文章编号　1004-924X(2024)03-0435-10

# 基于注意力交互的可见光红外跟踪算法

王　暐，付飞亚，雷　灏，唐自力*

(中国人民解放军63870部队，陕西 渭南 714299)

**摘要**：在可见光红外跟踪(RGB and Thermal Infrared Tracking，RGB-T)的研究中，为了在常规跟踪算法的基础上实现两个模态的有效融合，基于注意力机制提出了一种基于注意力交互的RGB-T跟踪算法。该算法引入注意力机制对可见光和红外两种模态的图像特征进行增强和融合，设计了自特征增强编码器对单一模态的特征进行增强，设计了互特征解码器对两个模态增强后的特征进行交互融合。编码器和解码器均采用两层注意力模块。为了减小算法模型的复杂度，对传统注意力模块进行简化，将全连接层改为1×1卷积。此外，该算法对多个卷积层的特征均进行分层融合，以充分挖掘各层卷积特征中的细节和语义信息。在GTOT，RGBT234和LasHeR三个数据集上进行对比测试。实验结果表明，所提算法性能优异，特别是在RGBT234和LasHeR这两个大规模数据集上取得了最优的跟踪结果，验证了注意力机制在RGB-T跟踪中的有效性。

**关　键　词**：可见光红外跟踪；注意力机制；多模态特征融合；特征增强

**中图分类号**：TP391.4　　**文献标识码**：A　　**doi**：10.37188/OPE.20243203.0435

# Attention interaction based RGB-T tracking method

WANG Wei，FU Feiya，LEI Hao，TANG Zili*

(*The* 63870 *Unit of PLA*，*Weinan* 714299，*China*)

* *Corresponding author*，*E-mail*：*tang_zili@qq.com*

**Abstract**：In visible and thermal infrared tracking(RGB-T)，to effectively merge these two modalities building on traditional tracking techniques，this study introduces an attention-based RGB-T tracking approach based on the attention mechanism. This method employs the attention mechanism to augment and integrate features from both visible and infrared images. It features a self-feature enhancement encoder to boost single modality features，and a cross-feature interaction decoder for merging the enhanced features from both modalities. Both the encoder and decoder incorporate dual layers of attention modules. To streamline the network，the traditional attention module is simplified by substituting fully connected layers with 1×1 convolutions. Moreover，it merges features from various convolutional layers to thoroughly explore details and semantic insights. Comparative experiments on three datasets—GTOT，RGBT234，and LasHeR—demonstrate that our method achieves superior tracking performance，underscoring the efficacy of the attention mechanism in RGB-T tracking.

**Key words**：RGB-T tracking；attention mechanism；feature fuse of multi-modality；feature enhancement

**收稿日期**：2023-07-19；**修订日期**：2023-09-04.

**基金项目**：XX识别性能评估技术研究

## 1 引　言

作为计算机视觉的一个重要研究方向，视觉跟踪长期以可见光图像为研究对象[1]，但可见光图像在光照较差及雨、雾、霾等气候条件下的成像效果不理想，导致跟踪算法性能下降[2]。热红外相机在上述恶劣环境中的成像质量更高[1-2]，因此，将可见光和热红外图像进行融合，利用二者信息的互补能够实现更稳定的跟踪。目前，可见光红外跟踪（RGB and Thermal infrared tracking，RGB-T）已成为一个新的研究热点，在无人驾驶、监控、军事等领域具有广泛的应用前景[2-3]。

RGB-T跟踪在传统可见光跟踪算法的基础上进行多模态扩展。Li等[4]较早进行了RGB-T的跟踪研究，在基于稀疏表示的跟踪框架下，将单模特扩展为多模态，引入自适应模态加权系数。Zhang等[5]通过计算不同模态的融合权重，探索了后融合在RGB-T跟踪中的作用，同时引入运动估计提高跟踪性能。

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。Li等[6]利用两个结构相同、参数经过微调的卷积神经网络分别提取可见光和红外图像的特征，进行在线特征选择，然后在核相关滤波[7]框架下跟踪。Zhang等[8]基于经典跟踪算法DiMP[9]，全面考察了像素级、特征级和决策级的融合策略，指出特征级融合对RGB-T跟踪的性能提升最为显著，但该方法使用了大量额外的训练数据。

基于Siamese网络的跟踪[10-11]是可见光跟踪算法中的主流，具有优异的跟踪性能。Zhang等[12]在SiamRPN＋＋[11]的基础上，提出一种特征互补和干扰物识别的RGB-T跟踪算法，设计专用的多模态特征融合模块对可见光特征和红外特征进行交互融合。但这类方法对RGB-T跟踪的性能提升并不明显，通常需要采用较多的额外数据对网络进行训练才能达到RGB-T跟踪的基线水平，同时两个模态各有两路卷积网络，使得模型较为繁琐复杂。

