文章编号　1004-924X(2025)12-1940-15

多模态特征融合的RGB-T目标跟踪网络

金　静，刘建琴*，翟凤文
(兰州交通大学 电子与信息工程学院，甘肃 兰州 730070)

摘要：近年来，RGB-T跟踪方法因可见光与热红外图像的互补特性而在视觉跟踪领域得到广泛应用。然而，现有方法在模态互补信息利用方面仍存在局限，特别是基于Transformer的算法缺乏模态间的直接交互，难以充分挖掘RGB和TIR模态的语义信息。针对这些问题，提出了一种多模态特征融合的RGB-T目标跟踪网络(Multi-Modal Feature Fusion Tracking Network for RGB-T，MMFFTN)。该网络首先在骨干网络提取初步特征后，引入通道特征融合模块(Channel Feature Fusion Module，CFFM)，实现RGB和TIR通道特征的直接交互与融合。其次，针对RGB和TIR模态差异可能导致的融合效果不理想问题，设计了跨模态特征融合模块(Cross-Modal Feature Fusion Module，CMFM)，通过自适应融合策略进一步融合RGB和TIR的全局特征，以提升跟踪的准确性。对本文提出的跟踪模型在GTOT，RGBT234和LasHeR三个数据集上进行了详细的实验评估。实验结果表明，与当前先进的基于Transformer的跟踪器ViPT相比，MMFFTN在成功率(Success Rate)和精确率(Precision Rate)上分别提升了3.0%和4.7%；与基于Transformer的跟踪器SDSTrack相比，成功率和精确率分别提升了2.4%和3.3%。

关　键　词：RGB-T目标跟踪；Transformer；通道特征融合；跨模态特征融合
中图分类号：TP391　　**文献标识码**：A
doi：10.37188/OPE.20253312.1940　　**CSTR**：32169.14.OPE.20253312.1940

RGB-T tracking network based on multi-modal feature fusion

JIN Jing，LIU Jianqin*，ZHAI Fengwen

(*School of Electronic and Information Engineering*，*Lanzhou Jiaotong University*，*Lanzhou* 730070，*China*)
* *Corresponding author*，*E-mail*：1970477938@*qq.com*

Abstract：In recent years，RGB-T tracking methods have been widely used in visual tracking tasks due to the complementarity of visible image and thermal infrared images. However，the existing RGB-T moving target tracking methods have not yet made full use of the complementary information between the two modalities，which limits the performance of the tracker. The existing Transformer-based RGB-T tracking algorithms are still short of direct interaction between the two modalities，which limits the full use of the original semantic information of RGB and TIR modalities. To solve this problem，the paper proposed a Multi-modal Feature Fusion Tracking Network for RGB-T (MMFFTN). Firstly，after extracting the preliminary features from the backbone network，the Channel Feature Fusion Module (CFFM) was introduced to realize the direct interaction and fusion of RGB and TIR channel features. Secondly，in order to

收稿日期：2024-11-25；**修订日期**：2025-02-10.
基金项目：甘肃省高校教师创新基金项目(No. 2025B-060)；宁夏自然科学基金资助项目(No. 2023AAC03741)；甘肃省科技计划项目重点研发计划-工业类(No. 23YFGA0047)

solve the problem of unsatisfactory fusion effect caused by the difference between RGB and TIR modality, a Cross-Modal Feature Fusion Module (CMFM) was designed and the global features of RGB and TIR were further fused through an adaptive fusion strategy to improve the tracking accuracy. The proposed tracking model was evaluated in detail on three datasets: GTOT, RGBT234 and LasHeR. Experimental results demonstrate that MMFFTN improves the success rate and precision rate by 3.0% and 4.7%, respectively compared with the current advanced Transformer-based tracker ViPT. Compared with the Transformer-based tracker SDSTrack, the success rate and accuracy are improved by 2.4% and 3.3%, respectively.

Key words: RGB-T tracking; transformer; channel feature fusion; cross-modal feature fusion

1　引　言

近年来，运动目标跟踪作为计算机视觉领域的基础任务之一，在机器人[1]、无人机任务、自动驾驶汽车、视频监控[2]等多个领域展现出了广泛的应用价值。当前的视觉跟踪技术主要基于RGB跟踪，但在面临严重遮挡、剧烈形变和光照变化等挑战时，RGB模态跟踪依然存在许多困难。RGB图像能够提供丰富的色彩信息，视觉效果优良，但是易受到天气和光照条件等因素的影响，抗干扰能力相对较弱，从而导致跟踪精度下降。相比之下，热红外(Thermal Infrared，TIR)图像基于热辐射成像，对温度变化敏感，在有烟雾等障碍物时表现出较强的穿透能力，即便在光线不足的情况下也能有效工作。但TIR图像容易受到热干扰的影响，并且缺乏丰富的纹理和颜色信息[3]。针对可见光和热红外图像的不同特点，一些研究人员提出将两种模态图像优势信息进行融合，利用两种模态融合后的互补信息进行目标跟踪，以克服仅使用单模态图像跟踪方法的一些局限性[4]。

随着深度学习在可见光跟踪中的成功应用，深度学习也被应用于RGB-T跟踪。MDNet[5]是一种相对较小的跟踪网络，将它扩展到可见光和红外两个模态时，对训练数据的量要求较低，符合该领域的发展现状，因此，较多的学者关注MDNet架构下的RGB-T跟踪算法。一些基于MDNet的RGB-T目标跟踪算法尝试利用数据集上不同场景的属性标注，使跟踪器学习不同属性下的鲁棒特征表示[6]。ADRNet[7]，CAT[8]，APFNet[9]均利用图像序列的属性提升跟踪性能，为每个属性类别的视频训练专门的网络，然后将基于属性的特征进行不同方式的融合。这些算法能够在一定程度上提高跟踪的精度和稳定性，但是在训练阶段需要额外提供图像的属性信息，同时专用网络也会增加算法的复杂性[10]。

2021年，ViT(Vision Transformer)[11]的提出标志着Transformer在图像识别任务中的正式应用。ViT借助自注意力机制，可以显式地对一个图像中的所有元素之间的关系进行建模，对位置进行编码，实现了卓越的性能[12]。同时，基于Transformer的RGB-T跟踪方法通过有效地建模全局依赖关系，展现了出色的性能。例如，Feng等人[13]开创了一种基于Transformer的端到端可训练跨模式跟踪器，通过加强长距离特征关联建模，显著提高了跟踪性能。Hui等人[14]提出了一种新的模板桥搜索区域交互(TBSI)模块，利用模板通过收集和分发目标相关对象和环境背景，在RGB和TIR搜索区域之间架起跨模态交互的桥梁。为充分利用预训练基本模型在大规模训练中的优势，Zhu等人[15]开发了视觉提示多模态跟踪(ViPT)，重点学习与模式相关的线索，使冻结的预训练基本模型能够适应各种下行多模态跟踪任务。Hou等人提出一种新颖的多模态视觉对象跟踪方法SDSTrack[16]，该方法通过对称框架和参数高效微调有效利用和融合多模态信息，在多模态跟踪场景中表现出色。然而，这些方法通过使用中间特征和时间信息来实现跨模态特征融合，它们没有探索RGB和TIR特征之间直接交互的有效性，这导致模型无法从RGB和TIR特征之间的原始语义交互中获得信息。

针对上述问题，本文拟提出一种基于多模态

特征融合的RGB-T目标跟踪网络(Multi-Modal Feature Fusion Tracking Network for RGB-T，MMFFTN)，具体而言，本研究将从以下几个方面展开：

(1)提出通道特征融合模块CFFM(Channel Feature Fusion Module)，该模块首先将RGB和TIR特征进行串联和线性融合，之后并行进行通道增强和局部多级特征增强，对输出特征进行求和，最后将混合特征与原始特征进行残差连接。CFFM有助于提升两种模态信息的互补性，使得最终的特征表达更加丰富，提高跟踪器的鲁棒性。

(2)引入跨模态特征融合模块CMFM(Cross-Modal Feature Fusion Module)。为提取更为精细的特征，该模块首先对RGB和TIR特征分别执行多次卷积操作，然后采用自适应融合策略，强调各模态下的优势局部区域(如边缘、纹理、颜色、热辐射特性)。这一设计使CMFM在复杂环境下能够充分利用各模态优势，提升目标跟踪的准确性和鲁棒性。

(3)基于以上特征融合模块构建了RGB-T跟踪器MMFFTN，并在三个具有挑战性的基准数据集上进行了全面测试和比较。预期结果表明，MMFFTN相比当前先进的基于Transformer的跟踪器ViPT，在SR和PR指标上均有所提升，分别提高了3.0%和4.7%。同时，与另一基于Transformer的跟踪器SDSTrack相比，MMFFTN在SR和PR指标上也展示了明显的优势，提升幅度分别为2.4%和3.3%。

