采用注意力与模板在线更新的可见光-红外目标跟踪网络

韩向东[1]，钟傲[2]，刘冲澳[2]，孙延鑫[1]，张向永[1]，徐淋智[2]
（1. 中国兵器工业试验测试研究院智能技术中心，710065，西安；
2. 西安电子科技大学电子工程学院，710071，西安）

摘要：针对当前可见光-红外目标跟踪算法对红外可见光双模态特征交互与目标动态变化建模不足的问题，结合卷积掩码自编码器模型，提出一种基于注意力与模板在线更新的可见光-红外双模态目标跟踪网络。以卷积掩码自编码器模型为骨干网络，通过采用双嵌入层配合共享权重的骨干网络结构提取可见光与红外特征，深入挖掘可见光与红外数据间的内在联系。通过强化模板与搜索图像的关联性，引入通道空间自注意力机制以增强模板和搜索图像间的交互，来提取模态间可区分的异质互补特征。提出模板在线更新模块，通过在线更新模板与设计模板分数头，利用置信度评分机制融合初始模板的稳定性与在线模板的适应性，解决目标随时间变化导致的模型漂移问题。实验结果表明，所提算法在GTOT和RGBT234公开数据集上的精确率和成功率分别达到93.3%/75.6%和87.2%/63.8%，可在目标不断变化情况下实现精确跟踪。可视化分析表明，所提算法在双模态热力图上可自适应互补，单一模态失效时仍能精准定位目标。

关键词：目标跟踪；可见光-红外；注意力机制；模板在线更新；卷积掩码自编码器
中图分类号：TP391.4 **文献标志码**：A
DOI：10.7652/xjtuxb202508018 **文章编号**：0253-987X(2025)08-0187-12

RGB-Thermal Object Tracking Network Based on Attention Mechanism and Online Template Update

HAN Xiangdong[1], ZHONG Ao[2], LIU Chongao[2], SUN Yanxin[1],
ZHANG Xiangyong[1], XU Linzhi[2]
(1. Intelligent Technology Center, China Ordnance Industry Testing and Research Institute, Xi'an 710065, China;
2. School of Electronic Engineering, Xidian University, Xi'an 710071, China)

Abstract: To address the insufficient feature interaction between RGB and thermal modalities and the inadequate modeling of dynamic target variations in current RGB-thermal object tracking algorithms, a dual-modal tracking network based on attention and online template updates is proposed, incorporating a convolutional masked autoencoder model. Using the convolutional masked autoencoder as the backbone network, the model extracts RGB and thermal features through a dual-embedding layer with shared-weight backbone architecture, deeply exploring the intrinsic relationships between RGB and thermal data. To enhance the correlation between the template and search images, a channel-spatial self-attention mechanism is introduced to strengthen their

收稿日期：2024-12-21。 作者简介：韩向东(1989—)，男，副研究员。 基金项目：国家自然科学基金资助项目(62173265)；航空科学基金资助项目(F024020021)；国防科工局资助项目。
网络出版时间：2025-04-07 网络出版地址：https://link.cnki.net/urlid/61.1069.T.20250407.1101.002

interaction and extract discriminative heterogeneous complementary features across modalities. An online template update module is proposed, which dynamically updates the template and incorporates a template scoring head. By leveraging a confidence-based fusion mechanism, it balances the stability of the initial template and the adaptability of the online template, mitigating model drift caused by target variations over time. Experimental results demonstrate that the proposed algorithm achieves precision and success rates of 93.3%/75.6% and 87.2%/63.8% on the GTOT and RGBT234 datasets, respectively, enabling accurate tracking under dynamic target conditions. Visualization analysis shows that the algorithm adaptively complements dual-modal heatmaps and maintains precise target localization even when one modality fails.

Keywords: visual object tracking; RGB-thermal; self-attention mechanism; online template update; convolutional masked autoencoder

视觉目标跟踪在视频序列中根据首帧目标位置准确预测目标在后续帧中的位置,在视频监控、自动驾驶以及军事侦察等方面都有广泛应用[1-3]。基于可见光-红外(RGB-thermal,RGBT)双传感器图像的跟踪算法充分利用两种模态数据的相关性和补充性,提高跟踪算法的精度和鲁棒性。可见光与红外数据的互补性[4]体现为:可见光图像的高分辨率特性与细节信息能辅助红外解决目标跟踪时的热交叉问题;红外的热特性和强穿透性可以辅助可见光解决目标跟踪时因光照不佳等导致的精度下降问题。

早期 RGBT 目标跟踪采用基于手工特征的传统机器学习算法,如基于相关滤波[5]与稀疏表示[6]等的算法。传统相关滤波算法高度依赖手工设计特征,如方向梯度直方图(histogram of oriented gradients,HOG)、颜色直方图等。这些手工特征在复杂环境下难以捕捉多层次语义信息。当目标发生非刚性形变时,HOG 特征无法有效描述局部细节,导致跟踪框漂移。在光照剧烈变化或背景存在动态干扰的情况下,手工特征的区分度会大幅下降,进而使得跟踪精度降低。基于深度学习的 RGBT 目标跟踪算法借助深度学习的强大特征提取能力,提高跟踪精度。RGBT 目标跟踪算法主要包含 3 类:基于多域网络、基于孪生网络和基于注意力机制的 RGBT 目标跟踪算法[4,7-10]。

多域网络系列算法以将目标和背景进行二分类的思路对候选框进行判别,选取分类分数最高的候选框作为跟踪结果。例如,Zhang 等[11]将多域网络(mutil-domain network, MDNet)拓展成双流结构分别提取可见光与红外的图像特征,再将二者拼接起来,送入多域分类网络中进行判别,确认当前帧的目标位置。Li 等[12]提出的多适配器卷积神经网络(multi-adapter convolutional network,MANet)采用多适配器的策略,将网络划分为全局适配器、模态适配器以及实例适配器共 3 个部分,分别学习双模态的共有特征、单模态的特有特征以及不同模态的实例特征。Wang 等[13]提出的跨模态模式传播(cross-modal pattern-propagation,CMPP),通过设计模态内交互模块加强可见光与红外模态之间的信息交互,同时提出了时空交互模块存储历史帧信息,引入时间信息加强网络的学习。然而,上述目标跟踪网络缺少合理的模态特征融合机制以充分利用两种模态数据的互补性。

基于孪生网络的 RGBT 跟踪算法以模板搜索图像对的思路对跟踪的目标进行匹配,选取置信度最高的地方预测跟踪框的位置。Zhang 等[14]提出的孪生融合跟踪(Siamese fusion tracking,SiamFT)网络通过使用两个并行的孪生网络分别提取可见光与红外两个模态的特征,并采用手工设计生成模态权重的方法用于多模态特征融合,最后通过互相关操作选取搜索特征响应最高的位置作为跟踪结果,SiamFT 保持了孪生网络在跟踪速度方面的优势,但是在精度上仍有待提高。Peng 等[15]提出的孪生红外与可见光融合网络(Siamese infrared and visible light fusion network,SiamIVFN)在可见光与红外单个模态融合模板分支与搜索分支,融合时按照不同的卷积层设置特征交互比例,最后通过通道注意力机制加强模板搜索分支的特征表征。然而,上述算法未考虑如何让跟踪器动态适应目标变化,因此导致算法的鲁棒性低。

