Modeling and Simulation 建模与仿真, 2025, 14(6), 99-109
Published Online June 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.146479

Hans汉斯

基于轻量孪生网络的无人机RGB-T目标跟踪算法

刘哲宇，魏　赟

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年5月12日；录用日期：2025年6月5日；发布日期：2025年6月12日

摘　要

随着无人机在目标跟踪领域的广泛应用，尤其是在复杂环境(如低光照、恶劣天气)下的追踪效果难以保障，可见光与热红外(RGB-T)双模态数据融合成为提升跟踪性能的关键手段。然而，这种融合面临异构特征高效交互、视角差异及计算资源受限等挑战。本文提出一种基于孪生网络的轻量化目标跟踪算法SiamTSA (Siamese Network with Temporal and Spatial Attention)。首先，采用改进的MobileNetV3-small作为主干网络，降低计算开销并适配无人机平台；其次，设计跨模态时空交互注意力模块，通过时间注意力建模视觉风格差异和空间注意力对齐视角差异，抑制冗余噪声并增强跨模态一致性特征表达；进一步提出双模态自适应惩罚选择模块，通过分析预测框的尺度与宽高比变化筛选更优输出框，提升了跟踪框的稳定性。在GTOT、RGBT234及VTUAV数据集上的实验表明，SiamTSA在跟踪成功率(VTUAV: 67.5%)与实时性(56.3 FPS)方面均优于主流算法，兼顾精度与效率。本文方法为复杂场景下的无人机多模态目标跟踪提供了轻量化解决方案。

关键词

目标跟踪，无人机，RGB-T融合，轻量化网络，时空注意力

RGB-T Tracking Algorithm for Unmanned Aerial Vehicles Based on Lightweight Siamese Network

Zheyu Liu, Yun Wei

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: May 12th, 2025; accepted: Jun. 5th, 2025; published: Jun. 12th, 2025

文章引用: 刘哲宇, 魏赟. 基于轻量孪生网络的无人机 RGB-T 目标跟踪算法[J]. 建模与仿真, 2025, 14(6): 99-109.
DOI: 10.12677/mos.2025.146479

Abstract

With the widespread application of unmanned aerial vehicles (UAVs) in object tracking, especially the increasing demand for robust performance in complex environments (e.g., low-light conditions, adverse weather), the fusion of visible and thermal infrared (RGB-T) multimodal data has become a critical approach to enhance tracking accuracy. However, this fusion faces challenges such as efficient interaction of heterogeneous features, perspective differences, and limited computational resources. This paper proposes a lightweight object tracking algorithm named SiamTSA (Siamese Network with Temporal and Spatial Attention). First, an improved MobileNetV3-small is adopted as the backbone to reduce computational costs and adapt to UAV platforms. Second, a cross-modal temporal spatial interaction attention module is designed to model visual style differences via temporal attention and align spatial discrepancies via spatial attention, thereby suppressing redundant noise and enhancing cross-modal consistent feature representation. Furthermore, a dual-modal adaptive penalty selection module enhances tracking stability by selecting optimal bounding boxes through analysis of scale and aspect ratio variations. Experiments on GTOT, RGBT234, and VTUAV datasets demonstrate that SiamTSA outperforms state-of-the-art methods in tracking success rate (VTUAV: 67.5%) and real-time performance (56.3 FPS), balancing accuracy and efficiency. The proposed method provides a lightweight solution for UAV-based multimodal object tracking in complex scenarios.

Keywords

Object Tracking, Unmanned Aerial Vehicle (UAV), RGB-T Fusion, Lightweight Network, Temporal Spatial Attention

1. 引言

随着计算机视觉技术的快速发展，目标跟踪在安防监控、自动驾驶、军事侦察等领域的应用需求日益增长。无人机具有低成本，操作灵活，体积小等优点，因此无人机跟踪[1]在近年来已经引起了广泛的关注，然而无人机跟踪面临着许多挑战：1) 无人机视角移动变化大，导致目标姿态尺度易发生较大变化。2) 应用场景较为复杂，极易受到天气光照变化的影响。3) 无人机平台计算资源十分有限，对算法性能要求高。针对第二点问题，现有的许多无人机平台引入了热红外摄像头来应对。在基于 RGB-T [2]的目标跟踪方向，SiamFT [3]采用了沿通道维度进行多模态特征拼接的方案，SiamCDA [4]提出了通过自适应权重进行跨模态特征残差连接的方案，但目前研究仍存在不足，比如多模态图像数据有着天然的异构性，图像特征难以高效融合，并且无人机通常采用两个不同的摄像头来进行拍摄，获取的双模态数据在空间视角上存在着差异性。

为了解决上述问题，本文提出一个基于孪生网络的轻量化目标跟踪算法 SiamTSA (Siamese Network with Temporal and Spatial Attention)，为了保证算法在无人机平台轻量化的需求，选择了基于 MobileNetV3-small 算法的特征提取主干网络。同时为了解决多模态图像特征难以高效融合，以及无人机多模态图像天然存在的空间视角差异问题，本文提出了一个跨模态时空交互注意力模块，将具有明显视觉风格差异的双模态图像特征视作具有时间差异的图像特征，计算其时间注意力分数与空间注意力分数，充分抑制因模态风格差异导致的冗余特征(如 RGB 中过曝区域与红外中低温噪声)，增强跨模态一致的特征表达，同