MDNet[13]是一种相对较小的跟踪网络（参数量约为4.4M），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。文献[14]提出了一种MANet算法，将MDNet中的卷积层定义为模态共享适配器，针对红外和可见光图像设计模态适配器，将MDNet中的全连接层定义为目标适配器，同时考虑了模态间共享特征、模态专有特征以及目标特征。ADRNet[15]，CAT[16]，APFNet[17]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。Wang等[18]利用相关性对MDNet中的卷积特征进行增强，建模得到了模态内、模态间和图像帧间的相关关系，建立了红外和可见光图像特征的相互作用机制，取得较好的跟踪结果。在该研究的基础上，Xu等[19]提出跨模态交互学习框架，将两个模态的特征进行逐像素的相关，尝试挖掘不同模态间更直接的相关关系，提升模态间的信息传播。

上述研究的焦点基本都在可见光特征与红外特征的有效融合上，融合方式主要以加权和串接为主。可见光和红外图像以相同的视角获取，不仅具有模态的差异性，而且在图像结构和内容上也具有共性，模态间的特征融合应同时考虑这种差异性和共性。文献[18-19]通过直观的相关操作对这种共性进行了探索和增强，但没有进行更深入的挖掘。近年来，注意力机制在自然语言处理和计算机视觉领域大放异彩[20-22]，特别是Transformer[20]。利用Transformer的编码器能够对特征进行增强，而解码器模块能够实现两种模态特征自然且充分的交互。

本文采用transformer的编码器和解码器对两个模态的特征进行有效融合，增强可见光和红外特征，并在融合后的特征中保留两个模态的互补和共性信息，最终提升RGB-T跟踪的性能。

## 2　原　理

### 2.1　基于 MDNet 的 RGB-T 跟踪算法框架

MDNet[13]是一种采用离线预训练和在线更新结合的视觉跟踪算法，其网络由三个卷积层（CNN）和三个全连接层（FC）组成。在 RGB-T 跟踪任务中，针对可见光和红外图像分别设置一路卷积，并将卷积特征进行融合，即可将 MDNet 扩展至多模态跟踪。其算法框架如图 1 所示，包括双路卷积层、融合网络和全连接层等三部分。

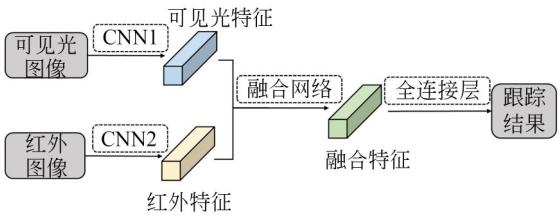



图 1　基于 MDNet 的 RGB-T 跟踪算法

Fig. 1　Basis framework of MDNet-based RGB-T tracking method

网络的输入为候选图像块，输出为该图像块为目标的概率，每一帧的跟踪结果为概率最大的图像块。MDNet 的卷积网络已经针对可见光跟踪进行了充分的训练，能够提取相应的卷积特征进行跟踪。在跟踪前仅需要对卷积层进行微调，并对融合网络和全连接层进行初始化和训练；在线跟踪过程中，仅对全连接层进行微调和训练。

### 2.2　Transformer

Transformer 首先在文献[20]中被提出，并应用于机器翻译任务。Transformer 由多个注意力模块相互串联构成，每个注意力模块的输入为整个句子，具有全局的表达能力。最初的 Transformer 中采用两种注意力模块，编码器和解码器的结构如图 2 所示。编码器中，输入序列 $z^{l-1}$ 首先映射生成 $q,k,v$，三者进行注意力（Attention）运算：

$$\hat{z}^l = softmax\left(\frac{qk^{\mathrm{T}}}{\sqrt{d}}\right)v. \tag{1}$$

该式通常采用矩阵化运算，$d$ 为 $k$ 的维度。对 $\hat{z}^l$ 进行残差操作（Add），经过映射和一个多层感知机（Multilayer Perception，MLP），即为编码器的输出 $z^l$。

解码器和编码器的结构基本相同，区别为解码器多进行一个自注意力操作，并在第二个注意力操作中将 $k$ 和 $v$ 替换为外部输入 $x$ 的映射。文献[20]中还采用位置编码、多头注意力等设置，本文不再详述。

