非对称的分层特征融合的 RGBT 跟踪网络

吴习惠[1,2)], 李婷[1,2)], 葛洪伟[1,2)*]

[1)] (江南大学人工智能与计算机学院 无锡 214122)
[2)] (江苏省模式识别与计算机智能工程实验室(江南大学) 无锡 214122)
(ghw@jiangnan.edu.cn)

摘 要: 为了解决可见光图像和热红外图像由于成像原理不同而导致的模态存在异质性的问题, 提出一种非对称的分层特征融合的 RGBT 跟踪网络. 首先通过双流网络分别提取可见光和热红外的特征; 然后通过模态特征提取模块挖掘不同模态特征, 并对获得的特征进行自适应聚合, 以获得有利于增强可见光模态的特征; 最后将各层获得的聚合特征与双流网络获得的可见光特征进行增强融合, 获得更具有鲁棒性的特征. 在 GTOT, RGBT234 和 LasHeR 数据集上的实验结果表明, 所提网络的跟踪精度(PR)和成功率(SR)分别达到 92.2%/77.2%, 82.9%/61.1%和 52.7%/40.3%, 与目前主流的 RGBT 目标跟踪网络相比, PR 和 SR 均有所提高, 验证了该网络的有效性.

关键词: 非对称结构; 分层特征融合; RGBT 目标跟踪; Transformer
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.20162

Asymmetric Hierarchical Feature Fusion Network for RGBT Tracking

Wu Xihui[1,2)], Li Ting[1,2)], and Ge Hongwei[1,2)*]

[1)] (*School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi* 214122)
[2)] (*Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi* 214122)

Abstract: In order to solve the problem of modal heterogeneity between visible light images and thermal infrared images due to different imaging principles, an asymmetric hierarchical feature fusion RGBT tracking network is proposed. Firstly, a two-stream network is used to extract visual light and thermal infrared features; then through the modal feature extraction module to mine different modal features and adaptively aggregate the obtained features to obtain features that are conducive to enhancing the visible light mode; finally, the aggregated features obtained by each layer and the visible light features obtained by the two-stream network perform enhanced fusion to obtain more robust features. Experimental results on GTOT, RGBT234 and LasHeR datasets show that the tracking precision rate (PR) and success rate (SR) of the network reach 92.2%/77.2%, 82.9%/61.1% and 52.7%/40.3%, compared with the current mainstream RGBT target tracking network, both PR and SR have been improved, which verifies the effectiveness of the network.

Key words: asymmetric structure; hierarchical feature fusion; RGBT target tracking; Transformer

视觉目标跟踪是计算机视觉的重要任务之一, 在自动驾驶、视频监控等领域都有着重要的应用. 目前, 单目标跟踪任务对于存在极端光照、烟雾等恶劣场景的跟踪效果较差, 这是由于单目标跟踪

收稿日期: 2023-02-21; 修回日期: 2023-05-11. **基金项目**: 国家自然科学基金(61806006); 江苏省高校优势学科建设工程资助项目; 111 引智计划(12018). **吴习惠**(1998—), 女, 硕士研究生, 主要研究方向为目标跟踪、深度学习; **李婷**(1977—), 女, 硕士, 讲师, 主要研究方向为机器学习、大数据分析、智能优化算法; **葛洪伟**(1967—), 男, 博士, 教授, 博士生导师, 论文通信作者, 主要研究方向为人工智能、模式识别、机器学习、图像处理与分析.

任务中仅使用可见光这一种模态. 随着热红外技术的发展, 鉴于红外成像具有对光照不敏感、可以穿透烟雾等优点, RGBT(RGB-thermal)目标跟踪被相关研究人员广泛的关注. RGBT 跟踪任务是给定视频序列的第 1 帧, 根据其目标状态预测目标在后续帧的状态的任务. 由于可见光和热红外图像具有互补优势, 因此跟踪器在低光照、雾天等极端条件下也可以很好地跟踪目标.

早期的 RGBT 跟踪算法基于传统的算法[1-2], 采用手工提取特征的方式, 效率较低, 随着深度学习技术的发展, 研究者将深度学习技术应用于 RGBT 跟踪中. Li 等[3]提出多种适配器的方法 MANet, 分别提取模态特定特征、模态共享特征和实例感知特征; DAFNet[4]通过计算所有特征图的通道权重, 有效地聚合来自 2 种模态的特征和前一层的特征; 对于任意 2 层卷积层, CBPNet[5]采用双线性池化操作聚合来自不同层的特征. 但是, 上述的算法都没有充分地利用可见光模态和热模态之间的异质性. 现有的大部分算法对于可见光图像和热红外图像都采用相同的网络结构进行处理, 没有充分地发挥 2 种模态的不同特性.

针对上述问题, 本文提出一种非对称的分层特征融合的跟踪网络(asymmetric hierarchical feature fusion network, AHFNet). 由于可见光图像和热红外图像的成像原理不同, 因此包含的信息也有所不同, 可见光图像包含颜色纹理等外观信息, 热红外图像则不包含这些信息, 但是它在光照变化和背景杂乱的情况下可以提供更多有效的信息. 针对 2 种模态的不同的特性, 受 AGMINet[6]的启发, 本文设计了一个非对称结构, 与热红外图像相比, 因为可见光图像包含了更多有利于跟踪的信息, 所以将可见光模态设计为主分支, 热红外模态设计为辅助分支, 提供特征增强的辅助信息; 考虑浅层特征包含更多的目标定位信息, 而深层特征包含更多有利于分类的语义信息, 所以设计分层特征融合, 在主干网络提取特征的每层对可见光特征进行增强融合; 为了进一步利用热红外模态中的语义信息, 在热红外模态特征的最后一层对它进行增强, 以获得更利于跟踪的特征. 因此, AHFNet 包括用于特征提取的骨干网络、模态特征提取(modal feature extraction, MFE)模块、特征自适应聚合(feature adaptive aggregation, FAG)模块, 以及对不同模态进行增强融合的特征增强融合(feature enhancement fusion, FEF)模块.

