynu.edu.cn

enables the acquisition of deep-level features of the target. Subsequently, a multimodal feature interactive learning module (FIM) is designed to leverage the discriminative information from one modality to guide the learning process of target appearance features in the other modality. By mining the cross-modal information within the feature space and channels, the module enhances the network's attention towards foreground information. Thereafter, a multimode feature fusion module (FAM) is designed, which calculates the degree of feature fusion between the input visible light image and the infrared image, enabling spatial fusion of significant features from different modalities to effectively eliminate redundant information and reconstructing multimodal images by employing a cascade fusion strategy. Finally, a feature space shift module (FSM) is designed, which divides the feature maps of the infrared modal branches and shifts them in four different directions to enhance the edge representation of the heat source target. Extensive experiments on two RGB-T datasets thoroughly validate the effectiveness of the proposed algorithm, while ablation experiments demonstrate the superiority of each designed module.

Key words: RGB-T tracking; multi mode feature fusion module; feature space shift module; feature interactive learning module

目标跟踪旨在确定视频序列第一帧中目标的位置,在后续帧中准确定位该位置.近年来,基于深度学习的目标跟踪技术取得了一些进展,如ATOM[1]、DiMP[2]和MDNet[3]等深度单模跟踪算法在跟踪精度和鲁棒性方面均有较好的效果.然而,单模跟踪算法用于低能见度、低光照和遮挡等复杂场景时,性能急剧下降.为了弥补单模跟踪算法的不足,研究者们提出了RGB-T(可见光及红外)跟踪算法.早期的RGB-T跟踪算法依赖人工提取特征[4-7],无法捕捉目标的多尺度特征,因此在目标外观变化、被遮挡和快速移动环境中的性能不佳.为了克服人工提取特征的缺点,文献[8]提出了一种基于深度学习的响应级融合算法,引入联合通道注意力模块,在特征提取阶段为不同特征通道重新分配对应的权重.文献[9]构建了双流级联深度学习网络,将跟踪任务建模为相似度量任务,该算法设计了多模区域子网络和候选区域选择策略,有效提高了跟踪精度.文献[10]设计了一种多适配性跟踪卷积网络,设计了不同模态特有特征的并行子网络结构,使网络能够充分学习到具有区分能力的特征.但上述算法的特征提取网络均采用浅层卷积神经网络,导致目标的运动模糊特征和形变特征丢失,限制了跟踪算法性能提升.

为了提取目标的深层次特征,文献[11]集成了深度网络[12],以获得可见光和红外两种模态的深层特征.但该方法设计的自适应融合策略未能有效挖掘跨模信息,导致跨模态互补信息利用率较低.为了充分利用不同模态间的跨模特征,文献[13]设计了多模态互补感知融合模块,利用不同模态的互补特征,提高网络的特征学习能力.文献[14]提出了一种新的多分支自适应融合网络,该网络在提取到多尺度特征后,通过多分支自适应融合模块,融合了不同模态之间的跨模互补特征.然而,上述两种方法未能有效消除不同模态之间的非互补冗余信息,导致目标特征表示能力较弱.针对此问题,文献[15]探索了不同的融合方案,提出了一种决策级动态加权融合方法,一定程度上消除了可见光和红外模态之间的冗余信息,减少了非互补信息的干扰,增强了可见光目标特征表示.但是,由于红外模态自身特征表示能力较弱,热源目标区域发散和目标边界模糊,因此该融合方法不能有效增强红外模态分支目标特征表示.

综上所述,目前大多数基于孪生网络的RGB-T跟踪算法都面临以下问题:1)如何集成深度网络提取深层目标特征? 2)如何挖掘红外及可见光的跨模态特征? 3)如何增强红外模态分支特征,消除不同模态之间的冗余信息,增强目标特征表示? 为了解决这些问题,本文提出了基于孪生网络的特征融合位移RGB-T目标跟踪算法.针对第一个问题:本文采用了ResNet50[12]作为特征提取网络,提取可见光和红外模态的深层特征.为了有效提取红外及可见光的深度特征,将特征提取网络第四层和第五层的步长统一调整为8像素.针对第二个问题,设计了特征交互学习模块有效获取红外模态特征,利用这些特征引导可见光模态的特征学习,并将学习到的可

见光模态特征引导红外模态特征学习.通过双向特征指导学习方式充分挖掘跨模态信息.针对第三个问题,本文设计了特征融合模块和特征空间位移模块.特征融合模块计算输入可见光图像和红外图像的特征融合度,将红外模态和可见光模态的重要特征进行空间融合,增强特有特征细节,同时消除非互补信息的干扰,增强目标特征表示.特征空间位移模块将红外模态特征分支按照通道数划分为四组,在四个方向上分组空间位移,使红外特征元素之间形成错位,以增强红外模态特征边缘细节.

1 基于孪生网络的特征融合位移 RGB-T 目标跟踪法

1.1 基线跟踪算法

孪生跟踪网络采用视频序列的第一帧作为模板帧,后续帧作为检测帧,并在整个跟踪过程中,模板帧保持不变.本文提出的跟踪算法由两个 SiamRPN++[16]网络组成,每个 SiamRPN++网络中包含一个模板分支和一个检测分支,它们共享相同的网络参数.在跟踪阶段,通过相同的特征提取网络分别对模板图像和检测图像进行分层特征提取.之后将获得的分层特征送入颈部模块进行通道数调整,得到尺寸为 7×7×256 和 33×33×256 的分层特征图.在多个 RPN 块中对分层特征图进行分类和回归计算,网络模型如图 1 所示.图 1 的左半部分是网络的整体结构图,模板分支和检测分支采用相同的特征提取网络.图 1 的右侧部分是对 X_n 模块的具体示意图,如蓝色部分所示.以图 1 检测分支为例,网络处理流程如下:可见光图像和红外图像首先通过多模特征融合模块进行去噪融合,消除非互补冗余信息的影响.随后,通过特征提取网络分别提取可见光模态和红外模态的分层特征.接着,将红外模态分层特征通过特征空间位移模块进行特征分组和空间位移,以增强热源目标边缘表示.最后,通过特征交互学习模块利用来自不同模态的互补特征,并与模板分支特征进行跨模态特征交互学习,最终生成用于目标分类和边界框回归的多维特征向量.跟踪网络的计算过程定义如下:

$$\begin{cases} S_i = \sigma_i^{2K}(H(f_i(T)), H(h_i(D))) \\ R_i = \sigma_i^{4K}(H(f_i(T)), H(h_i(D))) \end{cases} \tag{1}$$

$$\begin{cases} S = \sum_{i=3}^{5} \alpha_i S_i \\ R = \sum_{i=3}^{5} \beta_i R_i \end{cases} \tag{2}$$

