光学学报

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

陈建明[1,2]，李定鲤[1]，曾祥津[1,2]，任振波[3]，邸江磊[1*]，秦玉文[1,2**]

[1]通感融合光子技术教育部重点实验室，广东省信息光子技术重点实验室，广东工业大学信息工程学院，先进光子技术研究院，广东 广州 510006；

[2]南方海洋科学与工程广东省实验室（珠海），广东 珠海 519082；

[3]光场调控与信息感知工业和信息化部重点实验室，陕西省信息光子技术重点实验室，西北工业大学物理科学与技术学院，陕西 西安 710129

摘要 提出一种跨模态光学信息交互和模板动态更新的可见光和热红外（RGBT）跟踪方法，选取能够在跟踪速度和精度上取得平衡的Siamese跟踪器作为基本框架，并设计特征交互模块以重构不同模态的信息比例和增强模态间信息交流。在此基础上，基于无锚框的思想构建预测网络，以提升跟踪器的灵活性和通用性，同时提出一种模板动态更新的策略，通过动态更新跟踪模板增强模型对变化目标的适应能力。在GTOT等3个基准数据集上的对比实验表明，所提方法可显著提升跟踪器在复杂环境下的目标跟踪性能。

关键词 机器视觉；计算机视觉；目标跟踪；孪生网络；模板更新

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/AOS231907

1 引 言

视觉目标跟踪是计算机视觉领域的一项重要任务，旨在根据目标初始状态，估计目标的后续位置和状态，广泛应用于交通和医学等多个领域[1-4]。但当目标处于遮挡、尺度变化或低光照等复杂场景下时，基于可见光视觉的跟踪算法的性能往往快速下降。热红外目标跟踪技术可以抑制背景环境干扰，但无法反映目标的颜色和纹理信息，易受相似物体影响，并且对温度的变化敏感[5]。由于不同光学模态信息之间的互补性，可见光和热红外（RGBT）跟踪技术近年来发展迅速，相比单模态跟踪算法，在复杂环境下的目标跟踪性能得到极大提升[6-7]。

早期RGBT跟踪算法中的特征提取主要依赖于手工设计[8-9]，该算法可以应用于简单的跟踪场景，但缺乏泛化能力。得益于深度学习技术的迅速发展和大规模RGBT数据集的相继发布[10-13]，更加先进的RGBT跟踪算法专注于使用神经网络构建具有较强的特征表征能力的跟踪器。现阶段主流RGBT跟踪算法主要分为基于MDNet[14]和基于Siamese网络两类算法。基于MDNet的算法[15-17]侧重于多模态信息的融合，通过构建有效的融合模块实现目标的鲁棒跟踪，但这一方法的目标外观建模能力较差，为适应目标变化，需要频繁更新跟踪参数，跟踪效率相对较低。随着Siamese网络在单模态跟踪中的成功应用，其被用于提升RGBT跟踪算法的效率。Zhang等[18]将Siamese网络与RGBT跟踪技术相结合，解决了RGBT跟踪器的实时性问题，但该方法在复杂环境下的跟踪性能依然受限。Zhang等[19]在SiamRPN++[20]算法的基础上提出了SiamCDA，旨在提升基于Siamese网络的跟踪器的跟踪精度，而Guo等[21]则将注意力机制引入基于Siamese网络的RGBT跟踪器中，在保证跟踪精度的同时提升跟踪效率。但上述基于Siamese网络的RGBT跟踪器均基于RPN思想构建，跟踪过程会产生大量冗余信息且锚框参数依赖于手工设计，这限制了其泛化能力。与此同时，由于跟踪模板固定，基于Siamese网络的跟踪策略对于变化的目标适应能力较差，极易发生跟踪漂移。此外，无论是基于MDNet还是基于Siamese网络的RGBT跟踪算法，在多模态特征提取上都是通过孤立的分支进行，忽视了多模态特征之间信息的交互，从而影响了跟踪器在复杂环境下的跟踪性能。

收稿日期：2023-12-11；**修回日期**：2024-01-17；**录用日期**：2024-01-25；**网络首发日期**：2024-02-20

基金项目：国家自然科学基金（62075183，62275218）、广东省“珠江人才计划”引进创新创业团队（2021ZT09X044，2019ZT08X340）、中央高校基本科研业务费专项资金（D5000230117）

通信作者：*jiangleidi@gdut.edu.cn；**qinyw@gdut.edu.cn

为解决上述问题，提出一种跨模态光学信息交互和模板动态更新的RGBT跟踪方法(SiamCTU)。首先，使用Siamese网络作为跟踪器的基本框架，通过特征交互模块重构各模态的信息比例，增强模态间信息交互。其次，受单模态跟踪器SiamCAR[22]的启发，基于无锚框思想构建预测网络，直接在候选图像搜索区域的每个位置点对目标边界框进行分类和回归，使得跟踪器能够更加灵活和通用。同时，将中心度分支引入到回归网络中，通过计算搜索区域中每个位置点的中心度得分剔除异常位置，降低锚框回归难度。最后，提出一种模板动态更新策略，通过动态更新跟踪模板以解决基于Siamese网络的跟踪器在跟踪过程中目标和模板不匹配的问题。在GTOT等3个基准数据集上进行了对比实验，与传统基于Siamese网络的RGBT跟踪器相比，SiamCTU能够更加灵活地适应目标的外观变化，显著减少跟踪漂移现象的发生，显著提高复杂环境下的目标跟踪性能，跟踪速度达到30 frame/s，可满足实时性跟踪要求。

2　原理与方法

SiamCTU的整体框架如图1所示，由基于跨模态信息交互的特征提取网络、基于锚框自适应的预测网络和基于特征级的模板更新网络3部分组成。图1中RGB和T分别表示可见光模态和热红外模态图像，FIM表示特征交互，ACC包含按元素相加的add操作、按通道数拼接的concat操作和卷积操作Conv，CNN表示卷积神经网络。模板更新网络只用于模型的跟踪推理阶段。

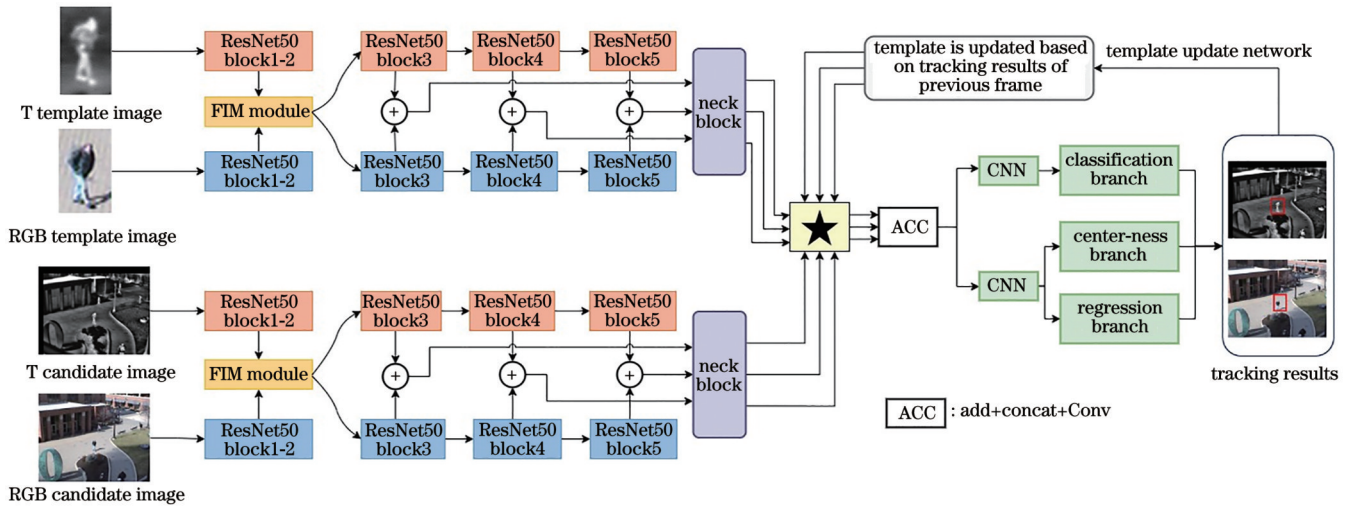


图1　SiamCTU整体网络结构

Fig. 1　Overall network structure of SiamCTU

2.1　基于跨模态信息交互的特征提取网络

2.1.1　基于Siamese网络的目标跟踪

Siamese网络是指具有两个相同结构的网络模型，该网络可以平衡目标跟踪的速度和精度，在跟踪领域获得广泛应用。基于Siamese网络的目标跟踪算法主要是将跟踪任务转化为相似度匹配问题[23]，利用神经网络学习目标与模板之间的相似性，进而实现跟踪，如图2所示。该方法将视频序列第一帧的目标模板z和后续帧的图像x作为输入，经过具有相同结构的特征提取网络$\varphi(\cdot)$映射到特征空间，使用度量函数$f(z,x)$

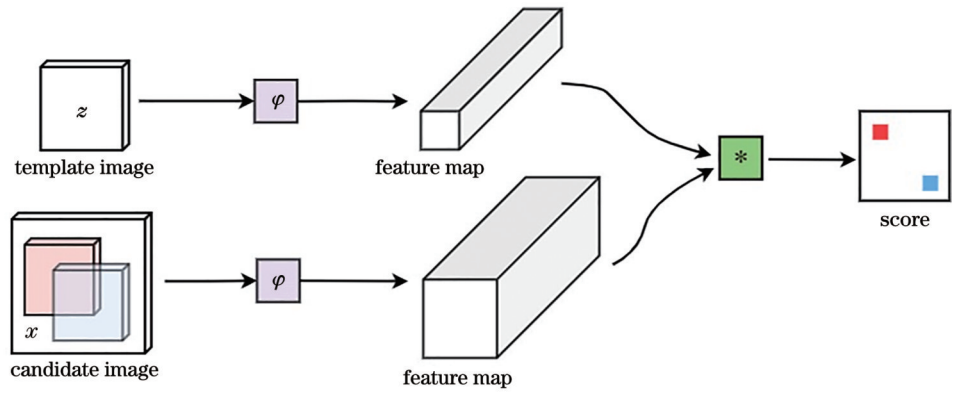


图2　基于Siamese网络的目标跟踪方法模型结构

Fig. 2　Model structure of target tracking method based on Siamese network

比较两者之间的相似度，生成响应得分图：

$$f(z,x)=\varphi(z)*\varphi(x)+b, \tag{1}$$

式中：*为互相关操作；b为响应图中每个位置点的相似度得分，得分越高代表二者越相似。为了处理多模态光学信息，SiamCTU包含两个完全相同的Siamese网络结构，输入来自两个模态视频序列帧裁剪出来的图像块，分别为T模板和RGB模板图像、T候选和RGB候选图像。这样的设计使得模型能够更灵活地适应不同传感器模态下的数据，进而提高模型的泛化性能。