1　相关工作

1.1　RGBT 目标跟踪

随着热红外技术的发展, RGBT 目标跟踪受到越来越多的关注. 深度学习技术的发展, 使得越来越多的 RGBT 跟踪算法采用深度学习的方法. 目前, RGBT 跟踪算法可以分为结合式融合算法和判别式融合算法. 结合式融合算法使用不同的融合方法将不同模态的信息进行结合. Zhang 等[7]提出一种自适应属性驱动感知网络, 充分地融合不同属性的特征; Xiao 等[8]提出一种基于 5 种挑战属性的渐进式融合网络; 王福田等[9]提出一种动态模态交互和特征自适应融合算法, 充分地挖掘可见光和热红外特征的互补性. 判别式融合算法通过挖掘不同模态的判别式信息, 并对判别式信息进行有效的融合. Zhu 等[10]提出特征自适应剪枝融合, 修剪所有模态的联合特征以减少冗余特征; Lu 等[11]提出对偶门控互条件网络, 通过对偶门控提取判别式特征. 然而, 这些 RGBT 跟踪算法大多采用的是对称结构, 没有充分地发挥可见光和热红外模态的不同特性.

1.2　基于 Transformer 的目标跟踪

早期的目标跟踪算法大多基于传统的算法[12-13], 伴随基于深度学习的目标跟踪算法高速发展[14-15], 近年来, Transformer[16]在目标检测中取得了非常好的效果, 很多算法将 Transformer 结构引入目标跟踪中. Chen 等[17]将 Transformer 结构引入单目标跟踪中; Wang 等[18]将 Transformer 结构应用于目标跟踪的时间线索, 利用视频跟踪时间特性来提高跟踪准确性. Cui 等[19]提出一种混合迭代的 Transformer 目标跟踪算法, 将特征提取和特征融合统一到一个架构中, 使模型更加简洁. 虽然 Transformer 结构在可见光跟踪中取得了非常好的效果, 但是其在 RGBT 目标跟踪中的应用还比较少.

1.3　MDNet 算法

MDNet[20]是一种基于多域学习的算法, 它将每个视频序列都看作一个单独的域, 其架构如图 1 所示. MDNet 的主干网络基于 VGG-M[21]架构, 其 conv1~conv3, fc4 和 fc5 都是共享层, 最后的 fc6 层是特定于域的层. MDNet 在共享层中进行特征提取和建模获得通用的特征表示, 对于 fc6 层, 训练一个用于区分目标和背景的分类器以达到跟踪的目的.

MDNet 将随机采样获得的样本裁剪为 107×107, 然后输入到图 1 所示的网络中, 获得样本的置信度分数. 在 MDNet 中, 采用难负样本挖掘策

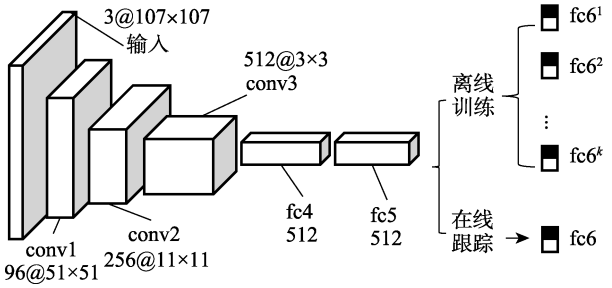


图 1 MDNet 架构

略解决偏移问题，每次迭代取分数最高的 96 个负样本和 32 个正样本，通过损失函数和随机梯度下降算法(stochastic gradient descent, SGD)对模型参数进行迭代更新. 损失函数的公式为

$$L_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1-y_i)\log(1-p_i) \quad (1)$$

其中；N 表示训练阶段样本的个数；y_i 表示第 i 个样本的真实标签；p_i 表示第 i 个样本的预测概率值.

2 AHFNet

2.1 整体结构

AHFNet 在 MDNet[20]的基础上构建，其结构如图 2 所示，包括双流主干网络、MFE 模块、FAG 模块和 FEF 模块. 首先使用 VGG-M 的前 3 层(conv1~conv3)作为主干网络，并将其扩展为双流网络，以便并行地提取可见光和热红外的特征，3 层卷积核大小分别 7×7，5×5 和 3×3；然后通过非对称分层融合结构对可见光特征进行增强融合，先使用 MFE 模块提取每层的可见光和热红外特征，再将提取到的特征输入 FAG 模块生成加权聚合后的特征；最后在 FEF 模块中，将每层聚合后的特征和主干网络每层提取到的可见光特征进行增强融合，得到融合后的特征，同时，主干网络第 3 层提取的热红外特征与本层的聚合特征进行增强融合，得到增强的热红外特征，将获得的 2 个增强融合特征级联后输入到 3 个全连接层，获得最终的分类和回归结果.

2.2 MFE 模块

为了增强双流网络每层获得的可见光特征，需要提取可用于增强的不同模态的特征. 由于深度神经网络计算的特征图存在冗余，为了减少计算负担，受到 Ghost[22]的启发，本文设计 MFE 模块，使用更少的参数获得更多的特征图，该模块的结构如图 3 所示. 首先输入特征图 $\boldsymbol{F}_{\text{p}}^{l}$；然后通过分组卷积得到另一个特征图 $\boldsymbol{F}_{\text{c}}^{l}$；将获得的 2 个特征图进行级联，得到最终的输出特征图

$$\boldsymbol{F}_{\text{out}}^{l} = \varepsilon\left(\boldsymbol{F}_{\text{p}}^{l}, \boldsymbol{F}_{\text{c}}^{l}\right) = \varepsilon\left(f_{\text{p}}^{l}\left(\boldsymbol{F}_{\text{in}}^{l}\right), f_{\text{c}}^{l}\left(\boldsymbol{F}_{\text{p}}^{l}\right)\right) \quad (2)$$

其中，$\boldsymbol{F}_{\text{in}}^{l} \in \mathbb{R}^{C_1\times H_1\times W_1}$；$\boldsymbol{F}_{\text{p}}^{l} \in \mathbb{R}^{(C_2/4)\times H_2\times W_2}$；$\boldsymbol{F}_{\text{c}}^{l} \in \mathbb{R}^{(C_2/4)\times 3\times H_2\times W_2}$；$\boldsymbol{F}_{\text{out}}^{l} \in \mathbb{R}^{C_2\times H_2\times W_2}$；$\varepsilon$ 表示级联操作；l 表示层数；f_{p}^{l} 表示卷积操作，f_{p}^{1}，f_{p}^{2} 和 f_{p}^{3} 的卷积核大小分别为 5×5，3×3 和 1×1；f_{c}^{l} 表示分组卷积操作，分组数为 $C_2/4$，卷积核大小为 3×3.