Ye 等提出的单流跟踪(one-stream tracking,OSTrack)算法[5]基于视觉变换器(vision transformer,ViT)架构,使用原始的模型框架进行特征提取,引入早期候选消除模块逐步识别与丢弃属于背景的候选者,最后使用基于分类与回归的 3 个损失函数对网络进行约束完成高效精确的目标跟踪。在目标跟踪

领域，OSTrack 充分利用了模型的优势，结合孪生网络目标跟踪算法，在单模态目标跟踪领域取得了巨大的成功。基于 ViT 的单模态目标跟踪器，借助模型本身强大的泛化能力及 ViT 天然的全局-局部注意力机制，显著提高了跟踪的精度和速度。Feng 等[16]提出了 RGBT 跟踪算法 RWTransT (reliable modal weight with transformer for robust RGBT tracking)，借助强大单模态基线模型，在网络浅层进行 RGBT 双模态融合，显著提升了网络的跟踪性能，在保持高精度跟踪的情况下满足了实时跟踪的需求。这些算法在网络嵌入层忽视了模板与搜索图像在通道间关联性，同时孪生网络通常不在线更新模板图像，导致目标随时间变化时跟踪失败。因此，针对模板与搜索图像模态信息交互不足与目标随时间发生动态变化的问题，本文提出一种基于通道空间自注意力与模板在线更新的 RGBT 目标跟踪算法，主要贡献如下。

(1)针对孪生网络模板与搜索图像之间的交互性不足的问题，采用双分支嵌入层与共享权重的卷积掩码自编码器(convolutional masked autoencoder，ConvMAE)模型作为特征提取网络，以加强可见光与红外图像的特征交互，并在嵌入层设计针对模板与搜索图像的通道空间自注意力机制，以有效提取双模态的互补信息。

(2)针对跟踪过程中目标动态变化导致的跟踪漂移问题，提出了一种高效的在线模板更新模块。通过引入在线模板与模板在线分类头获取当前目标的置信度，并在跟踪过程中动态的调整在线模板，适应目标在跟踪中的变化，从而提升网络的跟踪性能。公开基准数据集 GTOT 和 RGBT234 上的定量和定性实验证明了所提算法的有效性。

1　基于注意力与模板在线更新 ConvMAE 模型的 RGBT 目标跟踪

本文提出了一种基于通道空间自注意力与模板在线更新的 RGBT 目标跟踪算法，整体网络结构如图 1 所示。该网络包括基于通道空间自注意力的特征提取网络、目标跟踪框架和模板在线更新模块。

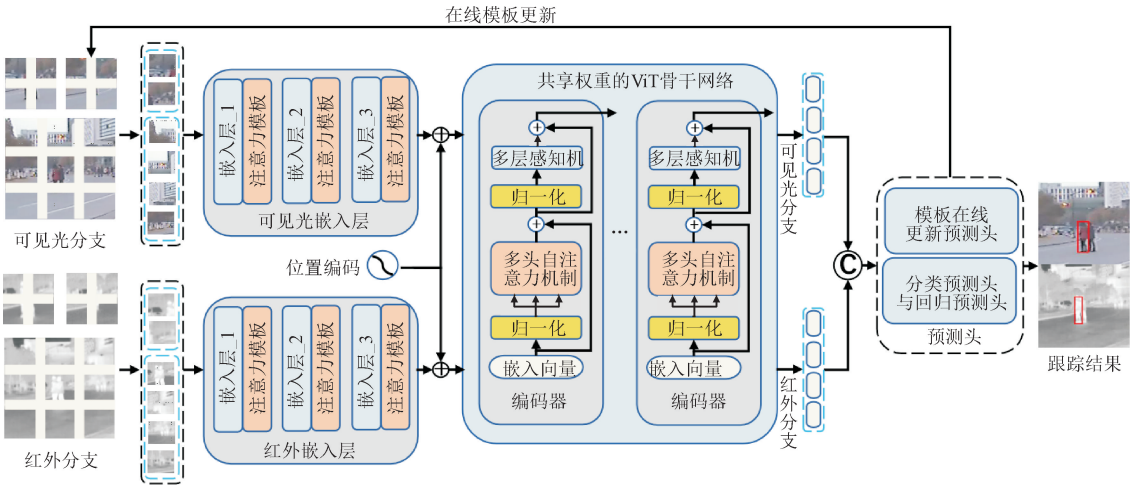


图 1　基于通道空间自注意力与模板在线更新的 RGBT 目标跟踪网络框架

Fig. 1　RGBT target tracking network framework based on channel space self-attention and online template updating

1.1　基于通道空间自注意力的特征提取网络

1.1.1　特征提取网络结构

ViT 在计算机视觉领域取得了巨大成功，但其在预训练阶段需大量标注数据，当前很多工作通过微调网络结构与改进网络的预训练策略来提升 ViT 的性能。ConvMAE 模型作为 ViT 模型的变体，结合掩码自编码器强大的无监督训练策略，显著提升了网络的泛化性与鲁棒性，其训练出来的网络性能甚至超越了有监督训练的原始 ViT 模型。

本文结合 ConvMAE 模型设计双分支嵌入层与共享权重的骨干网络，嵌入层包含可见光和红外两个分支，分别处理两种模态的图像嵌入，如图 1 所示。在每个嵌入层中，引入通道空间自注意力机制，以提高模板与搜索图像在通道上的交互性，同时保持空间上的独立性，从而获得更好的嵌入向量。此外，为模板和搜索分支的嵌入向量添加对应的二维正余弦位置编码，以增强各自的归纳偏置并加速网络收敛。最后，将模板和搜索分支的嵌入向量拼接

起来，送入骨干网络进行处理。

在骨干网络部分，采用共享权重的 ViT 编码器对可见光与红外拼接后的嵌入向量做多头自注意力，将两种模态的特征送入同一编码器，依次前向传播。这促进了双模态间的信息交互，使网络能同时学习可见光与红外图像的共有特征，充分利用双模态数据的互补性，提高某模态信息缺失时的网络跟踪性能。在网络末端，经过多层多头自注意力机制的处理，可见光与红外的搜索嵌入向量中均已包含足够多的模板信息，故将两种模态的搜索嵌入向量送入分类预测头与回归预测头做进一步的分析和预测。

在分类与回归头中，截取骨干网络输出的可见光红外搜索嵌入向量并转化为二维特征，经过一个 1×1 卷积进行线性融合得到包含双模态信息的搜索特征，再送入分类回归预测头和模板在线分类头。分类预测头与回归预测头包含置信度、偏移量、尺寸 3 个分支，每个分支由 4 层卷积层堆叠而成，在通道上进行降维，空间上大小保持不变，用于分别学习搜索嵌入向量每个像素位置上的置信度、偏移量以及目标框的大小信息。