2.1.2 骨干网络

深度神经网络模型凭借其高效的特征表征能力，在目标跟踪领域获得广泛关注[24-26]。ResNet50是He等[27]针对深层次网络退化问题，提出的一种残差网络模型，结构如图3所示。其中，ReLu是一种激活函数，用来提升模型的非线性拟合能力。残差网络通过恒等映射将模型的输入跨层引入到更深的网络层上作为输入，使得模型可以直接学习从输入到输出的映射，解决神经网络在反向传播过程中的梯度消失和爆炸问题。将残差模块的输入和输出分别用y和$I(y)$表示，则残差模块的实现原理可表示为

$$I(y)=F(y)+y, \tag{2}$$

式中：$F(y)$为神经网络模型学习的部分。

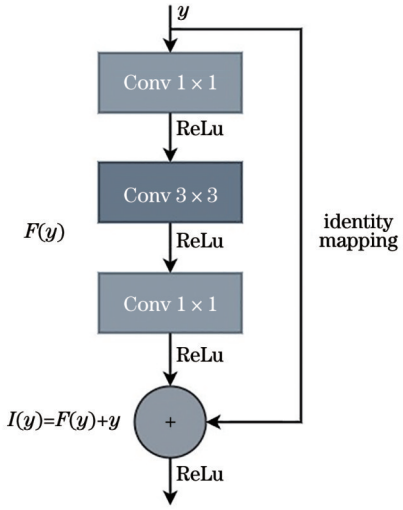


图3 ResNet的残差结构

Fig. 3 Residual structure of ResNet

不同深度的卷积神经网络可以学习到不同的目标特征。浅层特征包含物体更多的颜色、纹理和形状等信息，对目标精确定位必不可少，而深层特征则包含物体更多的语义信息，对目标分类至关重要。因此，只使用物体深层特征进行跟踪会降低跟踪器的性能。当前的跟踪算法通常综合利用目标浅层特征和深层特征来提高跟踪精度[19-20]。SiamCTU以ResNet50作为特征提取器，将网络最后3个块的特征图作为输出，摒弃包含大量背景噪声的前两个块，同时采用聚合多层次特征的方式进行跟踪。采用连续卷积步进的ResNet50可获得目标更多的抽象特征，但同时会降低特征分辨率，这不利于需要丰富的空间信息进行目标位置预测的Siamese网络跟踪器。为解决这一问题，SiamCTU删除了ResNet50网络最后两个块的下采样操作，以提升特征图的空间分辨率，并同时使用空洞卷积替代原本的卷积操作，增加网络的感受野。

为有效利用目标的多模态信息，跟踪器将从特征提取网络4个分支获取的T模板特征图(Z_{t})、RGB模板特征图(Z_{rgb})、T候选特征图(X_{t})、RGB候选特征图(X_{rgb})分层进行融合，并将融合后的模板图像特征用Z表示，融合后的候选图像特征用X表示，融合过程可以表示为

$$Z_i=Z_{\mathrm{t},i}+Z_{\mathrm{rgb},i}, \tag{3}$$

$$X_i=X_{\mathrm{t},i}+X_{\mathrm{rgb},i}, \tag{4}$$

式中：i=3，4，5分别为ResNet50的第3、4和5块输出的特征图。对Z_i和X_i执行互相关操作，生成响应得分图R_i(i=3，4，5)。互相关操作可描述为以模板特征图作为卷积核在候选特征图上进行滑动卷积并生成特征图的过程，在数学上等同于使用内积运算来组合特征图，如图4所示。最后，为降低模型计算复杂度，通过大小为1×1的卷积核聚合R_i，生成一张只包含256个通道的响应得分图R。

2.1.3 特征交互模块

基于Siamese网络的目标跟踪方法在进行特征提取时，都是针对单个模态数据独立进行。然而，孤立的分支不利于模态间的信息交流。因此，为加强网络对双模态数据有效互补特征的学习，基于SENet[28]设计了如图5所示的特征交互模块。其中：GAP表示全局平均池化运算，通过计算每一个通道特征图的所有像素的平均值获得通道的全局信息；CRC表示卷积操作Conv、非线性激活函数ReLu和卷积操作Conv。特征交互模块由热红外分支、共享分支和可见光分支3条分支构成，输入分别来自ResNet50第2块之后的T模态特征和RGB模态特征，通过共享分支学习模态间的关联信息，提高网络对每个模态重要区域的关注度。在共享分支中，首先将来自两个模态的特征图按对应通道进行元素加和并生成联合特征$\boldsymbol{U}$，通过全局平均池化运算计算每个通道的权重，以突出重要特征通道，减少信息冗余。然后，经过CRC模块重构不同模态的信息比例，生成共享权值矩阵，学习到模态间的关联信息。将输入的T模态特征和RGB模态特征分别表示为$\boldsymbol{F}_{\mathrm{t}}$和$\boldsymbol{F}_{\mathrm{rgb}}$，则上述过程可表示为

$$\boldsymbol{U}=\boldsymbol{F}_{\mathrm{t}}+\boldsymbol{F}_{\mathrm{rgb}}, \tag{5}$$

$$\boldsymbol{W}=\mathrm{CRC}[(\mathrm{GAP})\boldsymbol{U}], \tag{6}$$

式中：$\boldsymbol{W}$为共享分支生成的共享权重矩阵，携带了模态间的关联信息。最后，将其分别与热红外和可见光分支输出的原始模态特征$\boldsymbol{F}_{\mathrm{t}}$和$\boldsymbol{F}_{\mathrm{rgb}}$做乘法操作，分别

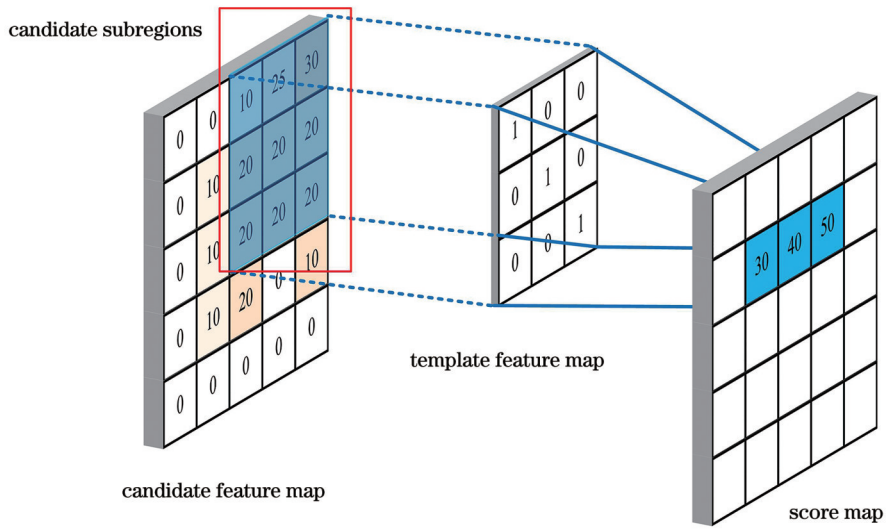


图 4　互相关运算示意图

Fig. 4　Schematic diagram of cross correlation operations

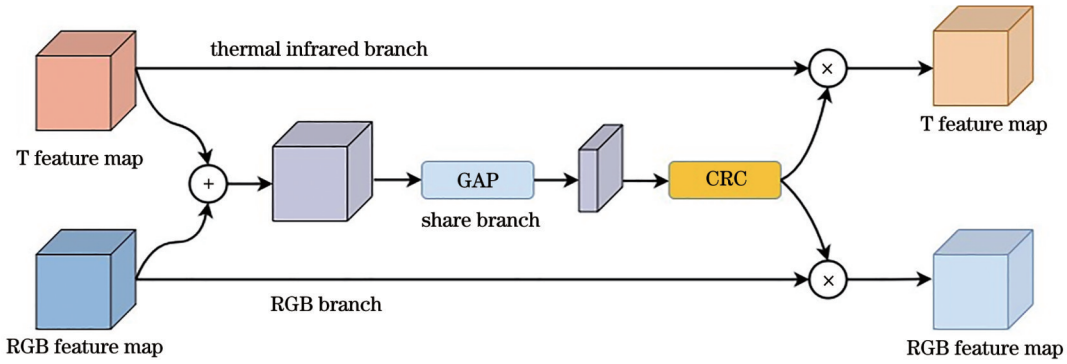


图 5　特征交互模块结构图

Fig. 5　Structure diagram of feature interaction module

生成重构后的热红外特征 $\boldsymbol{F}'_{\mathrm{t}}$ 和 $\boldsymbol{F}'_{\mathrm{rgb}}$：

$$\boldsymbol{F}'_{\mathrm{t}}=\boldsymbol{F}_{\mathrm{t}}\times \mathrm{p}(\boldsymbol{W}), \tag{7}$$

$$\boldsymbol{F}'_{\mathrm{rgb}}=\boldsymbol{F}_{\mathrm{rgb}}\times \mathrm{p}(\boldsymbol{W}), \tag{8}$$

式中：$\mathrm{p}(\boldsymbol{W})$ 为对共享权重矩阵 $\boldsymbol{W}$ 进行扩展操作，使其与输入特征的形状保持一致。模态间的交互作用使得输出特征内部信息更加具有互补性。后续消融实验也充分证明了所设计模块的有效性。

2.2　基于锚框自适应的预测网络

基于 RPN 的目标跟踪器通过预先设置不同比例的锚框来匹配目标位置，但当目标发生大尺度变化或者形变时，跟踪器性能表现不佳。与之不同的是，SiamCTU 基于无锚框思想设计预测网络，将跟踪任务分为目标分类和位置回归两个子任务，根据目标的尺寸自适应调整锚框的大小，使跟踪器更加灵活通用。通常位置点距离物体标注框中心点越近，则其锚框回归越容易，回归精度也越高[22]。因此，为进一步降低锚框回归难度，将中心度分支引入到预测网络，计算搜索区域中每个位置点的中心度得分，再利用中心度得分乘上锚点的分类得分，降低远离物体中心的锚点分数，并在后续过程过滤掉得分较低的锚框。

预测网络由分类、回归和中心度 3 条分支构成，每条分支均输出 1 个大小为 25×25 的特征图。分类特征图上的每个点都包含 1 个 2 维向量，用于区分位置点的类别。回归特征图上的点 (i,j) 都包含 1 个 4 维向量 $\boldsymbol{g}(i,j)=(l,m,r,n)$，其中：$l$ 为点 (i,j) 与标签框左边的距离；m 为点 (i,j) 与标签框上边的距离；r 为点 (i,j) 与标签框右边的距离；n 为点 (i,j) 与标签框下边的距离。中心度用于计算位置的中心度得分，计算规则如下：

$$C(i,j)=G(i,j)\times\sqrt{\frac{\min(l,r)}{\max(l,r)}\times\frac{\min(m,n)}{\max(m,n)}}, \tag{9}$$