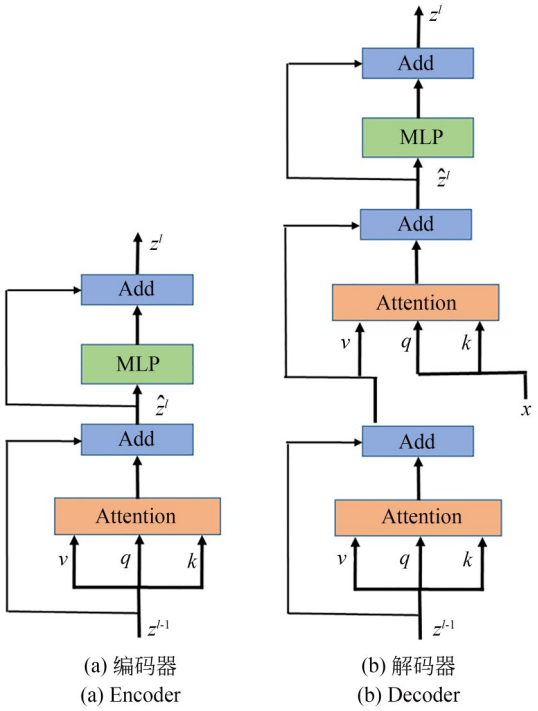



图 2　Transformer 结构

Fig. 2　Structure of transformer

Transformer 结构目前已应用在图像分类[21]、目标检测[23]和目标跟踪[24]等计算机视觉任务中，并取得了较好的效果，验证了编码器-解码器的注意力机制对视觉任务的有效性。APFNet[17]在 RGB-T 跟踪中首先引入 Transformer，但仅作为其融合网络的众多模块中的一个，且只设计两个编码器和一个解码器对特征进行增强。消融实验表明，该 Transformer 模块的引入并未大幅提升 RGB-T 跟踪性能。与 APFNet 不同，本文设计了一个完全由 Transformer 结构组成的融合网络，用以验证 Transformer 的注意力结构对多模态特征增强和融合的有效性。

## 3　本文算法

本文对可见光和红外两个模态的特征融合

进行深入挖掘，在基于MDNet的RGB-T跟踪框架下，提出一种基于注意力交互的RGB-T跟踪算法（Attention Interaction based RGB-T Tracking method，AIT）。AIT的网络结构如图3所示，包括卷积网络、全连接网络和融合网络，其中前两个网络与MDNet中相同。融合网络的核心是特征增强和交互模块（Feature Enhance and Interaction module，FEI），它利用自注意力特征增强编码器（Self-feature Enhance Encoder，SEE）对红外和可见光图像特征进行增强，利用互注意力特征交互解码器（Cross-feature Interaction Decoder，CID）对两个模态的图像进行交互融合。为了充分利用多层卷积特征，融合网络对每层卷积特征均采用FEI进行处理，并利用FEI对处理后的融合特征进行进一步的融合和增强，以有效提高跟踪算法的性能。

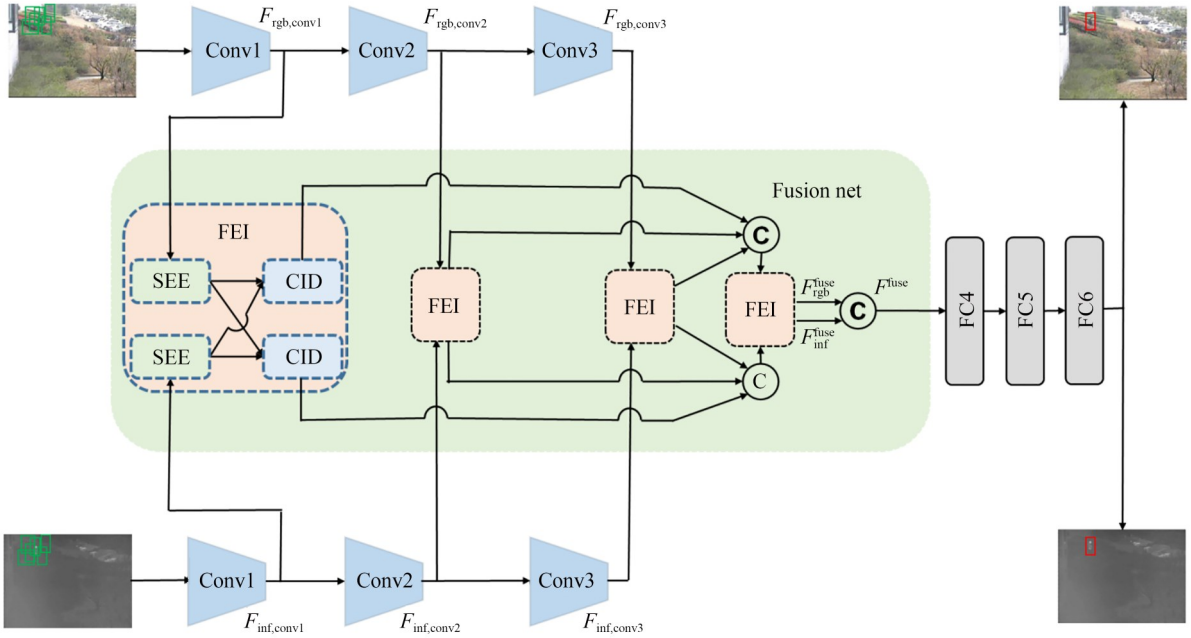



图3　基于自注意力交互的RGB-T跟踪算法总体结构

Fig. 3　Framework of attention interaction based RGB-T tracking method

### 3.1　特征增强和交互模块

原始的Transformer结构由编码器和解码器串联组成。为了实现更有效的特征增强和交互融合，这里采用相同的结构，利用两层编码器实现模态内的特征增强、两层编码器实现跨模态的特征交互。特征增强和交互模块的结构如图4所示，由两个自注意力特征增强编码器（SEE）和两个互注意力特征交互解码器（CID）组成，SEE和CID的细节分别如图5(a)和5(b)所示。