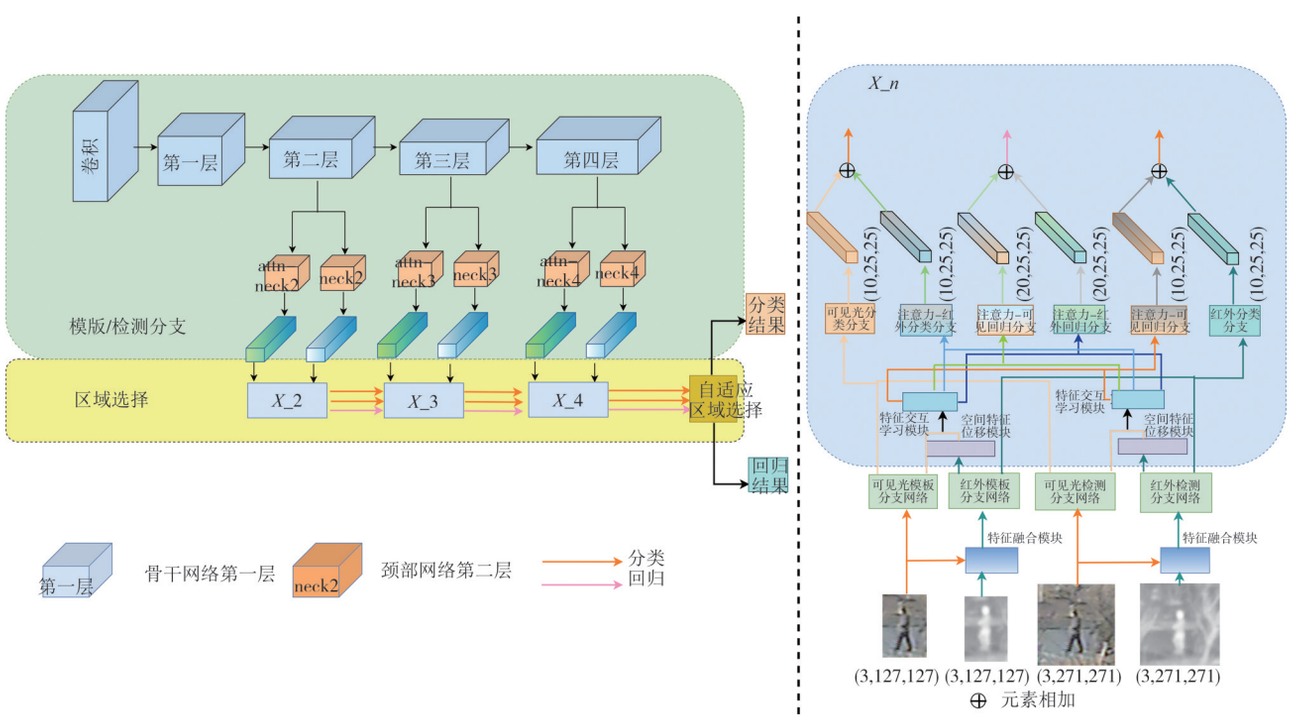


图 1 特征融合位移孪生网络

Fig.1 Feature fusion shift siamese network model

式中:T表示模板图像;D表示检测图像;f_i和h_i分别表示模板分支特征和检测分支特征;$H(\cdot)$表示颈部模块输出特征;$\sigma_i^{2K}(\cdot)$和$\sigma_i^{4K}(\cdot)$分别表示第i个RPN模块的分类结果和回归结果;α_i和β_i分别为第i层的分类线性权重和回归线性权重.

为了适应扩展出的红外模态分支,本文提出了新的区域选择网络,设计了一个新的分类分支和回归分支,前者用于分类目标和背景,后者用于预测锚框位置.计算过程如下:

$$\begin{cases} S_i^{\mathrm{RGB}} = \varphi_i^{2K}(H(f_i(T_{\mathrm{RGB}})), H(h_i(D_{\mathrm{RGB}}))) \\ R_i^{\mathrm{RGB}} = \varphi_i^{4K}(H(f_i(T_{\mathrm{RGB}})), H(h_i(D_{\mathrm{RGB}}))) \end{cases} \tag{3}$$

$$\begin{cases} S_i^{\mathrm{T}} = \varphi_i^{2K}(H(f_i(T_{\mathrm{T}})), H(h_i(D_{\mathrm{T}}))) \\ R_i^{\mathrm{T}} = \varphi_i^{4K}(H(f_i(T_{\mathrm{T}})), H(h_i(D_{\mathrm{T}}))) \end{cases} \tag{4}$$

$$\begin{cases} S^{\mathrm{RGB}} = \sum_{i=3}^{5} \alpha_i^{\mathrm{RGB}} S_i^{\mathrm{RGB}} \\ R^{\mathrm{RGB}} = \sum_{i=3}^{5} \beta_i^{\mathrm{RGB}} R_i^{\mathrm{RGB}} \end{cases} \tag{5}$$

$$\begin{cases} S^{\mathrm{T}} = \sum_{i=3}^{5} \alpha_i^{\mathrm{T}} S_i^{\mathrm{T}} \\ R^{\mathrm{T}} = \sum_{i=3}^{5} \beta_i^{\mathrm{T}} R_i^{\mathrm{T}} \end{cases} \tag{6}$$

式中:T_{RGB}和D_{RGB}分别表示可见光模板图像块和可见光检测图像块;T_{T}和D_{T}分别表示红外模板图像块和红外检测图像块;$\varphi_i^{2K}(\cdot)$和$\varphi_i^{4K}(\cdot)$分别表示第i个X_n模块输出的分类值和回归值;α_i^{RGB}和β_i^{RGB}分别表示第i层的可见光分类线性权值和可见光回归线性权值;α_i^{T}和β_i^{T}分别表示第i层的红外分类线性权值和红外回归线性权值.

1.2　多模特征融合模块

在特定场景下,如低分辨率和热交叉时,红外模态会存在大量噪声,可见光和红外模态之间存在大量非互补冗余信息,会减弱目标特征表示能力.为了最大限度地消除冗余信息,去除红外模态中的噪声干扰,增强目标的特征表示,设计了多模特征融合模块,该模块通过去噪和特征融合来提取两种模态下特征的互补信息,在消除非互补信息干扰的同时丰富了特有特征细节表现,其可视化效果如图2所示.其中(a)为可见光图像,(b)为红外图像,(c)为融合后图像,(d)为融合去噪后图像.