式中：$G(i,j)\in\{0,1\}$，当且仅当特征图点 (i,j) 对应搜索区域的点落在第一帧划定的标签框内时，$G(i,j)$ 取 1。

2.3　基于特征级的模板更新网络

模板更新分支只应用于跟踪阶段，不参与网络训练。基于 Siamese 网络的目标跟踪算法在跟踪开始

前，通常会将每个视频序列中第一帧的目标作为跟踪模板，但在跟踪过程中随目标的运动及目标所处环境的变化，目标的尺寸和外观不断发生改变，故从视频序列第一帧中裁剪得到的模板难以匹配到后续发生形变和外观改变的目标，使跟踪器性能下降。针对这一问题，设计了一种特征级的模板更新网络，在保证初始模板特征不变的情况下，递归更新在线模板以获取目标的实时状态，实现跟踪目标的动态更新。通常，第一帧中的固定模板提供了目标的初始表示，这种表示在目标变化较小的情况下是相对稳定的。通过保留这个固定模板，算法可以避免受到目标外观变化的干扰，提高目标表示的稳定性。具体来说，在第一帧中，初始模板和在线更新模板被初始模板取代。在随后的帧中，初始模板保持固定，而在线模板不断更新。正如图 1 中的模板更新分支所示，在线更新分支利用跟踪器的预测结果从上一帧的候选图像中分别获得 RGB 模态和 T 模态的在线模板，随后经过与初始模板一致的特征提取网络分别与候选图像进行深度互相关操作，得到在线响应得分图 R_i'（i=3,4,5）。利用 R_i' 对初始模板得到的响应得分图 R_i（i=3,4,5）进行更新，将更新后的响应表示为 R_i''（i=3,4,5），则响应得分图的更新过程可表示为

$$R_i''=\lambda\times R_i+(1-\lambda)R_i', \tag{10}$$

式中：$\lambda\in[0.5,1]$，为 R_i 对应的权重。λ 越大，代表 R_i'' 中包含的初始模板 R_i 的特征越多。由于目标在运动中可能会面临大量遮挡或外观变化的情况，跟踪器需要使用更多的初始模板特征进行修正。因此，λ 的取值范围为 0.5～1。最后将 R_i'' 送入预测模块中，生成预测结果。

2.4 网络训练

训练开始前，使用从 ImageNet[29] 数据集中预训练好的权重初始化的 ResNet50 参数，采用初始学习率为 10^{-3} 的 SGD 优化器对模型进行优化。模型共训练 30 轮，前 15 轮冻结模型中的 Siamese 网络，训练特征交互模块和预测网络。后 15 轮解冻 ResNet50 的最后 3 个块并对其进行微调。训练过程中，模板图像的输入尺寸统一设置为 127×127，候选图像的输入尺寸统一设置为 255×255。

模型的总损失函数为类别损失 L_{cls}、中心度损失 L_{cen} 和回归损失 L_{reg} 的线性加权：

$$L=L_{cls}+\lambda_1 L_{cen}+\lambda_2 L_{reg}, \tag{11}$$

式中：L_{cls} 为交叉熵损失函数；λ_1、λ_2 为超参数，通过网格搜索参数的方式得到。实验中沿用 Siamese 框架中的超参数设置[22]，令 λ_1=1，λ_2=3。中心度损失 L_{cen} 的定义如下：

$$L_{cen}=\frac{-1}{\sum G(i,j)}\sum_{G(i,j)=1}C(i,j)\log_2 V_{cen}(i,j)+[1-C(i,j)]\times\log_2[1-V_{cen}(i,j)], \tag{12}$$

式中：$V_{cen}(i,j)$ 为预测点 (i,j) 的中心度得分。回归损失函数用于调整预测框的大小，可表示为

$$L_{reg}=\frac{1}{\sum G(i,j)}\sum_{i,j}G(i,j)L_{iou}[g(i,j),V_{reg}(i,j)], \tag{13}$$

式中：$V_{reg}(i,j)$ 为预测点 (i,j) 到标签框四条边的距离；L_{iou} 为网络预测框和真实标签框之间的交并比，通过 $g(i,j)$ 和 $V_{reg}(i,j)$ 计算得到。

2.5 跟踪推理

训练后的模型最终生成一个 6 维向量（S_{cls}，S_{cen}，l，m，r，n），完成目标位置和尺寸的预测，其中 S_{cls} 和 S_{cen} 分别为位置点的分类得分和中心度得分。在跟踪过程中，由于视频序列中相邻帧之间目标的形状变化比较微小，所以通过在分类得分图中加入尺度惩罚 P 来抑制预测框产生大的形变，尺度惩罚 P 定义如下：

$$P=\exp\left[k\times\max\left(\frac{r}{r'},\frac{r'}{r}\right)\times\max\left(\frac{s}{s'},\frac{s'}{s}\right)\right], \tag{14}$$

式中：k 为超参数；r 为边界框的宽高比；r' 为最后一帧的宽高比；s 为边界框尺度；s' 为最后一帧的目标尺度。同理，相邻帧之间目标的位移变化也比较微小，使用余弦窗惩罚 H 和网络输出的中心度得分共同抑制预测框产生大的位移，最终的跟踪过程可表示为

$$q=\arg\max[P\times S_{cls}\times S_{cen}\times(1-\omega)+H\omega], \tag{15}$$

式中：q 为响应图上目标最大位置点的索引；ω 为超参数。

3 实验与分析

3.1 数据集和评估指标

GTOT 数据集[10] 是 RGBT 跟踪器最通用的评估数据集，由 50 对配准好的 RGB 视频序列和 T 视频序列构成，包含大尺度变化（LSV）、快速移动（FM）、热交叉（TC）、遮挡（OCC）、小目标（SO）、形变（DEF）和低光照（LI）等 7 种挑战属性。RGBT234 数据集[11] 更为复杂，由 234 个未对齐的 RGBT 视频对和 22 个目标类别组成，包含无遮挡（NO）、FM、LI、TC、尺度形变（SV）、DEF、部分遮挡（PO）、严重遮挡（HO）、相机移动（CM）、低分辨率（LR）、背景杂波（BC）和运动模糊（MB）等 12 种挑战属性。LasHeR 数据集[12] 的规模更大、挑战属性更为全面，由 1224 对配准好的 RGBT 视频序列构成，其中训练子集为 979 对，测试子集为 245 对。近期新提出的 VTUAV 数据集[13] 常用来进行模型训练。当模型对 GTOT 数据集和 RGBT234 数据集进行评估时，使用 LasHeR 数据集和 VTUAV 数据集训练跟踪模型。当模型在 LasHeR 数据集的测试子集上进行评估时，使用 LasHeR 数据集的训练子集和 VTUAV 数据集的训练网络。

精确率（PR）、成功率（SR）和跟踪器每秒视频处

理帧数(FPS)作为RGBT跟踪算法的评估指标。PR表示跟踪器预测的目标中心点位置与真实目标中心点位置的差值小于给定阈值的视频帧数占所有视频检测帧的百分比。SR表示跟踪器预测框与真实标签框的重叠面积小于给定阈值的视频帧数占所有视频检测帧的百分比。GTOT数据集中存在大量小目标,将其阈值设置为5,将RGBT234数据集和LasHeR数据集阈值设置为20。

3.2 定量分析

3.2.1 GTOT数据集性能评估

为了验证方法的有效性,将SiamCTU与HMFT[13]、CAT[16]、APFNet[17]、SiamCDA[19]、CMPP[30]、FANet[31]、ADRNet[32]、M5L[33]、DFNet[34]和DFAT[35]等10种先进的跟踪算法进行比较,实验结果如图6所示。SiamCTU的PR为0.94,相比当前PR性能最好的跟踪器CMPP[30]提升约1.5%;SiamCTU的SR值为0.756,相比当前SR性能最好的跟踪器HMFT[13]提升约0.9%;相比基于MDNet的RGBT跟踪器APFNet[17],SiamCTU的PR和SR值分别提升约3.9%和2.6%;相比同样基于Siamese网络的RGBT跟踪器SiamCDA[19],SiamCTU的PR和SR值分别提升约7.2%和3.3%。

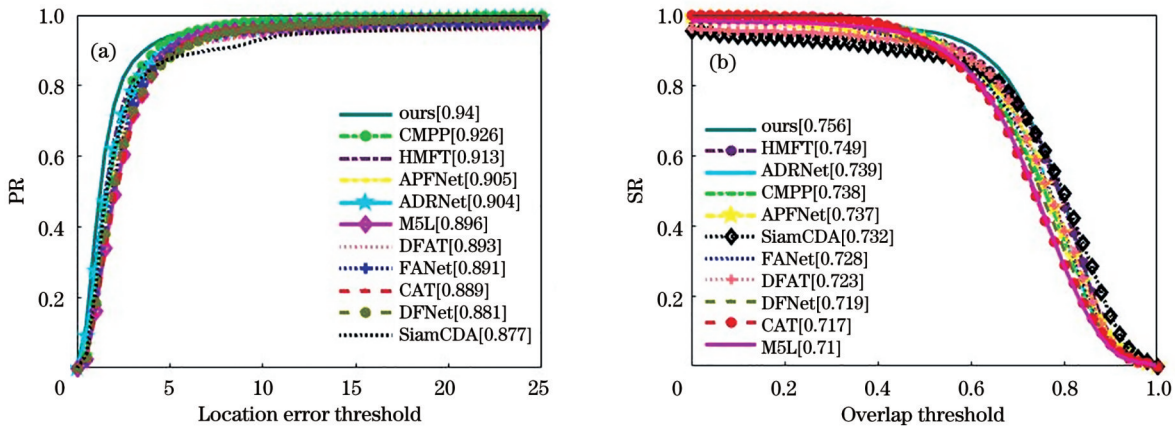


图6 SiamCTU与10种先进的跟踪器在GTOT数据集上的对比结果。(a)精确度;(b)成功率

Fig. 6 Comparison of SiamCTU with 10 advanced trackers on GTOT dataset. (a) PR; (b) SR

挑战属性的比较结果显示了跟踪器应对各种挑战属性的能力,实验结果如表1所示。由表1可见,跟踪器在GTOT数据集的7种挑战属性下都取得了较好的跟踪性能。特别地,在OCC、LSV、FM、LI和TC等5种挑战属性下所设计的跟踪器在PR和SR得分上都取得了前3的跟踪结果,这证明SiamCTU在应对这些挑战属性时是有效的。

表1 在GTOT数据集中,SiamCTU与10个最先进的目标跟踪器基于挑战属性的PR和SR(用PR/SR表示)

Table 1 PR and SR of the SiamCTU compared with the 10 most advanced target trackers based on challenge attributes in the GTOT dataset (represented by PR/SR) unit: %