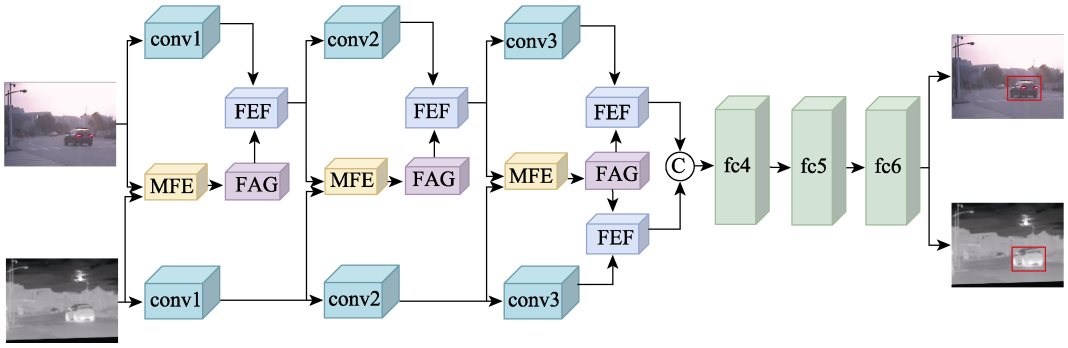


图 2 AHFNet 的结构

2.3 FAG 模块

为了聚合所提取的不同模态的特征，以获得有利于区分目标的特征，受 SKNet[23]的启发，本文设计的 FAG 模块可以自适应地选择来自不同分支的特征进行聚合，获得有助于增强可见光模态的特征. FAG 模块的结构如图 4 所示，运算过程如下：首先将每层输入的 2 个特征图 $\boldsymbol{F}_{\text{RGB}}^{l}$ 和 $\boldsymbol{F}_{\text{T}}^{l}$ 按元素相加；然后将相加获得的特征经过一个全局平均池化(global average pooling, GAP)操作、2 个卷积操作和归一化操作，获得 2 个注意力权重图 $\boldsymbol{w}_{\text{RGB}}^{l}$ 和 $\boldsymbol{w}_{\text{T}}^{l}$；再将输入的 2 个特征图与获得的 2 组注意力

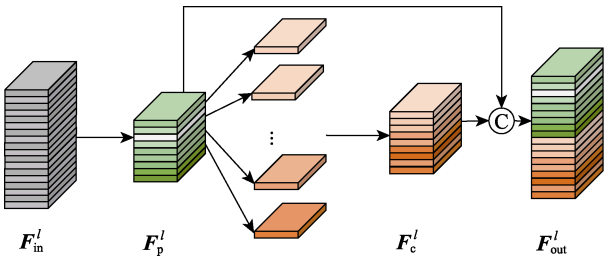


图 3　MFE 模块的结构

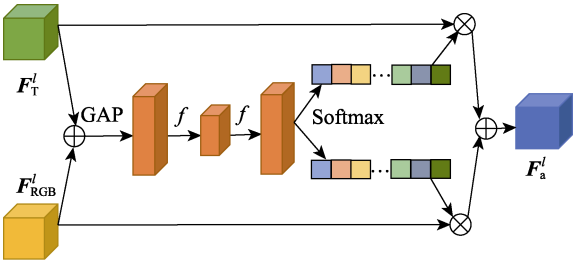


图 4　FAG 模块的结构

权重图按元素相乘，获得注意力加权后的特征图；最后将获得的特征图按元素相加，得到每层输出的特征图

$$\boldsymbol{X}_{\mathrm{a}}^{l}=\boldsymbol{w}_{\mathrm{RGB}}^{l}\times\boldsymbol{F}_{\mathrm{RGB}}^{l}+\boldsymbol{w}_{\mathrm{T}}^{l}\times\boldsymbol{F}_{\mathrm{T}}^{l} \tag{3}$$

其中，$\boldsymbol{w}^{l}=\mathrm{Softmax}\left(f\left(f\left(\mathrm{GAP}\left(\boldsymbol{F}_{\mathrm{RGB}}^{l}+\boldsymbol{F}_{\mathrm{T}}^{l}\right)\right)\right)\right)$；$f$ 表示1×1卷积操作，l 表示层数.

2.4　FEF 模块

为了更好地利用红外模态的辅助特性以及获得的增强可见光特征，受 Wang 等[18]和 Xiao 等[8]提出的将 Transformer 中的编码器和解码器进行拆分的启发，设计了 2 个编码器和 1 个解码器的结构. 编码器用于分别增强 2 种模态的特征，解码器用于交互融合 2 种特征，以充分地利用 2 种模态的互补特性，FEF 模块的结构如图 5 所示. (1) 使用1×1卷积将可见光特征 $\boldsymbol{X}_{\mathrm{RGB}}^{l}$ 和辅助特征 $\boldsymbol{X}_{\mathrm{a}}^{l}$ 进行降维，得到低维特征向量 $\boldsymbol{X}_{\mathrm{RGB}}^{l}\in\mathbb{R}^{d\times H\times W}$ 和 $\boldsymbol{X}_{\mathrm{a}}^{l}\in\mathbb{R}^{d\times H\times W}$；因为特征增强模块的编码器部分是以一组特征向量作为输入，所以将上面的低维特征向量展平获得长度为 d 的特征向量，将正弦函数的空间位置编码 P 与输入的一组特征向量相加作为输入. (2) 通过自注意力部分进行特征增强，增强后的特征输入到解码器部分，使用交叉注意力进行交互融合，得到最终输入的融合特征 $\boldsymbol{X}_{\mathrm{fuse}}^{l}$.

Transformer 的核心部分是注意力机制，本文使用自注意力机制和交叉注意力机制，为了简化网络，采用的是单头注意力机制. 注意力机制的计算公式为

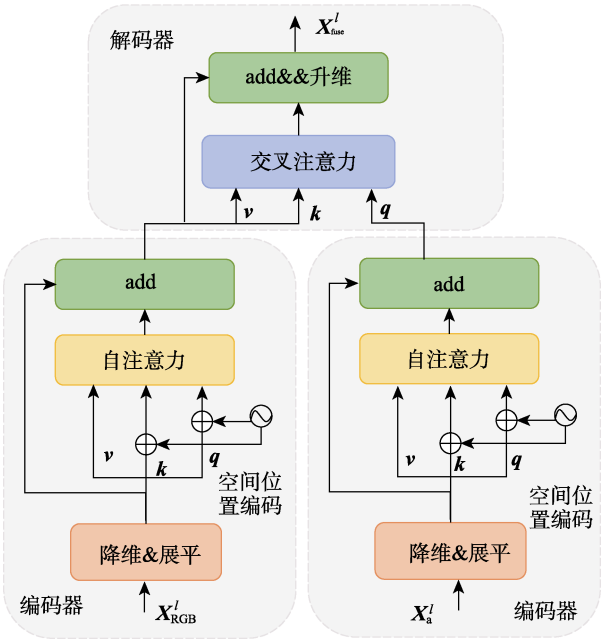


图 5　FEF 模块的结构

$$\boldsymbol{A}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v})=\mathrm{Softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}}{\tau}\right)\boldsymbol{v} \tag{4}$$

其中，$\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}$ 表示输入的一组特征向量；$\tau=30$.

