文章编号:1001-5078(2024)11-1767-10

·图像与信号处理·

可逆多分支的双模态自适应融合目标跟踪算法

耿礼智,周冬明,王长城,刘宜松,孙逸秋
(云南大学 信息学院,云南 昆明 650500)

摘　要:由于可见光和红外图像具有很强的互补性,越来越多的关注集中在通过这两种模态的联合信息进行跟踪。然而,在现有的跟踪算法中,不能有效地学习两者的互补信息并挖掘模态特定特征,这限制了跟踪器的性能。因此,本文提出了一种可逆多分支的双模态自适应融合跟踪网络。首先,设计了一个三分支结构网络,分别用于学习热红外、可见光以及它们的通用特征。这不仅充分利用了两种模态之间的共享信息,还保留了红外和可见光数据之间的差异特性以及丰富的细节信息。此外,还引入了一个模态特征交互模块,以自适应地挖掘模态之间的互补信息并滤除冗余信息。通过在多个公开数据集上进行大量实验证明了该跟踪器的有效性,尤其在面对尺度变化、镜头抖动、遮挡等环境时,表现出卓越的抗干扰能力。

关键词:热红外目标跟踪;多分支;自适应融合;可逆结构
中图分类号:TP391;TN219　　**文献标识码:**A　　**DOI:**10.3969/j.issn.1001-5078.2024.11.018

Enhanced target tracking algorithm with reversible multi-branch bimodal adaptive fusion

GENG Li-zhi, ZHOU Dong-ming, WANG Chang-cheng, LIU Yi-song, SUN Yi-qiu
(School of Information Science and Engineering, Yunnan University, Kunming 650500, China)

Abstract: Due to the strong complementarity between visible light and infrared images, more attention has been focused on tracking through the joint information of these two modalities. However, in existing tracking algorithms, hthe inability to effectively learn the complementary information of both and mine modality-specific features limits the performance of the tracker. In responseto this issue, a reversible multibranch bimodal adaptive fusion network for tracking is proposed. Firstly, a tri-branch structured network is designed for separate learning of thermal infrared, visible light, and their shared characteristics. This design not only maximizes the utilization of shared modal information, but also preserves the differential characteristics between infrared and visible data as well as the rich detail information. Furthermore, an adaptive module for modal feature interaction is introduced to efficiently mine complementary modal information and filter out redundant data. Extensive experiments conducted on multiple public datasets proves the effectiveness of this tracker, particularly showcasing remarkable anti-interference capabilities in scenarios involving scale changes, camera shakes, and occlusion.

Keywords: thermal infrared object tracking; multibranch; adaptive fusion; reversible structure

基金项目:国家自然科学基金项目(No. 62066047;No. 61966037);云南大学专业学位研究生实践创新基金项目(No. ZC-23234092)资助。
作者简介:耿礼智(2000-),男,硕士研究生,研究方向为计算机视觉、多模态目标跟踪技术。E-mail:lizhigeng@mail.ynu.edu.cn
通讯作者:周冬明(1963-),男,博士,教授,研究方向为计算机视觉、智能图像处理技术。E-mail:zhoudm@ynu.edu.cn
收稿日期:2023-12-22;**修订日期:**2024-01-30

1 引 言

视觉目标跟踪是在初始帧中指定一个任意对象，并在后续帧中预测该对象的位置和尺度。这项技术作为计算机视觉的一项重要任务，在智能监控、自动驾驶、人机交互等领域得到了广泛的应用[1-3]。在跟踪任务中，可见光图像因其高分辨率和富含纹理、颜色等细节信息而取得了显著的成果。然而，可见光图像是通过捕捉物体反射的电磁波成像的，因此在特殊场景下（如长时间遮挡、低光照、高曝光、雨雪天气等）跟踪器会难以正常工作[4-5]。相反，红外图像是通过捕捉物体发出的热辐射进行成像的，它对光不敏感且具有强大的穿透能力。由于可见光图像和红外图像（RGB and Thermal，RGBT）具有良好的互补特性，因此越来越多的研究关注了具有全天候鲁棒跟踪性能的 RGBT 跟踪[6]。与基于可见光图像的跟踪任务不同，RGBT 目标追踪算法旨在综合两种模态的视觉信息以提升目标跟踪性能。因此，合理地利用可见光模态和热红外模态的信息、充分挖掘两者之间的互补性，是 RGBT 视觉跟踪任务的核心问题。

由于稀疏表示可以抑制噪声和误差，早期的方法主要是基于稀疏表示的传统模型[7-9]。然而，这些方法仅使用手工制作特征，因此在复杂场景下往往无法有效进行跟踪。深度学习方法相较于传统模型具有更强的特征表示能力，因此在 RGBT 跟踪领域，基于深度学习的方法已经占据主导地位。这些方法在特征提取和融合两个关键阶段展开了不同的研究。首先，一些算法专注于特征提取阶段。例如，Li 等人[10]以 MDNet[11]为基准，设计了一个双分支网络分别提取可见光和热红外模态特征，然后将两者简单相加去除冗余后进行联合跟踪。Li 等人[12]提出了一种多适配器结构来有效提取多模态特征，并将两种模态特征直接拼接后跟踪。而 Lu 等人[13]则通过嵌入散度损失来升级多适配器网络，以实现更鲁棒的表示学习，但并未对两种模态特征融合方式做出改变。Yan 等人[14]引入了外部注意力对红外和可见光特征进行增强，也只是简单将不同特征进行拼接。尽管这些方法采用了不同的策略来提取特征取得很好的成效，但它们忽视了模态之间的互补特性，没有充分利用不同模态的优势。此外，还有一些算法专注于改进特征融合策略。例如，Gao 等人[15]将不同卷积层的特征加权融合，通过自适应融合模块有效地聚合了不同层的可见光和热红外模态特征。Zhang 等人[16]提出一个模态感知注意网络对原始数据进行整合，建立表征不同特征层重要性的注意力模型，然后引导特征融合。Hou 等人[17]提出了一种多阶段跟踪器，通过一个模态交互模块自适应地融合多模态特征，并对跟踪结果进行评估和调整。Wang[18]等人提出了一种动态模态感知滤波器生成模块通过自适应调整卷积核来增强可见光和热红外模态之间的信息交互。然而，这些方法虽然在公开的 RGBT 基准数据集上表现不错，但它们忽略了模态共享信息与特定信息的充分挖掘。例如，轮廓和边缘等模态共享特征提供了关于目标形状、位置和运动的重要信息，对目标定位相当重要。而纹理、亮度等其他模态特定的细节则提供了更多关于目标和背景的特征信息，这对提高跟踪的准确性和稳定性非常关键[19]。