时融合双模态视角信息，提升了目标跟踪过程中的鲁棒性。最后，本文设计了一个双模态自适应惩罚选择模块，通过预测框前后帧的比例变化以及尺度变化进行惩罚，抑制了尺度突变的预测框输出，缓解了目标跟踪中由于快速移动引起的跟踪框抖动以及目标形变导致的跟踪漂移问题。

2. 方法

文中提出的跟踪算法总体结构框架如图 1 所示，总共可以分为三个部分：1) 前端使用改进过的 MobileNetV3-small 作为主干网络对双模态的搜索图与模板图像分别进行特征提取。2) 由主干网络提取的双模态特征经过双模态时空交互注意力模块进行充分特征融合后再与原始特征相连接。3) 双模态图像特征分别进行深度互相关并由分类回归网络生成响应图，再经过双模态自适应惩罚选择模块得到最终输出结果。

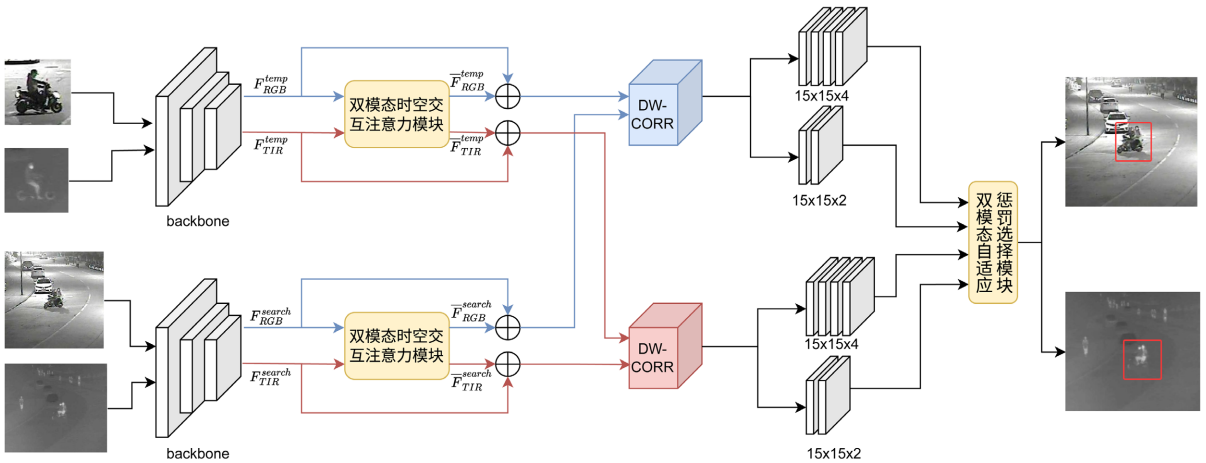

Figure 1. Overall framework of UAVs RGB-T object tracking algorithm based on lightweight Siamese network
图 1. 基于轻量孪生网络的无人机 RGB-T 目标跟踪算法总体框架

2.1. 特征提取网络

为了保证算法模型的轻量化以及对于无人机平台设备的适应性，本文使用 MobileNetV3-small [5]作为主干网络，相较于 MobileNetV2 [6]，MobileNetV3 引入轻量级注意力机制(SE 模块)与硬件感知的 h-swish 激活函数，在显著降低计算开销的同时，实现了多尺度特征的有效表征，其改进的倒残差结构进一步平衡了模型效率与特征提取能力的矛盾。为了使其适应目标跟踪任务，本文对网络结构做出了一定的调整。原始的 MobileNetV3 网络经过多次下采样得到目标图像的特征图，过大的总步长会导致输出图像特征分辨率过低，丢失部分特征信息，并且一定程度上会影响计算效率，受到论文[7]的启发，本文选择原始 MobileNetV3-small 的前 10 层作为主干网络，并且将最后一个 Bneck 层的步长由 2 调整为 1，进一步降低了网络参数量与计算量的同时将感受野控制在了一个合理的大小，避免了过大的感受野所导致的性能下降。表 1 为网络结构的具体参数，表中输入以搜索图像为例，输入尺寸大小为 255 × 255 × 3 的图像，经过主干网络后，得到特征尺寸为 16 × 16 × 96。

2.2. 双模态时空交互注意力模块

有效挖掘可见光与红外模态的互补特征并实现跨模态高效协同，是提升 RGB-T 目标跟踪算法鲁棒性、精度及环境适应能力的关键挑战，除了双模态特征本身存在视觉风格差异性之外，在无人机 RGB-T 目标跟踪中，还存在着双模态摄像头带来的空间视角差异问题。受论文[8]的启发，本文引入了一种时空交互注意力的双模态特征处理模块，通过视觉风格差异的时间注意力建模和空间视角差异的空间注意力