整个编码器部分的计算公式为

$$\boldsymbol{X}_{\mathrm{encoder}}^{l}=\boldsymbol{A}(\boldsymbol{X}+P,\boldsymbol{X}+P,\boldsymbol{X})+\boldsymbol{X} \tag{5}$$

其中，$\boldsymbol{X}$ 表示展平后的特征向量；P 表示空间位置编码.

将可见光特征经过编码器获得的 $\boldsymbol{k},\boldsymbol{v}$ 特征向量和辅助特征经过编码器获得的特征向量 $\boldsymbol{q}$ 作为一组特征向量输入解码器中，通过交叉注意力机制和残差相加得到新的特征，然后使用1×1卷积进行升维，得到最终的融合特征.

2.5　训练过程

本文采用端到端的方式进行训练，使用 MDNet 在 imageNetvid 上预训练的模型参数来初始化主干网络，包含 3 个卷积层(conv1~conv3)和 2 个全连接层 fc4 和 fc5，随机初始化其他卷积层和全连接层. 采用 SGD 对网络进行训练，每次训练时，首先从训练数据集中随机选取 8 帧相对应的可见光图像和热红外图像，将选取的图像裁剪为107×107；然后对每幅图像进行随机采样，获得 32 个正样本和 96 个负样本，参照 MDNet，将高斯采样的候选框和真实框的交并比(intersection over union, IoU)值大于 0.7 的作为正样本，IoU 值小于 0.5 的为负样本. 在训练过程中，设卷积层的学习率为 0.000 1，全连接层的学习率为 0.001，权重衰减为 0.000 5，动量

为 0.9, 网络共训练 200 次循环. 具体的训练过程可以参照 MDNet[20]. 在 RGBT234[24]和 LasHeR[25]数据集上进行测试时, 使用 GTOT 数据集进行训练; 在 GTOT 数据集上测试时, 使用 RGBT234 数据集进行训练.

2.6 在线跟踪过程

在线跟踪阶段加载训练好的模型, 固定除 fc4~fc6 层之外的所有参数, 将 fc6 层的多域分支用单个分支代替, 随机初始化 fc6 层的参数, 对 fc4 层和 fc5 层的参数进行微调. 跟踪时, 首先使用序列的第 1 帧对跟踪器进行初始化, 在第 1 帧上, 高斯采样 500 个正样本和 5 000 个负样本, 使用这些样本训练 50 轮, fc4 和 fc5 层的学习率为 0.000 1, fc6 层的学习率设为 0.001; 然后采样 1 000 个样本训练候选框回归器, 得到精确的边界框. 在后续的跟踪中, 以前一帧的跟踪结果为中心, 采样 256 个候选区域作为当前帧的候选区域, 计算这 256 个候选区域的分数, 选择分数最高的 5 个计算平均值作为当前帧的跟踪结果; 然后通过候选框回归器对跟踪结果的边界框进行处理; 最后使用 AR(alpha-refine)[26]模块对跟踪结果边界框进一步精细化处理, 得到更加精确的跟踪结果. 在线跟踪过程中包含短时更新和长时更新, 更新过程与 MDNet[20]一致.

AR 模块的结构如图 6 所示, 该模块可以充分利用空间信息优化跟踪目标边界框, 实现更高性能的跟踪. 首先使用第 1 帧图像和第 1 帧目标的真实边界框对参考分支进行初始化; 然后在当前帧的测试分支中, 以候选回归器获得的跟踪结果扩大 2 倍作为搜索区域来预测目标位置, 获得更准确的目标边界框.

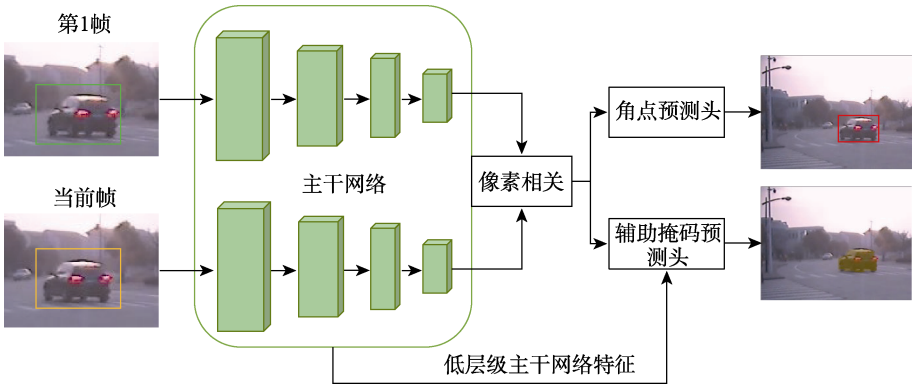


图 6 AR 模块的结构

AR 模块利用像素相关[27]操作, 对参考分支和测试分支经主干网络特征提取获得的模板特征 $\boldsymbol{F}_{\mathrm{t}} \in \mathbb{R}^{C\times H_{\mathrm{t}}\times W_{\mathrm{t}}}$ 和搜索特征 $\boldsymbol{F}_{\mathrm{s}} \in \mathbb{R}^{C\times H_{\mathrm{s}}\times W_{\mathrm{s}}}$ 进行信息交互; 使用角点预测头和辅助掩码预测目标的具体位置. 对于像素相关操作, 首先将搜索特征分成 $H_{\mathrm{t}}\times W_{\mathrm{t}}$ 个小的卷积核 $\boldsymbol{F}_i \in \mathbb{R}^{C\times 1\times 1}$; 然后计算这些卷积核与模板特征之间的相关性, 获得相关图

$$\boldsymbol{C}=\left\{\boldsymbol{C}_i \middle| \boldsymbol{C}_i=\boldsymbol{F}_i * \boldsymbol{F}_{\mathrm{s}}\right\}_{i\in\{1,2,\cdots,H_{\mathrm{t}}\times W_{\mathrm{t}}\}} \tag{6}$$

其中, $\boldsymbol{C} \in \mathbb{R}^{H_{\mathrm{t}}W_{\mathrm{t}}\times H_{\mathrm{s}}\times W_{\mathrm{s}}}$, $*$ 表示相关操作.