1.1.2 通道空间自注意力机制特征增强模块

本文使用通道空间注意力机制（convolutional block attention module，CBAM）[17]，结合模板和搜索图像在执行嵌入操作时采用共享权重嵌入层的关联，设计一种基于模板搜索图像的通道空间自注意力机制，保持各模态数据空间独立性。使用共享权重的嵌入层可以降低网络参数数量，加快目标跟踪速度，每一层嵌入层之后引入通道空间自注意力机制，增强模板图像与搜索图像间的信息交互。通道空间自注意力机制网络结构如图 2 所示。

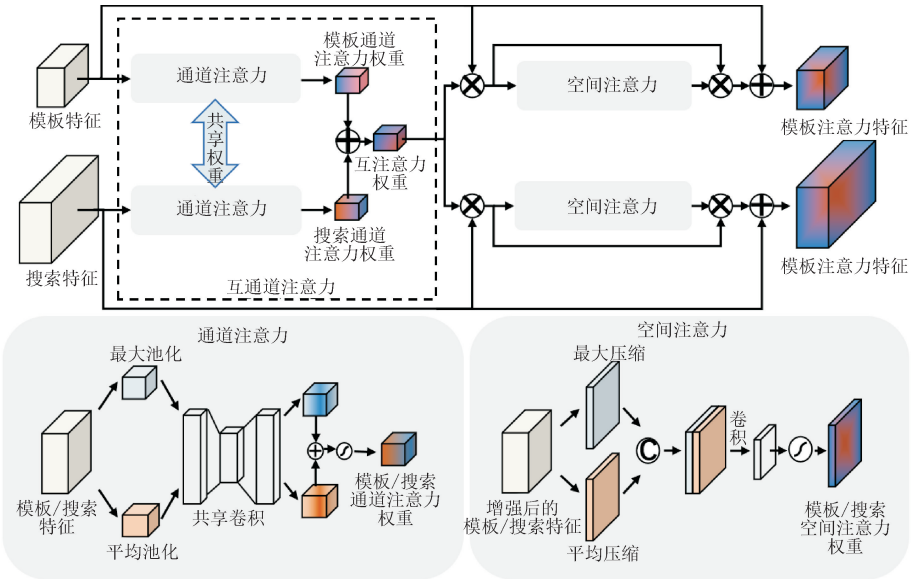


图 2 通道空间自注意力机制网络结构

Fig. 2 Network architecture of the channel space self-attention mechanism

处理嵌入层输出的模板和搜索特征时，两者在通道维度上的分布显示出相似性。将这两种特征输入到具有共享权重的互通道注意力模块，以获得模板和搜索特征的通道注意力信息，来探索它们的通道关联性。利用这两个信息的均值生成互通道注意力掩码，增强通道注意力的稳定性。最后，将掩码与原始模板和搜索特征相乘，获得融合搜索分支与模板分支的互通道注意力结果。具体的模板搜索通道注意力计算方法为

$$w_{CC} = (C(F_s) + C(F_t))/2 \tag{1}$$

$$F_s' = w_{CC} F_s \tag{2}$$

$$F_t' = w_{CC} F_t \tag{3}$$

式中：C 表示共享权重的通道注意力网络；w_{CC} 表示搜索模板的通道互注意力权重；F_s 与 F_t 分别表示初始的搜索图像与模板图像特征；F_s' 与 F_t' 分别表示添加通道注意力后搜索图像与模板图像特征。由于模板图像与搜索图像大小不同，因此再使用对应卷积核大小的空间注意力加强各自的空间信息，最后与原始模板和搜索特征进行残差连接。空间注意力的计算公式为

$$F_s'' = S(F_s')F_s' + F_s \tag{4}$$

$$F_t'' = S(F_t')F_t' + F_t \tag{5}$$

式中：S 表示空间注意力网络；F_s'' 与 F_t'' 分别表示添加空间注意力的搜索图像特征与模板图像特征。

1.2　模板在线更新模块

1.2.1　模板在线更新模块

视频帧的数量从几十到上千不等，目标外观也会随时间改变。常规跟踪算法使用第一帧的目标作为模板跟踪整个视频，难以适应目标外观变化。若对中间的跟踪结果不加判断就替换模板，会导致跟踪漂移[18-19]。为此，本文提出一种监督的模板在线更新机制，用在线分类头判断目标置信度，结合初始模板与在线更新模板的双重策略，应对常规模板更新策略的漂移问题。在线分类头能在跟踪过程中评估目标的可靠性并迭代优化，而双重模板策略则平衡了目标的初始特征与随时间变化的特征。模板在线更新模块的网络结构如图 3 所示。

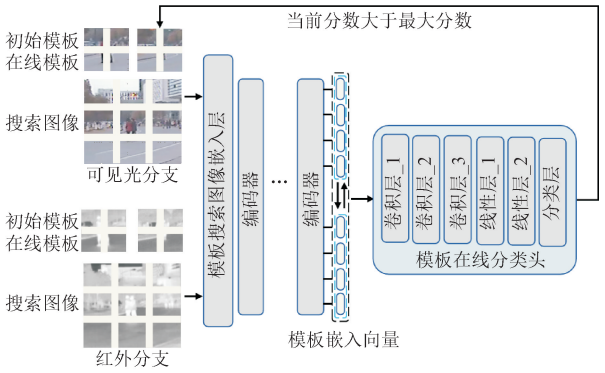


图 3　模板在线更新结构

Fig. 3　Structure of the online template updating

本文采用包括初始模板和在线模板的双重模板策略。输入阶段同时处理两个模板图像及一个搜索图像，经过 ViT 编码器处理，其中初始模板保留目标原始特征，而在线模板依据在线分类头评分动态更新。骨干网络提取两个模板的嵌入向量，并通过裁剪、拼接及 1×1 卷积融合后重构为二维特征，供在线分类头分析。在线分类头由 3 层卷积层和 3 层全连接层构成，逐步聚合语义信息并基于全局信息获取表示目标置信度的评分。若目标评分超过在线模板评分，则按固定间隔（本文设为 50 帧）更新在线模板，及时反映目标的变化，同时确保新模板经过多帧验证，提升跟踪的准确性和鲁棒性。

1.2.2　训练策略与在线跟踪

本文采用双阶段训练策略。在第一阶段，通过自适应权重的 Focal 分类损失[20]、gIoU 损失[21]和 L_1 回归损失[22]约束网络，使网络专注生成高置信度且坐标精确的目标框。在第二阶段，固定嵌入层、骨干网络和分类/回归预测头等参数，只训练模板在线分类模块，确保网络生成高质量目标框的同时，训练良好的模板在线分类头。在处理网络输出时，先使用 Sigmoid 函数归一化，再使用 BCE 损失[23]约束模板在线分类模块，最终输出表示当前目标置信度评分的一维向量。BCE 损失函数的公式为

$$L_{\mathrm{BCE}} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \ln(p_i) + (1-y_i)\ln(1-p_i)\right] \tag{6}$$