Publisher information	Trackers	OCC	LSV	FM	LI	TC	SO	DEF	All
ECCV 2020	CAT[16]	89.9/69.2	85.0/67.9	83.9/65.4	89.2/72.3	89.9/71.0	94.7/69.9	92.5/75.5	88.9/71.7
IEEE TIV 2020	FANet[31]	86.4/70.3	81.6/68.2	80.1/65.3	89.9/73.5	90.4/72.3	94.3/70.8	94.6/77.6	89.1/72.8
CVPR 2020	CMPP[30]	94.7/71.7	91.2/69.9	91.7/68.6	92.4/74.3	93.8/72.9	98.0/72.5	94.6/78.8	92.6/73.8
IJCV 2021	ADRNet[32]	88.5/69.6	86.1/70.6	83.4/67.1	92.2/75.9	91.1/73.6	94.7/72.1	94.5/77.5	90.4/73.9
IEEE TIP 2022	M5L[33]	87.1/66.6	91.0/70.2	89.4/68.5	91.7/73.0	89.2/69.5	96.0/70.2	92.2/74.6	89.6/71.0
AAAI 2022	APFNet[17]	90.3/71.3	87.7/71.2	86.5/68.4	91.4/74.8	90.4/71.6	94.3/71.3	94.6/78.0	90.5/73.7
IEEE TCSVT 2022	SiamCDA[19]	82.2/69.4	91.5/74.8	86.6/72.0	92.4/76.4	82.6/68.5	87.4/69.1	87.9/72.7	87.7/73.2
CVPR 2022	HMFT[13]	88.5/72.0	89.3/75.2	85.8/74.1	94.6/76.7	89.6/73.0	92.8/71.3	94.4/74.7	91.3/74.9
IEEE TITS 2023	DFNet[34]	88.7/68.9	84.2/69.7	81.4/64.4	89.6/73.3	88.6/71.5	94.3/71.3	92.8/74.8	88.1/71.9
Inf Fusion 2023	DFAT[35]	86.3/68.7	92.4/75.0	89.1/74.0	92.2/74.1	89.1/70.7	94.4/71.9	91.9/73.5	89.3/72.3
	Our	91.5/72.1	94.2/75.9	90.2/73.3	95.8/76.6	92.7/74.0	92.8/71.7	94.1/74.8	94.0/75.6

3.2.2　RGBT234 数据集评估结果

在进一步的实验中，在 RGBT234 数据集上将 SiamCTU 与 HMFT[13]、APFNet[17]、SiamCDA[19]、FANet[31]、ADRNet[32]、M5L[33]、DFNet[34]、DFAT[35]、DAPNet[36]和 HDINet[37]等 10 种先进的 RGBT 跟踪算法进行综合分析对比，比较结果如图 7 所示。相比于 ADRNet[32]和 M5L，SiamCTU 在 PR 得分上低于这两种先进的跟踪算法，但是有着更高的跟踪成功率。相比于 APFNet[17]，SiamCTU 在 PR 和 SR 得分上都低于这种算法，但跟踪速度约是 APFNet 算法的 21 倍。其中，APFNet 算法的跟踪速度为 1.4 frame/s。

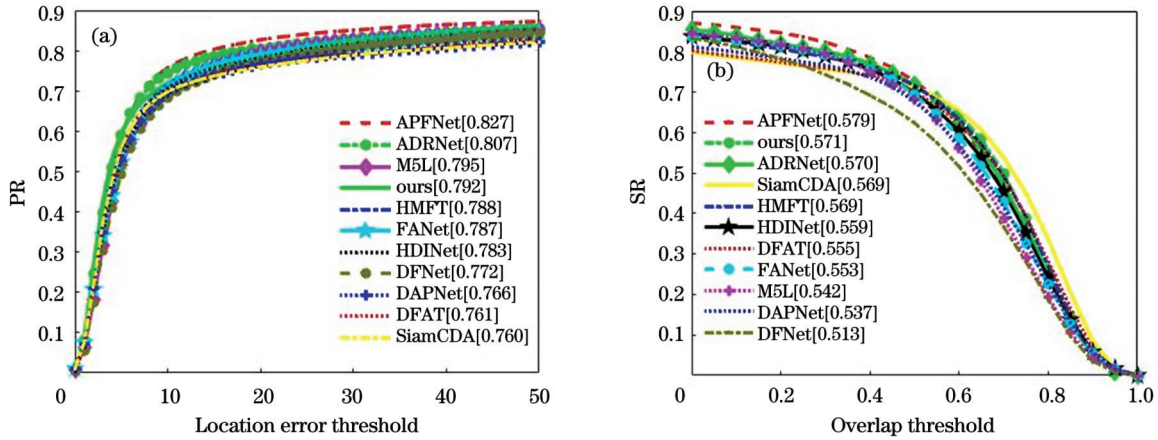
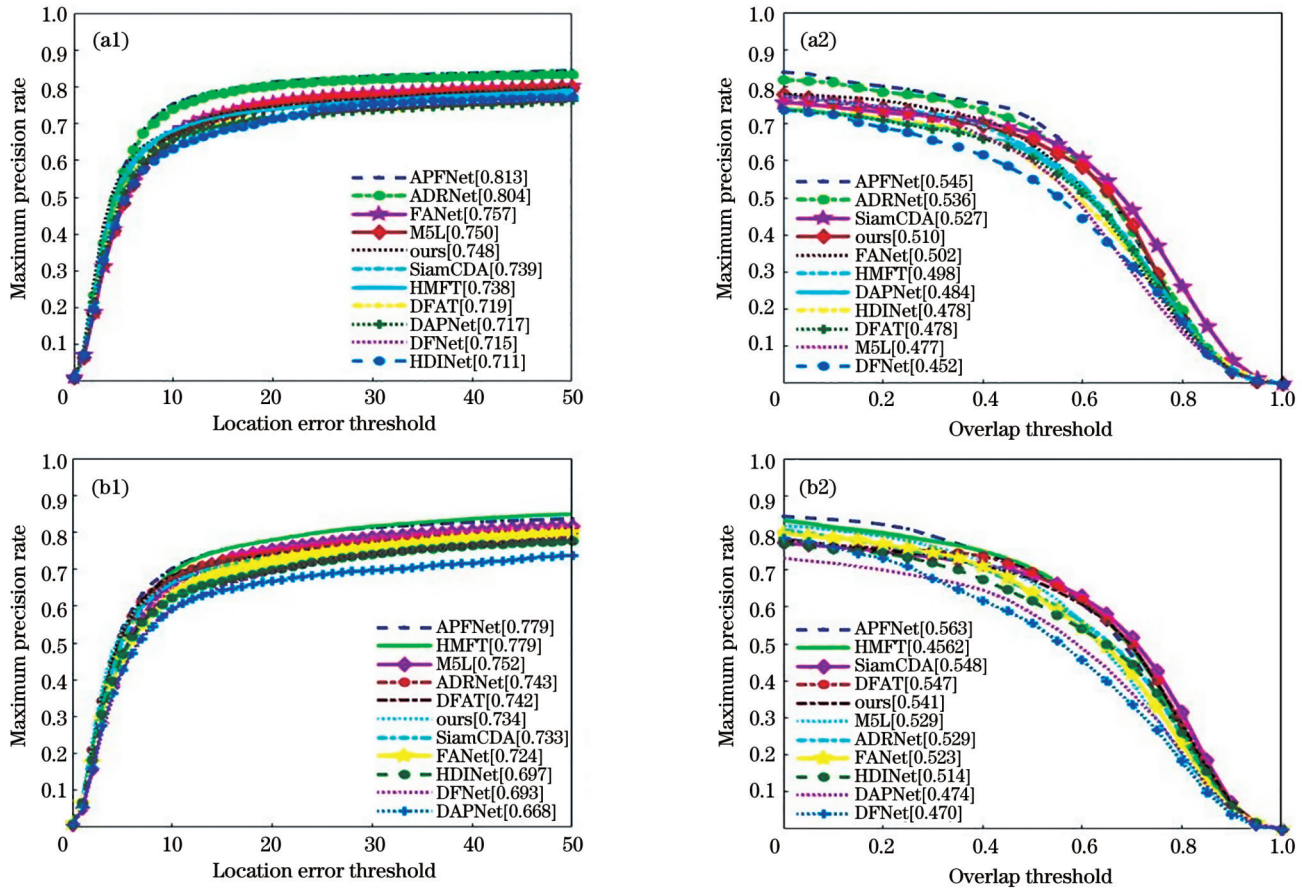
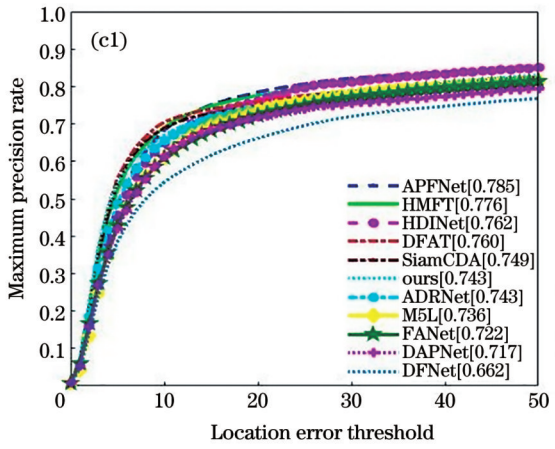


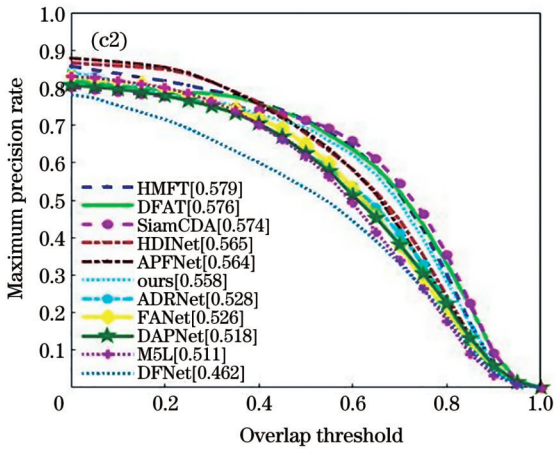
图 7　SiamCTU 与 10 种先进的跟踪器在 RGBT234 数据集上的对比结果。(a)精确率；(b)成功率

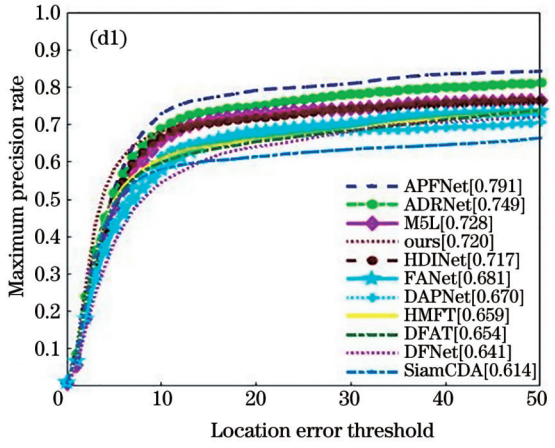
Fig. 7　Comparison of SiamCTU with 10 advanced trackers on RGBT234 dataset. (a) PR; (b) SR