SEE的输入为单一模态的卷积特征，与原始Transformer不同，本文没有对特征进行升维和降维处理。首先直接将输入特征$F_{rgb}$（以可见光模态为例，红外模态处理方式相同）按照卷积的空间维度进行变换，设$F_{rgb}$的维度为$H\times W\times C$，$H$和$W$分别为特征的空间高宽，$C$为卷积特征通道数，则索引$\boldsymbol{Q}$、键$\boldsymbol{K}$、值$\boldsymbol{V}$的维度为$HW\times C$，表示有$HW$个向量，每个向量维度是$C$，且$\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}$。然后进行自注意力运算，得到：

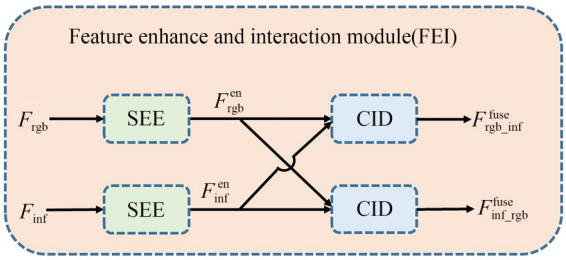



图4　特征增强和交互模块结构

Fig. 4　Structure of FEI module

$$F_{rgb}^{1}=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{C}}\right)\boldsymbol{V}. \tag{2}$$

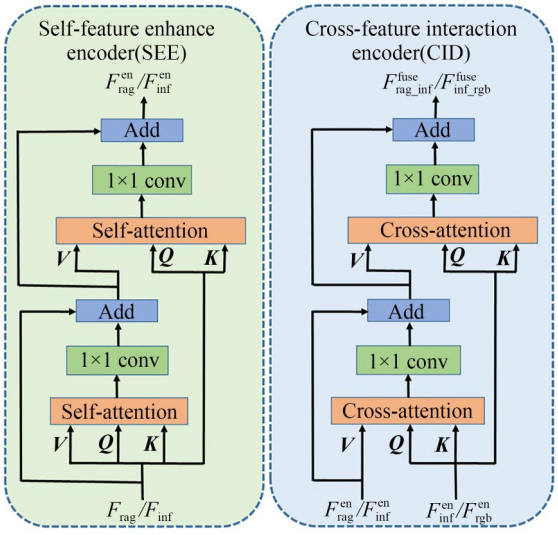
$F^{1}_{\mathrm{rgb}}$ 为经过自注意力增强的特征。式(2)的自注意力运算是输入值的直接计算操作，没有需要优化的参数。在原始 Transformer 中，对其进行残差和全连接层的处理，为了减少网络的复杂度、提高跟踪的实时性，将 $F^{1}_{\mathrm{rgb}}$ 的维度恢复到三维，记为 $F^{2}_{\mathrm{rgb}}$，并用 1×1 卷积层替换全连接层，然后进行残差操作，得到 SEE 第一层编码器的输出：

$$F^{\mathrm{en\_1}}_{\mathrm{rgb}}=\mathrm{Res}\left(F_{\mathrm{rgb}},Conv_{1\times1}\left(F^{2}_{\mathrm{rgb}}\right)\right),\tag{3}$$

其中：$Conv_{1\times1}$ 表示 1×1 卷积，Res 表示残差连接。

SEE 第二层编码器结构与第一层相同，但输入中，索引 $\boldsymbol{Q}$、键值 $\boldsymbol{K}$ 由最初的特征 $F_{\mathrm{rgb}}$ 维度变换得到，值 $\boldsymbol{V}$ 由第一层增强的特征 $F^{\mathrm{en\_1}}_{\mathrm{rgb}}$ 维度变换得到，经第二层编码器增强后的特征表示为 $F^{\mathrm{en}}_{\mathrm{rgb}}$。上述操作表述为：

$$F^{\mathrm{en}}_{\mathrm{rgb}}=SEE\left(F_{\mathrm{rgb}}\right),F^{\mathrm{en}}_{\mathrm{inf}}=SEE\left(F_{\mathrm{inf}}\right).\tag{4}$$




(a) SEE　　(b) CID

图 5　自注意力增强编码器和互注意力交互解码器构图

Fig. 5　Structure of Self-feature Enhanced Encoder (SEE) and Cross-feature Interaction Decoder (CID)

CID 在结构上与 SEE 类似。为了实现可见光和红外两个模态的交互融合，利用注意力机制进行两个模态特征的相互调节。以红外特征对可见光特征进行调节为例，CID 两层解码器的索引 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 均来自 SEE 增强后的红外特征 $F^{\mathrm{en}}_{\mathrm{inf}}$，第一层解码器的值 $\boldsymbol{V}$ 由 SEE 增强后的可见光特征 $F^{\mathrm{en}}_{\mathrm{rgb}}$ 维度变换得到，第二层解码器的值 $\boldsymbol{V}$ 来自第一层解码器调节交互后的特征。将上述操作表述为：

$$F^{\mathrm{fuse}}_{\mathrm{rgb\_inf}}=CID\left(F^{\mathrm{en}}_{\mathrm{rgb}},F^{\mathrm{en}}_{\mathrm{inf}}\right),\tag{5}$$

$$F^{\mathrm{fuse}}_{\mathrm{inf\_rgb}}=CID\left(F^{\mathrm{en}}_{\mathrm{inf}},F^{\mathrm{en}}_{\mathrm{rgb}}\right).\tag{6}$$