2　多模态特征融合跟踪网络

本文提出的多模态特征融合跟踪网络(MMFFTN)的总体框架如图1所示。首先，将模板图像和搜索区域图像经过Patch Embedding转换为token形式，随后将这些token输入到Transformer模型中。在Transformer的层级结构中，引入通道特征融合模块(Channel Feature Fusion Module，CFFM)和跨模态特征融合模块(Cross-Modal Feature Fusion Module，CMFM)，以实现特征的有效融合。具体而言，CFFM负责在通道层面上对特征进行精细整合，而CMFM则进一步促进RGB与TIR特征之间的跨模态交互与融合。最终，Transformer的最后一层输出的搜索区域RGB和TIR特征被送入预测头，以精确获取目标的状态信息。

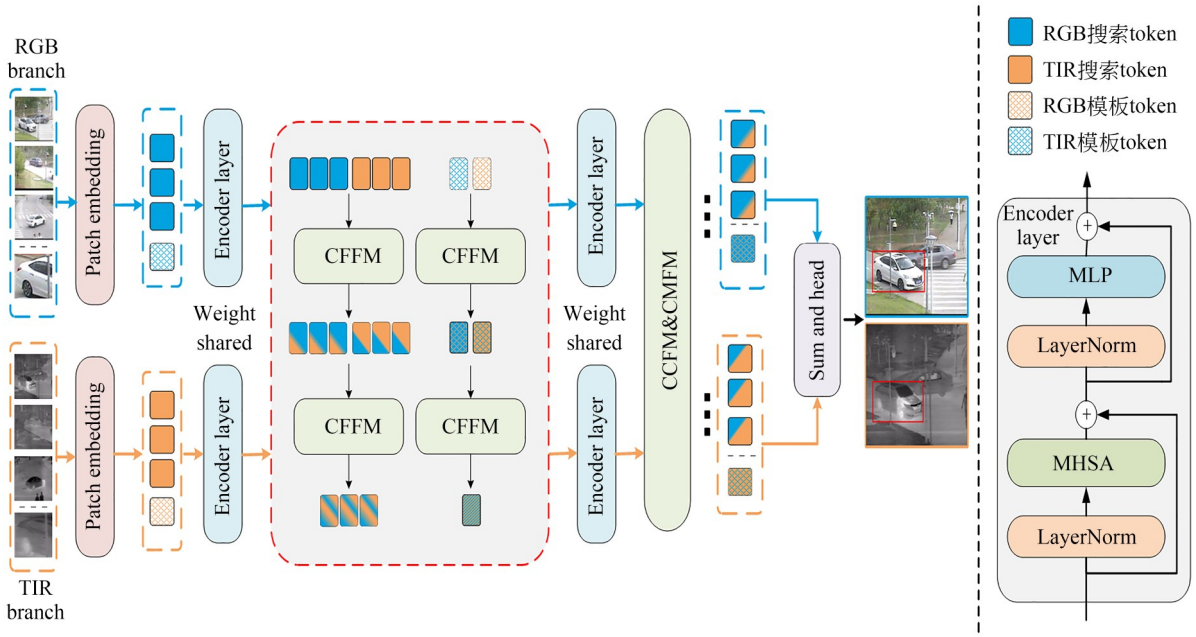


图1　MMFFTN网络总体结构图

Fig. 1　Overall structure diagram of the MMFFTN network

2.1　基于ViT的骨干网络

ViT通过自注意力机制能够捕捉图像中任意两个位置之间的长距离依赖关系，有效提取全局特征。基于这一优势，本文选择ViT作为骨干网络，将RGB和TIR模态的模板帧与搜索帧输入到Transformer中，构建多模态跟踪器。具体来说，输入RGB和TIR搜索区域图像对 $\boldsymbol{s}_r, \boldsymbol{s}_t \in \boldsymbol{R}^{H_s \times W_s \times 3}$ 及模板图像对 $\boldsymbol{m}_r, \boldsymbol{m}_t \in \boldsymbol{R}^{H_m \times W_m \times 3}$，裁剪为 $H_m \times W_m = 128 \times 128$ 和 $H_s \times W_s = 256 \times 256$ 的大小并分割为 $P \times P$ 的patch，得到的patch被扁平化为四个 *patch* 序列 $\boldsymbol{P}_{sr}, \boldsymbol{P}_{st} \in \boldsymbol{R}^{N_s \times 3P^2}$ 和 $\boldsymbol{P}_{mr}, \boldsymbol{P}_{mt} \in \boldsymbol{R}^{N_m \times 3P^2}$，其中 $N_s = H_s W_s / P^2$，$N_m = H_m W_m / P^2$ 表示搜索区域token和模板token的数量。然后，对这些序列应用具有线性投影的补丁嵌入层，以获得RGB和TIR搜索区域和模板的嵌入特征 $\boldsymbol{E}_{sr}, \boldsymbol{E}_{mr}, \boldsymbol{E}_{st}, \boldsymbol{E}_{mt}$。然后将 $\boldsymbol{E}_{sr}, \boldsymbol{E}_{mr}, \boldsymbol{E}_{st}, \boldsymbol{E}_{mt}$ 添加位置嵌入得到RGB和TIR搜索区域及模板的初始特征 $\boldsymbol{S}_r, \boldsymbol{S}_t, \boldsymbol{M}_r, \boldsymbol{M}_t$。

由于RGB和TIR模态的特征提取相同，本文仅讨论RGB模态。RGB模态的token连接方式为 $\boldsymbol{H}_r = [\boldsymbol{S}_r, \boldsymbol{M}_r] \in \boldsymbol{R}^{(N_s + N_m) \times C}$，然后将 $\boldsymbol{H}_r$ 送入一系列Transformer块。在每个Transformer块中，首先对 $\boldsymbol{H}_r$ 执行三个投影得到查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$。然后，对特征进行矩阵乘法聚合，得到注意力权值如下：

$$A = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{C}}\right), \tag{1}$$

其中：C 为通道数，$\sqrt{C}$ 为缩放因子；上标T表示矩阵转置。

2.2　通道特征融合模块（CFFM）

目前，基于Transformer的RGB-T目标跟踪的方法一般通过中间提示特征或使用部分融合的多模态标记来实现，但它们之间的间接特征交互导致了对多模态特征原始语义的利用不充分，这种不足限制了模型在多样环境中的适应性和鲁棒性。基于此，本文在TBSI[14]的骨干网络上引入了通道特征融合模块来直接对多模态特征进行交互融合。本文提出的通道特征融合模块主要由CBAM（Convolutional Block Attention Module）注意力[17]和局部聚合模块（Local Aggregation Module，LAM）构成。CBAM注意力机制能够捕捉和突出多模态特征的互补性，自动识别出在不同环境条件下最为关键的特征信息，确保关键信息得到优先关注和增强。此外，CFFM优化了特征融合策略，使模型能够灵活应对多种复杂场景。整体而言，通道特征融合模块显著增强了RGB-T目标跟踪模型的信息融合能力，从而有效提升了跟踪效果。

通道特征融合模块（CFFM）如图2所示。首先对从主干网络得到模板区域和搜索区域的RGB特征和TIR特征进行线性融合。以搜索区域的RGB和TIR特征融合为例说明，将搜索区域的RGB和TIR特征串联，使用线性层获得融合特征。

$$\boldsymbol{S}_c = Linear\left([\boldsymbol{S}_r; \boldsymbol{S}_t]\right) \in \boldsymbol{R}^{N_m \times C}, \tag{2}$$

其中，Linear表示线性层。然后，使用CBAM模块作为通道增强单元，实现 $\boldsymbol{S}_c$ 的通道增强。其中，CBAM由通道注意力模块和空间注意力模块组成，分别对特征图的通道维度和空间维度进行