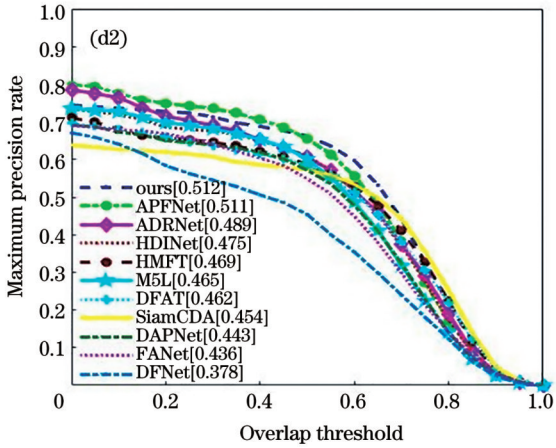
同理，为了更好地验证跟踪器性能，在 RGBT234 数据集上同样进行了挑战属性的比较。图 8 分别给出了对比算法在 RGBT234 数据集上 12 种挑战属性的 PR 和 SR 得分。从图 8 可以看出，与先进的跟踪算法相比，SiamCTU 在大多数挑战属性下都取得了较好的跟踪性能，这充分证明所提方法的有效性。

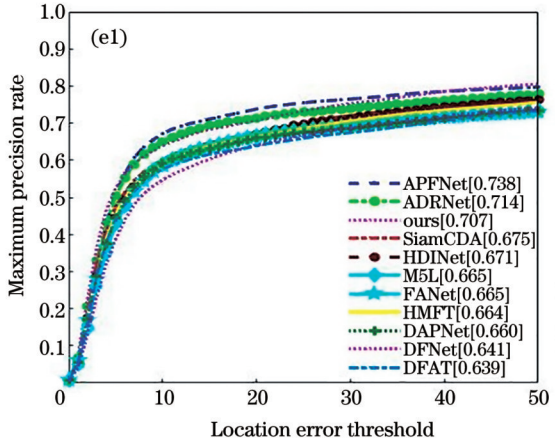


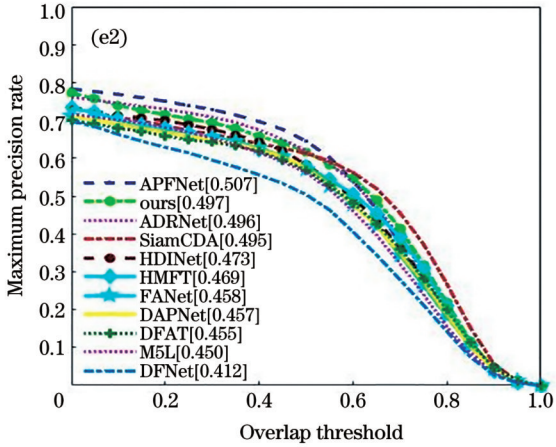

(c1)
Maximum precision rate
Location error threshold
APFNet[0.785]
HMFT[0.776]
HDINet[0.762]
DFAT[0.760]
SiamCDA[0.749]
ours[0.743]
ADRNet[0.743]
M5L[0.736]
FANet[0.722]
DAPNet[0.717]
DFNet[0.662]

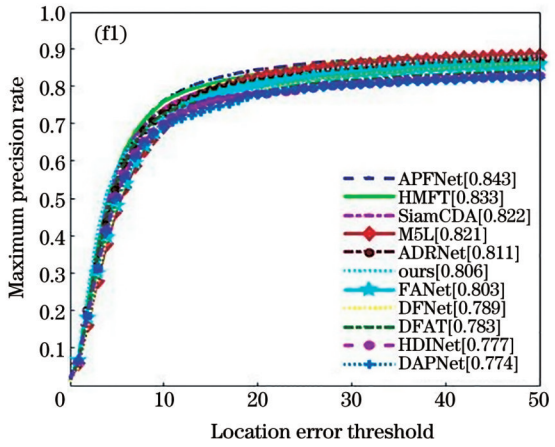

(c2)
Maximum precision rate
Overlap threshold
HMFT[0.579]
DFAT[0.576]
SiamCDA[0.574]
HDINet[0.565]
APFNet[0.564]
ours[0.558]
ADRNet[0.528]
FANet[0.526]
DAPNet[0.518]
M5L[0.511]
DFNet[0.462]

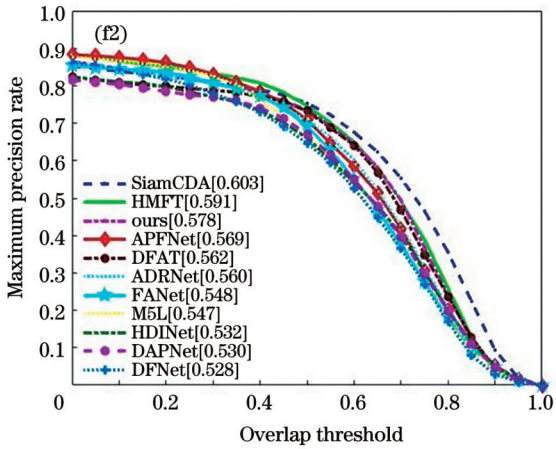

(d1)
Maximum precision rate
Location error threshold
APFNet[0.791]
ADRNet[0.749]
M5L[0.728]
ours[0.720]
HDINet[0.717]
FANet[0.681]
DAPNet[0.670]
HMFT[0.659]
DFAT[0.654]
DFNet[0.641]
SiamCDA[0.614]

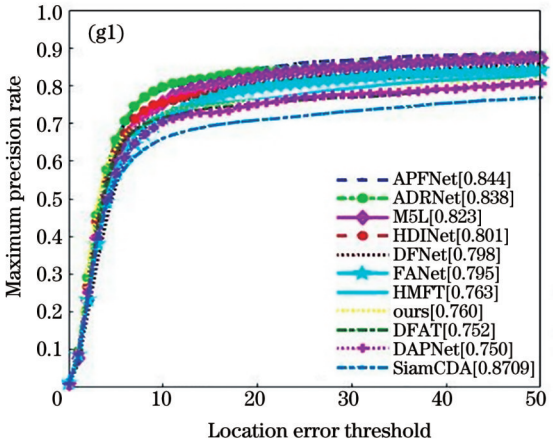

(d2)
Maximum precision rate
Overlap threshold
ours[0.512]
APFNet[0.511]
ADRNet[0.489]
HDINet[0.475]
HMFT[0.469]
M5L[0.465]
DFAT[0.462]
SiamCDA[0.454]
DAPNet[0.443]
FANet[0.436]
DFNet[0.378]

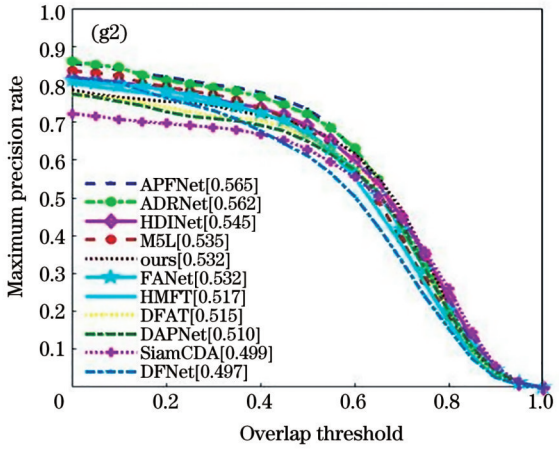

(e1)
Maximum precision rate
Location error threshold
APFNet[0.738]
ADRNet[0.714]
ours[0.707]
SiamCDA[0.675]
HDINet[0.671]
M5L[0.665]
FANet[0.665]
HMFT[0.664]
DAPNet[0.660]
DFNet[0.641]
DFAT[0.639]

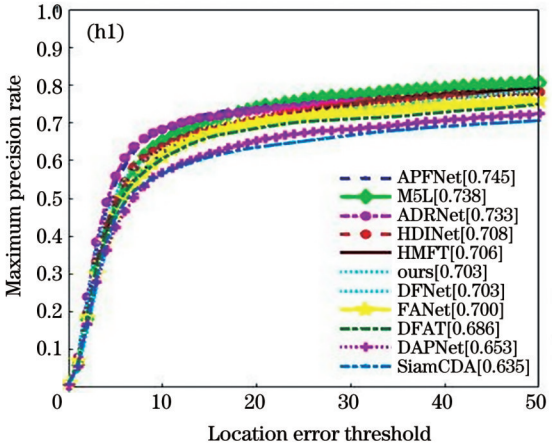

(e2)
Maximum precision rate
Overlap threshold
APFNet[0.507]
ours[0.497]
ADRNet[0.496]
SiamCDA[0.495]
HDINet[0.473]
HMFT[0.469]
FANet[0.458]
DAPNet[0.457]
DFAT[0.455]
M5L[0.450]
DFNet[0.412]

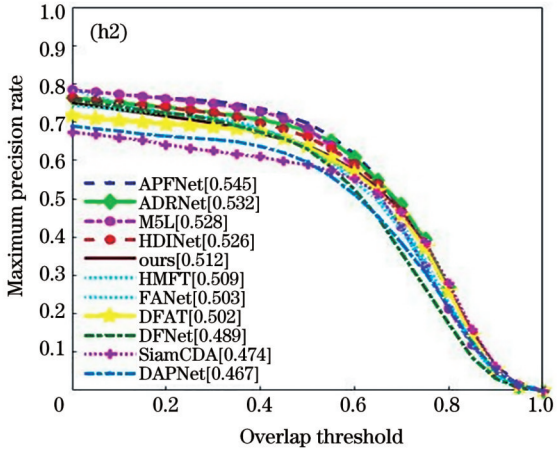

(f1)
Maximum precision rate
Location error threshold
APFNet[0.843]
HMFT[0.833]
SiamCDA[0.822]
M5L[0.821]
ADRNet[0.811]
ours[0.806]
FANet[0.803]
DFNet[0.789]
DFAT[0.783]
HDINet[0.777]
DAPNet[0.774]

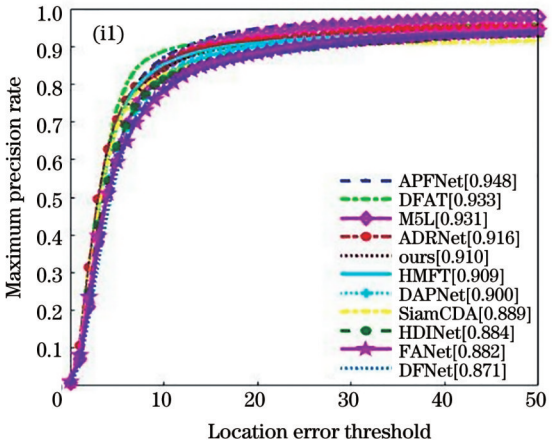

(f2)
Maximum precision rate
Overlap threshold
SiamCDA[0.603]
HMFT[0.591]
ours[0.578]
APFNet[0.569]
DFAT[0.562]
ADRNet[0.560]
FANet[0.548]
M5L[0.547]
HDINet[0.532]
DAPNet[0.530]
DFNet[0.528]