为了解决上述问题，本文提出了一种名为可逆多分支的双模态自适应融合跟踪网络（Reversible Multibranch Bimodal Adaptive Fusion Tracking Network，RMFNet）。这一方法设计了两个可逆学习模块（Reversible Learning Module，RLM），以充分挖掘各自模态特定信息，并通过不同层次特征的融合引入更多细节特征。此外，还提出了一个模态特征交互模块（Modal Feature Interactive Module，MFIM），其中包含了模态交互注意力（Modal Interactive Attention，MIA）和门控单元（Gate Unit，GU）两个组件，以实现对 RGB 和热红外模态特征的自适应融合。这有助于实现不同模态的互补学习，并有效地抑制特征冗余，充分发挥可见光数据与热红外数据之间的互补优势。为验证 RMFNet 的性能，通过在两个流行的 RGBT 跟踪基准数据集 GTOT[20]和 RGBT234[21]和进行了大量实验。结果表明，RMFNet 在准确性和稳定性方面表现出色，为 RGBT 跟踪任务带来了显著的性能提升。

2 可逆多分支的自适应融合 RGBT 跟踪算法

2.1 网络整体结构

适当的网络结构是跟踪性能提升的关键，为了有效提取所需模态特定特征和模态共享特征，并实现跨模态特征建模，本文设计的可逆多分支的双模态融合跟踪网络的具体网络结构如图 1 所示。RMFNet 主要由以下几部分组成：可见光模态分支、模

态共享分支、热红外模态分支、模态特征交互模块和实例分类层。

首先在特征提取阶段，模态共享分支使用在大型数据集 ImageNet[22] 上预训练的轻量级 VGG－M 网络的前三层，并将其扩展为共享参数的并行双流结构来提取 RGB 和热红外的特征。中第一层和第二层结构相似，都由卷积、ReLU 激活函数、层归一化和最大池化组成，但卷积核大小分别为 7×7 和 5×5。而第三层由 3×3 的卷积、ReLU 激活函数、层归一化组成。此外，可见光模态分支和红外模态分支通过引入可逆学习模块（Reversible Learning Module，RLM）来对可见光和热红外模态信息进一步挖掘。其次，通过包含了模态交互注意力和门控单元两个组件的模态特征选择模块来充分学习模态之间的互补性，自适应地聚合模态之间的有效特征。最后，使用由三个全连接层 FC4，FC5 和 FC6 组成的实例分类层来区分每一帧的目标和背景，从而对目标位置和尺度进行预测。

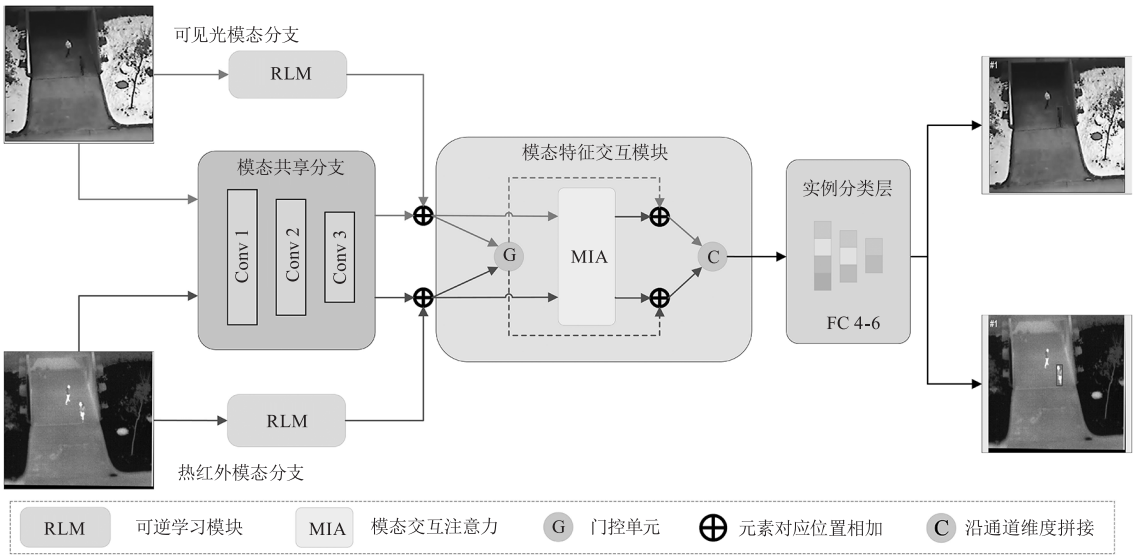

图 1　RMFNet 整体结构

Fig. 1 The structure of RMFNet

2.2　可逆学习模块

为了充分挖掘和利用不同模态特定特征，本文在 RMFNet 中设计了可逆学习模块。在提取可见光和红外模态各自特征的同时，对多层特征由浅到深进行融合。可逆学习模块是一种旨在通过可逆结构来较大程度保留原始信息的特征学习策略[23]，以此来增强目标物体的定位。浅层特征包含更多细节信息，而深层特征具有丰富的语义。信息在逐层传播过程中被逐渐压缩，可能造成有用信息的丢失。可逆结构使得信息在不同分支间无损传播，通过将不同层次特征进行融合来防止有用信息过多丢失。如图 2 所示，可逆学习模块由主干网络（STEM）和两个包含 4 层多级可逆单元（Multilevel Reversible Unit，MRU）的分支组成。