建模，自适应挖掘可见光模态与红外模态的时空依赖相关性，抑制跨模态噪声干扰，从而提升复杂场景下的跟踪鲁棒性，结构如图 2 所示。

Table 1. Structure parameters of backbone network
表 1. 主干网络结构参数

网络层	输入	输出	卷积核	步长	是否使用 SE
Conv1	255 × 255 × 3	128 × 128 × 16	3 × 3	2	无
Bneck1	128 × 128 × 16	64 × 64 × 16	3 × 3	2	是
Bneck2	64 × 64 × 16	32 × 32 × 24	3 × 3	2	否
Bneck3	32 × 32 × 24	32 × 32 × 24	5 × 5	1	否
Bneck4	32 × 32 × 24	16 × 16 × 40	5 × 5	2	是
Bneck5	16 × 16 × 40	16 × 16 × 40	5 × 5	1	是
Bneck6	16 × 16 × 40	16 × 16 × 40	5 × 5	1	是
Bneck7	16 × 16 × 40	16 × 16 × 48	5 × 5	1	是
Bneck8	16 × 16 × 48	16 × 16 × 48	5 × 5	1	是
Bneck9	16 × 16 × 48	16 × 16 × 96	5 × 5	1	是

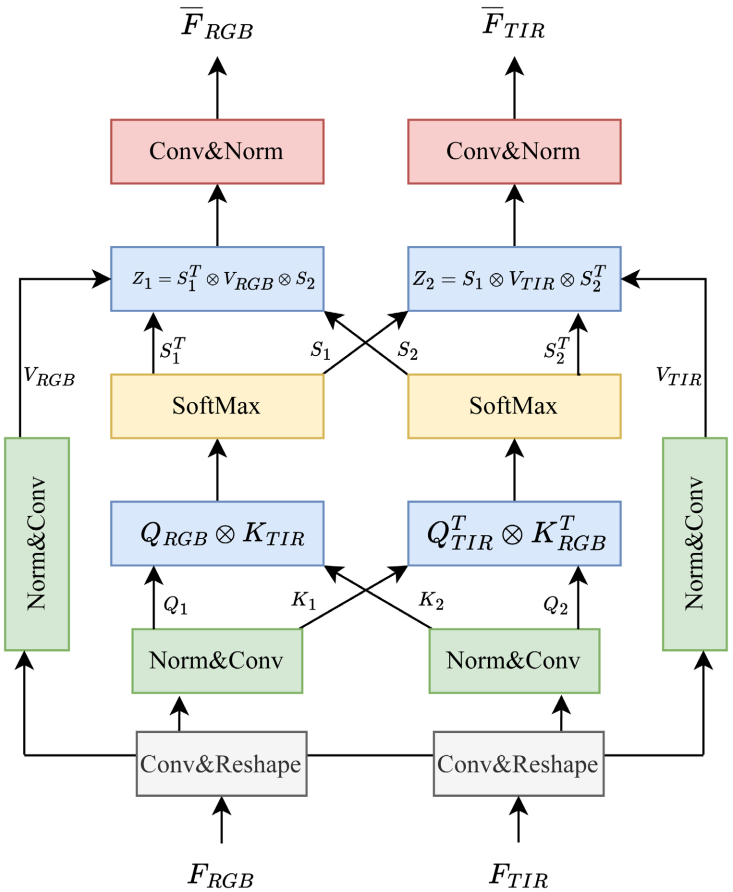

Figure 2. Dual-modality temporal-spatial interactive attention module
图 2. 双模态时空交互注意力模块

对于输入的特征$\{F_{RGB}|F_{TIR}\}\in\mathbb{R}^{C\times H\times W}$，首先通过1×1的卷积核进行通道压缩，再进行展平操作得到$\{F_{RGB}|F_{TIR}\}\in\mathbb{R}^{C'\times HW}$，其中$C'=C/2$。展平操作旨在降低计算复杂度，并使得特征矩阵适配注意力机制的序列化处理。接下来对于$\{F_{RGB}|F_{TIR}\}\in\mathbb{R}^{C'\times HW}$，分别进行归一化操作(Batch Normalization)和1×1卷积操作得到：

$$\left.\begin{cases}Q_i=\mathrm{Conv}_{(i)}\left(BN\left(F_i\right)\right)\\V_i=\mathrm{Conv}\left(BN\left(F_i\right)\right)\\K_i=Q^T\end{cases}\right|i\in\left(RGB,TIR\right)\tag{1}$$

其中$Q_i,V_i\in\mathbb{R}^{C'\times HW}$，$K_i\in\mathbb{R}^{HW\times C'}$，其中$Q_i$不共享参数，$V_i$共享参数。

然后分别计算视觉风格依赖时间注意力S_1与空间位置依赖空间注意力S_2，如式(2)所示。其中$S_1\in\mathbb{R}^{C'\times C'}$表征了可见光模态和红外模态在不同通道维度上视觉风格的相似性(将视觉风格上的差异视做时间风格的差异)，$S_2\in\mathbb{R}^{HW\times HW}$表征了具有视角差异的可见光模态和红外模态在空间维度上的空间位置相似性。

$$\begin{cases}S_1=\mathrm{softmax}\left(Q_{RGB}\otimes K_{TIR}\right)\\S_2=\mathrm{softmax}\left(Q_{TIR}^T\otimes K_{RGB}^T\right)\end{cases}\tag{2}$$