式中：y_i 表示样本标签；p_i 表示预测概率；i 表示当前帧数。在实际跟踪过程中，将训练好的模型载入网络，设置初始模板与在线模板为第一帧目标，初始分数 $s=1$，分数衰减公式为

$$s = s(1-d) \tag{7}$$

式中：衰减系数 $d=0.01$。每跟踪一帧视频，网络都会计算当前目标的分数，并与衰减后的在线模板分数进行对比。如果当前目标的分数更高，就记录当前目标及其分数，并在固定间隔更新在线模板，本文设置间隔为 50 帧。

2　实验结果与分析

2.1　实验环境与设置

2.1.1　实验环境

本文采用 Python 3.6 版本，选用 PyTorch 作为深度学习框架。所有实验均在 Ubuntu 18.04 操作系统上进行，CPU 为 AMD Ryzen Threadripper 2970WX，GPU 型号为 RTX3090。ConvMAE 采用在 ImageNet1K 数据集[24]上的预训练模型。模型训练算法伪代码如下。

输入：

```
    T_vis      //初始可见光模板图像
    T_ir       //初始红外模板图像
    Frames     //可见光和红外图像序列
    th_conf    //置信度更新阈值
    params     //模型及更新过程中的其他参数
```

输出：

```
    Tracking_Result   //每帧的目标定位结果
//初始化
T_template ← Fuse(T_vis, T_ir)   //初始模板融合
T_online ← T_template   //初始化在线模板
FeatureExtractor ← ConvMAE(weight shared)
Initialize ClassificationHead   //初始化在线模板分类头
//主跟踪循环
For each frame t=1 to T do:
    I_vis ← Frames[t].visible
```

I_ir ← Frames[t].infrared

S←ExtractSearchRegion(I_vis,I_ir,previous_target)

F_template ← FeatureExtractor(T_online)

F_search ← FeatureExtractor(S)

F_fused ← ChannelSpatialSelfAttention(F_template, F_search)

(target_box, conf_score) ← MatchingAndLocalization(F_fused)

//在线模板更新

If conf_score ＞ th_confthen:

T_online←OnlineTemplateUpdate(T_online, F_search, target_box)

End If

Tracking_Result[t] ← target_box

End For

Return Tracking_Result

2.1.2 实验设置

本文模型基于 PyTorch 框架，总训练批量为 16 个图像对。在 RGBT234 数据集训练，测试模型在 GTOT 数据集的性能指标；在 GTOT 数据集训练，测试模型在 RGBT234 数据集的性能指标。训练分为两个阶段：第 1 阶段训练骨干网络与回归头，共 15 轮；第 2 阶段训练模板在线更新模块，共 15 轮。每轮随机抽取 60 000 个图像对。ConvMAE 骨干网络与通道注意力机制模块部分的学习率设置为 4×10^{-5}，模型其他部分的学习率设置为 4×10^{-4}，在 10 轮后学习率衰减到 1/10 至训练结束；采用 AdamW 作为优化器，并设置 1×10^{-4} 的权重衰减。搜索区域设置为 256×256 像素，模板设置为 128×128 像素。

2.2 数据集与评价指标

本文在 GTOT 数据集[25]和 RGBT234 数据集[26]上进行对比实验。由于 RGBT234 数据集较大，所以消融实验在 RGBT234 上进行。GTOT 数据集[25]包含 50 个真实环境下的 RGBT 视频对，每帧都标有可见光和红外标签。视频长度从 40～376 帧不等，平均每个视频 157 帧，共 7.8×10^3 对图像，分辨率为 384×384 像素，按挑战性特征分为 7 个子集。RGBT234 数据集[26]包含 234 个真实场景的 RGBT 视频序列，同样每对图像都有精确标签。视频长度 40～4 140 帧，平均每个视频 498 帧，共约 116.7×10^3 对图像，分辨率为 630×460 像素，按挑战性属性分为 12 个子集。

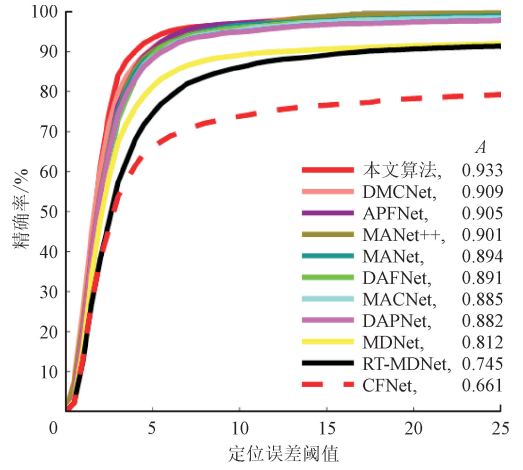
评估指标采用精确率(precision rate，PR)和成功率(success rate，SR)。PR：预测目标框与标准目标框中心点欧氏距离小于阈值的帧数比总帧数。GTOT 数据集因目标尺寸小，PR 阈值设为 5 像素；RGBT234 数据集 PR 阈值为 20 像素，作为跟踪器对比标准。SR：预测目标框与标准目标框交集面积与并集面积比值大于某阈值的帧数比总帧数。GTOT 和 RGBT234 数据集的 SR 阈值均为 0.5，作为算法对比标准。

2.3 实验部分

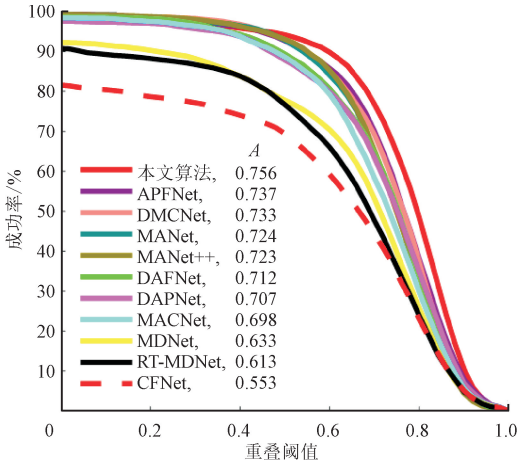
2.3.1 对比实验

为全面评估本文算法的性能，选取了 14 种跟踪算法做对比实验，包括 CFNet[27]、RT-MDNet[28]、MDNet[29]、DAPNet[30]、MANet[12]、MACNet[31]、DAFNet[32]、MANet＋＋[33]、APFNet[34]、DMCNet[35]、SGT[36]、KCF[37]、SDSTrack[38]和 STMT[39]。对比算法实验结果的获取方式如下：对于 DAPNet、MACNet、DAFNet、MANet ＋ ＋、APFNet、DMCNet、SGT、KCF、SDSTrack 和 STMT 10 种公开结果的算法，本文采用其官方发布的测试结果作为对比；针对 MDNet、RT-MDNet 和 MANet 提供模型的算法，在与本文一致的实验环境下，严格保持原论文参数配置，进行模型推理，使用模型推理结果作为对比；对于 CFNet 仅开源代码框架的算法，在与本文一致的实验环境下，依据原论文的训练设置完成模型训练与测试，得到测试结果。