使用角点预测头预测目标的左上角和右下角坐标, 以获得目标的边界框. 使用 2 个各包含 4 个堆叠了卷积、BN 层和 ReLU 激活函数的卷积层, 获取目标的 2 个角的热图; 然后使用 Soft-argmax 函数对热图进行处理, 获得准确的角点位置. 辅助掩码预测头是基于 U-Net[28]进行设计的, 首先对主干网络获得模板特征和搜索特征, 进行逐层的上采样操作; 然后将它和主干网络获得的特征进行融合, 直至分辨率与输入图像相同; 最后通过获取的最后一层特征图预测掩码. 在跟踪阶段, 为了提高跟踪速度, 辅助掩码预测头被禁用. 更多的细节见文献[22].

3 实验结果与分析

本文实验平台为 Intel i9-9900x 3.5 GHz, 64 GB 内存, NVIDIA GeForce RTX2080Ti GPU; 在 Pytorch 深度学习框架下, 使用 Python 编程语言实现.

3.1 数据集和评价指标

本文在目前主流的 3 种 RGBT 跟踪数据集 GTOT[1], RGBT234[24]和 LasHeR[25]上进行实验.

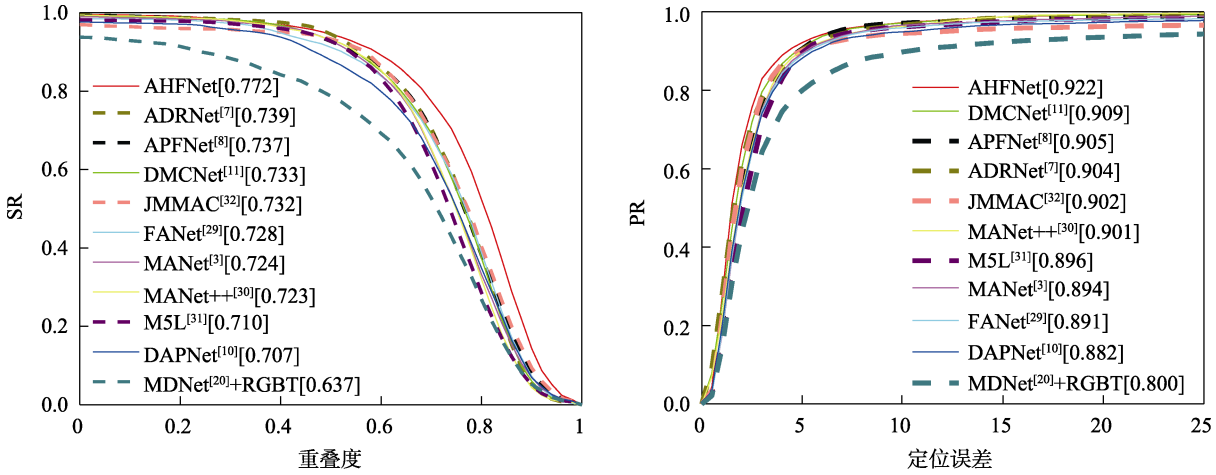
GTOT 数据集包含 50 个可见光和热红外视频序列, 有 7 个属性类别. RGBT234 数据集包含 234 个可见光和热红外的视频序列, 包含 12 个属性类别, 分别为背景杂乱(background clutter, BC), 相机运动(camera moving, CM), 形变(deformation, DEF), 快

速运动(fast motion, FM)，严重遮挡(heavy occlusion, HO)，低光照(low illumination, LI)，低分辨率(low resolution, LR)，运动模糊(motion blur, MB)，无遮挡(no occlusion, NO)，部分遮挡(partial occlusion, PO)，比例变化(scale variation, SV)和热交叉(thermal crossover, TC). LasHeR 数据集是一个大型的 RGBT 跟踪数据集，包含 1 224 个 RGBT 跟踪序列，选择其中的 245 个视频序列作为测试数据集.

本文采用精确率(precision rate, PR)和成功率(success rate, SR)作为客观评价指标. PR 表示跟踪结果中心点与真实值的中心点的距离小于阈值的帧数占所有帧数的百分比. 在 GTOT 数据集中，阈值设为 5 像素；在 RGBT234 和 LasHeR 数据集中，阈值设为 20 像素. SR 是预测边界框和真实边界框的 IoU 大于阈值的帧数占所有帧数的百分比.

3.2　在 GTOT 数据集上的实验

在 GTOT 数据集上，将 AHFNet 与 10 种 RGBT 跟踪算法进行实验，包括 MDNet[20]+RGBT, DAPNet[10], FANet[29], MANet[3], MANet++[30], M5L[31], JMMAC[32], ADRNet[7], APFNet[8]和 DMCNet[11]，客观评价指标结果如图 7 所示. 可以看出，AHFNet 的 PR 和 SR 分别为 0.992 和 0.772，比 MDNet+RGBT(0.800/0.637)提高 12.2 个百分点和 13.5 个百分点；与 APFNet(0.905)和 DMCNet (0.909)相比，AHFNet 的 PR 也提高 1.7 个百分点和 1.3 个百分点. 实验结果证明了 AHFNet 的有效性.