### 3.2　多层卷积特征

为了充分利用低层到高层的所有卷积特征，对 3 个卷积层的特征都进行了特征增强和交互，得到两个模态分层增强和交互后的特征 $\left\{F^{\mathrm{fuse}}_{\mathrm{rgb\_inf,convi}},F^{\mathrm{fuse}}_{\mathrm{inf\_rgb,convi}}|i=1,2,3\right\}$。为了进一步在层间进行特征的增强和融合，对同一模态的多层融合特征进行维度变换并串接，得到 $F^{\mathrm{fuse}}_{\mathrm{rgb\_inf}}$ 和 $F^{\mathrm{fuse}}_{\mathrm{inf\_rgb}}$，如图 3(b)所示。将得到的两个特征作为 FEI 的输入，进行多层特征总体的增强和融合，最后将 FEI 的两个输出特征串接，得到两个模态的融合特征 $F^{\mathrm{fuse}}$。

### 3.3　离线训练和在线跟踪

所提网络的离线训练需要考虑：(1)两路卷积层需要分别学习可见光和红外图像的特征；(2)全连接层重新初始化，以适应融合特征的输入；(3)多层的特征增强和交互模块具备特征增强和跨模特的交互能力。在 APFNet 等算法中，研究人员采用多阶段离线训练方式逐个模块进行训练，但训练过程较为繁琐。本文提出对各模块进行联合训练，同步优化卷积层、全连接层和融合模块。

通过分析，卷积层和全连接层的参数量与 MDNet 基本相同(卷积层参数量为 1.8M+1.8M，全链接层参数量为 5M)。在 FEI 结构下，解码器和编码器中采用传统的 MLP，融合网络的参数量将超过 100M，对其进行完全初始化的训练不利于网络的收敛，因此，本文提出将解码器和编码器中的 MLP 改为 1×1 卷积，将融合网络的参数量降至 6M，以实现联合训练。联合训练的学习率设置为：卷积层 0.000 01，全连接层 0.001，融合网络 0.000 1，训练的迭代次数为 1 500。

在线跟踪阶段，在第一帧初始化 FC6 层，微调 FC4 和 FC5 层，后续帧也仅对这 3 个全连接层进行微调，FC4 和 FC5 的学习率为 0.000 1，FC6 层学习率为 0.001，网络的其他层参数固定。其余细节请参考 MDNet[13] 和 APFNet[14]。

### 3.4　算法跟踪流程

综上，AIT 算法的在线跟踪流程如下。

初始化：人工或检测算法给定第一帧的目标状态，离线训练的网络参数，随机初始化FC6层
for $t=1$ to $T$（T为视频序列总帧数）
（第二帧开始由所提AIT确定目标状态）
if $t>1$ then
步骤1　根据上一帧得到的目标状态，采样候选图像块；
步骤2　利用卷积网络提取候选图像块的多层卷积特征；
步骤3　利用2.1节的EFI对各层卷积网络进行特征交互和增强，增强后的各层特征再经过一次EFI得到最终的融合特征；
步骤4　利用全连接层计算每个候选图像块为目标的概率，对概率最大的图像块做边框回归，并将其作为当前帧的跟踪结果；
（在每一帧中初始化或者更新网络参数）
步骤5　根据当前帧的目标状态，采样正负训练样本，并添加至正负样本集，如果样本集内样本数量超过预设值，则剔除帧数最小的图像中采集的样本；
步骤6　每间隔10帧，利用训练样本集对全连接层参数进行微调训练。

## 4　实验结果分析

实验硬件平台配置如下：CPU为Intel® Xeon E5-2630 v4，内存为32 GB，GPU为NVIDIA GeForce RTX 1080Ti，操作系统为Ubuntu 22.04 LTS，使用pytorch 1.0.1，python 3.7，CUDA 10.2 a环境。

### 4.1　数据集及测评方法

GTOT数据集[24]：该数据集包含50个可见光-红外视频序列，共有15 000帧图像，采集场景有道路、行人密集区域等，标注了尺度变化、快速运动、目标形变、热交叠、小目标、遮挡和低光照等7类对跟踪具有挑战性的标签。

RGBT234数据集[1]：这是目前应用最普遍的RGB-T跟踪数据集，包含234个可见光-红外视频序列，总帧数超过234 000帧，标注了12类属性标签和遮挡水平等。

LasHeR数据集[3]：该数据是目前最大的RGB-T跟踪数据集，包含1 224个可见光-红外视频序列，其中979个序列被划分为训练集，245个被划分为测试集。

本文采用精确率（Precision Rate，PR）和成功率（Success Rate，SR）作为各数据集的评价指标。对某个数据集，精确率定义为跟踪算法输出的目标位置与标注位置的像素距离小于或等于给定阈值的帧数同该数据集总帧数的百分比。通过给定不同的阈值得到算法对数据集的PR曲线。为了更直观地比较，在PR曲线中取一个典型PR值对算法进行排名。GTOT数据集中图像的分辨率较低，目标的像素尺寸小，取阈值为5对应的精确率为典型值，RGBT234和LasHeR数据集中，取阈值为20对应的精确率为典型值。此外，定义成功率曲线，曲线的横坐标为阈值，阈值范围为[0, 1]，曲线纵坐标值为跟踪算法输出的边界框与标注边界框之间的交并比（Intersection of Union，IoU）大于该阈值的帧数同数据集总帧数的百分比，成功率的典型值定义为该成功率曲线下的面积。后续分析中的具体数值对比均采用上述典型值。