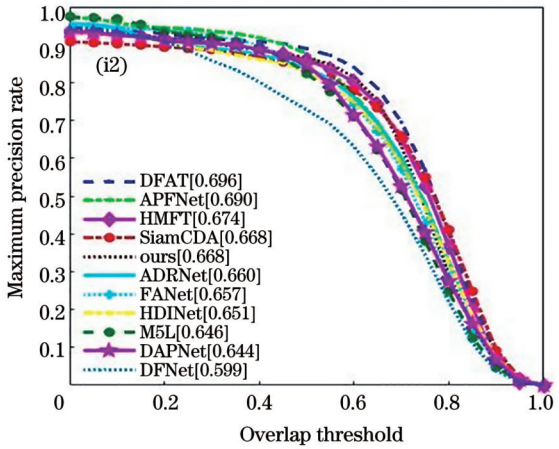

(g1)
Maximum precision rate
Location error threshold
APFNet[0.844]
ADRNet[0.838]
M5L[0.823]
HDINet[0.801]
DFNet[0.798]
FANet[0.795]
HMFT[0.763]
ours[0.760]
DFAT[0.752]
DAPNet[0.750]
SiamCDA[0.8709]

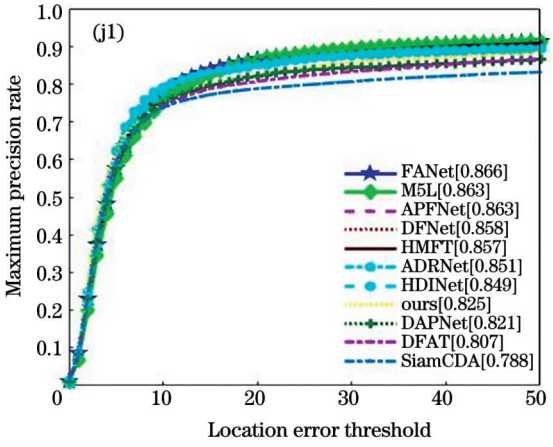

(g2)
Maximum precision rate
Overlap threshold
APFNet[0.565]
ADRNet[0.562]
HDINet[0.545]
M5L[0.535]
ours[0.532]
FANet[0.532]
HMFT[0.517]
DFAT[0.515]
DAPNet[0.510]
SiamCDA[0.499]
DFNet[0.497]

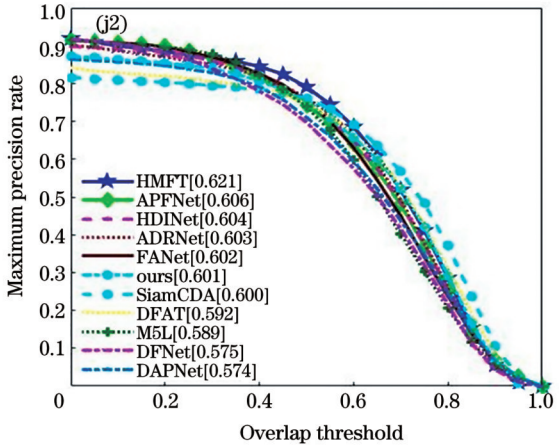

(h1)
Maximum precision rate
Location error threshold
APFNet[0.745]
M5L[0.738]
ADRNet[0.733]
HDINet[0.708]
HMFT[0.706]
ours[0.703]
DFNet[0.703]
FANet[0.700]
DFAT[0.686]
DAPNet[0.653]
SiamCDA[0.635]


(h2)
Maximum precision rate
Overlap threshold
APFNet[0.545]
ADRNet[0.532]
M5L[0.528]
HDINet[0.526]
ours[0.512]
HMFT[0.509]
FANet[0.503]
DFAT[0.502]
DFNet[0.489]
SiamCDA[0.474]
DAPNet[0.467]


(i1)
Maximum precision rate
Location error threshold
APFNet[0.948]
DFAT[0.933]
M5L[0.931]
ADRNet[0.916]
ours[0.910]
HMFT[0.909]
DAPNet[0.900]
SiamCDA[0.889]
HDINet[0.884]
FANet[0.882]
DFNet[0.871]


(i2)
Maximum precision rate
Overlap threshold
DFAT[0.696]
APFNet[0.690]
HMFT[0.674]
SiamCDA[0.668]
ours[0.668]
ADRNet[0.660]
FANet[0.657]
HDINet[0.651]
M5L[0.646]
DAPNet[0.644]
DFNet[0.599]


(j1)
Maximum precision rate
Location error threshold
FANet[0.866]
M5L[0.863]
APFNet[0.863]
DFNet[0.858]
HMFT[0.857]
ADRNet[0.851]
HDINet[0.849]
ours[0.825]
DAPNet[0.821]
DFAT[0.807]
SiamCDA[0.788]


(j2)
Maximum precision rate
Overlap threshold
HMFT[0.621]
APFNet[0.606]
HDINet[0.604]
ADRNet[0.603]
FANet[0.602]
ours[0.601]
SiamCDA[0.600]
DFAT[0.592]
M5L[0.589]
DFNet[0.575]
DAPNet[0.574]

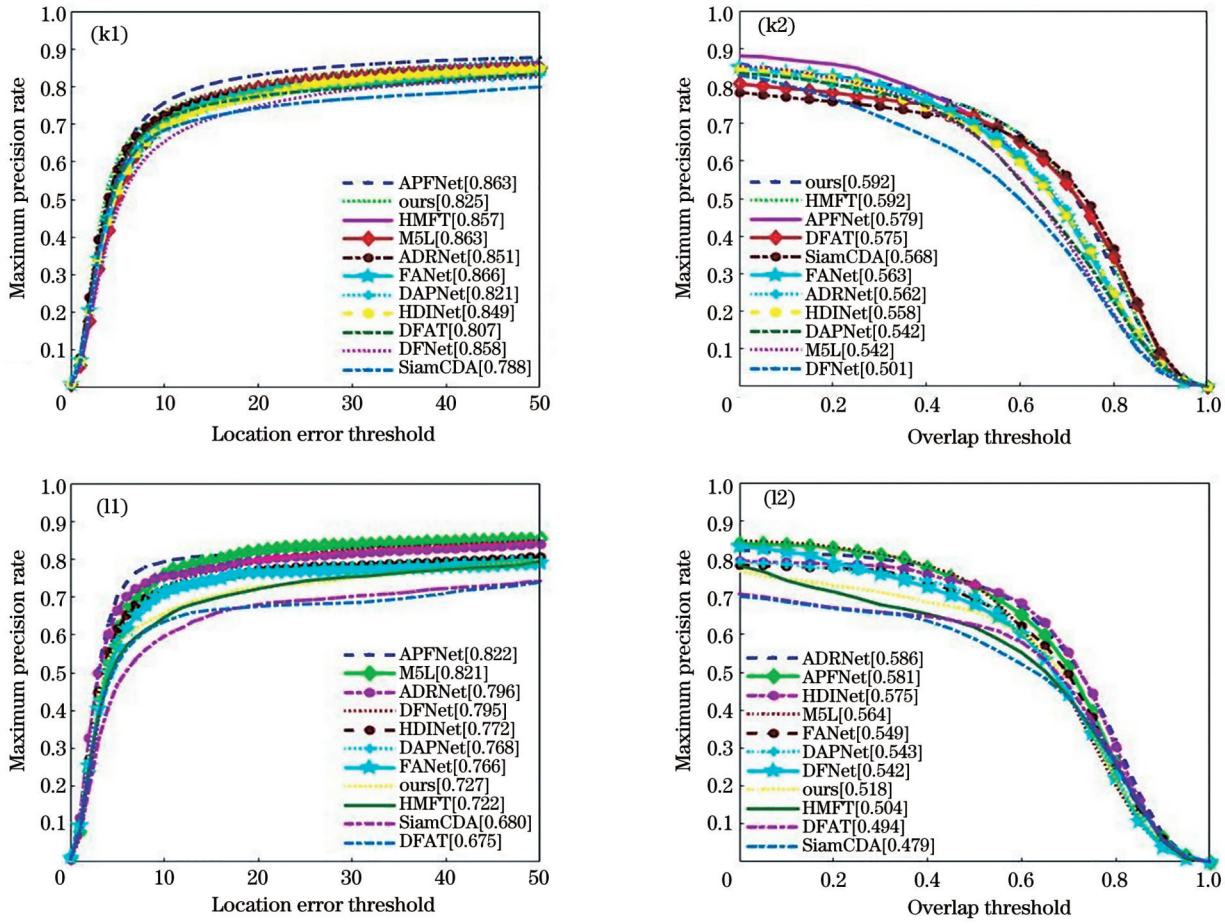


图 8　RGBT234 数据集上基于 12 种挑战属性的实验对比结果。(a1)(a2)背景杂波;(b1)(b2)相机移动;(c1)(c2)形变;(d1)(d2)快速移动;(e1)(e2)严重遮挡;(f1)(f2)低光照;(g1)(g2)低分辨率;(h1)(h2)移动模糊;(i1)(i2)无遮挡;(j1)(j2)部分遮挡;(k1)(k2)尺度形变;(l1)(l2)热交叉

Fig. 8　Experimental results on RGBT234 dataset based on 12 challenge attributes. (a1)(a2) Background clutter; (b1)(b2) camera moving; (c1)(c2) deformation; (d1)(d2) fast moving; (e1)(e2) heavy occlution; (f1)(f2) low illumination; (g1)(g2) low resolution; (h1)(h2) motion blur; (i1)(i2) no occlution; (j1)(j2) partial occlution; (k1)(k2) scale variation; (l1)(l2) thermal crossover

3.2.3　LasHeR 数据集性能评估

LasHeR 数据集的规模更大且更为复杂,能够全面展现跟踪器的跟踪性能。图 9 展示了 SiamCTU 与 MANet[15]、CAT[16]、FANet[31]、DFAT[35]、DAPNet[36]、

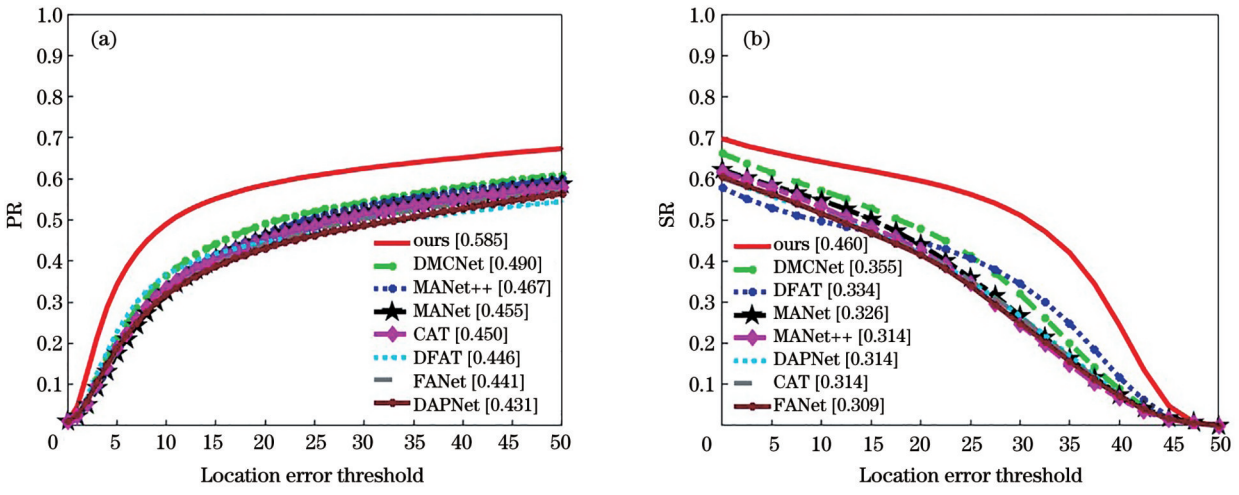


图 9　SiamCTU 与 7 种先进的跟踪器在 LasHeR 数据集上的对比结果。(a)精确率;(b)成功率

Fig. 9　Comparison of SiamCTU with 7 advanced trackers on LasHeR dataset. (a) PR; (b) SR