如式(3)所示，利用矩阵S_1与S_2对V_i分别进行重建可以得到$Z_i\in\mathbb{R}^{C'\times HW}\left|i\in\left(RGB,TIR\right)\right.$。

$$\begin{cases}Z_{RGB}=S_1^T\otimes V_{RGB}\otimes S_2\\Z_{TIR}=S_1\otimes V_{TIR}\otimes S_2^T\end{cases}\tag{3}$$

将重建结果通过重塑操作恢复到$Z_i\in\mathbb{R}^{C'\times H\times W}$，再通过$1\times1$的卷积核将通道由$C'$恢复至$C$并进行归一化操作，得到与原始特征大小相同的特征$\overline{F}_i=\mathrm{BN}\left(\mathrm{Conv}\left(Z_i\right)\right)\left|i\in\left(RGB,TIR\right)\right.$。

2.3. 分类回归网络

经过双模态时空交互注意力模块得到的特征$\overline{F}_{RGB}$与$\overline{F}_{TIR}$分别与原始特征F_{RGB}与F_{TIR}连接以保证原始特征信息的保留，共得到$\hat{F}_{RGB}^{\mathrm{temp}}$，$\hat{F}_{TIR}^{\mathrm{temp}}$，$\hat{F}_{RGB}^{\mathrm{search}}$，$\hat{F}_{TIR}^{\mathrm{search}}$四个特征，为获取更丰富的通道语义信息，本文采用了经典目标跟踪算法 SiamRPN++ [9]和 SiamMASK [10]等使用的深度互相关(Depth-Wise Correlation)操作，分别对可见光模态与红外模态上的搜索特征与模板特征进行深度互相关以得到分类响应图和回归响应图。

在分类分支中，分类响应图上的位置(i,j)映射到搜索图像上，对应搜索图像(x,y)在标注框内则代表是前景，否则为背景。分类损失采用二分类交叉熵损失函数进行训练，其数学表达式为：

$$L_{cls}=-\sum_{(i,j)}y_{(i,j)}\log\left(p_{(i,j)}\right)+\left(1-y_{(i,j)}\right)\log\left(1-p_{(i,j)}\right)\tag{4}$$

在回归分支中，目标框的回归精度会直接影响跟踪器的鲁棒性。本文在回归分支中采用 CIoU (Complete Intersection over Union)损失函数，相较于 IoU 损失函数，CIoU 同时考虑了重叠区域面积、中心点距离与宽高比，使得回归损失计算更加精确，在目标框部分重叠或完全包含等复杂场景下也可提供更准确的梯度方向，从而加速模型收敛，其公式定义为：

$$L_{reg}=1-\mathrm{IoU}+\frac{\rho^2\left(b,b^{gt}\right)}{c^2}+\alpha v\tag{5}$$

其中 IoU 代表了预测边框与真实边框的重叠度得分，b 与 b^{gt} 分别代表了预测框与真实框的中心位置，ρ 代表了预测框与真实框中心点之间的欧氏距离(Euclidean Distance)，c 代表了能够同时包围预测框和真实框的最小矩形框的对角线长度，ν 是用来衡量预测框与真实框长宽比的一致性参数，α 是用于平衡高宽比惩罚项 ν 的权重系数，α，ν 的计算公式分别如式(6)和式(7)所示：

$$\alpha = \frac{\nu}{(1-\mathrm{IoU})+\nu} \tag{6}$$

$$\nu = \frac{4}{\pi^2}\left(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h}\right)^2 \tag{7}$$

其中 ω^{gt}，h^{gt} 分别代表真实框的宽和高，ω，h 分别代表预测框的宽和高。

网络的整体损失函数由可见光损失和红外损失共同构成，总损失的定义可以表示为式(8)：

$$L_{\text{total}} = \partial_1 L_{RGB} + \partial_2 L_{TIR} \tag{8}$$

其中 L_{RGB} 代表可见光模态损失，L_{TIR} 代表红外模态损失，∂_1 与 ∂_2 为调整两个模态收敛速度的权重因子。对于其中任一模态的损失可分为分类损失和回归损失两个部分，公式可以表示为式(9)：

$$L_{\text{loss}} = \lambda_1 L_{cls} + \lambda_2 L_{reg} \tag{9}$$

其中 λ_1 与 λ_2 为调节回归损失与分类损失的权重因子。

2.4. 双模态自适应惩罚选择模块

为了从分类回归网络输出的双模态跟踪结果中得到最好的预测框，本文设计了一个双模态自适应惩罚选择模块。首先选择保留分类响应图中的前景通道 $A_{cls} \in \mathbb{R}^{H\times W}$ 作为目标得分图，同时考虑前景的最高得分与平均得分，确保目标空间位置的准确性的同时抵抗局部噪声干扰，具体如式(10)所示：

$$\mathrm{ConfScore} = \alpha \times \mathrm{PeakScore} + (1-\alpha) \times \mathrm{MeanScore} \tag{10}$$