多模特征融合模块采用级联融合策略,由去噪模块和图像重建模块组成,能全面融合红外和可见光图像,有效捕捉复杂的空间细节,最大程度消除冗余信息的影响.融合过程如下:



图2　多模特征融合模块

Fig.2　Multimode feature fusion module

$$\begin{cases} T_{\mathrm{f}} = \mathrm{Concat}(T_{\mathrm{RGB}}, T_{\mathrm{T}}) \\ D_{\mathrm{f}} = \mathrm{Concat}(D_{\mathrm{RGB}}, D_{\mathrm{T}}) \end{cases} \tag{7}$$

式中:T_{RGB}、T_{T}、T_{f}、D_{RGB}、D_{T}和D_{f}分别表示可见光模板图像、红外模板图像、融合模板图像、可见光检测图像、红外检测图像和融合检测图像;Concat(·)表示按照通道维度拼接.融合后的图像通过去噪模块消除图像中的噪声,其中$D_{\mathrm{s}}(\cdot)$是去噪器,计算过程如下:

$$\begin{cases} \overline{T_{\mathrm{f}}} = D_{\mathrm{s}}(T_{\mathrm{f}}) \\ \overline{D_{\mathrm{f}}} = D_{\mathrm{s}}(D_{\mathrm{f}}) \end{cases} \tag{8}$$

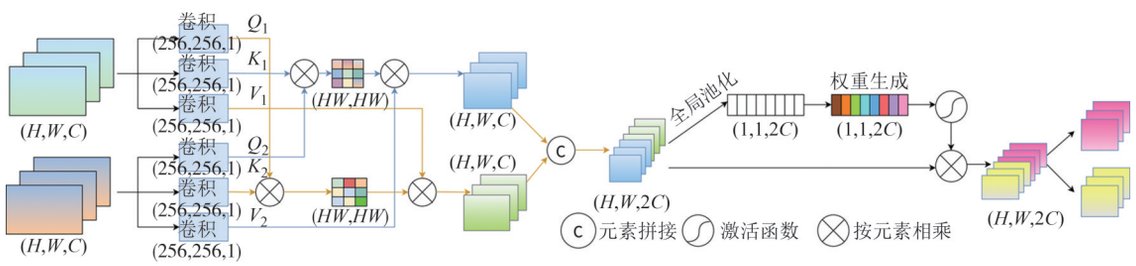


图3 特征交互学习模块

Fig.3 Feature interactive learning module

$$\begin{cases} Q^{\mathrm{T}} = \mathrm{Conv}(H_i^{\mathrm{T}}) \\ K^{\mathrm{R}} = \mathrm{Conv}(H_i^{\mathrm{RGB}}) \\ V^{\mathrm{R}} = \mathrm{Conv}(H_i^{\mathrm{RGB}}) \end{cases} \tag{9}$$

$$X_{\mathrm{R-T}}(H_i^{\mathrm{RGB}}, H_i^{\mathrm{T}}) = \mathrm{Atten}(Q^{\mathrm{T}}, K^{\mathrm{R}}, V^{\mathrm{R}}) = \mathrm{Softmax}\left(\frac{Q^{\mathrm{T}}(K^{\mathrm{R}})^{\mathrm{T}}}{\sqrt{d_x}}\right)V^{\mathrm{R}} \tag{10}$$

式中:H_i^{T}表示第i个颈部模块特征输出;Conv(·)表示1×1卷积;d_x是缩放因子.在Q^{T}和K^{R}之间建立了不同模态的交叉特征相关矩阵之后,采用Softmax(·)生成归一化的注意矩阵,最后与V^{R}相乘生成交叉模态特征相关矩阵.为了学习到不同模态之间更多的注意分布,提出算法采用多头注意力机制进行学习:

$$\begin{cases} \mathrm{Out}_{\mathrm{R-T}} = \mathrm{MultiHead}(Q^{\mathrm{T}}, K^{\mathrm{R}}, V^{\mathrm{R}}) = \mathrm{Concat}(H_1^{\mathrm{R-T}}, \cdots, H_n^{\mathrm{R-T}}) \\ \mathrm{Out}_{\mathrm{T-R}} = \mathrm{MultiHead}(Q^{\mathrm{R}}, K^{\mathrm{T}}, V^{\mathrm{T}}) = \mathrm{Concat}(H_1^{\mathrm{T-R}}, \cdots, H_n^{\mathrm{T-R}}) \end{cases} \tag{11}$$

$$\begin{cases} H_i^{\mathrm{R-T}} = \mathrm{Atten}(Q^{\mathrm{T}}W_i^{Q}, K^{\mathrm{R}}W_i^{K}, V^{\mathrm{R}}W_i^{V}) \\ H_i^{\mathrm{T-R}} = \mathrm{Atten}(Q^{\mathrm{R}}W_i^{Q}, K^{\mathrm{T}}W_i^{K}, V^{\mathrm{T}}W_i^{V}) \end{cases} \tag{12}$$

式中:W_i^{Q}为权重;n为头数;$\mathrm{Out}_{\mathrm{R-T}}$表示从红外模态到可见光模态的判别特征变换;$\mathrm{Out}_{\mathrm{T-R}}$表示从可见光模态到红外模态的判别特征变换.

为了增强网络特征学习能力,在多头跨模注意模块后引入通道注意机制为不同模态的通道级特征分配权重系数.计算过程如下,其中$\sigma(\cdot)$表示$\mathrm{Out}_{\mathrm{R-T}}$和$\mathrm{Out}_{\mathrm{T-R}}$通过通道注意机制的运算:

$$F_{\mathrm{RGB}}, F_{\mathrm{T}} = \sigma(\mathrm{Concat}(\mathrm{Out}_{\mathrm{R-T}}, \mathrm{Out}_{\mathrm{T-R}})) \tag{13}$$

式中:$\overline{T_{\mathrm{f}}}$表示互补模板特征,$\overline{D_{\mathrm{f}}}$表示互补检测特征.

1.3 特征交互学习模块

为了有效地融合可见光模态和红外模态的互补特征,充分利用潜在的隐藏信息,设计了特征交互学习模块,如图3所示.该模块由多头跨模注意机制和通道注意机制组成.多头跨模注意机制建立不同模态之间有效的关联,探索跨模态信息的全局空间特征.红外模态到可见光模态的判别特征转换表示为$X_{\mathrm{R-T}}$.转换过程如下:

1.4 特征空间位移模块

与可见光图像相比,红外图像通常缺乏重要的纹理信息.尽管多模特征融合模块能有效减少网络中冗余信息的干扰,提高目标特征细节表现.但由于热重叠,红外图像存在热源目标边缘模糊的现象.因此,设计了特征空间位移模块以有效增强热源目标边缘表示.该模块对颈部网络提取的特征进行分组空间位移,该位移操作不仅能提高目标边界的清晰度,同时能保持精简的参数结构,其实现过程为:首先,从颈部模块获得红外模板特征$H(T)$和红外检测特征$H(D)$,其中$H(T) \in \mathbb{R}^{C\times H_1\times W_1}$,$H(D) \in \mathbb{R}^{C\times H_1\times W_1}$沿通道维度将$H$分为四组:

$$\begin{cases} H_1 = H\left[1{:}\frac{C}{4}, :, :\right] \\ H_2 = H\left[\frac{C}{4}+1{:}\frac{C}{2}, :, :\right] \\ H_3 = H\left[\frac{C}{2}+1{:}\frac{3C}{4}, :, :\right] \\ H_4 = H\left[\frac{3C}{4}+1{:}C, :, :\right] \end{cases} \tag{14}$$