MANet++[38]和DMCNet[39]等7种先进的跟踪器的对比评估结果。由图9中的实验结果可见，SiamCTU在PR和SR指标上均获得了第一的排名，并且与排名第二的跟踪器DMCNet相比，分别在PR和SR得分上提升了约19.4%和29.6%。

3.2.4　跟踪速度分析

跟踪速度也是衡量跟踪器性能的重要指标之一。为全面评估SiamCTU的跟踪性能，在GTOT数据集上将其与CAT[16]、APFNet[17]、SiamCDA[19]、CMPP[30]和M5L[33]等5种先进跟踪器的跟踪速度进行了比较。由表2可见，与同样基于Siamese网络的跟踪器SiamCDA[19]相比，SiamCTU在跟踪速度减小的情况下，跟踪精度获得大幅度提升，且速度和精度均优于APFNet[17]等基于MDNet的先进跟踪器。上述实验结果充分证明了SiamCTU性能的优越性。

表2　基于GTOT数据集的跟踪效率对比实验结果

Table 2　Experimental results of tracking efficiency comparison based on GTOT dataset　　unit: %

Metric	CMPP	APFNet	SiamCDA	CAT	M5L	Ours
FPS	1.3	1.4	37.0	20.0	9.7	30.0
PR/SR	92.6/73.8	90.5/73.7	87.7/73.2	88.9/71.7	89.6/71.0	94.0/75.6

3.3　定性分析

为更直观地比较跟踪器的跟踪性能，在GTOT和RGBT234数据集上各选取2个具有挑战性的视频序列，将SiamCTU与FANet[31]、ADRNet[32]、SGT[40]等3种先进跟踪算法以及视频标注GT进行比较，并将跟踪结果进行可视化，可视化结果如图10所示。其中，图10(a)和图10(b)分别来自GTOT数据集中的LightOcc视频序列和Cycling视频序列，图10(c)和图10(d)分别来自RGBT234数据集中的night2视频序列和baketballwaliking视频序列。LightOcc序列上的跟踪结果展示了跟踪器在面临遮挡、低光照和小目标等多种挑战属性下的跟踪性能。如图10(a)所示，在跟踪的最后阶段只有SiamCTU和ADRNet跟踪器跟踪成功，且整个跟踪过程仅有SiamCTU能给出目标的准确位置。从图10(b)可以看出，针对目标发生大尺度变化等情况，SiamCTU较对比跟踪器的表现更佳，这充分证明了基于无锚框思想构建预测网络的优势。由图10(c)中night2序列的可视化结果可见，SiamCTU在面临低光照、低分辨率、相机移动和相似物干扰等多种挑战属性时，依然能取得良好的跟踪性

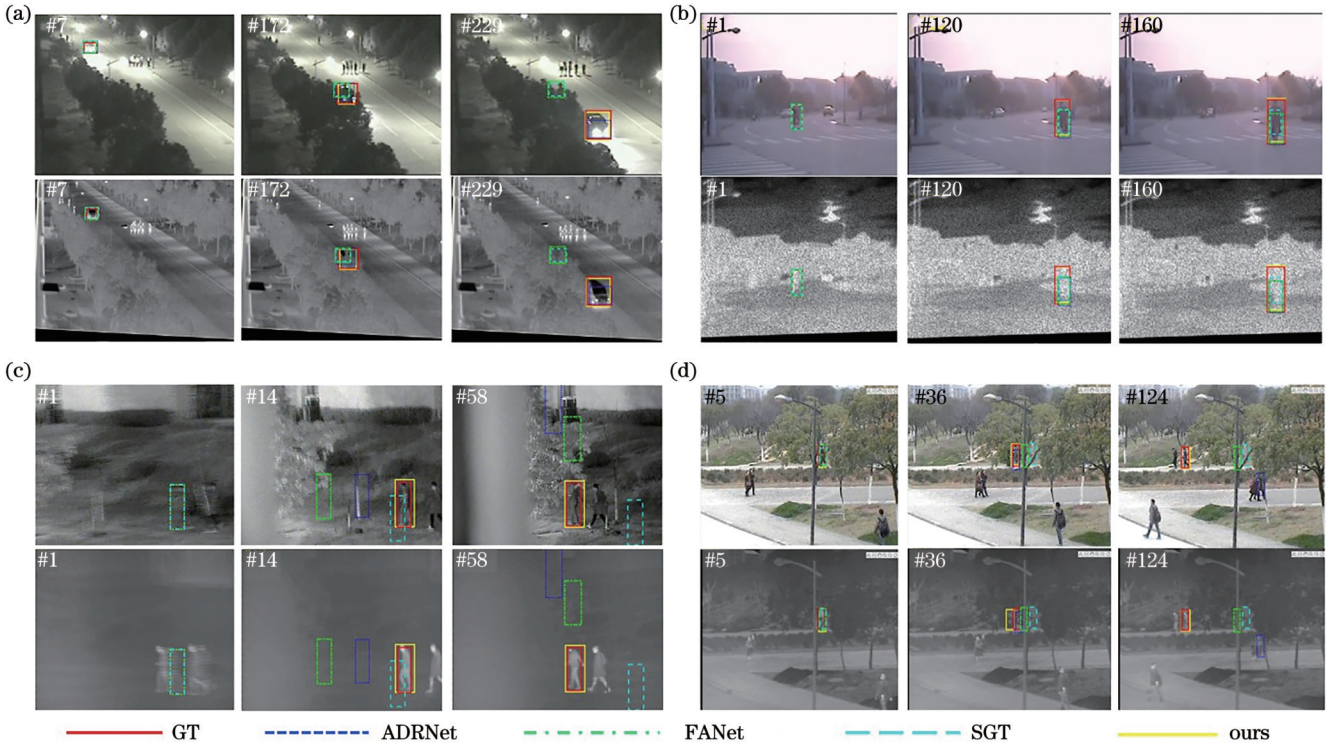


图10　在GTOT和RGBT234数据集上的可视化跟踪对比结果。(a) LightOcc序列；(b) Cycling序列；(c) night2序列；(d) baketballwaliking序列

Fig. 10　Visual tracking comparison results on GTOT and RGBT234 datasets. (a) LightOcc sequences; (b) cycling sequences; (c) night2 sequences; (d) baketballwaliking sequences

能。当出现图 10(d)中的严重遮挡并伴有相似物干扰时，SiamCTU 依然能取得较高的跟踪精度，这充分说明 SiamCTU 可以有效利用 RGB 模态和 T 模态信息的互补优势，提升复杂环境下的跟踪性能。

3.4 消融实验

3.4.1 网络模块

为了验证所设计的各个模块对算法产生的影响，设计如表 3 所示的消融实验。其中，Baseline 表示基于 Siamese 网络和无锚框的思想构建的基线跟踪器，CB(center-ness branch)表示中心度分支，FIM 表示所设计的特征交互模块，TU 表示模板更新。

表 3 在 3 个基准数据集上的消融实验结果(PR/SR)

Table 3 Results of ablation experiments on 3 baseline datasets (PR/SR) unit: %

Method	GTOT	RGBT234	LasHeR
Baseline	88.2/72.6	75.3/53.9	54.8/42.6
Baseline+CB	90.7/72.4	75.4/54.1	55.4/43.5
Baseline+CB+FIM	91.8/74.7	78.5/56.6	57.2/44.6
Baseline+CB+TU	91.9/73.1	76.1/54.0	57.1/44.4
Baseline+CB+FIM+TU(ours)	94.0/75.6	79.2/57.1	58.5/46.0

对比基线跟踪器：仅将中心度分支引入到跟踪器中，算法在 GTOT 数据集上的 PR 值获得约 2.8% 的提升，同时，在 RGBT234 和 LasHeR 数据集上的 PR 和 SR 均获得小幅度提升；添加特征交互模块，但不进行模板更新，跟踪器在 RGBT234 和 LasHeR 数据集上的 PR 值分别获得约 4.2% 和 4.4% 的提升，SR 值分别获得约 2.9% 和 4.7% 的提升；利用模板更新分支更新网络的跟踪模板，跟踪器在 3 个数据集上的性能均再次获得大幅度提升，其中，在 GTOT 数据集的 PR 得分提升了约 2.2%。同时，将 SiamCTU 与未添加特征交互模块的跟踪算法进行对比，由表 3 中的实验结果可见，SiamCTU 去除特征交互模块后在 3 个基准数据集上的跟踪性能均出现不同程度的下降，尤其是在较为复杂的 RGBT234 数据集上，其 PR 和 SR 值约下降了 3.1%，这充分验证了所设计的特征交互模块有效增强模态间信息交流并提升跟踪性能的能力。特别地，SiamCTU 跟踪性能远超基线跟踪器，这说明所设计的各个模块对跟踪性能有较大贡献，提升了跟踪器应对复杂跟踪场景的能力。

3.4.2 模板更新参数和更新方式

为研究不同模板更新参数对跟踪性能的影响，在 GTOT 数据集上设置了如表 4 所示的消融实验。其中，$\lambda=0$ 和 $\lambda=1$ 分别表示只使用在线模板和只使用初始模板进行跟踪。从实验结果可以看出，当 λ 的取值为 0.7 时，跟踪器取得最佳跟踪效果，且跟踪性能远超只使用在线模板和只用固定模板的跟踪方式。特别地，当只使用在线模板进行跟踪时，跟踪器在 GTOT 数据集上的 PR 和 PR 分别下降 23% 和 17.4%。上述实验结果表明，当 λ 取合适的更新参数时，基于特征级的模板更新方式可显著提升跟踪器的性能。

表 4 在 GTOT 数据集上不同模板更新权重的实验结果

Table 4 Experimental results of updating weights with different templates on GTOT dataset unit: %

Metric	$\lambda=0$	$\lambda=0.5$	$\lambda=0.6$	$\lambda=0.7$	$\lambda=0.8$	$\lambda=0.9$	$\lambda=1$
PR/SR	71.0/58.2	86.9/72.6	91.2/73.6	94.0/75.6	91.7/74.6	91.8/74.5	91.8/74.7

4 结 论

针对复杂环境下的目标跟踪问题，提出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。跟踪模型以 Siamese 网络作为基本框架，设计了特征交互模块，通过重构不同光学模态的信息比例增强模态间信息交流，从而降低复杂背景对跟踪性能的影响。通过分析跟踪器初始模板与在线模板的关系，提出一种模板动态更新策略，利用预测结果动态更新跟踪模板，获取目标的实时状态，进而提升算法的鲁棒性。在 GTOT、RGBT234 和 LasHeR 等 3 个基准数据集上的评估结果表明，与当前先进目标跟踪方法相比，所提方法的跟踪精度更高，同时能够满足实时的跟踪要求，该方法在复杂环境下的目标光学信息探测、感知和识别等方面具有广阔的应用前景。

参 考 文 献

[1] 陈汶铭，洪濡，盖绍彦，等. 基于特征融合和相似度估计网络的三维多目标跟踪[J]. 光学学报，2022，42(16)：1615001.
Chen W M，Hong R，Gai S Y，et al. Three-dimensional multi-object tracking based on feature fusion and similarity estimation network[J]. Acta Optica Sinica，2022，42(16)：1615001.