主干网络由 4×4 的卷积、层归一化（LayerNorm）和自适应池化（Adaptive Pooling）组成。首先主干网络对输入图像进行初步学习，将其映射为固定大小为 24×24 的特征图。然后将得到的特征图分别输入到分支 1 和分支 2 的 MRU 逐层加深提取特征，并将分支 1 的深层特征、分支 2 的浅层特征以及各自同层级特征逐层融合后，输出模态特定特征。在训练过程中，采用随机初始化参数。所以，分支 1 和分支 2 的结构一致，只是初始化参数不同。

对于分支 2 的前三层 MRU，使用当前分支的上一层浅层特征 x'_{t-1} 和分支 1 的下一层深层特征 x_{t+1} 作为输入，然后和分支 1 的同层特征 x_t 融合生成输出 x'_t。其中 x_t 到 x'_t 的映射是可逆的，即 x_t 可以由后验特征 x'_{t-1}，x'_t 和 x_{t+1} 重建。形式上，正向和反向计算如公式（1）和（2）所示：

$$x'_t = f_t(x'_{t-1}, x_{t+1}) + \gamma x_t, (t = 1,2,3) \tag{1}$$

$$x_t = \gamma^{-1}[x'_t - f_t(x'_{t-1}, x_{t+1})], (t = 1,2,3) \tag{2}$$

其中，f_t 表示类似于标准 ResNets[24] 中的残差函数的任意非线性运算；γ 是简单的可逆操作（例如通道缩放），其逆由 γ^{-1} 表示。由于梯度计算的需要，传统网络的训练需要占用大量内存来存储前向传播过程中的激活。而由于采用显式可逆结构，因此在

反向传播期间,可以动态地从分支 2 到分支 1 重建所需的激活,这意味在训练期间只需要存储分支 2 激活信息来节省内存。

每层 MRU 由跨层特征聚合模块(Cross Layer Feature Aggregation Module,CFAM)和一个 ConvNext Block[25] 串联构成,实现不同层次特征的融合。CFAM 包含上采样层和下采样层两个部分,对不同层次的输入特征进行处理。上采样层将输入的深层低分辨率特征通过线性层(Linear)、层归一化(LayerNorm)和最近邻插值(Interpolation)输出 2 倍上采样特征。同时,下采样层将输入的浅层高分辨率特征通过 2×2 大小的卷积和层归一化进行下采样 2 倍,得到与上采样层输出大小相同的特征。最后通过 ConvNext Block 输出下一层特征,其由 3×3 大小的逐通道卷积(DWConv)、层归一化、GELU 激活函数和两个逐点卷积(PWConv)组成。