式中:$H_1, H_2, H_3, H_4 \in \mathbb{R}^{\frac{4}{C}\times H\times W}$,令$\overline{H_1} = H_1$,$\overline{H_2} = H_2$,

$\overline{H_3}=H_3,\overline{H_4}=H_4$.之后,从宽度和高度两个维度进行空间位移,移动过程如图4所示,计算如下:

$$\begin{cases}\overline{H_1}[:,0:h-c,:]=H_1[:,c:h,:]\\ \overline{H_2}[:,c:h,:]=H_2[:,0:h-c,:]\\ \overline{H_3}[:,:,0:w-c]=H_3[:,:,c:w]\\ \overline{H_4}[:,:,c:w]=H_4[:,:,0:w-c]\end{cases}\tag{15}$$

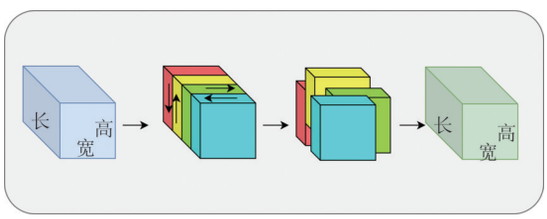


图4　特征空间位移模块

Fig.4　Feature space shift module

式(15)表示特征图分组空间位移策略,该策略使红外特征元素错位,增强了热源目标的边缘细节特征表示.最后,为了避免分组位移操作导致空间特征过度移动,将空间移动前后得到的特征图相加并沿通道维度串联.计算过程如下,其中$F\in\mathbb{R}^{C\times W\times H}$,Concat(·)表示按通道维度拼接:

$$\begin{cases}F_1=\overline{H_1}+H_1\\ F_2=\overline{H_2}+H_2\\ F_3=\overline{H_3}+H_3\\ F_4=\overline{H_4}+H_4\end{cases}\tag{16}$$

$$F=\mathrm{Concat}(F_1,F_2,F_3,F_4)\tag{17}$$

1.5　区域选择生成网络

为了提升前景和背景分类精度,更精确地确定锚框位置.本文将区域生成网络[17]扩展到多模态跟踪领域,设计了4个分类分支用来区分前景和背景信息,2个回归分支用来预测锚框位置.分类分支的输出表示为$\mathrm{Cls}\in\mathbb{R}^{g\times g\times 2N}$,其中$N$表示先验锚点的数量.$C(T)$和$C(D)$分别表示图1中$X_n$模块分类分支输出的模板特征和检测特征.分类分支表示为:

$$\mathrm{Cls}=C(T)*C(D)\tag{18}$$

其中*表示相关运算.回归分支可提取精确的目标边界框,其输出表示为$\mathrm{Reg}\in\mathbb{R}^{g\times g\times 4N}$,$R(T)$和$R(D)$分别表示回归分支输出的模板特征和检测特征.回归分支表示为:

$$\mathrm{Reg}=R(T)*R(D)\tag{19}$$

其中*表示相关运算.在获得6个分支的得分图后,将两种模态的得分图进行融合.

$$\begin{cases}S_{\mathrm{Cls}}^1=\mathrm{Cls}_R^{\mathrm{attn}}\oplus\mathrm{Cls}_T\\ S_{\mathrm{Cls}}^2=\mathrm{Cls}_R\oplus\mathrm{Cls}_T^{\mathrm{attn}}\end{cases}\tag{20}$$

$$S_{\mathrm{reg}}=\mathrm{Reg}_R^{\mathrm{attn}}\oplus\mathrm{Reg}_T^{\mathrm{attn}}\tag{21}$$

1.6　自适应区域选择模块

为了提高跟踪算法的鲁棒性并在跟踪推理阶段确定最佳区域,提出的算法设计了自适应区域选择模块.首先,采用余弦窗有效抑制锚框大幅位移,减少跟踪目标位置的偏移.此外,引入比例惩罚机制抑制锚框长宽比的变化,确保跟踪目标的形状和尺寸一致性.最后,对不同候选区域得分进行重新排序,确定最佳的候选区域.

I,J,L为索引集,变量i和j分别表示锚点的位置.$\mathrm{ANC}=\{(x_i^{\mathrm{an}},y_j^{\mathrm{an}},w_l^{\mathrm{an}},h_l^{\mathrm{an}})_{i\in I,j\in J,l\in L}\}$为锚框表示.边界框映射的点集得到的细化坐标表示为:

$\mathrm{REG}=\{(x_i^{\mathrm{reg}},y_i^{\mathrm{reg}},\mathrm{d}x_l^{\mathrm{reg}},\mathrm{d}y_l^{\mathrm{reg}},\mathrm{d}w_l^{\mathrm{reg}},\mathrm{d}h_l^{\mathrm{reg}})_{i\in I,j\in J,l\in L}\}$细化区域$\mathrm{PRO}=\{(x_i^{\mathrm{pro}},y_j^{\mathrm{pro}},w_l^{\mathrm{pro}},h_l^{\mathrm{pro}})\}$的计算表示为:

$$\begin{cases}x_i^{\mathrm{pro}}=x_i^{\mathrm{an}}+w_l^{\mathrm{an}}\times\mathrm{d}x_l^{\mathrm{reg}}\\ y_j^{\mathrm{pro}}=y_j^{\mathrm{an}}+h_l^{\mathrm{an}}\times\mathrm{d}y_l^{\mathrm{reg}}\\ w_l^{\mathrm{pro}}=w_l^{\mathrm{an}}\times\mathrm{e}^{\mathrm{d}w_l}\\ h_l^{\mathrm{pro}}=h_l^{\mathrm{an}}\times\mathrm{e}^{\mathrm{d}h_l}\end{cases}\tag{22}$$

对区域选择生成网络得到的S_{Cls}^1和S_{Cls}^2中的峰值进行索引,比较后选择最佳候选区域.将S_{Cls}形状调整为$[g\times g\times N,2]$,其中奇数通道代表前景信息,用PosScore来表示,奇数通道中的每个元素代表区域被分类为正样本的概率.接着,增加比例惩罚减轻目标大小和比例的显著变化,应用余弦窗函数抑制锚框大幅位移,得到预测区域:

$$\begin{cases}p=\mathrm{e}^{t*\left\{-\left[\max\left(\frac{s}{s'},\frac{s'}{s}\right)*\max\left(\frac{r}{r'},\frac{r'}{r}\right)-1\right]\right\}}\\ \mathrm{PredScore}_1=\mathrm{Softmax}(\mathrm{PosScore}_1)p(1-t)+\mathrm{cos}t\\ \mathrm{PredScore}_2=\mathrm{Softmax}(\mathrm{PosScore}_2)p(1-t)+\mathrm{cos}t\end{cases}\tag{23}$$

式中:t为超参数;s表示所有尺度的区域;s'表示区域的最后一帧;r表示区域的长宽比;r'表示区域最后一帧的长宽比;p为比例惩罚因子;cos表示余弦函数.从S_{Cls}^1和S_{Cls}^2中得到$\mathrm{PredScore}_1$和$\mathrm{PredScore}_2$,然后比较峰值得分并返回最大值:

$$\mathrm{Max}=\mathrm{Arg2max}(\mathrm{PredScore}_1,\mathrm{PredScore}_2)\tag{24}$$

式中:Max表示最大值元素的索引;Arg2max(·)表示两个数组中的最大值,并返回最大值元素的索引.最佳候选区域通过从Max到PRO的映射关系来确定.