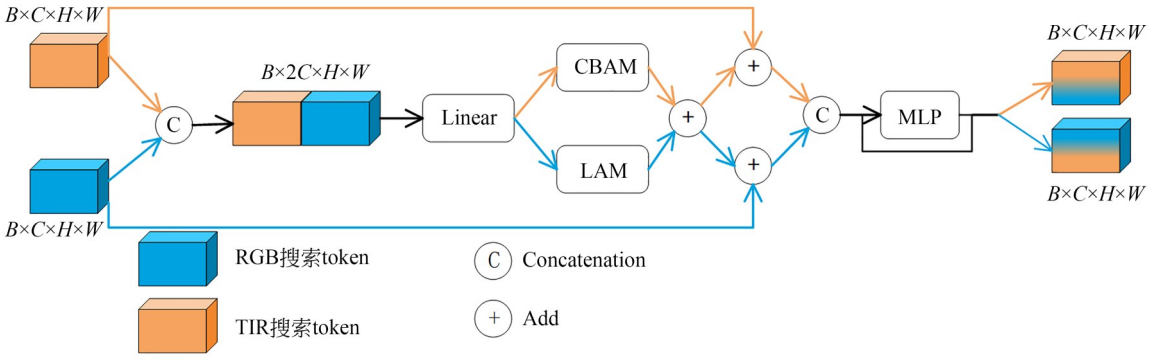


图2　通道特征融合模块结构图

Fig. 2　Structure diagram of channel feature fusion module

加权。

$$CBAM=\sigma(W_1F^c_{\text{avg}}+W_0F^c_{\max})\cdot \sigma(f_{7\times7}([F^s_{\text{avg}};F^s_{\max}])),\tag{3}$$

$$S_c^{\text{CBAM}}=CBAM(S_c),\tag{4}$$

其中：F^c_{avg}和$F^c_{\max}$分别表示通道维度的平均池化和最大池化结果，σ是Sigmoid激活函数；F^s_{avg}和$F^s_{\max}$分别表示空间维度的平均池化和最大池化结果，$f^{7\times7}$是一个7×7的卷积操作。

局部聚合模块(Local Aggregation Module, LAM)[18]用于有效整合不同尺度的通道特征。如图3所示，通过逐点卷积和深度卷积的组合，能够充分利用不同尺度的特征，提高模型对输入信息的理解能力。

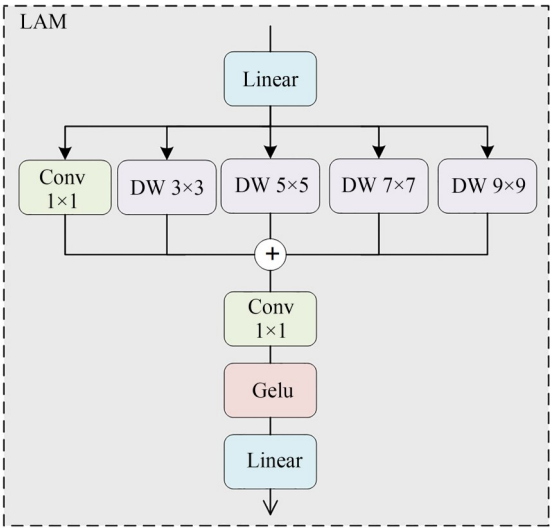


图3　局部聚合模块结构图

Fig. 3　Structure diagram of local aggregation module

具体来说，对特征S_c进行线性变换，得到$S_c^{\text{fc}}=Linear(S_c)$。其次，使用多个卷积块来提取不同尺度的特征，表示为：

$$S_c^{\text{sf}}=Conv_{1\times1}(S_c^{\text{fc}})+DWConv_{k\times k}(S_c^{\text{fc}})\quad(k=3,5,7,9),\tag{5}$$

其中：$Conv$和$DWConv$表示具有归一化(BN)的卷积和具有BN的深度可分离卷积。另一个线性层用于对融合特征进行线性变换，整合不同尺度的特征并提高一致性，表示为：

$$S_c^{\text{LAM}}=Linear\left(Gelu\left(Conv_{1\times1}(S_c^{\text{sf}})\right)\right)\in R^{N_m\times C},\tag{6}$$

其中：$Gelu$为$Gelu$激活函数。在CBAM和LAM模块后，特征被相加，表示为$S_c^{\text{add}}=S_c^{\text{CBAM}}+S_c^{\text{LAM}}$。此外，引入残差连接来补充多模态特征中丢失的原始语义信息，表示为$S_r^{\text{res}}=S_r+S_c^{\text{add}}$和$S_t^{\text{res}}=S_t+S_c^{\text{add}}$。

将RGB和TIR特征连接，并将其输入到残差连接的多层感知器(MLP)中。对拼接的特征进行拆分，最终得到RGB和TIR经过交互融合后的特征。该过程描述为：

$$S_r^{\text{cffm}},S_t^{\text{cffm}}=Split\left(\left[S_r^{\text{res}};S_t^{\text{res}}\right]+MLP\left(\left[S_r^{\text{res}};S_t^{\text{res}}\right]\right)\right),\tag{7}$$

其中：$Split$将特征从$R^{N_m\times2C}$映射到两个$R^{N_m\times C}$特征，MLP由三个线性层组成，S_r^{cffm}和S_t^{cffm}是CFFM的输出特征。

2.3　跨模态特征融合模块(CMFM)

尽管通道特征融合模块(CFFM)通过通道注意力机制和局部聚合策略实现了模态间特征的初步交互与融合，但RGB与TIR模态在物理属性上的显著差异，仍然可能引起两种模态在特征提取阶段的信息不对称问题。这种不对称性可能会对融合后的特征质量造成影响，进而降低跟踪算法的整体性能。因此，在通道特征融合模块之后，进一步引入跨模态特征融合模块(CMFM)。CMFM通过模态间的信息共享，进一步消除特征表达中的不对称性，增强了目标的鲁棒性表示。CMFM可以动态调整不同模态特征的权重，从而在不同环境条件下更灵活地利用具有判别力的模态信息。特别是在复杂环境或光照变化显著的情况下，CMFM能够有效缓解因模态差异带来的信息缺失问题，显著提升模型对目标的精准捕捉和跟踪能力。

CMFM的结构如图4所示，将经过通道特征融合模块之后的模板区域及搜索区域的RGB特征和TIR特征通过1×1卷积层对通道数进行调整。使用交叉增强策略，通过学习两种模态的增强特征来利用它们之间的相关性。将RGB特征和TIR特征送至具有Sigmoid函数的3×3卷积层中，可以获得归一化的特征图。以搜索区域的RGB和TIR特征融合为例说明，搜索区域的RGB和TIR特征表示为：

$$\omega_r=\sigma\left(Conv_{3\times3}(S_t^{\text{cffm}})\right)\in[0,1],\tag{8}$$

$$\omega_t=\sigma\left(Conv_{3\times3}(S_r^{\text{cffm}})\right)\in[0,1],\tag{9}$$

其中：$Conv_{3\times3}$表示3×3卷积操作，$\sigma()$表示Sig-

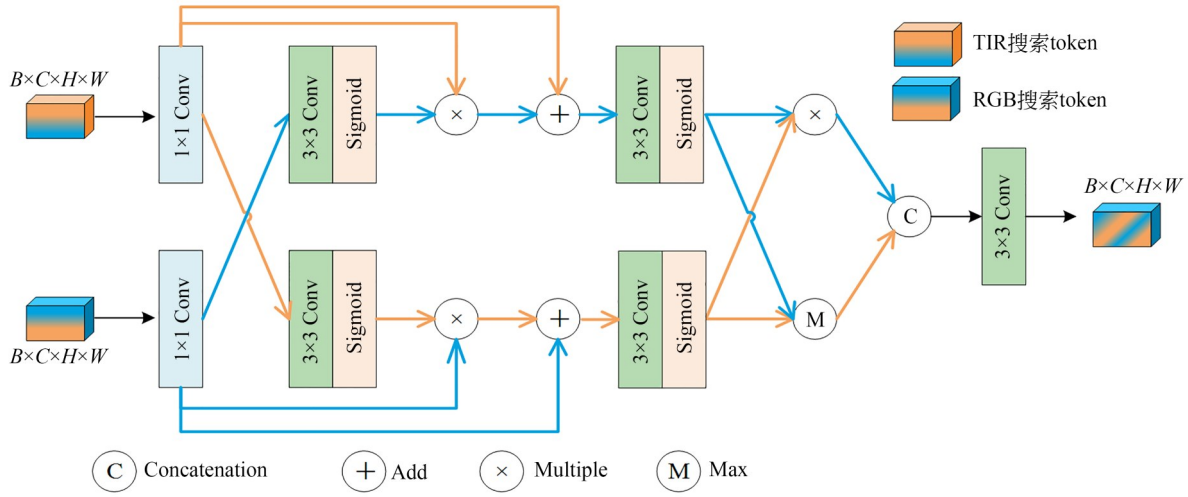


图 4　跨模态特征融合模块结构图

Fig. 4　Structure diagram of cross-modal feature fusion module

moid激活函数。为充分利用两种模态之间的相关性，归一化的特征图可以被视为特征级注意力图，以自适应地增强特征表示。此外，为了保留每个模态的原始信息，本文使用残差连接将增强的特征与其原始特征相结合。两种模态增强特征表示为：