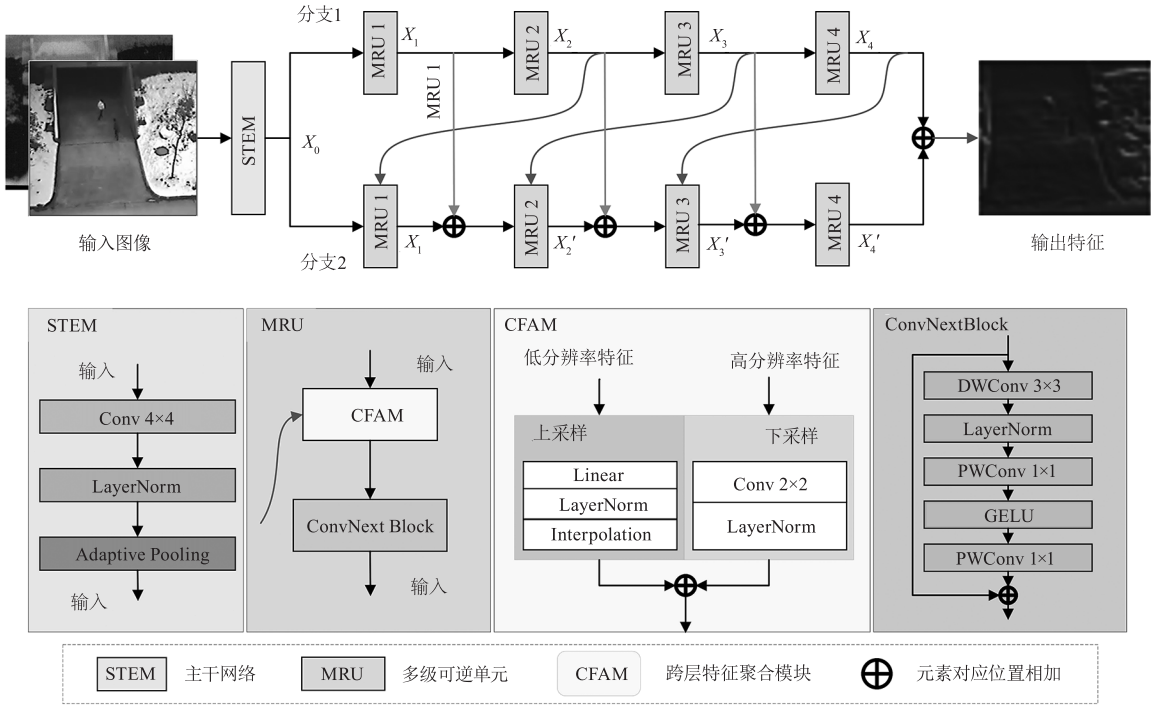

图 2 可逆学习模块结构

Fig. 2 The structure of reversible learning module

2.3 模态交互注意力

为了充分利用模态之间的互补性,本文设计了一个模态交互注意力来双向学习模态间潜在的相关性和全局感知信息,其结构如图 3 所示。

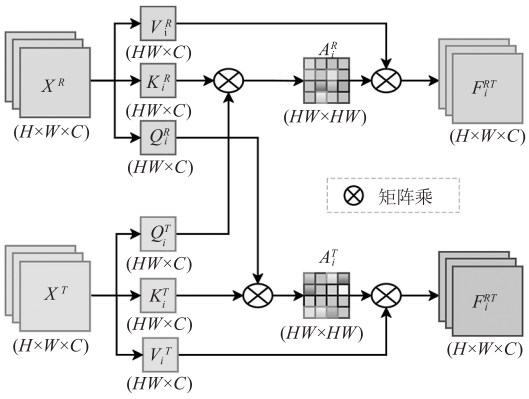

图 3 模态交互注意力结构

Fig. 3 The structure of modal interactive attention

首先将可见光模态特征 X^R 和红外模态的特征 X^T 分别通过三个 1×1 的卷积并进行维度变化($\mathbb{R}^{H\times W\times C}\rightarrow\mathbb{R}^{HW\times C}$),从而构建可见光模态和红外模态的查询向量、被查向量和内容向量,即 Q_i^R、K_i^R、V_i^R 和 Q_i^T、K_i^T、V_i^T,其在每个注意力头中是独立的。

然后对每个注意力头分别计算注意力分数,如公式:

$$\begin{cases} A_i^R = \dfrac{Q_i^R\,(K_i^T)^T}{\sqrt{d_k}} \\ A_i^T = \dfrac{Q_i^T\,(K_i^R)^T}{\sqrt{d_k}} \end{cases} \tag{3}$$

其中,A_i^R 和 A_i^T 分别表示可见光和红外模态的第 i 个注意力头的注意力分数矩阵;$\sqrt{d_k}$ 表示缩放因子,用来减小梯度爆炸的可能性;d_k 表示查询向量的维

度；( )^T 表示矩阵转置。

接着通过归一化指数函数 Softmax 计算每个注意力头的注意力权重，并对每个注意力头的内容向量进行加权求和，得到多头注意力的最终输出 FRT 和 FTR，计算过程如公式：

$$\begin{cases} F_i^{RT} = \text{Softmax}(A_i^R) \cdot V_i^T \\ F_i^{TR} = \text{Softmax}(A_i^T) \cdot V_i^R \end{cases} \tag{4}$$

最后，将所有注意力头的输出连接在一起，并通过一个线性变化得到从可见光到红外模态和从红外到可见光模态之间的全局关联 XRT 和 XTR，以此实现跨模态交互注意。如公式：

$$\begin{cases} X^{RT} = \text{Concat}(F_1^{RT}, F_2^{RT} \cdots, F_h^{RT}) \cdot W_{RT} \\ X^{TR} = \text{Concat}(F_1^{TR}, F_2^{TR} \cdots, F_h^{TR}) \cdot W_{TR} \end{cases} \tag{5}$$

其中，h 是注意力头的数量；W_{RT} 和 W_{TR} 是用于转换输出的权重矩阵。

2.4　门控单元

本文设计了一种门控单元，通过挤压和激励操作实现了跨模态通道注意力，从而建立了远程通道依赖，用于模态内部的相关性，控制两种模态的信息。门控单元的结构如图 4 所示。

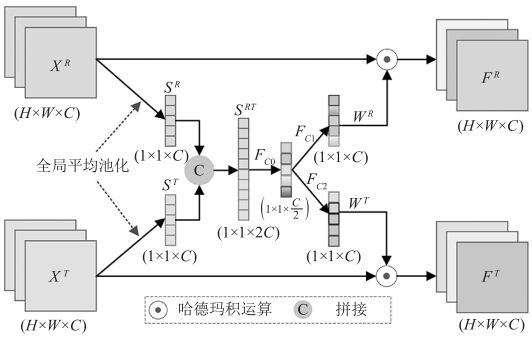

图 4　门控单元结构

Fig. 4 The structure of gate unit

首先，将输入可见光特征 X^R 和红外模态特征 X^T 经过全局平均池化(Global Average Pooling，GAP)操作，分别得到一维向量 S^R 和 S^T。如公式：

$$\begin{cases} S^R = \text{GAP}(X^R) = \dfrac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w} X^R(i,j,:) \\ S^T = GAP(X^T) = \dfrac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w} X^T(i,j,:) \end{cases} \tag{6}$$

其中，$(i,j,:)$ 表示特征图在位置 (i,j) 处的所有通道的数值；$\sum_{i=1}^{h}\sum_{j=1}^{w}$ 表示对整个特征图进行遍历并对每个位置的值进行累加；$\frac{1}{hw}$ 是缩放因子，用于计算平均值，其中 h 是高度；W 是宽度；hw 是总像素值。

然后，将得到的 S^R 和 S^T 拼接后将维度压缩为原来的四分之一从而得到融合后的特征 S^{RT}。如公式：

$$S^{RT} = F_{C0}(\text{Concat}(S^R, S^T)) \tag{7}$$

其中，F_{C0} 表示第一个全连接映射；Concat( )表示向量直接拼接操作；

然后通过两个独立的全连接层重新校准输入的模态特征，保持在同样维度大小，并预测每个模态通道的权重。如公式：

$$\begin{cases} W^R = F_{C1}(S^{RT}) \\ W^T = F_{C2}(S^{RT}) \end{cases} \tag{8}$$