其中 PeakScore 为最高得分，MeanScore 为平均得分，α 为超参数。同时为了抑制预测框幅度过大的比例与大小变化，本文受[11]启发设计了一个惩罚函数，根据前后帧目标框尺度的变化对目标得分进行调整，以此确保最终输出结果的稳定性，其定义如式(11)所示：

$$\mathrm{penalty} = \exp\left(-\left[\log^2\left(\frac{r}{r'}\right) + \log^2\left(\frac{s}{s'}\right)\right]\right) \tag{11}$$

其中 r 和 r' 分别表示当前帧候选框与上一帧预测框的宽高比，s 和 s' 分别表示当前帧与上一帧的尺度。当宽高比或尺度发生显著变化时，惩罚系数会相应降低目标得分，若宽高比与尺度变化微小，惩罚系数趋近于 1，目标得分基本保持不变。通过对数平方项的设计，该函数对宽高比和尺度的增减变化具有对称惩罚特性，确保不同变化方向的影响权重一致。最终目标得分由原始得分经惩罚函数修正后得出：

$$\mathrm{PredScore} = \mathrm{ConfScore} \times \mathrm{penalty} \tag{12}$$

对于可见光模态与红外模态的输出结果，分别使用上述方法进行处理，最终可决定输出分支为：

$$\mathrm{Output} = \max\left(\mathrm{PredScore}_{RGB}, \mathrm{PredScore}_{TIR}\right) \tag{13}$$

3. 实验分析

3.1. 实验细节

实验环境：本文在 Ubuntu18.04 操作系统下，使用的硬件环境配置为 CPU 型号 AMD Ryzen 7 5800H，

显卡 Nvidia GeForce RTX3080laptop，32GBRAM。软件环境为 python 版本 3.7.1，pytorch1.7.1，CUDA11.6 以及 CUDNN8.6。

训练细节：模板图像大小设置为 127 × 127 × 3，搜索图像大小设置为 255 × 255 × 3，孪生网络初始加载论文[5]中提供的 MobileNetV3 预训练模型，在 LasHeR [12]数据集与 VTUAV [13]数据集的训练集上进行训练，共训练 50 个 epoch，采用随机梯度下降法 SGD 进行训练，动量为 0.9，BatchSize 设置为 32，学习率初始为 0.001，在前 5 个 epoch 中上升至 0.005，后 45 个 epoch 中衰减至 0.0005，前 10 个 epoch 冻结主干网络参数，后 40 个 epoch 解冻主干网络参数进行整体训练。

3.2. 对比实验

将本文算法在主流的 RGB-T 数据集 GTOT 数据集，RGBT234 数据集以及针对无人机的 VTUAV 数据集上分别进行对比实验，图 3 展示了本文算法在具有挑战性的场景下展现出了显著的鲁棒性。

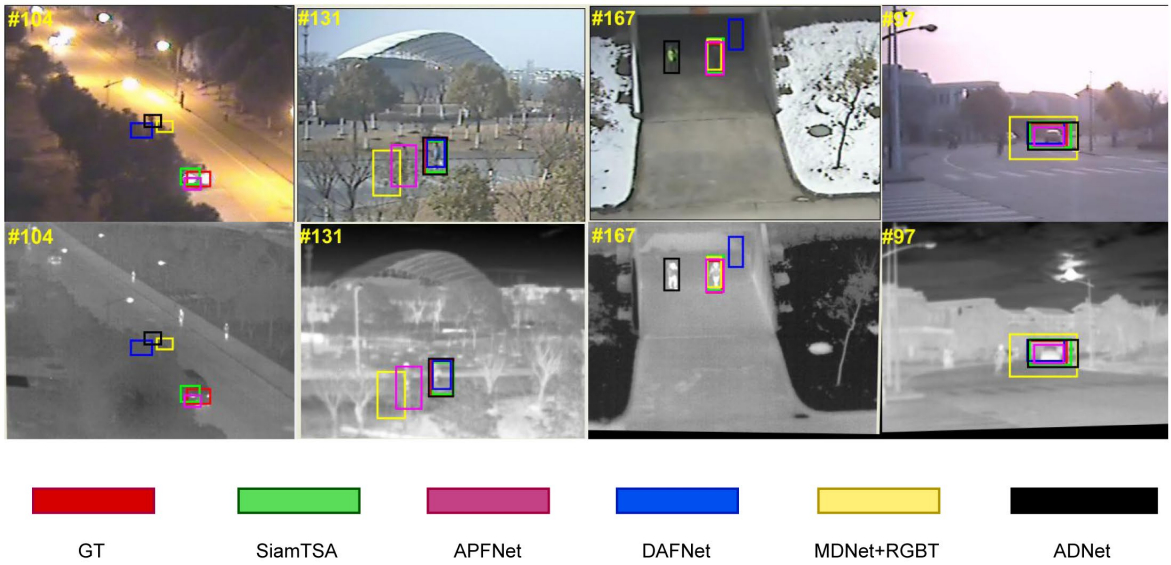

Figure 3. Performance of SiamTSA in challenging Scenarios
图 3. SiamTSA 在具有挑战性场景的表现