$$S_r' = S_r^{\text{cffm}} + S_r^{\text{cffm}} \times \omega_t, \tag{10}$$

$$S_t' = S_t^{\text{cffm}} + S_t^{\text{cffm}} \times \omega_r, \tag{11}$$

其中：×表示逐元素相乘。CMFM使用自适应融合策略，将获得的交叉模态增强特征进行有效融合。具体地，两种模态增强特征（S_r'和S_t'）被送入$Conv_{3\times3}$中，以学习它们在融合过程中的权重ω_r和ω_t，学习到的权重ω_r和ω_t通过Sigmoid激活函数被归一化到[0,1]区间内，确保了融合过程中的权重分配是合理的且能够根据输入数据的特性动态调整。其次，应用元素乘法和最大化策略来融合这些加权后的特征，将结果连接在一起。将获得的两个模态增强特征送到卷积层中，以适应不同模态特征的动态权重分布。然后进行逐个元素相乘和最大化，表示为：

$$\omega_{r1} = \sigma\left(Conv_{3\times3}(S_r')\right) \in [0,1], \tag{12}$$

$$\omega_{t1} = \sigma\left(Conv_{3\times3}(S_t')\right) \in [0,1], \tag{13}$$

$$S_r'' = S_r' \times \omega_{r1}, \tag{14}$$

$$S_t'' = S_t' \times \omega_{t1}, \tag{15}$$

$$p_{\text{mul}} = S_r'' \times S_t'', \tag{16}$$

$$p_{\max} = \left\{S_r''; S_t''\right\}, \tag{17}$$

其中：$Conv_{3\times3}$表示3×3卷积操作，max()表示最大化操作，p_{mul}和$p_{\max}$表示逐个元素相乘和最大化的结果。将逐个元素相乘和最大化的结果连接为$p = [p_{\text{mul}}, p_{\max}]$，再通过$Conv_{3\times3}$进一步提升特征表达，获得输出$M_{\text{cmfm}}^{s}$。

2.4　损失函数

最后一个Transformer模块输出的搜索区域的RGB和TIR特征分别表示为X_r^{final}和X_t^{final}，将其相加得到混合特征，描述如下：

$$X_{rt}^{\text{final}} = X_r^{\text{final}} + X_t^{\text{final}}. \tag{18}$$

本文采用全卷积中心头来预测目标的状态，使用Focal Loss训练分类分支。

对于正样本（$target=1$）：

$$FL_{\text{pos}} = \log(p)(1-p)^{\alpha}. \tag{19}$$

对于负样本（$target=0$）：

$$FL_{\text{neg}} = \log(1-p)\, p^{\alpha}(1-target)^{\beta}. \tag{20}$$

最后，将正样本和负样本的损失加和，并根据正样本的数量进行归一化，得到最终的Focal Loss。在本文中，将α设置为2，β设置为4，而p是模型预测该样本为正类的概率。

使用L1*loss*和*giou loss*训练回归分支。

$$L1loss = MAE = \left| f(x_i) - y_i \right|, \tag{21}$$

$$IOU = \frac{|A \cap B|}{|A \cup B|}, \tag{22}$$

$$giou = IOU - \frac{|C - (A \cup B)|}{|C|}, \tag{23}$$

$$giou\ loss = 1 - giou, \tag{24}$$

其中：$f(x_i)$和y_i分别表示第i个预测框的值及相应真实框的值。公式(23)中，C为包围A和B的最小矩形框的面积。

可得MMFFTN的总训练损失为：

$$L_{\text{total}}=L_{\text{cls}}+\lambda_{\text{giou}}L_{\text{giou}}+\lambda_{l1}L_1, \tag{25}$$

其中：$\lambda_{\text{giou}}=2$，$\lambda_{l1}=5$。

3 实　验

3.1 实验配置

MMFFTN基于64位Linux操作系统，使用Python 3.8和Pytorch 1.13.0，在两片RTX3080 GPU上进行训练。本文在LasHeR的训练集上对MMFFTN进行了50个epoch的训练，每个epoch由60 K个图像对组成。MMFFTN使用Vision Transformer作为特征提取网络，在SOT上进行预训练以实现骨干网络的初始化。在实验参数的设置中，模板和搜索区域的输入尺寸分别设置为128×128和256×256。利用AdamW优化器对模型进行更新，初始学习率设为2×1e-5，权值衰减均设为1×10^{-4}。CFFM和CMFM模块被级联并插入ViT的第4层、第7层和第10层。使用RGB-T追踪器[14]的ViT模型权重作为预训练权重。在GTOT和RGBT234和LasHeR数据集直接评估MMFFTN。

3.2 数据集

GTOT[19]是RGB-T跟踪器最通用的评估数据集，包含50个对齐的RGB和红外视频序列，涵盖7种挑战属性，如LSV(大尺度变化)、FM(快速移动)和LI(低光照)等，为评估跟踪器性能提供了丰富的测试场景。

RGBT234[20]是GTOT的扩展版本，包括234对多模态视频，并采用了与GTOT相同的指标。它提供12种属性，包括LI(低光照)、遮挡(OCC)，DEF(变形)、运动等。

LasHeR[21]是RGB-T跟踪领域规模较大、使用广泛的基准，包含1224个序列，其中测试包括245对视频。它分为19个具有挑战性的属性。LasHeR基准测试采用PR、SR和归一化精确(NPR)。各数据集的详细信息如表1所示。

在本文实验中，使用LasHeR为训练集，GTOT，RGBT234以及LasHeR为测试集。

表1 主要数据集

Tab. 1 Main datasets

数据集	序列	分辨率	最小帧数	最大帧数	平均帧数	总帧数	类别	属性
GTOT[19]	50	384×288	40	376	157	7.8K	9	7
RGBT234[20]	234	630×460	40	4 140	498	114.7 K	22	12
LasHeR[21]	1 224	630×480	57	12 862	600	734.8 K	32	19

3.3 评价指标

GTOT，RGBT234和LasHeR基准都使用相同的评估指标，即精确率(Precision Rate，PR)和成功率(Success Rate，SR)。PR表示跟踪器预测的目标中心点位置与真实目标中心点位置的差值小于给定阈值的视频帧数占所有视频检测帧的百分比。SR表示跟踪器预测框与真实标签框的重叠面积小于给定阈值的视频帧数占所有视频检测帧的百分比。GTOT数据集中存在大量小目标，将其阈值设置为5，将RGBT234数据集和LasHeR数据集阈值设置为20。

3.4 实验结果及分析

3.4.1 整体性能

GTOT数据集：在GTOT数据集上，将MMFFTN与其他10个跟踪器进行了比较，包括3个RGB跟踪器(SiameseFC[22]，ECO[23]和MDNet[5])和7种基于融合的RGB-T跟踪器(DFAT[24]，SiamCDA[25]，APFNet[9]，MACNet[26]，HMFT[27]，ADRNet[7]和TBSI[14])。结果如图5所示。本文所提出的MMFFTN的准确率和成功率分别为91%和75.8%。与TBSI相比，该方法的PR提高了2.7%，SR提高了1.8%。与

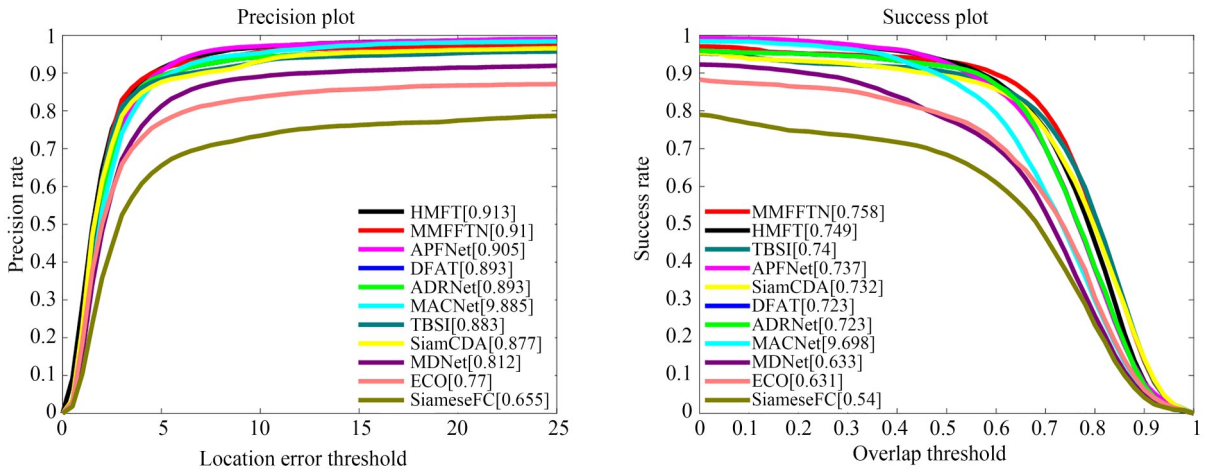


图5　MMFFTN及对比算法在GTOT数据集上的精确率图和成功率图

Fig. 5　Precision plot and Success plot of MMFFTN and compared algorithms on GTOT dataset