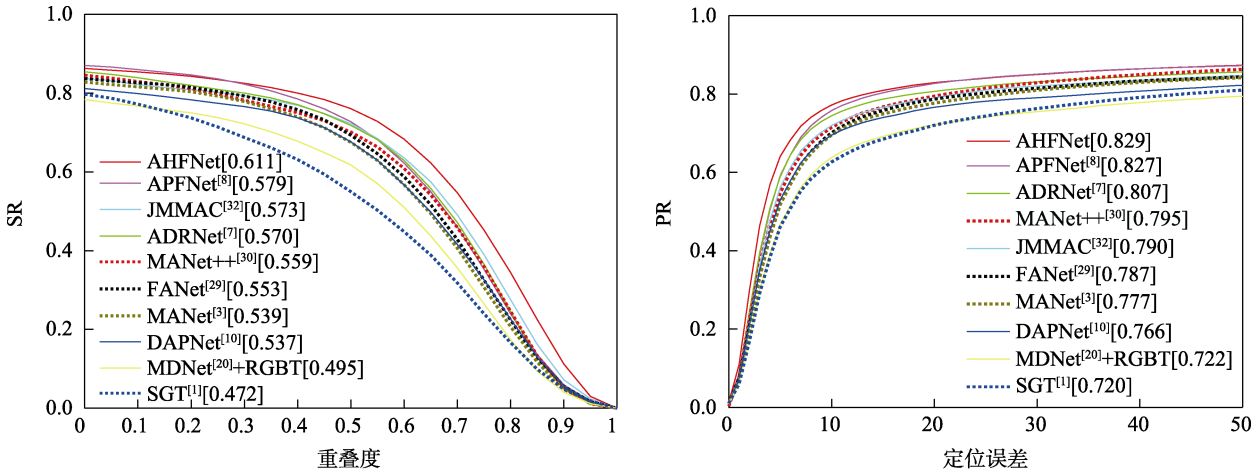


a. SR　　b. PR

图 7　在 GTOT 数据集上 11 种算法的客观评价指标对比

3.3　在 RGBT234 数据集上的实验

在 RGBT234 数据集上，将 AHFNet 与 SGT[1], MDNet[20]+RGBT, DAPNet[10], MANet[3], FANet[29], MANet++[30], JMMAC[32], ADRNet[7]和 APFNet[8]这 9 种 RGBT 跟踪算法进行实验，结果如图 8 所示. 可以看出，AHFNet 的 PR=0.829，SR=0.611，比 MDNet+RGBT(0.722/0.495)的 PR 和 SR 分别提高 10.7 个百分点和 11.6 个百分点；与 APFNet(0.827/



a. SR　　b. PR

图 8　在 RGBT234 数据集上 10 种算法的客观评价指标对比

0.579)相比, AHFNet 的 PR 和 SR 分别提高 0.2 个百分点和 3.2 个百分点. 实验结果验证了 AHFNet 的有效性.

在 RGBT234 数据集上, 12 个属性类别下, 将 AHFNet 与 MDNet[20]+RGBT, FANet[29], DAPNet[10], MANet++[30], ADRNet[7]和 APFNet[8]这 6 种算法进行实验, 结果如表 1 所示. 可以看出, 在大部分属性类别下, AHFNet 都能取得很好的结果, 优于其他 RGBT 跟踪算法, 但是对于 MB, AHFNet 的 PR 和 SR 都低于 APFNet; 对于 CM, FM, HO, LR 和 TC 这 5 个属性类别, AHFNet 的 PR 低于 APFNet, 但 SR 高于 APFNet, 这是因为 APFNet 中设计了基于属性的分支, 在训练时使用相对应属性的数据集来训练相应的分支, 而 AHFNet 中没有针对属性分支的设计. 但是与其他 RGBT 跟踪算法相比, AHFNet 取得了很好的结果.

表 1 在 RGBT234 数据集上不同跟踪算法对 12 个属性类别的客观评价指标 PR/SR 对比

属性类别	跟踪算法						
	MDNet[20]+RGBT	FANet[29]	DAPNet[10]	MANet++[30]	ADRNet[7]	APFNet[8]	AHFNet
背景杂乱	0.644/0.432	0.757/0.502	0.717/0.484	0.763/0.492	0.804/0.536	0.813/0.545	**0.839/0.567**
相机运动	0.640/0.454	0.724/0.523	0.668/0.474	0.741/0.520	0.743/0.529	**0.779**/0.563	0.756/**0.568**
形变	0.668/0.473	0.722/0.526	0.717/0.518	0.787/0.562	0.743/0.528	0.785/0.564	**0.807/0.611**
快速运动	0.586/0.363	0.681/0.436	0.670/0.443	0.694/0.455	0.749/0.489	**0.791**/0.511	0.767/**0.526**
严重遮挡	0.619/0.421	0.665/0.458	0.660/0.457	0.711/0.482	0.714/0.496	**0.738**/0.569	0.725/**0.513**
低光照	0.670/0.455	0.803/0.548	0.775/0.530	0.796/0.543	0.811/0.560	0.843/0.569	**0.848/0.593**
低分辨率	0.759/0.515	0.795/0.532	0.750/0.510	0.789/0.520	0.838/0.562	**0.844**/0.565	0.838/**0.579**
运动模糊	0.654/0.463	0.700/0.503	0.653/0.467	0.733/0.518	0.733/0.532	**0.745/0.545**	0.709/0.538
无遮挡	0.862/0.611	0.882/0.657	0.900/0.644	0.902/0.664	0.916/0.660	0.948/0.680	**0.952/0.725**
部分遮挡	0.761/0.518	0.866/0.602	0.821/0.574	0.830/0.591	0.851/0.603	0.863/0.606	**0.879/0.658**
比例变化	0.735/0.505	0.785/0.563	0.780/0.542	0.792/0.573	0.786/0.562	0.831/0.579	**0.832/0.625**
热交叉	0.756/0.517	0.766/0.549	0.768/0.543	0.772/0.558	0.796/0.581	**0.822**/0.581	0.803/**0.597**
全部	0.722/0.495	0.787/0.553	0.766/0.537	0.795/0.559	0.807/0.570	0.827/0.579	**0.829/0.611**

注. 粗体表示最优值.

3.4 在 LasHeR 数据集上的实验

在 LasHeR 数据集上, 将 AHFNet 与 SGT[5], DAPNet[10], FANet[29], mfDiMP[33], CAT[34], MANet[3], MANet++[30], MaCNet[35], DMCNet[11]和 APFNet[8]这 10 种跟踪算法进行实验, 结果如表 2 所示. 可以看出, AHFNet 的 PR 和 SR 优于其他算法, 与 SGT 相比, PR 和 SR 分别提高 20.0 个百分点和 17.1 个百分点; 与 APFNet 相比, PR 和 SR 分别提高 2.7 个百分点和 4.1 个百分点; 与 DMCNet 相比, PR 和 SR 分别提高 3.7 个百分点和 4.8 个百分点. 实验结果充分证明了 AHFNet 的有效性.