为了展示所提算法的有效性，实验选择现有的主流RGB-T跟踪算法作为对比算法，包括APFNet[17]，FANet[25]，MANet[14]，HDINet[26]，M5L[27]，CMPP[18]，JMMAC[5]，CAT[16]，DAFNet[28]，DAPNet[29]，MaCNet[30]，以及基础MDNet在RGB-T跟踪任务上简单处理扩展后的MDNet+RGBT。其中，APFNet和CMPP具有最好的跟踪性能。

### 4.2　GTOT数据集测评

采用RGBT234数据集对所提算法进行离线训练，然后测评GTOT数据集，结果如图6所示。相较于基础的MDNet+RGBT算法，AIT的跟踪精确率和成功率分别提升了11.6%和10.1%。图6中，多数算法在GTOT数据集上具有较高的跟踪性能，趋近饱和，表明该数据在规模和难度上已经不足以充分评估RGB-T跟踪算法。

### 4.3　RGBT234数据集测评

采用GTOT对所提AIT算法进行离线训练，图7为所提算法与对比算法在RGBT234数据集上的跟踪结果。如图7所示，AIT算法取得最优的跟踪结果，与原始MDNet+RGBT相比，AIT在精确率和成功率上分别提升了11.4%和8.6%，说明特征的交互融合对RGB-T跟踪的重

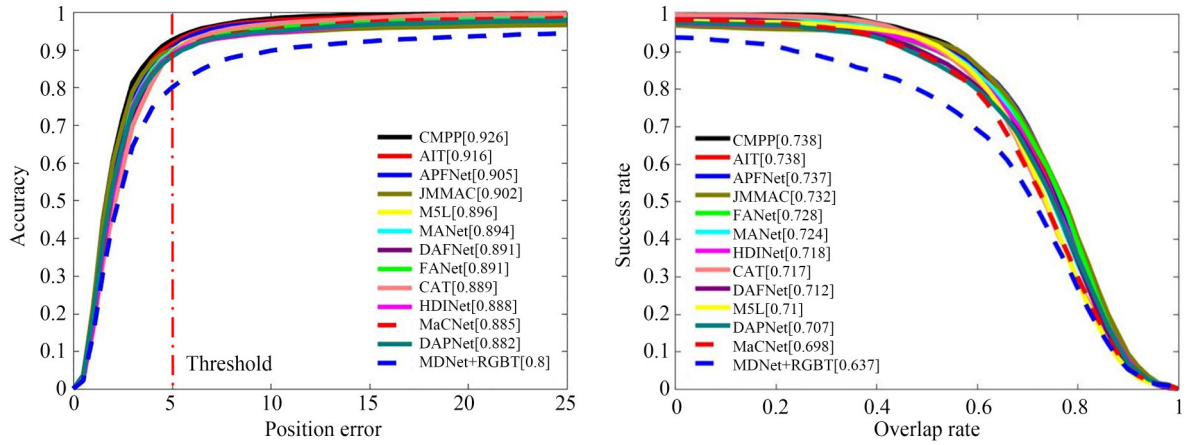



图 6　AIT 及对比算法在 GTOT 数据集上的跟踪结果

Fig. 6　Evaluation results of AIT and compared algorithms on GTOT dataset

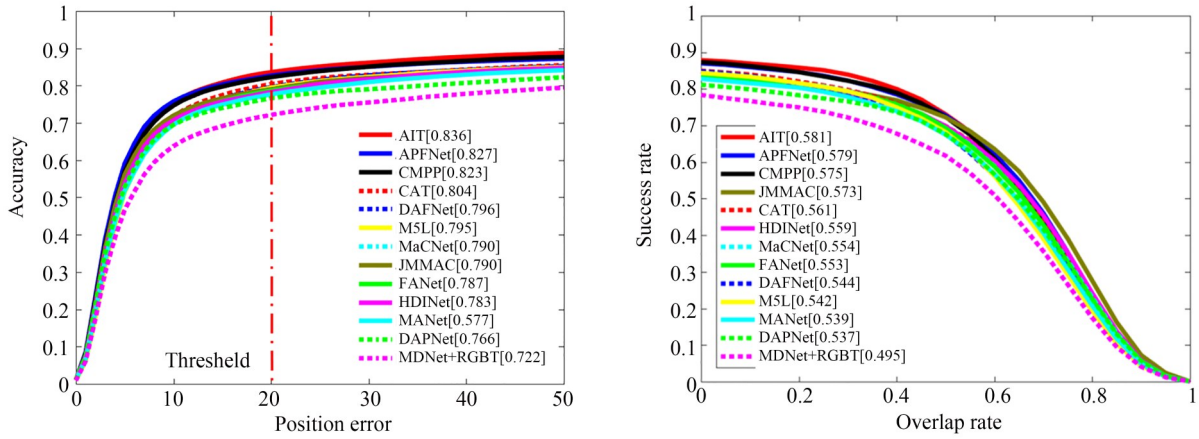



图 7　AIT 及对比算法在 RGBT234 数据集上的跟踪结果

Fig. 7　Evaluation results of AIT and compared algorithm on RGBT234 dataset

要性。与已有算法中性能最优的 APFNet 算法相比，AIT 在精确率和成功率上分别高出 0.9% 和 0.2%。需要注意的是，APFNet 在离线训练阶段利用视频的属性信息和多阶段的训练策略，而 AIT 仅仅是在 MDNet 的基础上利用 FEI 进行特征增强和融合，进一步说明了 FEI 的有效性。