APFNet相比，该方法的PR提高了0.5%，SR提高了1.9%。

RGBT234数据集：在RGBT234数据集上，将其他10个高性能跟踪器与MMFFTN进行了比较，包括2个RGB追踪器（ECO[23]和C-COT）以及8种RGB-T跟踪器（EANet[28]，APFNet[9]，DFAT[24]，SiamCDA[25]，HMFT[27]，ADRNet[7]，ViPT[15]和SDSTrack[16]）。评估结果如图6所示。MMFFTN的PR分数比SDSTrack高1.2%，比ViPT高2.5%，SR比SDSTrack高2.1%，比ViPT高2.9%。与SDSTrack相比，MMFFTN的性能具有一定的优势。

LasHeR数据集：在LasHeR数据集上，将MMFFTN与其他8个RGB-T跟踪器（GMMT[29]，APFNet[9]，DAFNet[30]，CAT[7]，MANet[31]，TBSI[14]，ViPT[15]，SDSTrack[16]）进行了比较。评估结果如图7所示。MMFFTN的PR和SR分别为69.8%和55.5%，在PR指标上，MMFFTN比SDSTrack和ViPT分别高出3.3%和4.7%；在NPR指标上，分别提升3.4%和4.1%；在SR指标上，分别提高3.1%和3.0%。结果表明，MMFFTN实现了更好的跨模态特征交互融合，性能更佳。

loss曲线：如图8所示，训练Loss随着迭代次数的增加而持续下降，这表明模型在训练集上不断学习并逐渐优化其参数。

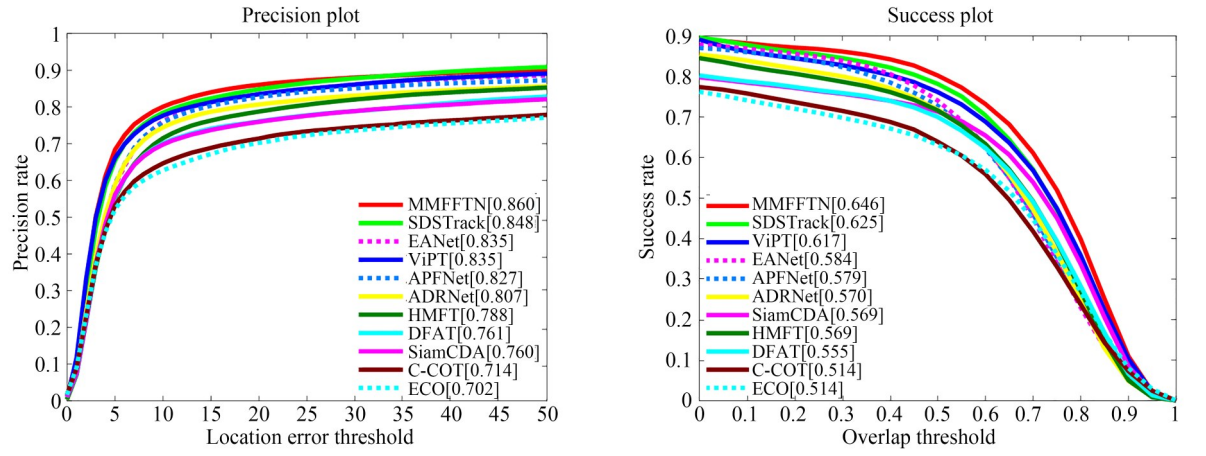


图6　MMFFTN及对比算法在RGBT234数据集上的精确率图和成功率图

Fig. 6　Precision plot and Success plot of MMFFTN and compared algorithms on RGBT234 dataset

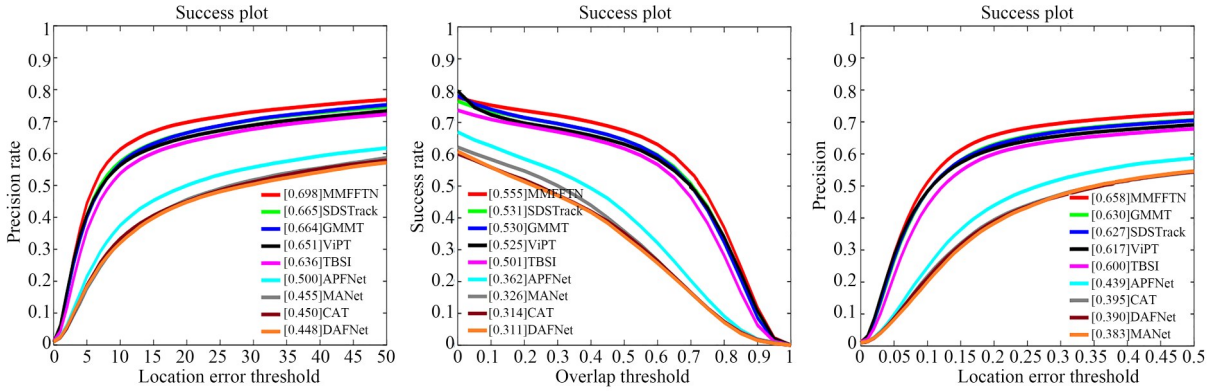


图 7　MMFFTN 及对比算法在 LasHeR 数据集上的精确率图、成功率图和归一化精确率图

Fig. 7　Precision, success, and normalized precision plots of MMFFTN and compared algorithms on LasHeR

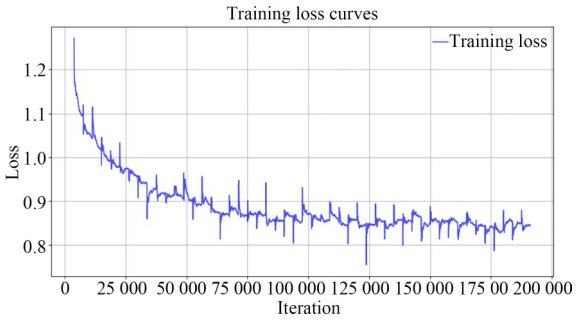


图 8　训练过程中的 loss 曲线

Fig. 8　Loss curves during training

3.4.2　基于属性挑战的评价

GTOT 数据集：本文对 MMFFTN 与其他 RGB-T 跟踪器（MACNet，SiamCDA，APFNet，DFAT，TBSI）在不同挑战属性的子集上进行了综合评估，结果如表 2 所示，最佳、第二和第三 PR/SR 用加粗、下划线和上划线表示。由结果可见，在大多数挑战中，MMFFTN 方法始终优于其他 RGB-T 跟踪器，证明了该跟踪器的有效性。

除了在 PR 和 SR 方面领先外，MMFFTN 在遮挡（OCC）、大尺度变化（LSV）和低照度（LI）方面也优于其他网络。这表明本文提出的特征融合模块有效地融合了 RGB 和 TIR 的特征信息。另一方面，MMFFTN 在形变（DEF）属性上的精确率和成功率均比较低，说明 MMFFTN 网络存在模板更新问题，导致网络无法适应运动过程中的形状变化。

表 2　GTOT 数据集上基于属性挑战的 PR/SR 分数

Tab. 2　PR/SR scores based on attribute challenges on the GTOT dataset　（%）

属性	OCC	LSV	FM	LI	TC	SO	DEF	ALL
MACNet	86.7/67.6	84.2/66.2	84.4/63.3	90.1/70.7	89.3/68.6	93.2/67.8	92.2/74.4	88.5/69.8
SiamCDA	82.2/69.4	91.5/74.8	86.2/71.9	92.4/96.4	82.6/68.5	87.9/**72.7**	**94.7/76.5**	87.7/73.2
APFNet	90.3/71.3	87.7/71.2	86.3/68.4	91.4/74.8	**90.4**/71.6	**94.6**/71.3	94.5/78	90.5/73.7
DFAT	86.3/68.7	92.4/75	**89.1/74**	92.2/74.1	89.1/70.7	94.4/71.0	91.9/73.5	89.3/72.3
TBSI	87.5/72.5	94.3/78	85.3/72	89.9/74.9	85.4/71.3	86.4/68.6	85.7/71.5	88.3/74
MMFFTN	**90.7/74.2**	**94.8/78.5**	86.4/73.7	**93.0/76.8**	88.7/**73.1**	89.7/71.1	90.1/74.1	**91/75.8**