其中，W^R 和 W^T 分别表示可见光和红外模态每个通道的权重；F_{C1} 和 F_{C2} 分别表示第二个和第三个全连接映射。

最后，计算对应的激励和抑制信号后再与原始输入的模态特征相乘，以此实现对可见光和红外模态的通道特征进行抑制或激活得到 F^R 和 F^T。具体计算如公式：

$$\begin{cases} F^R = 2 \times \sigma(W^R) \odot S^R \\ F^T = 2 \times \sigma(W^T) \odot S^T \end{cases} \tag{9}$$

其中，σ( )表示 Sigmoid 函数；$\odot$ 表示点乘操作。

2.5　实例分类层与损失函数

实例分类层由三个全连接层 FC4，FC5 和 FC6 组成，来区分每一帧的目标和背景，从而对目标位置和尺度进行预测。FC4，FC5 和 FC6 的输出通道数分别设置为 512，512 和 2。此外，为了防止过拟合，在每层前面添加了舍弃率为 0.5 的随机失活(dropout)。

使用常规二进制交叉熵(BCE)作为损失函数，首先计算正负样本的二值分类误差，然后将正负样本误差相加后除以样本总数进行归一化得到总损失，最终通过训练使总损失越来越小从而对模型进行优化。BCE 损失计算如公式：

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log_2 p_i + (1 - y_i)\log_2(1 - p_i) \tag{10}$$

其中，N 为样本个数；p_i 表示第 i 个样本预测为目标的概率值。y_i 表示样本的类别标签，其中正样本的标签为 1，负样本的标签为 0。

2.6 实验细节

在本文的方法中,分为离线训练阶段和在线跟踪阶段。为了公平,通过在数据集 RGBT234 和 GTOT 上进行交叉离线训练和在线跟踪验证。

在离线训练阶段,从每个序列随机选择 8 帧图像,从每帧中提取 32 个正样本和 96 个负样本用于网络的训练。RMFNet 采用随机梯度下降(SGD)算法,并分为两个阶段进行训练。在第一阶段,首先训练模态共享分支,使用 VGG-M 的预训练模型对其参数进行初始化。在此阶段,模态共享分支、分类层和其他部分的学习率分别被设置为 0.0001、0.001 和 0.0002,并进行 150 轮的训练。接着,仅加载第一阶段得到的共享模态分支的模型参数,并将其固定,对特定模态分支和模态特征交互模块进行额外的 150 轮训练。

在线跟踪阶段中,冻结了除实例分类层之外的所有网络参数,并在跟踪过程中初始化了新的 FC6 分支。根据第一帧的标签,裁剪了 500 个正样本和 5000 个负样本,用于微调 FC4 - 6,进行了 50 轮的微调训练。FC4 和 FC5 的学习率为 0.0004,FC6 的学习率为 0.001。在跟踪阶段,对于每一帧图像,提取了 256 个候选区域的样本集,用于预测下一帧目标的位置和尺度。从这些候选样本中选择置信度得分最高的前 5 个候选样本,取其平均值作为下一帧目标的跟踪结果。

实验环境配置如下:处理器 Intel i7 - 12700,显卡 Nvidia RTX 3060,内存 32 GB DDR4 3200 MHz,PyTorch 1.13,Python 3.8。

3 实验结果与分析

3.1 数据集与评估指标

本文在两个公开的 RGBT 数据集 GTOT[20] 和 RGBT234[21] 上进行实验验证。

GTOT 数据集[20]:GTOT 数据集包含 50 对 RGB 和热红外视频序列,并根据目标的状态划分了 7 类挑战属性(包括遮挡(OCC)、大尺度变化(LSV)、快速运动(FM)、低照度(LI)、热交叉(TC)、小目标(SO)和变形(DEF))。

RGBT234 数据集[21]:RGBT234 数据集包含 234 对 RGB 和热红外视频序列,并分为 12 个挑战属性(包括无遮挡(NO)、部分遮挡(PO)、重遮挡(HO)、低照度(LI)、低分辨率(LR)、热交叉(TC)、变形(DEF)、快速运动(FM)、尺度变化(SV)、运动模糊(MB)、像机偏移(CM)和背景干扰(BC))。

使用两个经典指标准确率(Precision Rate,PR)和成功率(Success Rate,SR)来验证跟踪器的有效性。

PR 是通过先计算所有帧中跟踪器预测的边界框与标签之间中心点的欧式距离,当距离在给定阈值的范围内则认为该帧被正确跟踪,再计算跟踪器在整个跟踪序列中正确跟踪目标的百分比得到。采用数据集公开标准,将 GTOT 的阈值设置为 5 像素,RGBT234 的阈值设置为 20 像素。其具体计算过程如公式:

$$PR = \frac{1}{M}\sum_{i=1}^{M} p_i; p_i = \begin{cases} 1, \Omega(bb, gt_i) \leqslant T_{\mathrm{PR}} \\ 0, \mathrm{other} \end{cases} \tag{11}$$

其中,M 表示总帧数;p_i 表示第 i 帧是否被正确跟踪;$\Omega(\quad)$ 表示计算两者中心点的欧氏距离;bb_i 表示预测边界框;gt_i 表示标签值;T_{PR} 表示计算 PR 的阈值。

SR 表示在整个跟踪序列中,预测的边界框与标签之间的重叠率大于给定阈值的帧所占百分比。其中,阈值统一设置为 0.6,计算过程如公式:

$$SR = \frac{1}{M}\sum_{i=1}^{M} S_i; S_i = \begin{cases} 1, IOU(bb, gt_i) \leqslant T_{\mathrm{SR}} \\ 0, \mathrm{other} \end{cases} \tag{12}$$