3.2.1. GTOT 数据集对比实验

GTOT [14]数据集包含 50 组 RGB 与热红外视频对，总帧数约 7.8 K，覆盖实验室、道路等 16 类场景。目标以行人、车辆为主。将文中提出算法与 APFNet [15]，TFNet [16]，MANet [17]，DAPNet [18]，ADNet [19]，SiamDW + RGBT [7]等 RGBT 跟踪算法进行对比。在 GTOT 跟踪基准上，本文算法与对比算法的准确率和成功率如图 4 所示。从实验结果可知，本文提出的 SiamTSA 算法跟踪成功率达到了 0.735，准确率达到了 0.907，分别排在第三和第一，在成功率和准确率具有竞争力的同时，本文算法在运算速度上有着明显的优势。

3.2.2. RGBT234 数据集对比实验

RGBT234 [20]由 RGBT210 拓展而来，包含 234 组 RGB 与热红外视频序列对，总帧数达 234 K，共有 12 种挑战属性(如遮挡、快速运动、尺度变化等)。在 RGBT 数据集上进行实验可以更好地评估算法在面对相对复杂场景时的鲁棒性，实验结果如图 5 所示，本文算法的成功率为 0.578，准确率为 0.814，排在第二，相较于 APFNet [15]，本文算法在成功率与准确率上稍逊，但 APFNet 是基于 MDNet [21]改进的

算法，依赖于在线更新机制，对于计算资源消耗大，不适合在低计算开销的无人机平台进行部署，而本文算法在较低的计算成本下实现了接近的跟踪性能，在 FPS 指标远超其他算法。

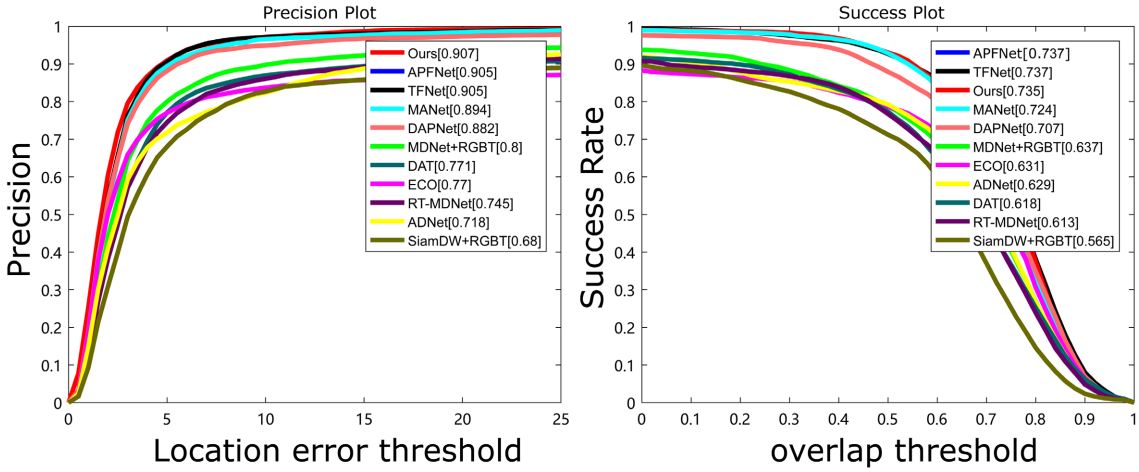

Figure 4. Comparative experimental results on the GTOT dataset
图 4. GTOT 数据集对比实验结果

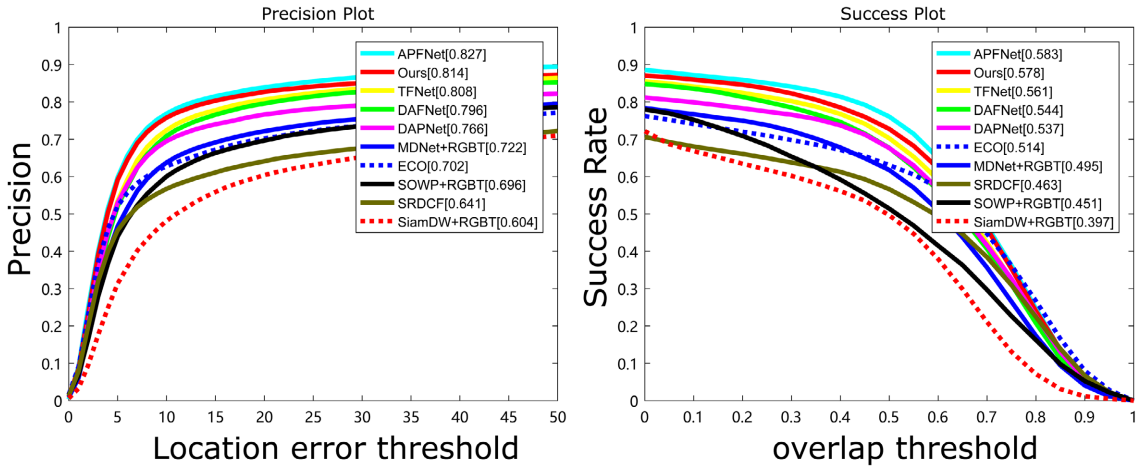

Figure 5. Comparative experimental results on the RGBT234 dataset
图 5. RGBT234 数据集对比实验结果