RGBT234 数据集：本文对 MMFFTN 和 2 个 RGB 追踪器（ECO 和 C-COT）以及 8 种 RGB-T 跟踪器（EANet，APFNet，DFAT，SiamCDA，HMFT，ADRNet，ViPT 和 SDSTrack），在 RGBT234 数据集进行了性能评估，评估对象包括 12 个具有挑战性的属性子集，评估结果如图 9 所示。

在 12 个属性中，本文方法在部分遮挡（PO）和尺度变换（SV）的挑战中表现出色。主要得益于采用了 ViT 和特征融合策略。ViT 通过自注意力机制实现全局特征建模，能够有效捕捉目标的长距离依赖关系，在遮挡场景中定位未被遮挡的关键区域，并适应尺度变化的动态调整。同

时，本文提出的特征融合模块充分利用了RGB和TIR模态的互补特性。在其他属性挑战中，MMFFTN的表现优于大多数RGB-T跟踪器，这表明该方法有效地利用了双模态中的互补特征。然而，MMFFTN在低分辨率(LR)、无遮挡(NO)和热交叉(TC)属性上未能取得令人满意的结果。分析原因是CFFM模块无法动态处理双模态上下文中的噪声。

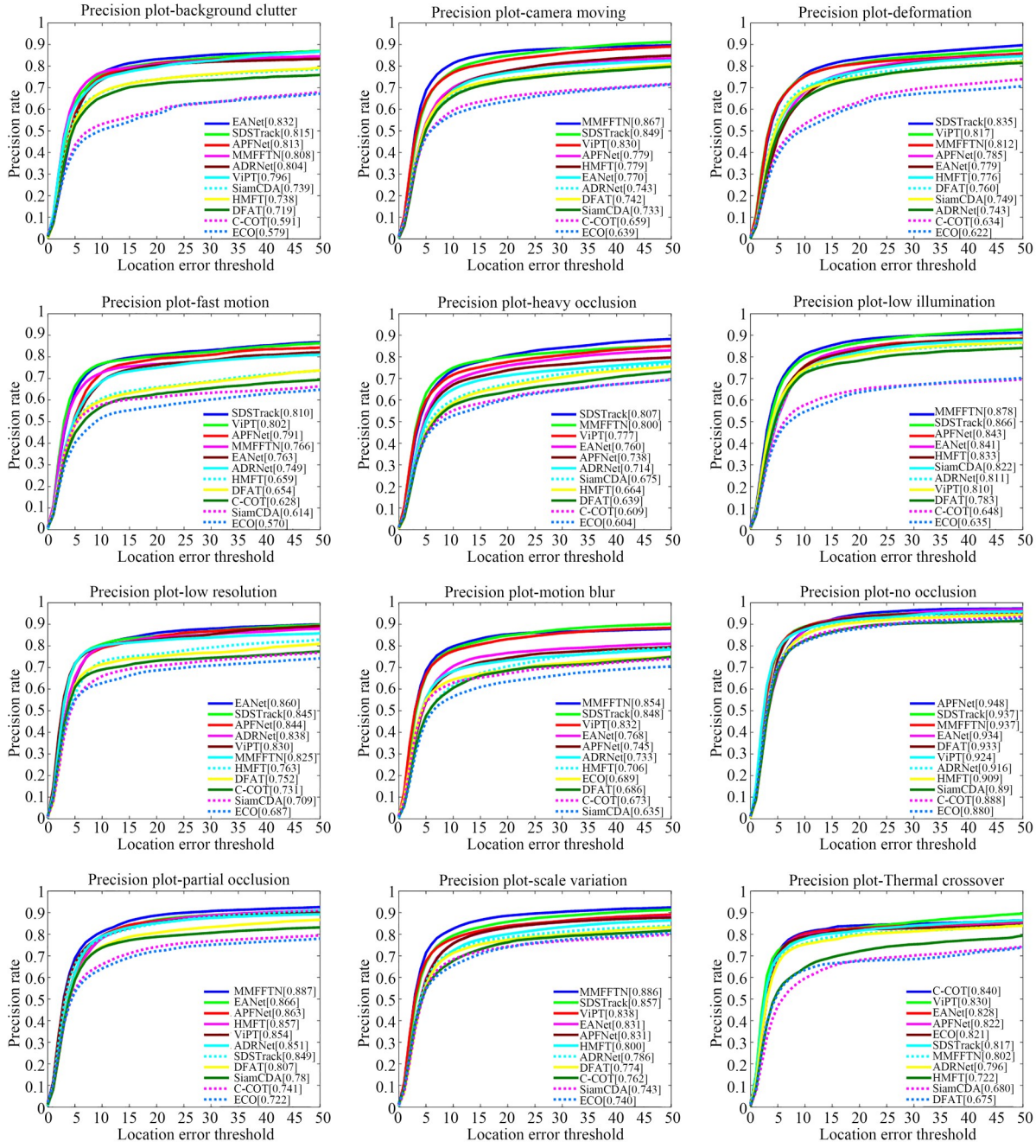


图9　在RGBT234数据集上不同属性下的精确率图

Fig. 9　Precision rate plot under various attributes on the RGBT234 dataset

LasHeR数据集：本文在LasHeR数据集的19个属性上评估了MMFFTN，ViPT和SDSTrack。结果如图10所示。在许多属性上，MMFFTN都优于其他跟踪器。与ViPT相比，MMFFTN在HI，DEF和ARC等挑战中的性能具有竞争力。与SDSTrack相比，MMFFTN在