其中,S_i 表示第 i 帧是否被成功跟踪;$IOU(\quad)$ 表示计算的重叠率;T_{PR} 表示计算 SR 的阈值。

3.2 评估结果与分析

本文将 RMFNet 与 8 个先进的 RGBT 跟踪器 L1 - PF[7]、MDNet[11] + RGBT、MANet + +[13]、DAFNet[15]、MaCNet[16]、SiamCSR[26]、DAPNet[27] 和 M5l[28] 进行比较。如表 1 所示,RMFNet 在 GTOT 数据集上评估的 PR 和 SR 分别达到了 90.5 % 和 72.7 %,超过第二名 0.4 % 和 0.4 %,优于其他对比算法。

此外,RMFNet 在 RGBT234 数据集上的 PR 和 SR 分别为 80.1 % 和 57.5 %,超过第二名 0.5 % 和 1.6 %,也达到最佳指标。这些结果验证了多分支结构,以及所提出的可逆学习模块和模态特征交互模块的有效性。也说明了 RMFNet 可以很好的学习可见光与热红外模态的各自特征和互补特性,从而提高跟踪的准确性和鲁棒性。

表 1 GTOT 和 RGBT234 整体对比结果

Tab. 1 Overall comparison results of GTOT and RGBT234

算法	GTOT		RGBT234	
	PR/%	SR/%	PR/%	SR/%
LI-PF[7]	55.1	42.7	43.1	28.7
MDNet + RGBT[11]	80.0	63.7	72.2	49.5
M5L[28]	89.6	71.0	79.5	54.2
SiamCSR[26]	88.2	70.9	75.4	53.2
DAPNet[27]	88.2	70.7	76.6	53.7
MaCNet[16]	88.6	71.2	78.1	53.9
DAFNet[15]	89.1	71.2	79.6	54.4
MANet++[13]	90.1	72.3	79.5	55.9
RMFNet	**90.5**	**72.7**	**80.1**	**57.5**

为了更具体有效地评估跟踪器性能,进一步在不同的挑战性场景下进行比较。在 GTOT 数据集中的七种属性对比结果如表 2 所示,其中加下划线和加粗的数值为指标第一,只进行加粗的数值为指标第二。RMFNet 的 PR 指标在场景 OCC、FM、LSV、LI、TC 和 SO 下都位列第一,在场景 DEF 下位列第二。此外,RMFNet 的 SR 指标在场景 LSV、LI、TC、SO 和 DEF 下位列第一,在场景 FM 和 OCC 下位列第二。

在 RGBT234 数据集中的 12 类属性对比结果如表 3 所示。RMFNet 的 PR 指标在场景 HO、FM 和 SV 下都位列第一,在 LI 和 DEF 场景下位列第二。此外,RMFNet 的 SR 指标在场景 NO、PO、HO、LI、DEF、FM、SV 和 CM 下位列第一。

表 2 GTOT 数据集分属性对比结果

Tab. 2 Comparison results of sub-attributes on the GTOT dataset

属性	算法								
	LI-PF[7]	SiamCSR[26]	MDNet[11] + RGBT	M5l[28]	DAPNet[27]	MaCNet[16]	DAFNet[15]	MANet++[13]	**RMFNet**
OCC	61.2/48.0	86.8/67.1	82.9/64.1	**89.2**/68.4	87.3/68.4	87.6/68.7	87.3/68.4	89.0/**70.1**	**89.7**/**69.9**
LSV	68.7/52.9	87.6/67.1	77.0/57.3	85.2/66.9	84.7/64.8	84.6/67.3	82.2/66.4	**86.6**/**69.3**	**89.5**/**69.4**
FM	45.5/39.7	82.4/64.1	80.5/59.8	85.4/64.7	82.3/61.9	82.3/65.9	80.9/64.2	**86.7**/**70.3**	**87.1**/**69.2**
LI	55.3/39.7	85.9/69.9	79.5/60.3	90.5/71.9	90.0/72.2	89.4/73.1	89.9/72.7	**91.7**/**73.1**	**92.0**/**73.7**
TC	59.2/43.9	87.7/68.6	79.5/60.9	89.7/70.2	89.3/69.0	89.2/69.7	89.8/70.3	**89.9**/**70.7**	**90.6**/**71.2**
SO	56.5/41.7	88.7/66.9	87.0/62.2	93.5/69.4	93.7/69.2	**95.0**/69.5	93.9/69.8	93.9/**69.9**	**95.8**/**70.8**
DEF	49.2/34.0	86.4/71.0	81.6/68.8	92.6/77.1	91.9/77.1	92.6/76.5	**94.7**/76.5	94.0/**77.3**	**94.3**/**77.3**
ALL	55.1/42.7	88.2/70.9	80.0/63.7	89.6/71.0	88.2/70.7	88.6/71.2	89.1/71.2	90.1/72.3	**90.5**/**72.7**