在 GTOT 数据集上，本文算法与 10 种算法进行比较，CFNet[27]、RT-MDNet[28]、MDNet[29]、DAPNet[30]、MANet[12]、MACNet[31]、DAFNet[32]、MANet＋＋[33]、APFNet[34]以及 DMCNet[35]，其中 CFNet、RT-MDNet 以及 MDNet 是 RGB 目标跟踪算法，其他均为 RGBT 目标跟踪算法。图 4 展示了本文算法与其他算法在 GTOT 数据集上的精确率和成功率，其中 A 为平均精确率与成功率之比。可以看出，本文算法在 PR 与 SR 上达到了 93.3% 和 75.6%，均取得了最好的跟踪性能，具体而言，本文算法比 DMCNet 与 APFNet 在精确率/成功率上分别取得了 2.4%/2.3%与 2.8%/1.9%的提升。与其他先进的跟踪器相比，本文算法能保持最优的跟踪性能。实验结果说明，本文提出的基于通道空间自注意力与模板在线更新的 RGBT 目标跟踪算法能有效地实现对搜索图像和模板之间的深度交互。同时，算法的骨干网络部分能够高效地学习可见光与红外两种模态的信息，而模板在线更新模块则能够有效应对目标在跟踪过程中可能出现的变化，提升跟踪的鲁棒性。



(a)精确率



(b)成功率

图 4　本文算法与 10 种对比算法在 GTOT 数据集上的准确率和成功率

Fig. 4　Accuracy rate and success rate of the proposed algorithm and ten comparative algorithms on the GTOT dataset

GTOT 数据集中包含 7 个挑战属性，用于评估不同场景下算法的跟踪性能，分别为：遮挡、大尺度变化、快速运动、低照度、热交叉、小目标和形变。本文统计了算法在这些挑战下的 PR 和 SR，并与 10 种对比算法进行比较，结果如表 1 和 2 所示。可以看出：本文算法除遮挡和小目标外的挑战条件下，均取得最优精确率；除小目标和形变外的挑战条件下，均取得最优成功率。这些结果充分证明了本文算法在多挑战条件下的优越性和有效性。

在 RGBT234 数据集上，本文算法与 12 种先进的跟踪算法进行比较，包括 SDSTrack[38]、STMT[39]、DMCNet[35]、APFNet[34]、MANet＋＋[33]、MANet[12]、DAFNet[32]、DAPNet[30]、MDNet[29]＋RGBT、SGT[36]、RT-MDNet[28] 以及 MDNet[29]。其中，MDNet、RT-MDNet 为 RGB 跟踪算法，其余均为 RGBT 跟踪算法，MDNet＋RGBT 算法是将可见光与红外在图像输入网络前沿通道拼接再送入 MDNet 算法的跟踪结果。图 5 展现了本文算法在 RGBT234 数据集上的精确率和成功率。本文算法以 87.2%的精确率与 63.8%的成功率达到最优性能。具体而言：在精确率指标上，本文算法较 STMT 算法(85.4%)提升 1.8 个百分点；在成功率指标上，本文算法与 STMT 算法(63.8%)并列，较 SDSTrack 算法(62.5%)提升 1.3 个百分点。在应对运动模糊和摄像头移动等挑战时，本文算法显示出优异的适应性，并在处理大量复杂场景的数据集上取得了最优性能。

表 1　11 种算法在 GTOT 数据集上 7 种挑战条件下的 PR 对比

Table 1　PR comparison of eleven algorithms under seven challenging conditions on the GTOT dataset　%

算法	遮挡	大尺度变换	快速运动	低照度	热交叉	小目标	形变	全部
CFNet	67.2	69.9	66.1	65.7	70.2	79.2	62.4	66.1
RT-MDNet	73.3	79.1	78.1	77.2	73.7	85.6	73.1	74.5
MDNet	77.2	81.7	78.2	82.8	79.9	87.9	83.1	81.2
DAPNet	87.3	84.7	82.3	90.0	89.3	93.7	91.9	88.2
MANet	85.2	84.7	85.8	89.9	86.9	91.4	93.0	88.3
MACNet	86.7	84.2	84.4	90.1	89.3	93.2	92.2	88.5
DAFNet	87.3	82.2	80.9	89.9	89.8	93.9	94.7	89.1
MANet＋＋	88.6	86.2	86.0	91.5	89.5	93.6	93.3	89.7
APFNet	**90.3**	87.7	86.5	91.4	90.4	94.3	94.6	90.5
DMCNet	90.1	89.6	86.9	92.8	91.7	**95.5**	94.0	90.9
本文	90.2	**92.5**	**88.0**	**96.1**	**92.2**	92.6	**95.8**	**93.3**

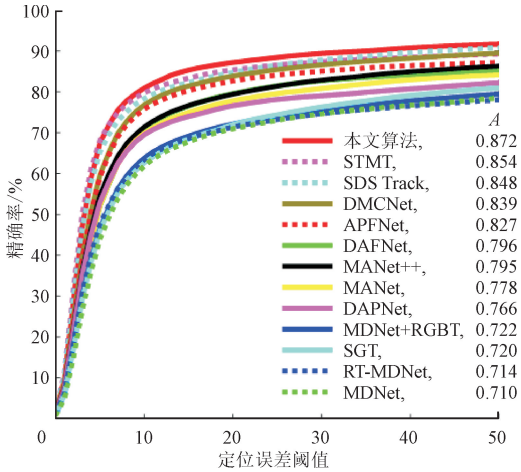
注：表中加粗数据为最优值。

表 2 11 种算法在 GTOT 数据集上 7 种挑战条件下的 SR 对比

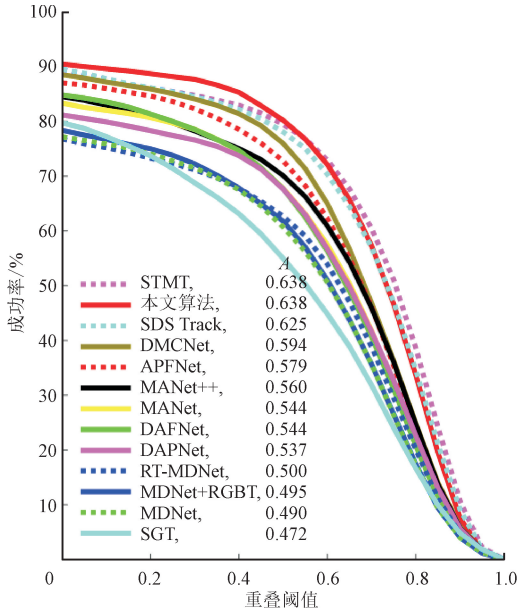
Table 2 SR comparison of eleven algorithms under seven challenging conditions on the GTOT dataset %