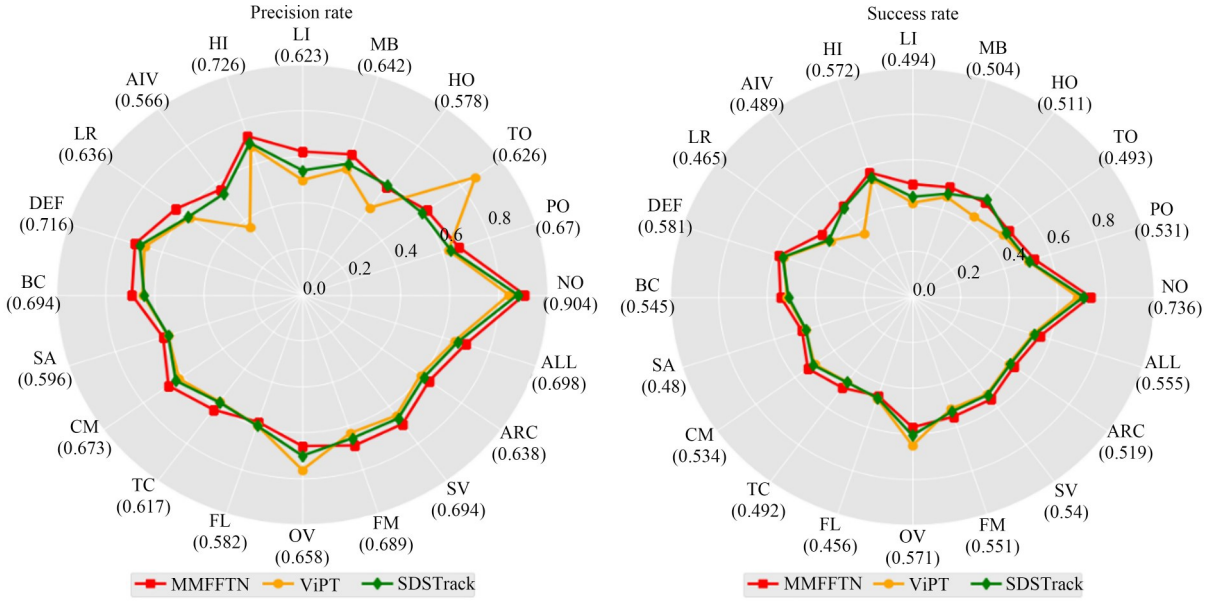


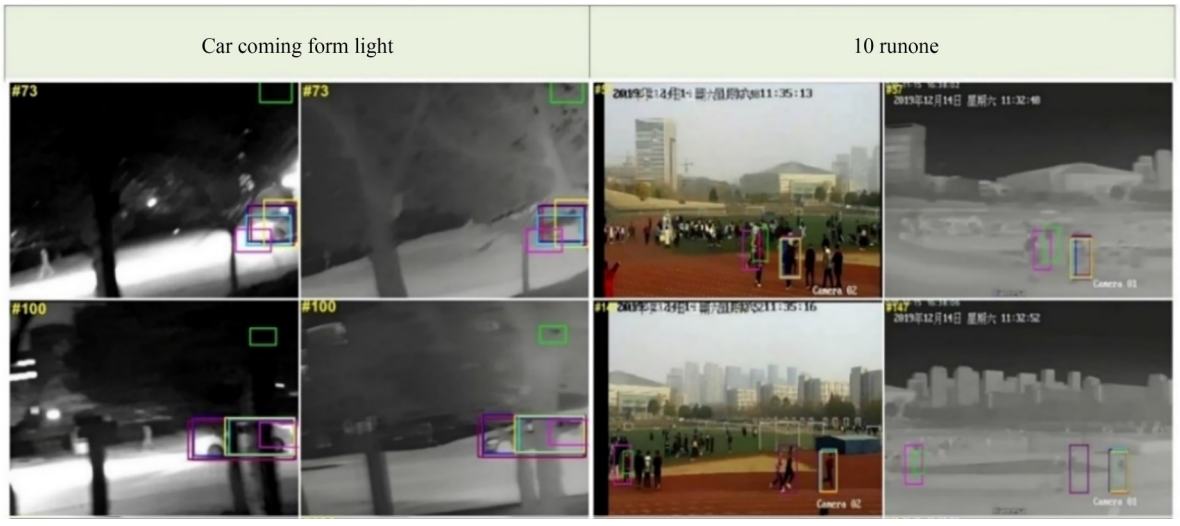
图 10　LasHeR 数据集上不同属性下的精确率和成功率图

Fig. 10　Precision and success rate plot in different attributes on the LasHeR dataset

NO、光照突变(AIV)和背景杂波(BC)等属性方面表现更好。然而，从图中可以看出，ViPT 在 TO 挑战中性能突出，这是由于 ViPT 引入了视觉提示的增强机制以及时间上下文建模能力，有效缓解了遮挡导致的跟踪丢失问题。相比之下，MMFFTN 在 TO，HO，FL 和 OV 属性上的 PR 和 SR 表现较为不足，这说明其模型的鲁棒性仍有待进一步提升。

3.4.3　可视化分析

为了更直观地比较跟踪器的跟踪性能，在 LasHeR 数据集上选取 2 个具有挑战性的视频序列，将 MMFFTN 与 ViPT，SDSTrack，TBSI，MANet，CAT5 种先进跟踪算法以及视频标注 GT 进行比较，并将跟踪结果进行可视化，可视化结果如图 11 所示。其中，在 carcomingfromlight 序列中，当跟踪目标部分遮挡时，由第 73 帧和 100 帧可知，在光照变化的条件下，除了 MMFFTN 和 ViPT 能继续跟踪目标，其余算法的跟踪框均发生了不同程度的偏移。除此之外，由第 345 帧可知，当跟踪目标被大部分遮挡时，只有 MMFFTN 可以跟踪目标，其余算法的跟踪框均发生偏移。在 10runone 序列中，在目标快速移动以及背景杂波



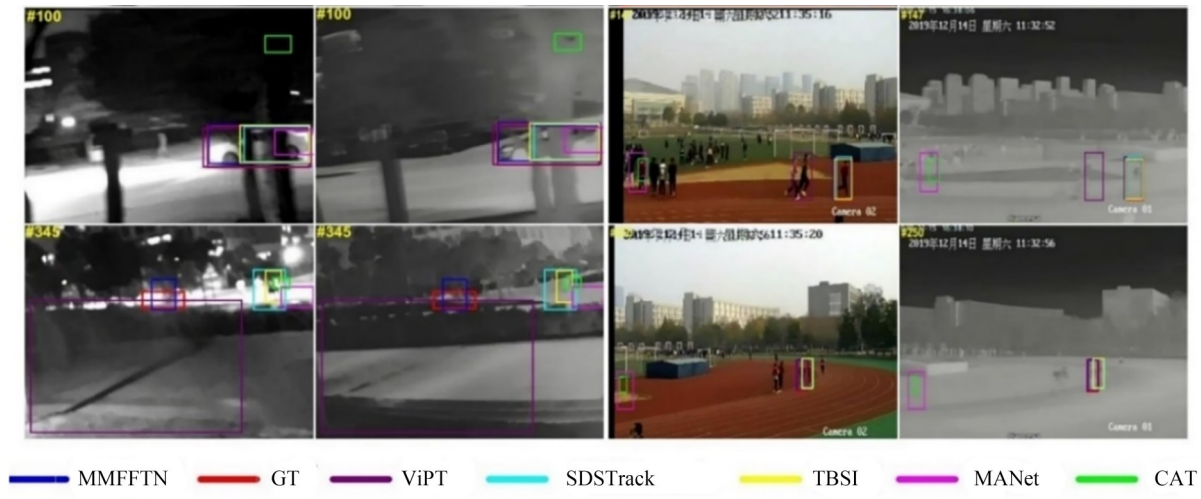


图11　LasHeR数据集上的跟踪可视化比较结果

Fig. 11　Comparison results of tracking visualization on the LasHeR dataset

等的挑战下，由第57帧和147帧可知，MANet和CAT均丢失了跟踪目标；由第250帧可知，MMFFTN与GT的重叠程度最高。通过可视化分析可知MMFFTN的综合跟踪性能最好。

3.4.4　消融实验

MMFFTN组件分析：为了研究不同组件对MMFFTN跟踪性能的影响，证明各个组件的有效性，本文在LasHeR上进行了消融实验。

首先评估RGB和TIR特征直接拼接以及CFFM和CMFM模块对跟踪器性能的影响。

Baseline表示将共享ViT骨干输出的搜索区域的RGB和TIR特征直接进行拼接。消融实验的结果如表3所示，通过CFFM模块和CMFM模块的贡献，PR分别提高了2.2%和2.2%，NPR分别提高了2.2%和2.1%，SR分别提高了2.4%和1.3%。此外，它们还分别在PR，NPR和SR方面实现了4.4%，4.3%和3.7%的提高。实验结果表明，本文提出的CFFM模块和CMFM模块对于跟踪器跟踪性能的提高都是十分有效的。

CFFM组件分析：本文评估了CFFM各组成部分对模型性能的具体贡献。如表4所示，当从模型中删除CFFM模块时，PR和SR分别为66.5%和52.5%。添加CBAM模块后，PR和SR分别提高了1.0%和1.5%。然后，使用LAM模块整合全局特征的结果是PR和SR分别提高了2.3%和1.5%。总体而言，CFFM模块中的两个关键组成部分均对提升模型的整体性能起到了积极作用。

不同损失权重对性能的影响：为探究损失函数中IOU损失与L1损失的权重分配对跟踪性能的影响，本文在LasHeR数据集上进行了不同损失权重对性能的影响分析。结果如表5所示。

表3　增加不同模块时LasHeR数据集上的指标值

Tab. 3　Increase in metrics on the LasHeR dataset with the addition of different modules　(%)

方法	PR	NPR	SR
Baseline	65.4	61.5	51.8
Baseline+CFFM	67.6	63.7	54.2
Baseline+CFFM+CMFM	69.8	65.8	55.5

表4　在LasHeR数据集上CFFM不同模块的消融

Tab. 4　Ablation of different modules of CFFM on LasHeR dataset　(%)

CFFM		PR	NPR	SR
CBAM	LAM			
		66.5	62.6	52.5
√		67.5	64.8	54.0
√	√	69.8	65.8	55.5

表5　不同权重对模型性能的影响

Tab. 5　Effect of different weights on model performance　(%)

λ_{giou}	λ_{l1}	PR	SR
1	1	65.0	51.3
2	3	67.5	53.4
5	2	66.2	52.3
2	5	69.8	55.5

特征可视化：为了验证CFFM和CMFM对特征表征的增强作用，通过对可见光与热红外图像的特征融合过程进行可视化分析，验证了CFFM与CMFM模块的核心作用。如图12所示，结果表明，两模块协同作用显著提升了模型的跟踪效果。