[2] 杨静，马龙. 基于位置感知的热红外目标跟踪方法[J]. 激光与光电子学进展，2023，60(12)：1210007.
Yang J，Ma L. Thermal infrared object tracking method based on positional perception[J]. Laser & Optoelectronics Progress，2023，60(12)：1210007.

[3] Liang Z Q，Wang J S，Xiao G，et al. FAANet：feature-aligned attention network for real-time multiple object tracking in UAV videos[J]. Chinese Optics Letters，2022，20(8)：081101.

[4] 蔡旺，王栋梁，冯伟，等. 基于激光传感的水下声学目标高分辨跟踪方法[J]. 中国激光，2022，49(18)：1810004.
Cai W，Wang D L，Feng W，et al. High-resolution acoustic tracking method for underwater target using laser-based sensor [J]. Chinese Journal of Lasers，2022，49(18)：1810004.

[5] 赵丹露，张永安，何光辉，等. 透烟雾红外数字全息像的亮度

增强算法[J]. 中国激光, 2023, 50(18): 1809001.
Zhao D L, Zhang Y A, He G H, et al. Brightness enhancement algorithm for infrared digital holographic image through smoke [J]. Chinese Journal of Lasers, 2023, 50(18): 1809001.

[6] Wang F T, Wang W Q, Liu L, et al. Siamese transformer RGBT tracking[J]. Applied Intelligence, 2023, 53(21): 24709-24723.

[7] Li M Y, Zhang P, Yan M, et al. Dynamic feature-memory transformer network for RGBT tracking[J]. IEEE Sensors Journal, 2023, 23(17): 19692-19703.

[8] Conaire C Ó, O'Connor N E, Smeaton A. Thermo-visual feature fusion for object tracking using multiple spatiogram trackers[J]. Machine Vision and Applications, 2008, 19(5): 483-494.

[9] Li C L, Sun X, Wang X, et al. Grayscale-thermal object tracking via multitask Laplacian sparse representation[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017, 47(4): 673-681.

[10] 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.

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

[12] 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.

[13] Zhang P Y, Zhao J, Wang D, et al. Visible-thermal UAV tracking: a large-scale benchmark and new baseline[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 8876-8885.

[14] 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. New York: IEEE Press, 2016: 4293-4302.

[15] 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, Republic of Korea. New York: IEEE Press, 2019: 2262-2270.

[16] Li C L, Liu L, Lu A D, et al. Challenge-aware RGBT tracking [M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12367: 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] Zhang X C, Ye P, Peng S Y, et al. SiamFT: an RGB-infrared fusion tracking method via fully convolutional Siamese networks [J]. IEEE Access, 2019, 7: 122122-122133.

[19] 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.

[20] 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. New York: IEEE Press, 2019: 4277-4286.

[21] 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, July 17-22, 2022, Kuala Lumpur, Malaysia. New York: IEEE Press, 2022: 803-806.

[22] Guo D Y, Wang J, Cui Y, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6268-6276.

[23] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[M]//Hua G, Jégou H. Computer vision-ECCV 2016 workshops. Lecture notes in computer science. Cham: Springer, 2016, 9914: 850-865.

[24] 张立国, 马子荐, 金梅, 等. 基于非局部感知网络的运动目标跟踪方法[J]. 激光与光电子学进展, 2023, 60(4): 0415007.
Zhang L G, Ma Z J, Jin M, et al. Nonlocal neural network-based moving target tracking method[J]. Laser & Optoelectronics Progress, 2023, 60(4): 0415007.

[25] Chen X, Yan B, Zhu J W, et al. Transformer tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 8122-8131.

[26] Zhang Z P, Peng H W. Deeper and wider Siamese networks for real-time visual tracking[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 4586-4595.

[27] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[28] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[29] 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, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 248-255.

[30] 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. New York: IEEE Press, 2020: 7062-7071.

[31] 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.

[32] 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.

[33] 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.

[34] Peng J C, Zhao H T, Hu Z W. Dynamic fusion network for RGBT tracking[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(4): 3822-3832.

[35] Tang Z Y, Xu T Y, Li H, et al. Exploring fusion strategies for accurate RGBT visual object tracking[J]. Information Fusion, 2023, 99: 101881.

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

[37] 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.

[38] 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.

[39] Lu A D, Qian C, Li C L, et al. Duality-gated mutual condition network for RGBT tracking[EB/OL]. (2022-04-29)[2023-11-09]. https://ieeexplore.ieee.org/document/9737634.
[40] Li C L, Zhao N, Lu Y J, et al. Weighted sparse representation regularized graph learning for RGB-T object tracking[C]// Proceedings of the 25th ACM international conference on Multimedia, October 23-27, 2017, Mountain View, California, USA. New York: ACM Press, 2017: 1856-1864.

Cross-Modal Optical Information Interaction and Template Dynamic Update for RGBT Target Tracking Method

Chen Jianming[1,2], Li Dingjian[1], Zeng Xiangjin[1,2], Ren Zhenbo[3], Di Jianglei[1*], Qin Yuwen[1,2**]

[1]*Key Laboratory of Photonic Technology for Integrated Sensing and Communication, Ministry of Education, Guangdong Provincial Key Laboratory of Information Photonics Technology, School of Information Engineering of Guangdong University of Technology, Institute of Advanced Photonics Technology, Guangzhou* 510006, *Guangdong, China*;

[2]*Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai* 519082, *Guangdong, China*;

[3]*Key Laboratory of Light-Field Manipulation and Information Acquisition, Ministry of Industry and Information Technology, Shaanxi Key Laboratory of Photonics Technology for Information, School of Physical Science and Technology, Northwestern Polytechnical University, Xi'an* 710129, *Shaanxi, China*

Abstract

Objective RGB and thermal infrared (RGBT) tracking technology fully leverages the complementary advantages of different optical modalities, providing effective solutions for target tracking challenges in complex environments. However, the performance of many tracking algorithms is constrained due to the neglect of information exchange between modalities. Simultaneously, as the tracking template remains fixed, existing tracking methods based on Siamese networks face limitations in adapting to variations in target appearance, resulting in tracking drift. Therefore, enhancing the performance of target trackers in complex environments remains challenging.

Methods The proposed algorithm adopts the Siamese network tracker as its foundational framework and introduces a feature interaction module to enhance inter-modal information exchange by reconstructing information proportions of different modalities. Based on the anchor-free concept, a prediction network is directly constructed to perform classification and regression on the target bounding box at each position point in the search region. To address the mismatch between the target and template during the tracking of the Siamese network tracker, we propose a template update strategy, which dynamically updates the tracking template using the predicted results from the previous frame.

Results and Discussions Qualitative and quantitative experiments are carried out on SiamCTU and advanced RGBT target tracking models, with ablation experiments analyzed. Meanwhile, comparative experiments are conducted by evaluating the proposed target tracker against state-of-the-art target trackers on three benchmark datasets (GTOT, RGBT234, and LasHeR) to assess the tracking performance of the algorithm. Figs. 6, 7, and 9 respectively display the quantitative comparison results between SiamCTU and advanced RGBT tracking algorithms on the three benchmark datasets. Compared with advanced RGBT target tracking algorithms, the experimental results on three baseline datasets demonstrate outstanding tracking performance of SiamCTU, fully exhibiting the effectiveness of the proposed method. Specifically, on the GTOT and LasHeR datasets, the proposed tracking algorithm secures top rankings in both PR and SR. Fig. 8 and Table 1 respectively present the experimental results based on challenge attributes for the tracking algorithm on the GTOT and RGBT234 datasets. The experimental results show that SiamCTU exhibits excellent tracking performance under various challenging attributes, suggesting that the proposed tracker is effective in handling complex target tracking scenarios. To provide a more intuitive demonstration of the tracker's tracking performance, we visualize the tracking results in Fig. 10. In the LightOcc sequence [Fig. 10(a)], the proposed tracking algorithm utilizing the template update strategy maintains continuous and stable tracking of the target even under such challenges as occlusion and low illumination. For scenarios involving significant scale variations [Fig. 10(b)], the proposed tracker outperforms the comparative tracker, demonstrating the advantages of constructing a prediction network based on the anchor-free concept. The visual results in Figs. 10(c) and 10(d) reveal that the proposed tracker can leverage the complementary advantages of

RGB and T modalities, reducing interference from similar objects. Meanwhile, the comparative tracking efficiency analysis of the tracker on the GTOT dataset (Table 2) indicates that SiamCTU significantly improves tracking accuracy with minimal tracking speed loss. Furthermore, the proposed tracker exhibits higher speed and precision advantages over the advanced MDNet-based tracker. In further ablation experiments (Table 3), the performance of the proposed tracker surpasses that of the baseline tracker, which underscores the substantial contributions of various modules designed in the algorithm and collectively enhances the tracker's ability to handle complex tracking scenarios. Specifically, when the feature interaction module is removed, the overall performance of SiamCTU decreases by 3.1% on the more complex RGBT234 dataset. Additionally, by varying template update parameters to study their influence on tracking performance, experimental results (Table 4) indicate that with an appropriate value of λ as the update parameter, the feature-level template update method can significantly enhance the tracker's performance.

Conclusions To address the target tracking challenges in complex environments, we propose a cross-modal optical information interaction method for RGBT target tracking. The tracking model adopts the Siamese network as its foundational framework and incorporates a feature interaction module. This module enhances the inter-modal information exchange by reconstructing information proportions of different optical modalities, mitigating the effect of complex backgrounds on tracking performance. Subsequently, by dealing with the relationship between the tracker's initial template and the online template, we introduce a template dynamic updating strategy. This strategy dynamically updates the tracking template using predicted results, capturing the real-time status of the target and improving the algorithm's robustness. Evaluation results on three benchmark datasets including GTOT, RGBT234, and LasHeR demonstrate that the proposed method surpasses current advanced RGBT target tracking methods in terms of tracking accuracy. Additionally, it meets real-time tracking requirements and holds potential for broad applications in optical information detection, perception, and recognition of targets in complex environments.

Key words machine vision; computer vision; object tracking; Siamese network; template update