算法	遮挡	大尺度变换	快速运动	低照度	热交叉	小目标	形变	全部
CFNet	55.9	60.5	60.5	54.4	58.2	60.5	50.3	55.3
RT-MDNet	57.6	63.7	64.1	63.8	59.0	63.4	61.0	61.3
MDNet	58.3	59.4	56.0	64.7	59.7	61.9	68.9	63.3
DAPNet	67.4	66.4	61.9	72.2	69.0	69.2	77.1	70.7
MANet	68.2	69.1	69.0	72.9	70.1	68.4	75.5	72.1
MACNet	67.6	66.2	63.3	70.7	68.6	67.8	74.4	69.8
DAFNet	68.4	66.4	64.2	72.7	70.3	69.8	76.5	71.2
MANet++	70.3	69.5	70.5	73.4	71.0	70.2	74.7	72.5
APFNet	71.3	71.2	68.4	74.8	71.6	**71.3**	**78.0**	73.7
DMCNet	70.5	71.5	68.6	74.6	72.0	70.4	76.5	73.3
本文	**73.3**	**75.2**	**74.3**	**77.0**	**74.6**	71.0	76.0	**75.6**

注：表中加粗数据为最优值。



(a)精确率



(b)成功率

图 5 本文算法与对比算法在 RGBT234 数据集上的准确率和成功率

Fig. 5 Accuracy rate and success rate of the proposed algorithm and comparative algorithms on the RGBT234 dataset

2.3.2 消融实验

为了全面评估本文算法的有效性，在 RGBT234 数据集上进行了消融实验，实验在 OSTrack 基线模型上逐一增加通道空间自注意力模块和模板在线更新模块，以系统地验证每个模块的有效性。消融实验结果如表 3 所示。

表 3 消融实验结果

Table 3 Ablation experiment results

基线模型	通道空间自注意力模块	模板在线更新模块	精确率/%	成功率/%
√			83.2	60.5
√	√		84.1	61.4
√	√	√	**87.2**	**63.8**

注：表中加粗数据为最优值。

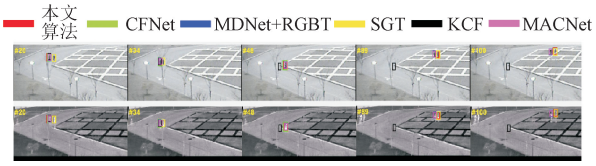
本文算法以 ConvMAE 模型为基础，采用双嵌入层配合共享权重的骨干网络结构，深入挖掘可见光与红外数据间的联系。在基线实验中实现较高的跟踪性能，精确率和成功率分别达到了 83.2% 和 60.5%。添加通道空间自注意力模块后，模板图像与搜索图像间的交互得到了加强，使得精确率和成功率相较于基线实验各自提高了 0.9%。添加模板在线更新模块后，算法能够在维持模板初始特征的同时动态适应目标的变化，实现更为稳定且精确的跟踪，精确率和成功率再度提升了 3.1% 和 2.4%，证明其有效缓解了模型漂移问题。提出的通道空间自注意力模块通过在网络的嵌入层进行注意力交互，能够获取两个模态各自的异质特征，同时还能加强模板与搜索图像的交互，增强了网络对可见光与红外的特征提取能力。

通道空间自注意力机制有效增强了特征交互能力。双分支结构允许模型独立学习可见光与红外的模态特异性特征，避免特征混淆。共享权重设计使两个模态在高层语义特征上对齐，增强模态互补性。通道空间自注意力机制通过通道注意力和空间注意力动态调整权重。通道维度上，在热交叉场景中，模型能抑制可见光中的干扰，增强红外中目标热特征通道的响应；空间维度上，目标快速运动时，通过空间注意力聚焦目标运动轨迹，忽略无关背景。

同时，提出的模板在线更新模块通过对当前目标的置信度判别，动态地选出稳定的在线模板，并结合共享权重的骨干网络有效的融合可见光与红外的信息，提升跟踪算法的精度与鲁棒性。置信度评分机制通过在线分类头输出目标置信度评分，反映当前跟踪结果的可靠性。当目标被遮挡时，置信度评分下降，模型倾向于保留历史模板；当目标外观显著变化且置信度评分高于阈值时，触发模板更新。双重模板策略中，初始模板保留目标的原始特征，在线模板捕捉动态变化。二者通过加权融合平衡鲁棒性与适应性。在长时跟踪中，初始模板可纠正短期跟踪误差，而在线模板逐步适应目标的长期形变。

2.3.3　跟踪结果定性实验

在 RGBT 数据集中，本文选用热交叉、快速运动和大尺度变化场景下的视频序列，选取 CFNet[27]、MDNet[29] + RGBT、SGT[36]、KCF[37]、MACNet[31]共 5 种跟踪算法进行可视化实验分析，如图 6 所示。



(a)Running

(b)FastMotorNig

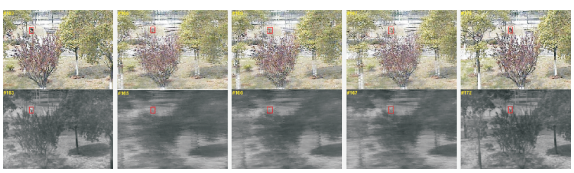
(c)Minibus

图 6　本文算法与 5 种对比算法的可视化实验结果对比

Fig. 6　Comparison of visualization experimental results between the proposed algorithm and five comparative algorithms

图 6(a)中，目标周围有相似背景和重叠现象，这在可见光条件下构成干扰，在红外条件下形成热交叉问题。本文算法在该复杂条件下仍保持稳定的性能，而其他算法在目标重叠后出现跟踪漂移，展现了本文算法能加强可见光与红外模态的学习，提取更鲁棒的特征。图 6(b)中，目标快速运动并遭遇强光遮挡，此时初始模板发挥作用，纠正跟踪框位置，保证跟踪稳定性，实现精确跟踪，而其他算法在强光附近出现了抖动或漂移，这展现了本文算法通过共享权重的骨干网络和自注意力机制学习到可见光和红外模态间的互补性，可见光模态不可靠时用红外模态的目标信息进行跟踪，显示出其在快速运动和强光遮挡中的鲁棒性。图 6(c)中，目标经历尺度的显著变化，本文算法能自适应调整跟踪框大小以适应目标尺度变化，而其他算法虽能跟上目标，但跟踪框不能很好地贴合。这展现了本文算法通过置信度评分机制和模板在线更新模块，能在跟踪过程中选择适合的在线模板，与初始模板结合，动态适应目标尺寸变化，保持跟踪框位置和尺寸与目标的贴合度，展现了本文算法在大尺度变化挑战中的鲁棒性。综上分析，通道空间自注意力机制加强了模板搜索图像间的交互，共享权重的 ViT 骨干网络加强了可见光与红外间的信息交互，模板在线更新模块帮助网络有效适应目标变化，通过这些，本文算法成功实现了在热交叉、快速运动和尺度变化等多重挑战场景下的精确跟踪。