3.2.3. VTUAV 数据集对比实验

VTUAV [13] (Visible-Thermal UAV Tracking Benchmark)是由大连理工大学卢湖川团队于 2022 年构建的大规模可见光－热红外无人机视角目标跟踪数据集，其数据规模与场景多样性显著超越传统 RGB-T 数据集。该数据集包含 500 个高分辨率(1920 × 1080 像素)视频序列，总计 170 万对的 RGB-T 图像，划分为 250 个训练序列与 250 个测试序列，支持端到端模型训练与评估。在场景多样性方面，VTUAV 覆盖白天(325 序列)与夜间(175 序列)环境下的 5 类目标超类(行人、车辆、动物等)及 13 个子类，采集高度为 5~20 米，目标尺寸动态范围从图像面积的 1%至 30%以上，模拟无人机低空动态追踪的复杂场景；同时标注 13 种挑战属性(如热交叉、极端光照、遮挡等)。相较于早期数据集，VTUAV 以高分辨率、长序列、多任务标注和严苛场景设计，为无人机视角下的算法鲁棒性评估提供了更贴近实际应用的基准平台。本文算法在 VTUAV-short 测试集上与 FSRPN [22]，mfDimp [23]，DAFNet [24]，ADRNet [25]，HMFT [13]

进行对比实验，得到结果如表 2 所示。从结果可以得知，本文算法在成功率与准确率领先的情况下，在性能方面有着较大的优势，可以满足无人机跟踪的实时性要求。

Table 2. Comparative experimental results on the VTUAV-short dataset
表 2. VTUAV-short 数据集对比实验结果

跟踪算法	成功率(%)	准确率(%)	速度(FPS)
FSRPN	54.4	65.3	34.7
mfDimp	55.4	67.3	32.6
DAFNet	45.8	62.0	17.5
ADRNet	46.6	62.2	21.4
HMFT	62.7	75.8	29.3
Ours	67.5	78.1	56.3

4. 消融实验

为了确保本文算法模块的有效性，在 VTUAV-short 上进行消融实验。第一组设置为单纯以 MobileNetV3-small 作为主干网络的孪生网络目标跟踪算法，第二组引入双模态时空交互注意力模块，第三组在第二组的基础上再引入双模态自适应惩罚选择模块。得到结果如表 3 所展示。可以看出双模态时空交互注意力模块显著地提升了算法模型的成功率和准确率，同时双模态自适应惩罚选择模块的引入进一步提升了算法模型的跟踪性能。

Table 3. Comparative results of the ablation study
表 3. 消融实验对比结果

分组	成功率(%)	准确率(%)	速度(FPS)
I	57.4	68.1	62.1
II	64.6	74.9	57.9
III	67.5	78.1	56.3

5. 结论

本文针对无人机平台下 RGB-T 双模态目标跟踪的挑战，提出了一种轻量化算法 SiamTSA。通过改进 MobileNetV3-small 主干网络，显著降低了模型计算复杂度；跨模态时空交互注意力模块有效融合了双模态特征，抑制了视角差异与噪声干扰；自适应惩罚选择模块进一步提升了跟踪框的稳定性。实验表明，SiamTSA 在 GTOT、RGBT234 和 VTUAV-short 数据集上均取得领先性能，尤其在无人机场景下的 VTUAV-short 数据集中，跟踪成功率和实时性分别达到 67.5%与 56.3 FPS，验证了算法的鲁棒性与实用性。

参考文献

[1] 卓力, 张时雨, 张辉, 等. 无人机影像单目标跟踪综述[J]. 北京工业大学学报, 2021, 47(10): 1174-1187.

[2] 张天路, 张强. 基于深度学习的 RGB-T 目标跟踪技术综述[J]. 模式识别与人工智能, 2023, 36(4): 327-353.

[3] Zhang, X., Ye, P., Peng, S., Liu, J., Gong, K. and Xiao, G. (2019) SiamFT: An RGB-Infrared Fusion Tracking Method via Fully Convolutional Siamese Networks. *IEEE Access*, **7**, 122122-122133. https://doi.org/10.1109/access.2019.2936914

[4] Zhang, T., Liu, X., Zhang, Q. and Han, J. (2022) SiamCDA: Complementarity and Distractor-Aware RGB-T Tracking

Based on Siamese Network. *IEEE Transactions on Circuits and Systems for Video Technology*, **32**, 1403-1417. https://doi.org/10.1109/tcsvt.2021.3072207

[5] Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., *et al*. (2019) Searching for MobileNetV3. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 1314-1324. https://doi.org/10.1109/iccv.2019.00140

[6] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4510-4520. https://doi.org/10.1109/cvpr.2018.00474

[7] Zhang, Z. and Peng, H. (2019) Deeper and Wider Siamese Networks for Real-Time Visual Tracking. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4586-4595. https://doi.org/10.1109/cvpr.2019.00472

[8] Wei, J., Sun, K., Li, W., Li, W., Gao, S., Miao, S., *et al*. (2024) Robust Change Detection for Remote Sensing Images Based on Temporospatial Interactive Attention Module. *International Journal of Applied Earth Observation and Geoinformation*, **128**, Article 103767. https://doi.org/10.1016/j.jag.2024.103767

[9] Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J. and Yan, J. (2019) SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4277-4286. https://doi.org/10.1109/cvpr.2019.00441

[10] Hu, W., Wang, Q., Zhang, L., Bertinetto, L. and Torr, P.H. (2023) SiamMask: A Framework for Fast Online Object Tracking and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 3072-3089.