2　实验结果与分析

2.1　实验细节

在英伟达RTX 3090 GPU及酷睿i9 13900k平台

进行实验,PyTorch 版本为 1.8.0.采用端到端训练,训练过程中使用多个损失函数加权组合,其中分类损失采用交叉熵损失函数,回归损失采用平滑 L1 损失函数.本文提出的架构可学习参数为 0.36×10^6 个,训练分为两个阶段:第一阶段,冻结红外模态分支和设计的模块权重,仅对单模可见光 SiamRPN++[16]跟踪网络进行训练.基础训练数据集包括 ImageNet VID[18]、YouTube-BB[19]、COCO[20]和 ImageNet Det [18].初始学习率设定为 10^{-3},终止学习率设定为 10^{-4},对网络进行端到端训练 20 轮,采用英伟达 GeForce RTX 3090 GPU 共计训练 46 h.第二阶段,使用 RGB-T 数据集训练.基础训练集包括 GTOT[6]、RGBT234[21]和 LasHeR[22].为了避免过拟合,用 GTOT 数据集作为测试集时,选择 RGBT234 和 LasHeR 作为训练集.用 RGBT234 作为测试集时,选择 GTOT 和 LasHeR 作为训练集.初始学习率为 10^{-2},终止学习率为 10^{-5},训练 50 轮,共计训练 28 h.

2.2 跟踪数据集评估

为了评估跟踪算法的整体性能,分别在 GTOT 数据集、RGBT234 数据集上与最先进的跟踪算法进行比较.如表 1 所示,提出的跟踪算法在满足实时跟踪速率要求的同时,跟踪性能优于最先进的跟踪算法,提出算法在 GTOT 数据集上获得了 90.5% 的精度(PR)和 72.8% 的成功率(SR),在 RGBT234 数据集上获得了 78.7% 的精度和 56.8% 的成功率,同时获得了 32 帧/s 的跟踪速率.

表 1 在 GTOT 数据集和 RGBT234 数据集上获得不同跟踪算法基于挑战属性的比较

Tab. 1 Comparison with other trackers on the GTOT and RGBT234 datasets

Trackers	MANet++ (2019)	MANet (2018)	DAFNet (2022)	GCMP (2021)	SiamCDA (2021)	SiamDL (2022)	TAAT (2022)	DFAT (2023)	SiamCSR (2023)	Ours
GTOT	90.1/72.3	89.4/72.4	89.1/71.2	86.9/71.1	87.7/73.2	88.8/73.1	85.8/69.6	89.3/72.3	88.2/70.9	90.5/72.8
RGBT234	78.1/54.0	77.7/53.9	79.6/54.4	75.1/49.1	76.0/56.9	—	—	76.1/54.4	75.4/53.2	78.7/56.8
FPS	27	2	23	35	37	45	34	20	140	32

2.2.1 GTOT 数据集

GTOT 包含 50 个成对的 RGB-T 视频序列,标注了包括遮挡(OCC)、大尺度变化(LSV)、快速运动(FM)、热交叉(TC)、低照明(LI)、变形(DEF)和小物体(SO).在 GTOT 数据集上,将提出算法与 TAAT、SiamCSR[9]、MANet[10]、SiamCDA[13]、DFAT[15]、DAFNet[23]、SiamDW+RGBT[24]、MANet++[25]和 SGT[26]进行比较,结果如图 5 和图 6 所示,红色、蓝色和绿色分别表示排名第一、第二和第三的精度和成功率指标.表 2 是基于挑战属性的比较结果.在 TC 子集中,提出算法在精度和成功率指标上都优于其他算法,表明提出算法在应对热交叉方面性能优越.在应对 LI 挑战时,提出的算法精度排名第一,这表明设计的特征交互学习模块能充分利用不同模态之间的互补信息,具有优秀的模态交互性能,在 LI 场景中优于对比算法.在 DEF 场景中,提出的算法在精度和成功率上都排名第二,但在基于孪生网络的 RGB-T 跟踪框架中,排名第一,与 SiamCDA[13]相比,提出算法的精度提高了 6.2 个百分点,成功率提高了 2.1 个百分点,这些结果表明提出算法处理 DEF 场景的有效性,在精度和成功率方面优于现有孪生跟踪网络算法.在应对 LSV 挑战时,基于孪生网络的跟踪算法性能要优于

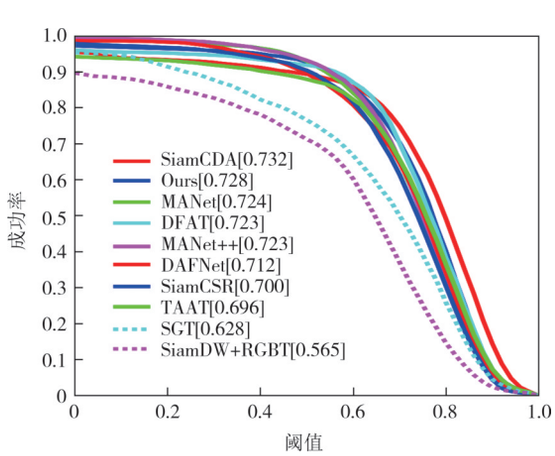


图 5 GTOT 数据集成功率比较结果

Fig.5 Comparison results of success rates in the GTOT dataset

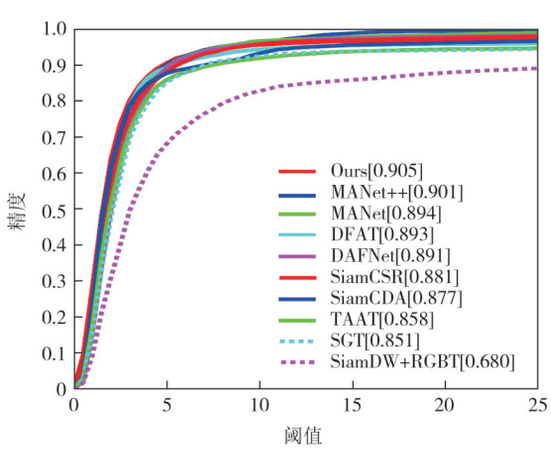


图 6 GTOT 数据集精度比较结果

Fig.6 Comparison results of precision rate in GTOT dataset

基于MDNet[3]的跟踪算法.与DAFNet[23]相比,提出算法精度指标提高了7.2个百分点,成功率指标提高了4.6个百分点,这证明了基于孪生网络的跟踪算法在应对LSV挑战时具有较强的鲁棒性.