在运动模糊与相机运动挑战场景下，其他算法表现不佳，本文算法选取 RGBT234 数据集中的两个运动模糊与相机移动挑战场景视频进行跟踪结果的可视化分析，结果如图 7 所示。可以看出，本文算法成功克服了因运动模糊和相机运动导致的跟踪失败问题。具体而言，本文算法通过通道空间自注意力机制，能够更有效地捕捉目标的关键特征，同时采用模板在线更新模块对在线模板进行动态更新，以适应目标在跟踪过程中的变化，可视化结果进一步证实了该算法在处理具有运动模糊和相机移动等挑战性条件下的目标跟踪任务时的优越性和鲁棒性。



(a)oldwoman

(b)cycle4

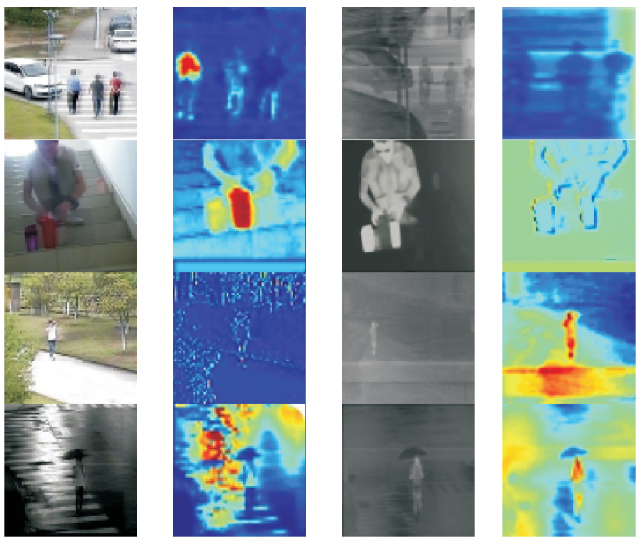
图 7 本文算法运动模糊与相机运动跟踪结果

Fig. 7 Tracking results under motion blur and camera motion of proposed algorithm

从算法在 GTOT 和 RGBT234 数据集上的表现看,本文提出的算法利用了 ConvMAE 模型、动态模板更新以及孪生网络匹配技术,不仅保持在简单场景中的跟踪性能,还能显著提升对复杂情况如运动模糊和相机运动的跟踪能力,有效提升算法的鲁棒性。

2.3.4 热力图可视化分析

为直观展示本文算法在 RGBT 目标跟踪的模态互补性,本文对若干关键帧进行可视化处理,结果如图 8 所示。可以看出,当模型在红外模态下未能有效识别目标特征时,可见光模态下仍能准确突出目标区域,提供精准的特征信息,实现对目标的准确跟踪。反之,当模型在可见光模态下的特征识别效果不佳时,红外模态下的热力图显示出清晰的目标信息,确保跟踪的稳定性和准确性。这证明了两种模态之间的互补性:在一模态特征提取受限的情况下,另一模态能提供有效信息,使模型具备更强的鲁棒性和适应性。热力图中目标区域的高亮显示也验证了自注意力机制在聚焦目标关键特征、抑制背景干扰方面的有效性,为模型决策提供了直观的解释依据。



可见光图像 | 可见光图像对应的热力图 | 红外图像 | 红外图像对应的热力图

图 8 本文算法热力图特征可视化结果

Fig. 8 Heat map feature visualization results of proposed algorithm

2.3.5 计算量分析

本文算法与其他 2 种算法的计算量对比结果如表 4 所示。可以看出,本文算法基于 ViT 架构的基线模型 OSTrack[5],在保持相近计算量和参数量的同时,将精确率从 82.3%提升至 87.2%,成功率从 60.5%提升至 63.8%。对比同样采用注意力机制的 SDSTrack 算法[38],本文算法尽管计算量较 SDSTrack 增加较大,但实现了更显著的性能增益,取得了更优的跟踪性能,这得益于本文算法提出的通道空间自注意力模块和模板在线更新模块。

表 4 本文算法与其他算法计算量对比结果

Table 4 Comparison results of calculation amount of proposed algorithm with other algorithms

算法	精确率,成功率	浮点运算次数/$10^9\ s^{-1}$	参数量/10^6
OStrack	0.832,0.605	96.72	92.12
SDSTrack	0.848,0.625	54.39	102.18
本文算法	0.872,0.638	94.57	114.27

2.4 局限性分析

本文提出的通道空间自注意力机制以及在线模板更新模块,在应对复杂场景时展现出了卓越的性能。然而,经过深入研究与分析发现,该算法仍具有局限性——多尺度特征协同效能瓶颈。预测头的单尺度回归策略制约了多层级特征融合。在处理剧烈尺度变化目标时,高层语义与底层细节特征未能形成有效协同,影响跟踪框的尺寸预测精度。对此,可构建轻量化特征金字塔网络,增强多尺度特征的融合能力,使模型在处理不同尺度目标时,能够更精准地预测跟踪框尺寸,提升跟踪性能。

3 结 论

(1)本文提出了一种基于通道空间自注意力与模板在线更新的 RGBT 目标跟踪算法。该算法以 OSTrack 作为基础跟踪框架,并选用 ConvMAE 作为主要的特征提取网络。

(2)本文提出的通道空间自注意力机制在网络的嵌入层加强了模板图像与搜索图像之间的交互,提升了对可见光和红外双模态特征差异性的捕捉能力,确保了对两种模态独有特征的有效学习。

(3)利用模板在线更新模块获取当前目标置信度评分来动态选择最具鲁棒性的模板,以适应目标变化,实现更高的跟踪精准度。两个模块的合作显著提高了算法的整体性能。此外,本文在 RGBT 基准数据集上开展了一系列实验,在 GTOT 数据集

上，本文算法在精确率和成功率上较次优算法取得了 2.4%和 1.9%的提升，充分显示了本文算法相比现有技术在目标跟踪方面的优越性和鲁棒性。

参考文献：

[1] 闵志方，杜虎，朱雪琼，等. 单目标跟踪研究综述 [J]. 光学与光电技术，2023，21(4)：1-14.
MIN Zhifang，DU Hu，ZHU Xueqiong，et al. Survey of single target tracking research [J]. Optics & Optoelectronic Technology，2023，21(4)：1-14.

[2] 马玉民，钱育蓉，周伟航，等. 基于孪生网络的目标跟踪算法综述 [J]. 计算机工程与科学，2023，45(9)：1578-1592.
MA Yumin，QIAN Yurong，ZHOU Weihang，et al. A survey of target tracking algorithms based on Siamese network [J]. Computer Engineering & Science，2023，45(9)：1578-1592.