[11] Guo, C. and Xiao, L. (2022) High Speed and Robust RGB-Thermal Tracking via Dual Attentive Stream Siamese Network. 2022 *IEEE International Geoscience and Remote Sensing Symposium*, Kuala Lumpur, 17-22 July 2022, 803-806. https://doi.org/10.1109/igarss46834.2022.9883659

[12] Li, C., Xue, W., Jia, Y., Qu, Z., Luo, B., Tang, J., *et al*. (2022) Lasher: A Large-Scale High-Diversity Benchmark for RGBT Tracking. *IEEE Transactions on Image Processing*, **31**, 392-404. https://doi.org/10.1109/tip.2021.3130533

[13] Zhang, P., Zhao, J., Wang, D., Lu, H. and Ruan, X. (2022) Visible-Thermal UAV Tracking: A Large-Scale Benchmark and New Baseline. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 8876-8885. https://doi.org/10.1109/cvpr52688.2022.00868

[14] Li, C., Cheng, H., Hu, S., Liu, X., Tang, J. and Lin, L. (2016) Learning Collaborative Sparse Representation for Grayscale-Thermal Tracking. *IEEE Transactions on Image Processing*, **25**, 5743-5756. https://doi.org/10.1109/tip.2016.2614135

[15] Xiao, Y., Yang, M., Li, C., Liu, L. and Tang, J. (2022) Attribute-Based Progressive Fusion Network for RGBT Tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 2831-2838. https://doi.org/10.1609/aaai.v36i3.20187

[16] Zhu, Y., Li, C., Tang, J., Luo, B. and Wang, L. (2022) RGBT Tracking by Trident Fusion Network. *IEEE Transactions on Circuits and Systems for Video Technology*, **32**, 579-592. https://doi.org/10.1109/tcsvt.2021.3067997

[17] Li, C.L., Lu, A., Zheng, A.H., Tu, Z. and Tang, J. (2019) Multi-Adapter RGBT Tracking. 2019 *IEEE/CVF International Conference on Computer Vision Workshop* (*ICCVW*), Seoul, 27-28 October 2019, 2262-2270. https://doi.org/10.1109/iccvw.2019.00279

[18] Zhu, Y., Li, C., Luo, B., Tang, J. and Wang, X. (2019) Dense Feature Aggregation and Pruning for RGBT Tracking. *Proceedings of the* 27*th ACM International Conference on Multimedia*, Nice, 21-25 October 2019, 465-472. https://doi.org/10.1145/3343031.3350928

[19] Yun, S., Choi, J., Yoo, Y., Yun, K. and Choi, J.Y. (2017) Action-Decision Networks for Visual Tracking with Deep Reinforcement Learning. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1349-1358. https://doi.org/10.1109/cvpr.2017.148

[20] Li, C., Liang, X., Lu, Y., Zhao, N. and Tang, J. (2019) RGB-T Object Tracking: Benchmark and Baseline. *Pattern Recognition*, **96**, Article 106977. https://doi.org/10.1016/j.patcog.2019.106977

[21] Nam, H. and Han, B. (2016) Learning Multi-Domain Convolutional Neural Networks for Visual Tracking. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 4293-4302. https://doi.org/10.1109/cvpr.2016.465

[22] Kristan, M., Matas, J., Leonardis, A., *et al*. (2019) The Seventh Visual Object Tracking VOT 2019 Challenge Results. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, 27-28 October 2019, 2206-2241.

[23] Zhang, L., Danelljan, M., Gonzalez-Garcia, A., van de Weijer, J. and Shahbaz Khan, F. (2019) Multi-Modal Fusion for End-to-End RGB-T Tracking. 2019 *IEEE/CVF International Conference on Computer Vision Workshop* (*ICCVW*), Seoul, 27-28 October 2019, 2252-2261. https://doi.org/10.1109/iccvw.2019.00278

[24] Gao, Y., Li, C., Zhu, Y., Tang, J., He, T. and Wang, F. (2019) Deep Adaptive Fusion Network for High Performance

RGBT Tracking. 2019 *IEEE*/*CVF International Conference on Computer Vision Workshop* (*ICCVW*), Seoul, 27-28 October 2019, 91-99. https://doi.org/10.1109/iccvw.2019.00017

[25] Zhang, P., Wang, D., Lu, H. and Yang, X. (2021) Learning Adaptive Attribute-Driven Representation for Real-Time RGB-T Tracking. *International Journal of Computer Vision*, **129**, 2714-2729. https://doi.org/10.1007/s11263-021-01495-3