表2　在GTOT数据集上获得不同跟踪算法基于挑战属性的比较

Tab.2　Attribute-based precision and success rate (PR/SR) obtained by using different trackers on GTOT

Method	SiamDW+RGBT(2019)	SiamRPN++(2019)	DFAT(2023)	MANet(2018)	DAFNet(2022)	MANet++(2019)	SiamCDA(2021)	SiamCSR(2023)	Ours
OCC	67.5/53.6	70.3/587	86.3/68.7	88.2/69.6	87.3/68.4	89.0/70.1	82.2/69.4	86.9/67.2	86.4/70.2
LSV	68.9/56.5	76.5/64.3	92.4/75.0	86.9/70.6	82.2/66.4	86.6/69.3	91.5/74.8	88.1/67.2	89.4/71.0
FM	71.1/57.6	75.9/65.9	89.1/74.0	87.9/69.4	80.9/64.2	86.7/69.4	86.6/72.0	82.9/64.3	85.3/69.0
LI	70.0/58.8	68.9/58.3	92.2/74.1	91.4/73.0	89.9/72.7	91.7/73.1	92.4/76.4	88.9/72.0	92.6/73.0
TC	63.5/51.7	76.6/64.0	89.1/70.7	88.9/70.2	89.8/70.3	89.9/70.7	82.6/68.5	87.7/68.5	90.3/72.2
SO	76.4/58.5	82.2/64.7	94.4/71.9	93.2/70.0	93.8/69.8	93.9/69.9	87.4/69.1	89.0/67.0	91.5/70.8
DEF	69.1/58.2	71.0/59.3	91.9/73.5	92.3/74.2	94.7/76.5	93.7/74.4	87.9/72.7	90.0/73.3	94.1/74.8
ALL	68.8/55.0	72.5/61.7	89.3/72.3	89.4/72.4	89.1/71.2	90.1/72.3	87.7/73.2	88.1/70.0	90.5/72.8

与先进的跟踪方法相比,提出的算法在保证实时性的同时,达到了90.5%的准确率和72.8%的成功率.与MANet[10]相比,准确率指标提高了1.1个百分点,成功率指标提高了0.4个百分点,跟踪速率提高了30 帧/s.这表明提出跟踪算法缩小甚至消除了与最先进的RGB-T跟踪算法的性能差距.

2.2.2　RGBT234数据集

RGBT234是一个大型的RGB-T跟踪数据集,包含234个视频集,共计233 000 帧.一共包括了12个挑战属性,其中包括无遮挡(NO)、部分遮挡(PO)、重度遮挡(HO)、低照明(LI)、低分辨率(LR)、热交叉(TC)、形变(DEF)、快速运动(FM)、尺度变化(SV)、运动模糊(MB)、摄像机移动(CM)和背景杂波(BC).在RGB-T234数据集上,将提出算法与MANet[10]、SiamCDA[13]、DFAT[15]、DAFNet[23]、SiamDW+RGBT[24]、SGT+RGBT[26]、SOWP+RGBT[27]等进行了比较,如图7、图8所示,排名第一、第二和第三的数据分别用红色、蓝色和绿色标出.基于挑战属性的比较结果如表3所示.首先,在DEF、HO和SV三个挑战中,提出算法的准确率和成功率最高.在应对DEF和HO的挑战时,其他跟踪算法的精度急剧降低,而提出算法仍然能够保持很高的精度,说明提出的算法具有很好的跨模态信息挖掘能力,能够充分利用不同模态之间的互补信息.其次,在应对LR和PO的挑战时,提出算法仍然优于大多数跟踪算法,这一方面得益于对低分辨率红外图像进行空间特征位移增强了目标边缘表示,另一方面得益于多模特征融合模块使集成图像包含丰富的细粒度空间细节.在应对所有挑战时,提出算法在性能上明显优于RGB跟踪算法[16],这表明多模态跟踪的性能优于单模态跟踪.此外,与其他基于孪生网络框架的跟踪算法相比,提出的算法在应对FM和MB的挑战时性能优越.

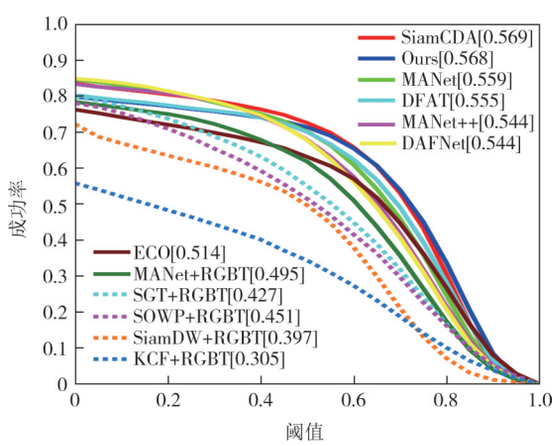


图7　RGBT234数据集成功率比较结果

Fig.7　Comparison results of success rates in the RGBT234 dataset

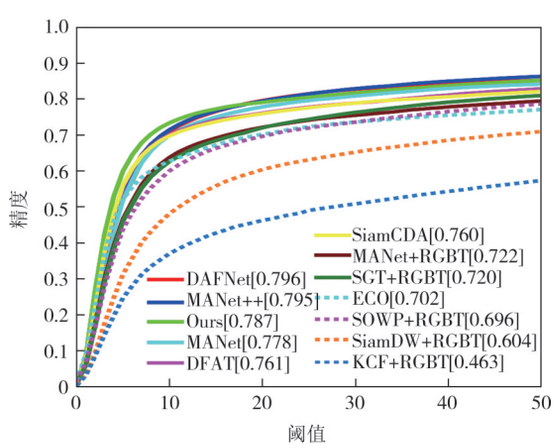


图8　RGBT234数据集精度比较结果

Fig.8　Comparison results of precision rate in the RGBT234 dataset

表3 在RGBT234数据集上获得不同跟踪算法基于挑战属性的比较

Tab.3 Attribute-based precision and success rate (PR/SR) obtained by using different trackers on RGBT234