[3] 韩瑞泽，冯伟，郭青，等. 视频单目标跟踪研究进展综述 [J]. 计算机学报，2022，45(9)：1877-1907.
HAN Ruize，FENG Wei，GUO Qing，et al. Single object tracking research：a survey [J]. Chinese Journal of Computers，2022，45(9)：1877-1907.

[4] ZHANG Xingchen，YE Ping，LEUNG H，et al. Object fusion tracking based on visible and infrared images：a comprehensive review [J]. Information Fusion，2020，63：166-187.

[5] YE Botao，CHANG Hong，MA Bingpeng，et al. Joint feature learning and relation modeling for tracking：a one-stream framework [C]//Computer Vision-ECCV 2022. Cham，Switzerland：Springer Nature Switzerland，2022：341-357.

[6] 朱杰，翟素兰，刘磊. 基于自适应上下文感知相关滤波的 RGB-T 目标跟踪 [J]. 阜阳师范大学学报(自然科学版)，2022，39(3)：65-72.
ZHU Jie，ZHAI Sulan，LIU Lei. RGB-T object tracking based on adaptive context-aware correlation filter [J]. Journal of Fuyang Normal University (Natural Science)，2022，39(3)：65-72.

[7] 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 Tianlu，ZHANG Qiang. A survey of RGB-T object tracking technologies based on deep learning [J]. Pattern Recognition and Artificial Intelligence，2023，36(4)：327-353.

[8] 李成龙，鹿安东，刘磊，等. 多模态视觉跟踪方法综述 [J]. 中国图象图形学报，2023，28(1)：37-56.
LI Chenglong，LU Andong，LIU Lei，et al. Multimodal visual tracking：a survey [J]. Journal of Image and Graphics，2023，28(1)：37-56.

[9] 张兆宇，田春娜，周恒，等. 联合在线分类的双注意力 RGBT 孪生网络跟踪 [J]. 西安电子科技大学学报，2022，49(6)：76-85.
ZHANG Zhaoyu，TIAN Chunna，ZHOU Heng，et al. Online classification jointed RGBT tracking based on the dual attention Siamese network [J]. Journal of Xidian University，2022，49(6)：76-85.

[10] ZHANG Pengyu，WANG Dong，LU Huchuan. Multimodal visual tracking：review and experimental comparison [J]. Computational Visual Media，2024，10(2)：193-214.

[11] ZHANG Xingming，ZHANG Xuehan，DU Xuedan，et al. Learning multi-domain convolutional network for RGB-T visual tracking [C]//2018 11th International Congress on Image and Signal Processing，BioMedical Engineering and Informatics (CISP-BMEI). Piscataway，NJ，USA：IEEE，2018：1-6.

[12] LI Chenglong，LU Andong，ZHENG Aihua，et al. Multi-adapter RGBT tracking [C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway，NJ，USA：IEEE，2019：2262-2270.

[13] WANG Chaoqun，XU Chunyan，CUI Zhen，et al. Cross-modal pattern-propagation for RGB-T tracking [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway，NJ，USA：IEEE，2020：7062-7071.

[14] ZHANG Xingchen，YE Ping，PENG Shengyun，et al. SiamFT：an RGB-infrared fusion tracking method via fully convolutional Siamese networks [J]. IEEE Access，2019，7：122122-122133.

[15] PENG Jingchao，ZHAO Haitao，HU Zhengwei，et al. Siamese infrared and visible light fusion network for RGB-T tracking [J]. International Journal of Machine Learning and Cybernetics，2023，14(9)：3281-3293.

[16] FENG Mingzheng，SU Jianbo. Learning reliable modal weight with transformer for robust RGBT tracking [J]. Knowledge-Based Systems，2022，249：108945.

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

[18] 陈建明，李定鳘，曾祥津，等. 一种跨模态光学信息交互和模板动态更新的 RGBT 目标跟踪方法 [J]. 光学学报，2024，44(7)：101-115.
CHEN Jianming，LI Dingjian，ZENG Xiangjin，et al. Cross-modal optical information interaction and template dynamic update for RGBT target tracking method [J]. Acta Optica Sinica，2024，44(7)：101-115.

[19] HU Weiming, WANG Qiang, ZHANG Li, et al. SiamMask: a framework for fast online object tracking and segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3072-3089.

[20] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 2999-3007.

[21] LI Xiang, WANG Wenhai, WU Lijun, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2020: 21002-21012.

[22] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: a metric and a loss for bounding box regression [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 658-666.

[23] JADON S. A survey of loss functions for semantic segmentation [C]//2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). Piscataway, NJ, USA: IEEE, 2020: 1-7.

[24] DENG Jia, DONG Wei, SOCHER R, et al. ImageNet: a large-scale hierarchical image database [C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2009: 248-255.

[25] LI Chenglong, HUI Cheng, HU Shiyi, et al. Learning collaborative sparse representation for grayscale-thermal tracking [J]. IEEE Transactions on Image Processing, 2016, 25(12): 5743-5756.

[26] LI Chenglong, LIANG Xinyan, LU Yijuan, et al. RGB-T object tracking: benchmark and baseline [J]. Pattern Recognition, 2019, 96: 106977.

[27] VALMADRE J, BERTINETTO L, HENRIQUES J, et al. End-to-end representation learning for correlation filter based tracking [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 5000-5008.

[28] JUNG I, SON J, BAEK M, et al. Real-time MDNet [C]//Computer Vision-ECCV 2018. Cham, Switzerland: Springer International Publishing, 2018: 89-104.

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

[30] ZHU Yabin, LI Chenglong, LUO Bin, et al. Dense feature aggregation and pruning for RGBT tracking [C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2019: 465-472.

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

[32] GAO Yuan, LI Chenglong, ZHU Yabin, et al. Deep adaptive fusion network for high performance RGBT tracking [C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway, NJ, USA: IEEE, 2019: 91-99.

[33] LU Andong, LI Chenglong, YAN Yuqing, et al. RGBT tracking via multi-adapter network with hierarchical divergence loss [J]. IEEE Transactions on Image Processing, 2021, 30: 5613-5625.

[34] XIAO Yun, YANG Mengmeng, LI Chenglong, et al. Attribute-based progressive fusion network for rgbt tracking [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2022: 2831-2838.

[35] LU Andong, QIAN Cun, LI Chenglong, et al. Duality-gated mutual condition network for RGBT tracking [J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(3): 4118-4131.

[36] LI Chenglong, ZHAO Nan, LU Nan, et al. Weighted sparse representation regularized graph learning for RGB-T object tracking [C]//Proceedings of the 25th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2017: 1856-1864.

[37] TANG Ming, FENG Jiayi. Multi-kernel correlation filter for visual tracking [C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2015: 3038-3046.

[38] HOU Xiaojun, XING Jiazheng, QIAN Yijie, 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). Piscataway, NJ, USA: IEEE, 2024: 26541-26551.

[39] SUN Dengdi, PAN Yajie, LU Andong, et al. Transformer RGBT tracking with spatio-temporal multimodal tokens [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 12059-12072.

(编辑 陶晴)