表 2 LasHeR 数据集上 11 种算法客观评价指标对比 %

算法	PR	SR	算法	PR	SR
SGT[5]	0.327	0.232	MANet++[30]	0.467	0.314
DAPNet[10]	0.431	0.311	MaCNet[35]	0.482	0.350
FANet[29]	0.441	0.309	DMCNet[11]	0.490	0.355
mfDiMP[33]	0.447	0.343	APFNet[8]	0.500	0.362
CAT[34]	0.450	0.314	AHFNet	**0.527**	**0.403**
MANet[3]	0.455	0.326			

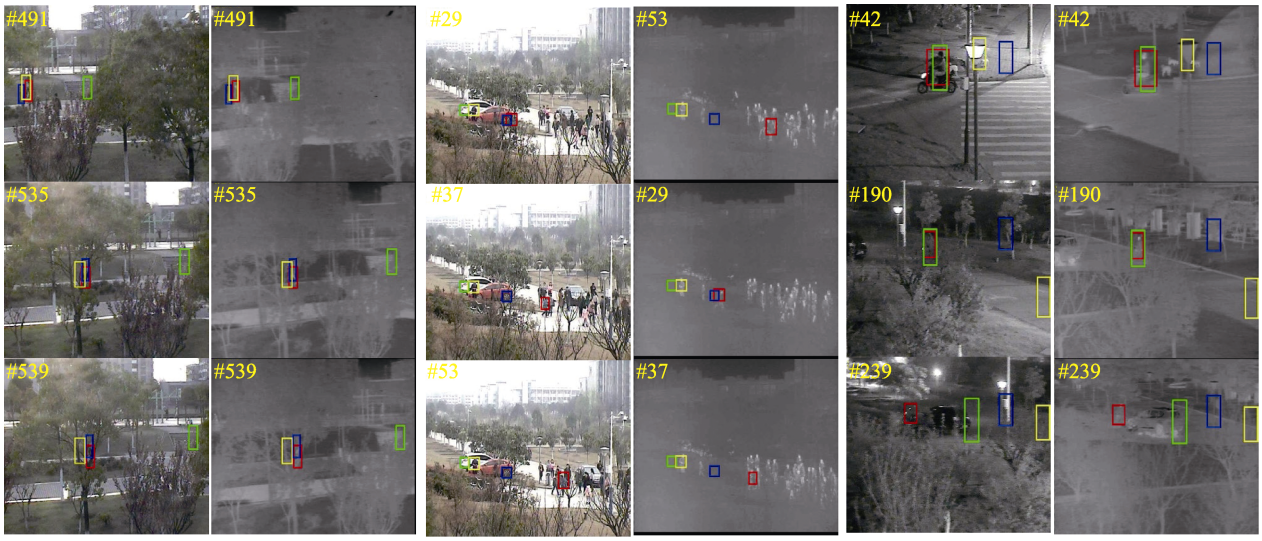
注. 粗体表示最优值.

3.5 定性分析

在 car37, children4 和 elecbike10 这 3 种序列下, 对 AHFNet, MANet[3], MaCNet[35]和 MDNet[20]+RGBT 这 4 种跟踪算法进行实验, 可视化结果如图 9 所示. 可以看出, 图 9a 中, 在树木遮挡后, 其他 3 种算法都发生了不同程度的偏移, AHFNet 还能够准确地跟踪目标; 图 9b 中, 在目标快速运动且存在遮挡的情况下, 只有 AHFNet 很好地跟踪了目标, 其他 3 种算法发生了不同程度的偏移或者丢失目标的情况; 图 9c 中存在光照不足的情况, 在目标被车辆完全遮挡后, 其他 3 种算法完全丢失目标, AHFNet 还能很好地跟踪目标. 实验结果表明, 与其他 3 种跟踪算法相比, AHFNet 的跟踪结果更加精确, 充分证明其有效性.

3.6 失败案例

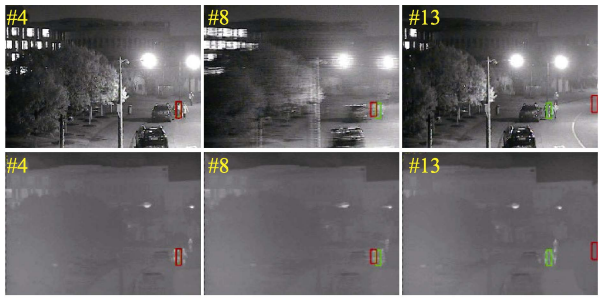
图 10 所示为 AHFNet 跟踪失败的几个案例, 其中, 红色框为 AHFNet 的跟踪结果, 绿色框为真实边界框. 可以看出, 图 10a 中, AHFNet 不能准确地跟踪目标, 跟踪失败的原因可能是相机突然运动导致背景模糊, 从而发生跟踪漂移; 图 10b 中,



a. car37　　b. children4　　c. elecbike10

AHFNet; MANet[3]; MaCNet[35]; MDNet[20]+RGBT

图 9　4 种跟踪算法的定性比较



a. threewoman2



b. child3

图 10　跟踪失败案例

在目标被完全遮挡且遮挡物与目标的相似度较高的情况下，AHFNet 无法准确地跟踪目标. 实验结果表明，对于多种复杂挑战同时存在的情况下，AHFNet 还不能很好地跟踪目标.

3.7　消融实验

为了进一步验证 AHFNet 的有效性，在 RGBT 234 数据集上进行消融实验，结果如表 3 所示. 其中，√表示采用，采用 Baseline 表示基线方法 MDNet+RGBT. 消融实验如下：(1) Baseline+MFE. 去除 FAG 模块、FEF 模块和 AR 模块，使用相加操作代替 FAG 和 FEF 模块；(2) Baseline+MFE+FEF. 去除 FAG 和 AR 模块，使用相加操作代替 FAG 模块；(3) Baseline+MFE+FEF. 去除 FAG 和 AR 模块，使用相加操作代替 FAG 模块；(4) Baseline+MFE+FAG+FEF. 去除 AR 模块，在得到跟踪结果后不使用 AR 模块进行精确处理；(5) Baseline+MFE+FAG+FEF+AR. 即为 AHFNet. 可以看出，Baseline+MFE+FAG 的结果优于 Baseline+MFE，说明自适应地选择 2 种模态的聚合特征比直接将 2 种模态的特征相加效果更好，自适应聚合模块可以选择更利于跟踪的特征，证明了 FAG 模块的有效性；Baseline+MFE+FAG+FEF 的结果比 Baseline+MFE+FAG 的结果更好，说明与简单的相加操作相比，使用 FEF 模块将特征进行增强的效果更好，证明了 FEF 模块的有效性；将 Baseline+MFE+FAG, Baseline+MFE+FEF 和 Baseline+MFE+FAG+FEF 进行比较，单独使用 FAG 模块或 FEF 模块的性能都有所下降，再次说明了 FAG 模块和 FEF 模块的有效性；将 Baseline+MFE+FAG+FEF 与 AHFNet 进行比较，AHFNet 的结果更好，说明在跟踪阶段，使用 AR 模块进一步精确跟踪结果是有效的.

表 3　在 RGBT234 数据集上的消融实验结果

Baseline	MFE	FAG	FEF	AR	PR	SR
√	√				0.809	0.567
√	√	√			0.821	0.575
√	√		√		0.811	0.573
√	√	√	√		0.825	0.576
√	√	√	√	√	**0.829**	**0.611**

注. 粗体表示最优值.