表 3 RGBT234 数据集分属性对比结果

Tab. 3 Comparison results of sub-3attributes on the RGBT234 dataset

属性	算法								
	LI-PF[7]	SiamCSR[26]	MDNet[11] + RGBT	M5l[28]	DAPNet[27]	MaCNet[16]	DAFNet[15]	MANet++[13]	**RMFNet**
NO	56.5/37.9	87.7/55.5	86.2/61.1	**90.4**/64.6	90.0/64.4	89.9/64.6	90.0/63.6	**90.2**/**66.4**	89.9/**66.5**
PO	47.5/31.4	77.9/51.3	76.1/51.8	82.1/58.9	82.1/57.5	82.7/58.7	**85.9**/58.8	83.0/**59.1**	**84.6**/**60.7**
HO	33.2/22.2	59.2/39.4	61.9/42.1	66.5/45.0	66.0/45.7	**71.0**/**48.6**	68.6/45.9	**71.1**/48.2	**71.1**/**48.9**
LI	40.1/26.0	70.5/46.2	67.0/45.5	**82.1**/**54.7**	77.5/53.0	78.7/52.6	81.2/54.2	79.6/54.3	**81.1**/**55.6**
LR	46.9/27.4	75.1/47.6	75.9/51.5	**82.3**/53.5	75.0/51.0	79.9/**54.5**	**81.8**/**53.8**	78.9/52.0	79.2/53.5
TC	37.5/23.8	76.0/47.0	75.6/51.7	**82.1**/56.4	76.8/54.5	78.4/**57.0**	**81.1**/**58.3**	77.2/55.8	78.3/56.1
DEF	36.4/24.4	68.5/47.4	66.8/47.3	73.6/51.1	71.7/51.9	73.6/52.1	74.1/51.6	**78.7**/**56.2**	**76.1**/**56.3**
FM	32.0/19.6	67.7/40.2	58.6/36.3	**72.8**/46.5	67.0/44.5	68.4/44.1	**74.0**/**46.5**	69.4/45.5	**74.0**/**49.1**
SV	45.5/30.6	69.2/43.4	73.5/50.5	**79.6**/54.2	78.0/54.2	79.0/56.0	79.1/54.4	79.2/**57.3**	**80.6**/**58.8**
MB	28.6/20.6	64.7/43.6	65.4/46.3	**73.8**/**52.0**	65.3/46.7	71.5/**52.4**	70.8/50.0	73.3/51.8	70.9/51.1
CM	31.6/22.5	66.7/45.2	64.0/45.4	**75.2**/**52.9**	66.8/47.4	70.2/50.4	72.3/50.6	**74.1**/52.0	72.7/**53.1**
BC	34.2/22.0	65.8/41.8	64.4/43.2	75.0/47.7	71.7/48.5	77.6/50.9	**79.1**/**49.3**	**76.3**/**49.2**	74.6/49.1
ALL	43.1/28.7	75.4/53.2	72.2/49.5	79.5/54.2	76.6/53.7	78.1/53.9	**79.6**/54.4	79.5/55.9	**80.1**/**57.5**

综合两个数据集上的验证结果可以看出本文的跟踪器 RMFNet 在大多数情况下都能达到较好性能。特别在面对遮挡,尺度变化,极端光照和相机抖动场景时,实现了最优的跟踪效果。这得益于两个可逆学习模块通过多层特征融合学习提取到了更加丰富的细节特征以及多尺度信息,增强目标定位,从而能更好应对遮挡,尺度变化和相机抖动挑战。此外,模态特征交互模块自适应融合可见光和热红外特征,使跟踪器在极端光照场景下对目标更加敏感,实现稳定跟踪。

3.3 可视化结果与分析

为了直观分析跟踪器的性能,将本文提出的 RMFNet 与 L1 – PF[7]、MDNet[11] + RGBT、MANet ++[13]、M5L[28]在四种具有挑战性的场景下的部分序列进行定性比较,并将结果可视化。

从图 5(a)可以看出,在高曝光场景下,除了 RMFNet 外的其他算法都会丢失跟踪目标。在图 5(b)中,当镜头抖动产生模糊和偏移,且目标较小时,只有 RMFNet 较为精准的跟踪到了目标。如图 5(c)所示,面对目标尺度发生变化时,RMFNet 可以及时对调整目标框做出调整。根据图 5(d)可以观察到,当目标被部分遮挡时,RMFNet 始终可以做到连续准确跟踪,很好地应对这一挑战。实验证明 RMFNet 在高曝光,相机抖动,尺度变化,部分遮挡等情况下具有较强的抗干扰能力。

(a)高曝光

(b)相机抖动

(c)尺度变化

(d) 部分遮挡

RMFNet L1-PF[7] MDNet[11]+RGBT MANet++[13] M51[28]

图 5 可视化对比结果

Fig. 5 The result of visual comparison

3.4 消融实验结果与分析

为了验证跟踪器所提模块的有效性,本文对完整 RMFNet 的三分支结构特征提取模块和模态特征交互模块进行裁剪构造了 RMFNet-noRLM 和 RMFNet-noMFIM 两个变体网络进行消融实验以及和 Baseline 进行比较。其中,RMFNet-noRLM 是对两个模态特定分支裁剪得到,RMFNet-noMFIM 是将模态特征交互模块进行裁剪,Baseline 是将本文设计模块都去除后的基础模型。将上述三个网络在 RGBT234 数据集上训练后并在 GTOT 数据集上验证。

如图 6 所示,RMFNet-noRLM 相对于 RMFNet 的 PR 和 SR 指标分别下降了 1.7 % 和 1.8 %,证明两个模态特定分支在提升跟踪器的准确性和稳定性方面是有用的,能很好提取不同模态特定信息增强特征学习。RMFNet-noMFIM 在 PR 和 SR 两个指标上比 RMFNet 低了 1.3 % 和 1.6 %,这验证了所提出的模态特征交互模块的有效性,这在很大程度上得益于其能够充分挖掘到可见光和红外模态的互补信息。此外,RMFNet 相较于基础模型在准确率和成功率上分别提高了 6.3 % 和 5.0 %,它结合了以上两种变体网络的优点,充分学习两种模态信息的同时还能挖掘两者的差异性,从而提高了跟踪的准确性和鲁棒性。