Method	SiamDW+RGBT(2019)	SiamRPN++(2019)	SGT+RGBT(2017)	DFAT(2023)	SiamCDA(2021)	MANet(2018)	DAFNet(2022)	Ours
BC	51.9/32.3	57.8/39.3	65.8/41.8	71.9/47.8	73.9/52.7	75.6/49.8	79.0/49.3	72.0/51.2
CM	56.2/38.2	66.4/49.9	66.7/45.2	74.2/54.7	73.3/54.8	69.9/50.5	72.3/50.6	66.9/49.0
DEF	55.8/39.0	69.5/53.2	68.5/47.4	76.0/57.5	74.9/57.4	72.3/52.4	74.1/51.6	77.7/57.5
FM	45.4/29.9	65.3/46.9	67.7/40.2	65.4/46.2	61.4/45.4	71.1/45.5	74.0/46.5	64.0/44.5
HO	52.0/33.7	59.4/43.4	59.2/39.4	63.9/45.5	67.5/49.5	66.5/45.9	68.6/45.9	71.9/51.8
LI	60.0/39.9	59.3/42.4	70.5/46.2	78.3/56.2	82.2/60.3	81.7/56.0	81.2/54.2	78.7/56.3
LR	60.5/37.0	66.4/46.5	75.1/47.6	75.2/51.5	70.9/49.9	78.5/51.3	81.8/53.8	75.8/54.0
MB	49.4/34.0	64.5/48.7	64.7/43.6	62.6/50.2	63.5/47.4	66.5/48.2	70.8/50.0	64.9/47.7
NO	78.3/53.4	83.8/64.2	87.7/55.5	93.3/69.6	88.9/66.8	91.4/64.9	90.0/63.6	84.1/61.8
PO	60.8/39.6	73.5/54.3	77.9/51.3	80.7/59.2	78.8/60.0	83.1/58.4	85.9/58.8	83.3/60.1
SV	60.9/40.5	72.7/55.5	69.2/43.4	77.4/57.3	74.3/56.8	77.8/54.8	79.1/54.4	79.3/57.3
TC	56.9/36.8	70.6/53.0	76.0/47.0	67.5/49.4	68.0/47.9	72.9/52.9	81.1/58.3	72.9/53.2
ALL	60.4/39.7	69.7/51.7	72.0/47.2	76.1/54.4	76.0/56.9	77.8/54.4	79.6/54.4	78.7/56.8

2.3 有效性分析

图9显示了提出算法与其他先进的跟踪算法(包括MDNet[3]、MANet[10]、SGT[26]和ECO[28])在Manlight、Hotglass和Caraftertree三个视频序列上的可视化比较.在遮挡严重、光照不足、物体较小且分辨率较低的场景中,提出算法明显优于其他算法.例如,在图9(a)中的低照度、低分辨率场景中,提出算法的跟踪效果明显优于其他跟踪算法.在第108帧中可见光图像无法确定目标,但是提出算法能够充分利用跨模态互补信息和增强目标边缘特征表示从而准确地跟踪目标,而其他算法在第128帧中都丢失了目标.在图9(b)中,提出算法在应对小物体场景挑战时表现出色,在第200帧大多数跟踪算法丢失目标时,本文提出的算法依然能准确跟踪目标.最后,如图9(c)所示,在应对目标的大尺度变换和遮挡时,提出算法通过对不同模态的信息进行特征融合增强,同时消除冗余信息和噪声干扰,获得了最佳跟踪性能.

2.4 消融实验

在GTOT数据集上测试了5个变体网络,以验证设计的模块的有效性.如表4所示,多模特征融合模块、特征交互学习模块和特征空间位移模块对提出算法的跟踪性能有不同程度的贡献.具体来说,双流主干去掉了多模特征融合模块、特征交互学习模块和特征空间位移模块,只保留了双流主干网络部分.变体-v1在双流主干的基础上增加了多模特征融合模块.变体-v2在变体-v1的基础上集成了特征交互学习模块.变体-v3在双流主干的基础上加入了特征交互学习模块和特征空间位移模块.通过变体-v1与双流主干、完整网络与变体-v3的对比,可以看出多模特征融合模块对跟踪性能的影响.通过集成多模特征融合模块,变体-v1相比于双流主干,精度指标提高了0.6个百分点,成功率指标提高了1.1个百分点.完整网络相比于变体-v3,精度指标提高了2.2个百分点,成功率指标提高了1.7个百分点.变体-v2与变体-v1的比较显示了特征交互学习模块的有效性,变体-v2精度指标提高了8个百分点,成功率指标提高了8.1个百分点.与变体-v2相比,完整网络加入了特征空间位移模块后精度指标提高了1.1个百分点,成功率指标提高了0.2个百分点.

为了进一步评估特征交互学习模块组件的有效性,本文在GTOT数据集上比较了两个变体组件,如表5所示.双流主干是指只保留基础跟踪框架,去掉了本文提出的多模特征融合模块、特征交互学习模块和特征空间位移模块.变体组件-v1是指在双流主干网络中只采用通道注意模块.变体组件-v2是指在双流主干网络中只采用多头跨模注意模块.完整组件是指在双流主干网络中只加入特征交互学习模块.从结果可以看出,双流主干网络添加通道注意模块后精度指标提高了6.6个百分点,成功率指标提高了7.4个百分点.添加多头跨模注意模块后成功率指标提高了6.1个百分点,精度指标提高了6.9个百分点.与表4中的完整网络相比,在双流骨干网添加特征交互学习模块后精度指标提高了7.7个百分点,成功率指标提高了8.1个百分点.通过对比实验,充分验证了特征交互学习模块以及通道注意模块组件和多头跨模注意模块组件的有效性.