### 4.4　LasHeR 数据集测评

虽然 LasHeR 数据集提供了专用的训练数据，但为了与已有算法保持一致，依然采用 GTOT 进行离线训练，图 8 为 AIT 算法在 LasHeR 数据集的测试集上的跟踪结果。与前两个数据集不同，该数据集的发布较晚，仅提供了部分较新算法的跟踪结果，因此在该数据集上的对比算法有所不同，包括 APFNet，DMCNet[31]，DAPNet，CAT，DAFNet，FANet 和 CMR[32]。如图 8 所示，LasHeR 数据集的跟踪难度较大，但 AIT 算法依然取得了最优的跟踪结果，在精确率上比已有的最好算法 APFNet 和 DMCNet 分别高 0.9% 和 1.9%，在成功率上分别高 0.1% 和 0.8%，说明了 AIT 算法的有效性。

### 4.5　消融实验

为了验证对所提融合网络设计的合理性，进行了消融实验。AIT 算法的几个变种包括：(1) AIT_ch，这个版本中将编码器和解码器的注意力按照通道 (channel) 维度进行操作；(2) AIT_spch，将编码器/解码器中的第二层注意力改为通道维度操作，第一层依然为空间维度；(3) AIT_conv3，仅利用第三层卷积层进行跟踪；(4) AIT_nolastEFI，在三个卷积层特征分别增强和融合后，直接进行维度变换和串接，不增加最后

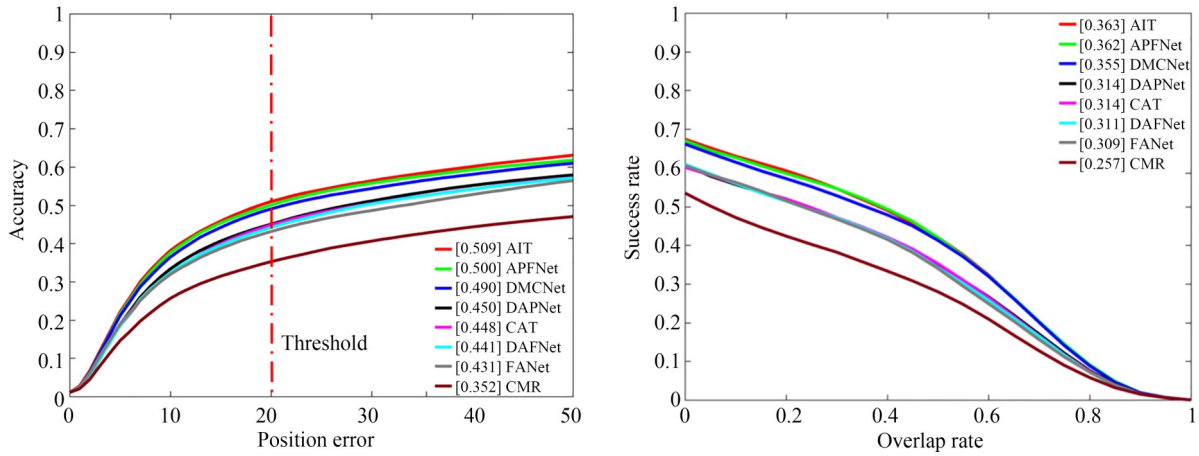



图 8　AIT 及对比算法在 LasHeR 测试集上的跟踪结果

Fig. 8　Evaluation results of AIT and compared algorithms on LasHeR testing set

一个 EFI 模块。表 1 为 AIT 与这 4 个变种在 RGBT234 数据集上的跟踪结果。

**表 1　AIT 及其变种在 RGBT234 数据集上的跟踪结果**

Tab. 1　PR/SR scores of AIT and its variants on RGBT234 dataset

| | PR | SR |
|---|---|---|
| AIT | 0.836 | 0.581 |
| AIT_ch | 0.822 | 0.579 |
| AIT_spch | 0.820 | 0.573 |
| AIT_conv3 | 0.802 | 0.560 |
| AIT_nolastEFI | 0.819 | 0.571 |

表 1 中，AIT 优于 AIT_ch 和 AIT_spch，表明在本文融合网络结构中，选用特征空间维度注意力操作优于特征通道维度。在特征方面，AIT 明显优于 AIT_conv3 和 AIT_nolastEFI，表明多层特征的有效性和最后一个 EFI 模块的有效性。