4 结 语

针对可见光和热红外模态的异质性问题，本文提出一种非对称分层特征融合的 RGBT 跟踪网络 AHFNet. 通过非对称的分支挖掘可见光模态和热红外模态的不同特性，充分发挥两者的优势，同时使用改进的 Transformer 结构来进一步增强模态的特征，提高跟踪网络的性能. 在 3 种数据集上，与主流算法的实验结果表明，AHFNet 优于其他 RGBT 跟踪网络. 但是, AHFNet 也存在不足，与实时的 RGBT 目标跟踪算法相比，其实时性较差，可以考虑进一步精简网络以提高实时性. 目前有很多成熟的精简网络的方法，包括剪枝、量化和轻量化网络等，未来可考虑使用这些方法提高网络的实时性. 另外，本文虽然使用了一个非对称结构，但是针对的是可见光模态的特征进行增强，对于热红外特征的利用不够充分，今后要进一步挖掘和利用热红外模态.

参考文献(References):

[1] 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

[2] Li Chenglong, Lu Andong, Liu Lei, *et al*. Multi-modal visual tracking: a survey[J]. Journal of Image and Graphics, 2023, 28(1): 37-56(in Chinese)
(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)

[3] Li C L, Lu A D, Zheng A H, *et al*. Multi-adapter RGBT tracking[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop. Los Alamitos: IEEE Computer Society Press, 2019: 2262-2270

[4] Gao Y, Li C L, Zhu Y B, *et al*. Deep adaptive fusion network for high performance RGBT tracking[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop. Los Alamitos: IEEE Computer Society Press, 2019: 91-99

[5] Xu Q, Mei Y M, Liu J P, *et al*. Multimodal cross-layer bilinear pooling for RGBT tracking[J]. IEEE Transactions on Multimedia, 2022, 24: 567-580

[6] Mei J T, Liu Y Y, Wang C C, *et al*. Asymmetric global-local mutual integration network for RGBT tracking[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: Article No.5017417

[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] Xiao Y, Yang M M, Li C L, *et al*. Attribute-based progressive fusion network for RGBT tracking[C] //Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 2831-2838

[9] Wang Futian, Zhang Shuyun, Li Chenglong, *et al*. RGBT tracking based on dynamic modal interaction and adaptive feature fusion[J]. Journal of Image and Graphics, 2022, 27(10): 3010-3021(in Chinese)
(王福田, 张淑云, 李成龙, 等. 动态模态交互和特征自适应融合的 RGBT 跟踪[J]. 中国图象图形学报, 2022, 27(10): 3010-3021)

[10] 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. New York: ACM Press, 2019: 465-472

[11] Lu A D, Qian C, Li C L, *et al*. Duality-gated mutual condition network for RGBT tracking[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022: 1-14

[12] Xiao Yiqing, Ge Hongwei. Adaptive correlation filtering algorithm suitable for long-term tracking[J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(1): 121-129(in Chinese)
(肖逸清, 葛洪伟. 适合长时跟踪的自适应相关滤波算法[J]. 计算机辅助设计与图形学学报, 2020, 32(1): 121-129)

[13] Yang Jinlong, Chen Xiaoping, Tang Yu, *et al*. Visual tracking algorithm based on label consistent *K*-SVD sparse coding[J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(2): 262-272(in Chinese)
(杨金龙, 陈小平, 汤玉, 等. 标签一致 *K*-SVD 稀疏编码视频跟踪算法[J]. 计算机辅助设计与图形学学报, 2018, 30(2): 262-272)

[14] Shao Jiangnan, Ge Hongwei. Siamese object tracking algorithm combining residual connection and channel attention mechanism[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(2): 260-269(in Chinese)
(邵江南, 葛洪伟. 融合残差连接与通道注意力机制的 Siamese 目标跟踪算法[J]. 计算机辅助设计与图形学学报, 2021, 33(2): 260-269)

[15] Fan Ying, Song Xiaoning. Siamese progressive attention-guided fusion network for object tracking[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(2): 199-206(in Chinese)
(范颖, 宋晓宁. 用于目标跟踪的孪生渐进注意引导融合网络[J]. 计算机辅助设计与图形学学报, 2021, 33(2): 199-206)

[16] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 6000-6010

[17] Chen X, Yan B, Zhu J W, *et al*. Transformer tracking[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 8122-8131

[18] Wang N, Zhou W G, Wang J, *et al*. Transformer meets tracker: exploiting temporal context for robust visual tracking[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 1571-1580

[19] Cui Y T, Jiang C, Wang L M, *et al*. MixFormer: end-to-end tracking with iterative mixed attention[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 13598-13608

[20] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 4293-4302

[21] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2023-02-21]. https://arxiv.org/abs/1409.1556

[22] Han K, Wang Y H, Tian Q, *et al*. GhostNet: more features from cheap operations[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 1577-1586

[23] Li X, Wang W H, Hu X L, *et al*. Selective kernel networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 510-519

[24] Li C L, Liang X Y, Lu Y J, *et al*. RGB-T object tracking: benchmark and baseline[J]. Pattern Recognition, 2019, 96: Article No.106977

[25] 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, 2022, 31: 392-404

[26] Yan B, Zhang X Y, Wang D, *et al*. Alpha-refine: boosting tracking performance by precise bounding box estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 5285-5294

[27] Wang Z Q, Xu J, Liu L, *et al*. RANet: ranking attention network for fast video object segmentation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 3978-3987

[28] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] //Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer Press, 2015: 234-241

[29] 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

[30] Lu A D, Li C L, Yan Y Q, *et al*. RGBT tracking via multi-adapter network with hierarchical divergence loss[J]. IEEE Transactions on Image Processing, 2021, 30: 5613-5625

[31] Tu Z Z, Lin C, Zhao W, *et al*. M^5L: multi-modal multi-margin metric learning for RGBT tracking[J]. IEEE Transactions on Image Processing, 2022, 31: 85-98

[32] Zhang P Y, Zhao J, Bo C J, *et al*. Jointly modeling motion and appearance cues for robust RGB-T tracking[J]. IEEE Transactions on Image Processing, 2021, 30: 3335-3347

[33] Zhang L C, Danelljan M, Gonzalez-Garcia A, *et al*. Multi-modal fusion for end-to-end RGB-T tracking[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop. Los Alamitos: IEEE Computer Society Press, 2019: 2252-2261

[34] Li C L, Liu L, Lu A D, *et al*. Challenge-aware RGBT tracking[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 222-237

[35] 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): Article No.393