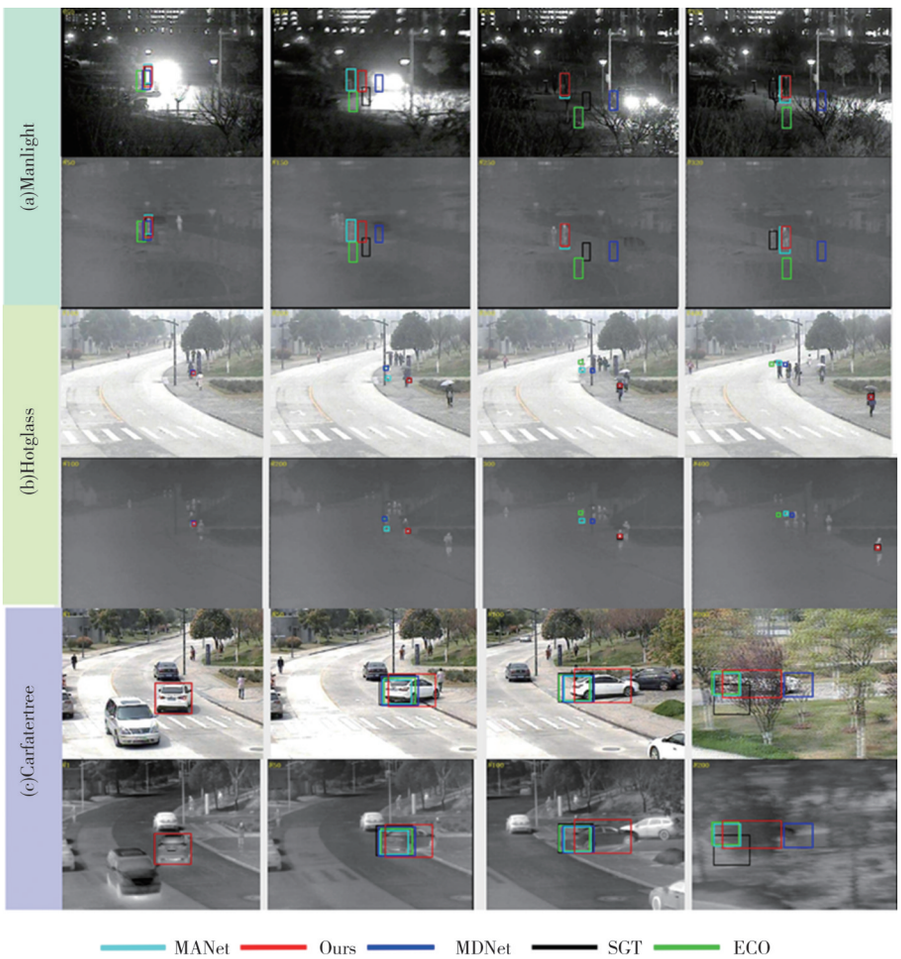


图 9　跟踪效果可视化对比图

Fig.9　Comparative visualization of tracking results

表 4　本文所提网络的变体网络精度和成功率比较

Tab.4　Comparison of accuracy and success rates for different variants of the proposed network in this paper

	多模特征融合模块	特征交互学习模块	特征空间位移模块	PR/SR
双流主干				80.8/63.4
变体-v1	∎			81.4/64.5
变体-v2	∎	∎		89.4/72.6
变体-v3		∎	∎	88.3/71.1
完整网络	∎	∎	∎	90.5/72.8

表 5　特征交互学习模块变体结构精度和成功率比较

Tab.5　Comparison of accuracy and success rates for different variants of the feature interaction learning module structure

	通道注意模块	多头跨模注意模块	PR/SR
双流主干			80.8/63.4
变体组件-v1	∎		87.4/70.8
变体组件-v2		∎	86.9/70.3
完整组件	∎	∎	88.5/71.5

3　结 论

本文提出了基于孪生网络的特征融合位移RGB-T目标跟踪.首先,设计特征交互学习模块挖掘跨模信息,指导不同模态间的特征学习,并通过双向特征指导学习方式充分挖掘跨模态信息.随后,设计多模特征融合模块,通过去噪和特征融合的方式消除非互补信息对网络的干扰,增强特有特征细节和目标特征细节.最后,特征空间位移模块在不增加计算复杂度的情况下对颈部网络提取的特征进行分组空间位移,增强目标边缘特征表示.在两个RGB-T跟踪数据集上进行的大量对比实验证实了提出算法的优越性.消融实验评估了不同模块的有效性.

参考文献

[1]　DANELLJAN M, BHAT G, KHAN F S, et al.　ATOM: accurate tracking by overlap maximization [C]//2019 IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. IEEE, 2019: 4660 - 4669.

[2] BHAT G, DANELLJAN M, VAN GOOL L, et al. Learning discriminative model prediction for tracking[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2019: 6182-6191.

[3] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 4293-4302.

[4] LAN X Y, YE M, SHAO R, et al. Learning modality-consistency feature templates: a robust RGB-infrared tracking system [J]. IEEE Transactions on Industrial Electronics, 2019, 66(12): 9887-9897.

[5] LI C L, ZHU C L, HUANG Y, et al. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking[C]// Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 831-847.

[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(12): 5743-5756.

[7] LI C L, SUN X, WANG X, et al. Grayscale-thermal object tracking via multitask Laplacian sparse representation[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017, 47(4): 673-681.

[8] GUO C, YANG D D, LI C, et al. Dual Siamese network for RGBT tracking via fusing predicted position maps [J]. The Visual Computer, 2022, 38(7): 2555-2567.

[9] GUO C Y, XIAO L. High speed and robust RGB-thermal tracking via dual attentive stream Siamese network [C]//IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium. Kuala Lumpur, Malaysia. IEEE, 2022: 803-806.

[10] LI C L, LU A D, ZHENG A H, et al. Multi-adapter RGBT tracking [C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South). IEEE, 2019.

[11] TANG Z Y, XU T Y, WU X J. Temporal aggregation for adaptive RGBT tracking[EB/OL]. 2022: 2201.08949.https://arxiv.org/abs/2201.08949v2.

[12] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 770-778.

[13] ZHANG T L, LIU X R, 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.

[14] LI Y D, LAI H C, WANG L J, et al. Multibranch adaptive fusion network for RGBT tracking[J]. IEEE Sensors Journal, 2022, 22(7): 7084-7093.

[15] TANG Z Y, XU T Y, LI H, et al. Exploring fusion strategies for accurate RGBT visual object tracking[EB/OL].2022: 2201.08673. https://arxiv.org/abs/2201.08673v1.

[16] LI B, WU W, WANG Q, et al. SiamRPN: evolution of Siamese visual tracking with very deep networks [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. IEEE, 2019: 4277-4286.

[17] GIRSHICK R. Fast R-CNN [C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. IEEE, 2015: 1440-1448.

[18] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[19] REAL E, SHLENS J, MAZZOCCHI S, et al. YouTube-BoundingBoxes: a large high-precision human-annotated data set for object detection in video [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 7464-7473.

[20] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context [C]// Computer Vision-ECCV 2014. Zurich, Switzerland. Springer, 2014: 740-755.

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

[22] 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.

[23] 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, Korea (South). IEEE, 2019: 91-99.

[24] 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, CA, USA. IEEE, 2019: 4591-4600.

[25] ZHANG H, ZHANG L, ZHUO L, et al. Object tracking in RGB-T videos using modal-aware attention network and competitive learning[J]. Sensors, 2020, 20(2): 393.

[26] LI C L, ZHAO N, LU Y J, et al. Weighted sparse representation regularized graph learning for RGB-T object tracking [C]// Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, California, USA. ACM, 2017: 1856-1864.

[27] KIM H U, LEE D Y, SIM J Y, et al. SOWP: spatially ordered and weighted patch descriptor for visual tracking [C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. IEEE, 2015: 3011-3019.

[28] DANELLJAN M, BHAT G, KHAN F S, et al. ECO: efficient convolution operators for tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 6931-6939.