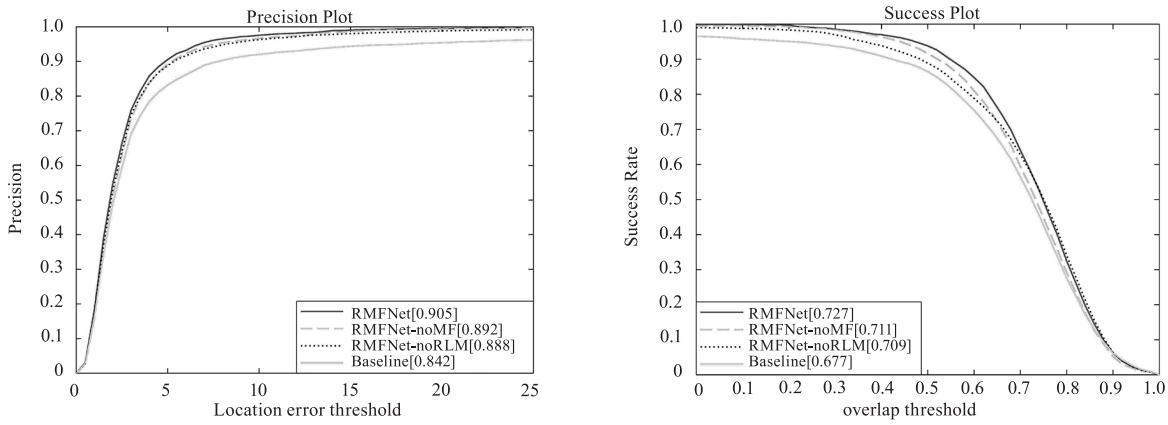

图 6 消融实验结果

Fig. 6 The results of the ablation experiment

4 结束语

本文提出了一种可逆多分支的双模态融合目标跟踪网络,它由多模态特征提取分支、模态特征交互模块和实例分类层组成。多分支结构的设计巧妙地将聚合分支和单模态分支结合,有助于充分学习可见光和红外模态特征。模态特征交互模块则实现了自适应融合两种模态的特征,有效地挖掘模态之间的互补信息。最后,通过实例分类层将可靠的融合特征用于实现鲁

棒的跟踪。在两个 RGBT 数据集上的实验结果证明了本文方法的出色跟踪效果。

参考文献:

[1] 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.

[2] Zhang Fangfang, Cao Jiahui, Wang Haijing, et al. Anti-occlusion moving target tracking algorithm based on multifeatured self-adaptive fusion[J] Infrared Technology, 2023, 45(2): 150 – 160. (in Chinese)
张方方, 曹家晖, 王海静, 等. 基于多特征自适应融合的抗遮挡目标跟踪算法[J]. 红外技术, 2023, 45(2): 150 – 160.

[3] Wu Jie, Ma Xiaohu. UAV infrared target tracking algorithm based on feature fusion and channel awareness[J]. Laser & Infrared, 2023, 53(4): 626 – 632. (in Chinese)
吴捷, 马小虎. 基于特征融合与通道感知的无人机红外目标跟踪算法[J]. 激光与红外, 2023, 53(4): 626 – 632.

[4] Li Y D, Lai H C, Wang L J, et al. Multibranch adaptive fusionnetwork for RGBT tracking[J]. IEEE Sensors Journal, 2022, 22(7): 7084 – 7093.

[5] Yan K X, Wang C C, Zhou D M, et al. RGBT tracking viamulti-stage matching guidance and context integration[J]. Neural Processing Letters, 2023: 55(8): 11073 – 11087.

[8] Li C, Xue W, Jia Y, et al. LasHeR: a large-scale high-diversity benchmark for RGBT tracking[J]. IEEE Transactions on Image Processing, 2022, 31: 392 – 404.

[9] Wu Y, Blasch E, Chen G S, et al. Multiple source data fusionvia sparse representation for robust visual tracking[C]//14th International Conference on Information Fusion. IEEE, 2011: 1 – 8.

[10] Lanx A Y, YE M, Zhang S P, et al. Modality-correlation-aware sparse representation for RGB-infrared object tracking[J]. Pattern Recognition Letters, 2020, 130: 12 – 20.

[11] Huang Y P, Li X F, Lu R T, et al. RGB-T object tracking via sparse response-consistency discriminative correlation filters [J]. Infrared Physics & Technology, 2023, 128: 104509.

[12] Li C L, Wu X, Zhao N, et al. Fusing two-stream convolutionalneural networks for RGB-T object tracking[J]. Neurocomputing, 2018, 281: 78 – 85.

[13] Nam H, Hamb. Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, 2016: 4293 – 4302.

[14] 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 Workshops, 2019.

[15] 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.

[16] Yan K X, Mei J T, Zhou D M, et al. External-attention dual-modality fusion network for RGBT tracking[J]. The Journal of Supercomputing, 2023, (15): 79.

[17] Gao Y, Li C L, Zhu Y B, et al. Deep adaptive fusion networkfor high performance RGBT tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[18] Zhan 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.

[19] Hou R C, Ren T W, Wu G S. MIRNet: A robust RGBT tracking jointly with multi-modal interaction and refinement[C]//2022 IEEE International Conference on Multimedia and Expo(ICME). IEEE, 2022: 1 – 6.

[20] Wang X, Shu X J, Zhang S, et al. MFGNet: Dynamic modality-aware filter generation for RGB-T tracking[J]. IEEE Transactions on Multimedia, 2023, 4335 – 4348.

[21] Zhao Z X, Bai H W, Zhang J S, et al. CDDFuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 5906 – 5916.

[22] 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 – 56.

[23] Li C L, Liang X Y, Lu Y J, et al. RGB-T object tracking: benchmark and baseline[J]. Pattern Recognition, 2019, 96: 106977.

[24] Deng J, Dong W, Socher R, et al. Imagenet: a large-scale-hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 248 – 255.

[25] Cai Y X, Zhou Y Z, Han Q, et al. Reversible column networks[C]//Proceedings of the International Conference on Learning Representations, 2023.

[26] He K M, Zhang X Y, Ren S Q, et al. Deep residual learningfor image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770 – 778.

[27] Liu Z, Mao H Z, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 11976 – 11986.

[28] 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, IEEE, 2022: 803 – 806.