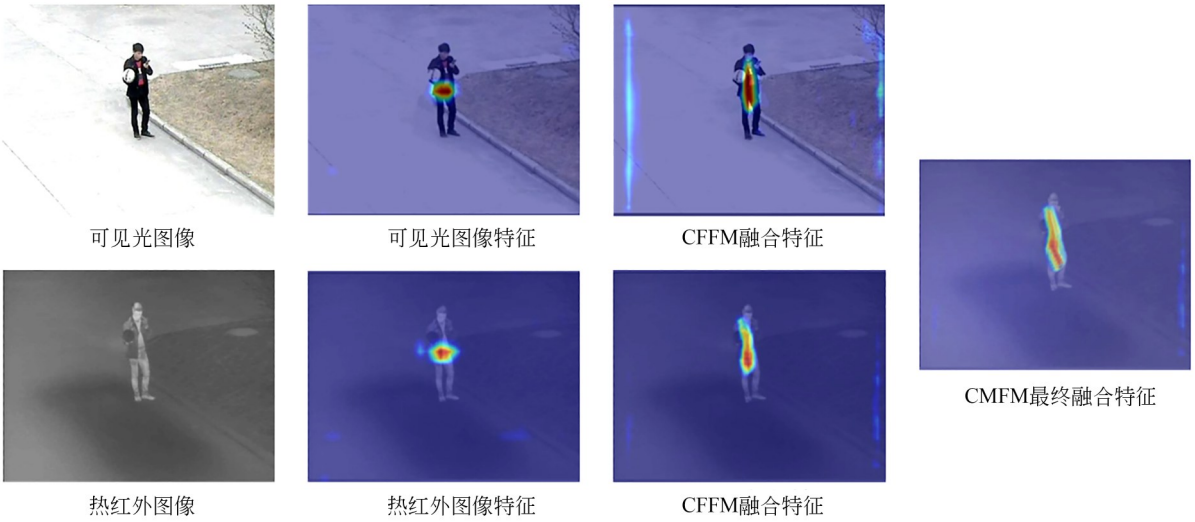


图12　跨模态图像特征融合可视化热力图

Fig. 12　Cross-modal image feature fusion visualization of heat maps

4　结　论

本文提出了一种鲁棒的RGB-T跟踪网络MMFFTN，该网络主要由特征提取网络、通道特征融合模块和跨模态特征融合模块组成。模型通过ViT分别提取RGB和TIR的特征；设计通道特征融合模块对RGB和TIR模态特征进行直接交互，从而增强了局部特征在两模态之间的协同作用，有效提升了模态间的互补性。同时，针对模态融合效果不佳的问题，设计了跨模态特征融合模块以有效融合RGB与TIR的全局特征。该模块利用残差连接生成交叉增强特征，并进一步采用自适应融合策略对这些增强特征进行优化融合，提高跟踪器的鲁棒性。在GTOT，RGBT234和LasHeR这三个RGB-T数据集上的实验结果充分表明，相较于现有方法，本文方法展现出了较为优越的性能。后续研究将继续探索更加高效的多模态融合方案和模板更新策略，以期进一步提升RGB-T跟踪器的整体性能。

作者贡献说明：

金静：提供资源，监督和领导实验的策划，论文审核；

刘建琴：实验的设计及数据整理和分析，论文构思和撰写，数据的可视化展现；

翟凤文：论文审核。

参考文献：

[1] ZHANG Z T, WANG Y C, MEI D Q, *et al*. Highly sensitive and stretchable ultrasonic transducer array for object internal characteristics detection in robotics[J]. *IEEE Transactions on Instrumentation and Measurement*, 2023, 72: 7503909.

[2] VENNILA T J, BALAMURUGAN V. A rough set framework for multihuman tracking in surveillance video[J]. *IEEE Sensors Journal*, 2023, 23(8): 8753-8760.

[3] WANG G R, JIANG Q, JIN X, *et al*. SiamTDR: time-efficient RGBT tracking *via* disentangled representations[J]. *IEEE Transactions on Industrial Cyber-Physical Systems*, 2023, 1: 167-181.

[4] 杜东丽. 基于Transformer的高性能RGB-T目标跟踪研究[D]. 金华：浙江师范大学，2023.

DU D L. *Research on High Performance RGB-T Target Tracking based on Transformer*[D]. Jinhua: Zhejiang Normal University, 2023. (in Chinese)

[5] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 27-30, 2016, *Las Vegas*, *NV*, *USA*. IEEE, 2016: 4293-4302.

[6] 张天路, 张强. 基于深度学习的RGB-T目标跟踪技术综述[J]. 模式识别与人工智能, 2023, 36(4): 327-353.
ZHANG T L, ZHANG Q. A survey of RGB-T object tracking technologies based on deep learning [J]. *Pattern Recognition and Artificial Intelligence*, 2023, 36(4): 327-353. (in Chinese)

[7] ZHANG P Y, WANG D, LU H C, *et al*. Learning adaptive attribute-driven representation for real-time RGB-T tracking[J]. *International Journal of Computer Vision*, 2021, 129(9): 2714-2729.

[8] LI C L, LIU L, LU A D, *et al*. *Challenge-aware RGBT Tracking* [M]. Computer Vision-ECCV2020. Cham: Springer International Publishing,

[9] XIAO Y, YANG M M, LI C L, *et al*. Attribute-based progressive fusion network for RGBT tracking [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(3): 2831-2838.

[10] 王暐, 付飞亚, 雷灏, 等. 基于注意力交互的可见光红外跟踪算法[J]. 光学 精密工程, 2024, 32(3): 435-444.
WANG W, FU F Y, LEI H, *et al*. Attention interaction based RGB-T tracking method[J]. *Opt. Precision Eng.*, 2024, 32(3): 435-444. (in Chinese)

[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An Image Is worth 16×16 words: transformers for image recognition at scale[C]. *International Conference on Learning Representations*, 2021. 1.

[12] 刘万军, 梁林林, 曲海成. 利用Transformer的多模态目标跟踪算法[J]. 计算机工程与应用, 2024, 60(11): 84-94.
LIU W J, LIANG L L, QU H C. Trans-RGBT: RGBT object tracking with transformer[J]. *Computer Engineering and Applications*, 2024, 60(11): 84-94. (in Chinese)

[13] FENG M Z, SU J B. Learning reliable modal weight with transformer for robust RGBT tracking [J]. *Knowledge-Based Systems*, 2022, 249: 108945.

[14] HUI T R, XUN Z Z, PENG F G, *et al*. Bridging search region interaction with template for RGB-T tracking [C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 13630-13639.

[15] ZHU J, LAI S, CHEN X, *et al*. Visual prompt multi-modal racking[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 9516 - 9526.

[16] HOU X J, XING J Z, QIAN Y J, *et al*. SDSTrack: self-distillation symmetric adapter learning for multi-modal visual object tracking [C]. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 16-22, 2024, *Seattle*, *WA*, *USA*. IEEE, 2024: 26541-26551.

[17] WOO S, PARK J, LEE J Y, *et al*. CBAM: Convolutional block attention module[C]. *Computer Vision-ECCV* 2018. *Cham*: *Springer International Publishing*, 2018: 3-19.

[18] LI Y F, WANG B, LI Y, *et al*. Transformer-based RGB-T tracking with channel and spatial feature fusion[J]. *arXiv preprint* arXiv: 2405.03177, 2024.

[19] 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.

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

[21] 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.

[22] BERTINETTO L, VALMADRE J, HENRIQUES J F, *et al*. *Fully-Convolutional Siamese Networks ffor Object Tracking*[M]. Computer Vision - ECCV 2016 Workshops. Cham: Springer International Publishing, 2016: 850-865.

[23] 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*). *July* 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 6931-6939.

[24] TANG Z Y, XU T Y, LI H, *et al*. Exploring fusion strategies for accurate RGBT visual object tracking [J]. *Information Fusion*, 2023, 99: 101881.

[25] ZHANG T, LIU X, ZHANG Q, *et al*. SiamCDA: Complementarity-and distractor-aware RGBT tracking based on Siamese network [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, 32(3): 1403-1417.

[26] 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.

[27] 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*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 8876-8885.

[28] TÜRKOĞLU A, AKAGUNDUZ E. EANet: Enhanced attribute-based RGBT tracker network [C]. *Sixteenth International Conference on Machine Vision* (*ICMV* 2023). *November* 15-18, 2023. *Yerevan*, *Armenia*. SPIE, 2024.

[29] TANG Z Y, XU T Y, WU X J, *et al*. Generative-based fusion mechanism for multi-modal tracking [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(6): 5189-5197.

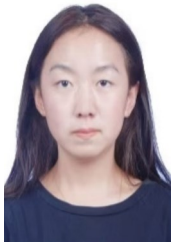
[30] GAO Y, LI C L, ZHU Y B, *et al*. Deep adaptive fusion network for high performance RGBT tracking [C]. 2019 *IEEE/CVF International Conference on Computer Vision Workshop* (*ICCVW*). *October* 27-28, 2019. *Seoul*, *Korea* (*South*). IEEE, 2019.

[31] 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*). *October* 27-28, 2019. *Seoul*, *Korea* (*South*). IEEE, 2019.

作者简介：

金　静(1982－),女,甘肃礼县人,博士,副教授,2020年于兰州交通大学获得博士学位,现为兰州交通大学电子与信息工程学院软件工程系教师。主要研究方向为模式识别、图像处理、人工智能。E-mail: mailto: 28089092@qq.com

通讯作者：



刘建琴(2000－),女,山西朔州人,硕士研究生,主要研究方向为计算机视觉。E-mail: mailto:1970477938@qq.com