## 5　结　论

本文提出了一种基于注意力交互的 RGB-T 跟踪算法，从可见光和红外两种模态图像的特征融合出发设计融合网络，引入注意力机制实现了特征增强和跨模态的特征交互。在传统 Transformer 注意力网络的基础上，通过利用 1×1 卷积替换全连接层等方式减小网络规模。考察了不同层卷积特征对跟踪性能的影响，提出了多层卷积融合的网络结构。AIT 算法在 GTOT，RGBT234 和 LasHeR 三个数据集上进行了验证，跟踪结果优于文献[17-18]中提出的基线算法，验证了在 RGB-T 跟踪中注意力机制对多模态特征融合的有效性。

**参考文献：**

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

[2] ZHANG X C, YE P, LEUNG H, *et al*. Object fusion tracking based on visible and infrared images: a comprehensive review [J]. *Information Fusion*, 2020, 63: 166-187.

[3] 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.

[4] 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.

[5] 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.

[6] LI C L, WU X H, ZHAO N, *et al*. Fusing two-stream convolutional neural networks for RGB-T ob-

ject tracking [J]. *Neurocomputing*, 2018, 281: 78-85.

[7] HENRIQUES J F, CASEIRO R, MARTINS P, *et al*. High-speed tracking with kernelized correlation filters[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37 (3): 583-596.

[8] ZHANG L C, DANELLJAN M, GONZALEZ-GARCIA A, *et al*. Multi-modal fusion for end-to-end RGB-T tracking[C]. 2019 *IEEE/CVF International Conference on Computer Vision Workshop* (*ICCVW*). *October* 27-28, 2019. *Seoul*, *Korea* (*South*). IEEE, 2019: 2252-2261.

[9] BHAT G, DANELLJAN M, VAN GOOL L, *et al*. Learning discriminative model prediction for tracking[C]. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 27-*November* 2, 2019. *Seoul*, *Korea* (*South*). IEEE, 2019: 6181-6190.

[10] YAN B, PENG H W, FU J L, *et al*. Learning spatio-temporal transformer for visual tracking[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021. *Montreal*, *QC*, *Canada*. IEEE, 2021: 10428-10437.

[11] LI B, WU W, WANG Q, *et al*. SiamRPN++: evolution of Siamese visual tracking with very deep networks [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019. *Long Beach*, *CA*, *USA*. IEEE, 2019: 4277 - 4286.

[12] ZHANG T L, LIU X R, ZHANG Q, *et al*. SiamCDA: complementarity- and distractor-aware RGB-T tracking based on Siamese network [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, 32(3): 1403-1417.

[13] 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.

[14] 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: 2262 - 2270.

[15] 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.

[16] LI C L, LIU L, LU A D, *et al*. *Challenge-aware RGBT Tracking* [M]. Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 222-237.

[17] 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.

[18] WANG C Q, XU C Y, CUI Z, *et al*. Cross-modal pattern-propagation for RGB-T tracking [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020. *Seattle*, *WA*, *USA*. IEEE, 2020: 7062-7071.

[19] XU C Y, CUI Z, WANG C Q, *et al*. Learning cross-modal interaction for RGB-T tracking [J]. *Science China Information Sciences*, 2022, 66(1): 119103.

[20] VASWANI A, SHAZEER N M, PARMAR N, *et al*. Attention is all you need [C]. *Conference on Neural Information Processing Systems*, 2017: 5998-6008.

[21] 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.

[22] LIU Z, LIN Y T, CAO Y, *et al*. Swin Transformer: hierarchical Vision Transformer using Shifted Windows[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021. *Montreal*, *QC*, *Canada*. IEEE, 2021: 9992- 10002.

[23] ZHU X Z, SU W J, LU L W, *et al*. Deformable DETR: deformable transformers for end-to-end object detection [C]. *International Conference on Learning Representations*, 2021: 1.

[24] CHEN B Y, LI P X, BAI L, *et al*. Backbone is all your need: a simplified architecture for visual object tracking[C]. *European Conference on Computer Vision*, 2022: 375-392.

[25] 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.

[26] MEI J T, ZHOU D M, CAO J D, *et al*. HDINet: hierarchical dual-sensor interaction network for RGBT tracking[J]. *IEEE Sensors Journal*, 2021, 21(15): 16915-16926.

[27] 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.

[28] GAO Y, LI C L, ZHU Y B, *et al*. Deep adaptive fusion network for high performance RGBT tracking[C]. *IEEE International Conference on Computer Vision Workshop*, 2019: 91 - 99.

[29] 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. October* 21 - 25, 2019, *Nice*, *France*. ACM, 2019: 465-472.

[30] 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.

[31] 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, PP. DOI: 10.1109/TNNLS.2022.3157594.

[32] LI C L, ZHU C L, HUANG Y, *et al*. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking[C]. *European Conference on Computer Vision*, 2018: 831-847.

**作者简介：**

**王　晖**(1989—),男,甘肃靖远人,博士,工程师,主要从事计算机视觉、光学测试方面的研究。E-mail: wang_wei.buaa@163.com

**通讯作者：**

**唐自力**(1975—),女,湖南永州人,博士,正高级工程师,主要从事计算机视觉、靶场光学测试方面的研究。E-mail: tang_zili@qq.com