时序信息引导跨视角特征融合的多无人机多目标跟踪方法

伍 瀚,孙 浩,计科峰*,匡纲要
(国防科技大学电子科学学院电子信息系统复杂电磁环境效应国家重点实验室,湖南长沙 410073)

摘 要: 多无人机多目标跟踪旨在从多架无人机同时捕获的视频中预测所有目标的轨迹和身份标识,以解决单个无人机视频受遮挡和杂乱背景等干扰时跟踪性能衰退的问题. 然而,不同无人机捕获的图像视角和尺度差异通常较大,导致对齐和融合不同无人机图像特征困难. 针对该问题,本文提出一种通过时序信息引导跨视角特征融合的跟踪算法——TCFNet. 该算法首先设计一种目标感知的对齐网络(Object-aware Alignment Network,OAN),利用跟踪过程中的目标轨迹先验估计先前时刻不同视角无人机视频帧间的转换关系. 其次,构建一种时序感知的对齐网络(Temporal-aware Alignment Network,TAN),探索前后时刻同一架无人机捕获图像的信息对不同视角图像的转换关系进行精调. 最后,基于OAN和TAN估计的不同无人机图像间的转换关系,设计一个跨机特征融合网络(Cross-drone Feature Fusion Network,CFFN)对不同无人机捕获的视觉信息进行融合,解决复杂场景下模型跟踪性能衰退的问题. 在MDMT数据集上的实验结果表明,所提出的TCFNet相比其他主流的跟踪方法更具竞争力,在跟踪准确率、识别F1值和多机目标关联分数上超出当前的先进算法2.23、1.67和2.15个百分点.

关键词: 多无人机多目标跟踪;时序信息;轨迹先验;跨视角特征融合;准确跟踪
基金项目: 国家自然科学基金(No.61971426)
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2025)03-0728-16
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240727

Temporal-Guided Cross-View Feature Fusion Network for Multi-Drone Multi-Object Tracking

WU Han, SUN Hao, JI Ke-feng*, KUANG Gang-yao
(*State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System*, *College of Electronic Science and Technology*, *National University of Defense Technology*, *Changsha*, *Hunan* 410073, *China*)

Abstract: Multi-drone multi-object tracking aims to predict the tracklets and identities of all targets from videos simultaneously captured by multiple drones, which alleviates the tracking performance degradation when individual drone videos suffer from challenges such as occlusion and cluttered backgrounds. However, the differences in viewpoints and scales of images captured by different drones are usually large, resulting in significant difficulties for aligning and fusing cross-drone features. To address this problem, we propose a novels tracker based on cross-view feature fusion guided by temporal information. It first designs an object-aware alignment network (OAN) that utilizes the tracklet prior during tracking to estimate the transformation relationships between cross-drone frames at previous moments. Then, a temporal-aware alignment network (TAN) is constructed to explore the information of single-drone images in the before-and-after moments to fine-tune the transformation relationship across the images. Finally, based on the cross-drone image transformation relationship estimated by OAN and TAN, this paper presents a cross-drone feature fusion network (CFFN) to fuse the visual information captured by multiple drones, which mitigates the tracking performance degradation in complex scenes. Experimental results on the MDMT dataset show that the proposed TCFNet is more competitive than existing mainstream trackers, exceeding current state-of-the-art model by 2.23, 1.67, and 2.15 percentage points in terms of tracking accuracy, identification F1 score, and multi-device association score.

收稿日期:2024-08-01;修回日期:2024-10-30;责任编辑:朱梅玉
*通讯作者:计科峰

Key words: multi-drone multi-object tracking; temporal information; tracklet prior; cross-view feature fusion; accurate tracking

Foundation Item(s): National Natural Science Foundation of China (No.61971426)

1 引言

无人机集群因其成本低、机动性强、覆盖范围广以及可低空作业等优势引起了科研人员广泛的研究兴趣. 随着人工智能和通信技术的发展,无人机集群已经在智慧城市、搜索救援和军事安全等领域[1]展现出巨大的应用潜力. 目标跟踪技术作为无人机视觉领域的重要研究方向,是实现自主导航、智能决策以及侦察和打击等军事任务的基础. 目前,大多数目标跟踪方法仅从单个无人机拍摄的单一视图中捕捉目标的运动轨迹[2]. 然而,无人机拍摄角度、飞行姿态和速度的变化带来了目标尺寸变化、遮挡严重和模糊成像等挑战,造成了目标跟踪困难. 通过无人机集群中多架无人机的协同感知,可以充分利用多视图的互补信息进行准确跟踪. 因此,研究多无人机多目标跟踪(Multi-Drone Multi-Object Tracking,MDMOT)算法对于无人机的实际应用具有重要意义.

相比于自然场景下的多目标跟踪(Multi-Object Tracking,MOT)和MOT技术,MDMOT技术尚处于起步阶段. 使用多架无人机对目标进行跟踪时,目标将同时存在于多个视角的监控范围内,由于背景杂乱、目标遮挡以及不同无人机捕获的图像视角和尺度差异大,目标在多个无人机中的有效特征难以有效协同. 如图1(a)所示,为了协同跨视角图像的信息、提升跟踪性能,当前以MIA-Net[3]为代表的MDMOT算法均采用决策级融合的策略,其首先通过关联检测结果完成单机内跟踪,再对不同视角下的跟踪轨迹进行跨机关联. 然而,这些方法在定位和关联目标的过程中忽略了多机间跨视角信息的交互,无法高效协同多视角的互补信息,导致复杂场景下的跟踪性能下降. 为实现跟踪过程中多机间的相互引导,一种直观的思路是将来自不同无人机的图像特征在空间维度上对齐后融合. 然而,在实际的跟踪过程中不同无人机捕获的重叠区域不确定且成像尺度和观测视角差异大,依靠图像的视觉信息对齐跨机的视频帧往往存在明显的误差,导致跨视角特征被错误融合. 这种不可靠的融合结果极大地限制了模型的跟踪性能和真实应用价值.

当前的图像融合方法[4~6]大多是在整张图像中搜索对应的关键特征点以计算图像间的空间转换关系,进而实现特征融合. 然而,不同无人机捕获的图像之间可能仅存在较小的重叠区域,且跨视角图像的特征差异较大. 因此,从跨视角的无人机图像中准确提取用于特征对齐的特征点往往非常困难. 事实上,跟踪过程中公共目标在不同图像中的位置可以为跨视角特征对齐提供更加可靠的特征点. 基于相同目标在不同视角中的位置,能更准确地计算图像之间的转换关系. 此外,相比于对齐跨视角的图像,对齐相同视角下前后时刻的图像往往更简单. 因此,在跟踪过程中结合先前时刻基于目标轨迹先验预测的跨视角图像转换关系以及同一视角相邻帧图像的转换关系,可以对齐当前图像和另一视角的前一帧图像,从而融合不同视角图像的互补特征,提升复杂场景下模型的跟踪性能.

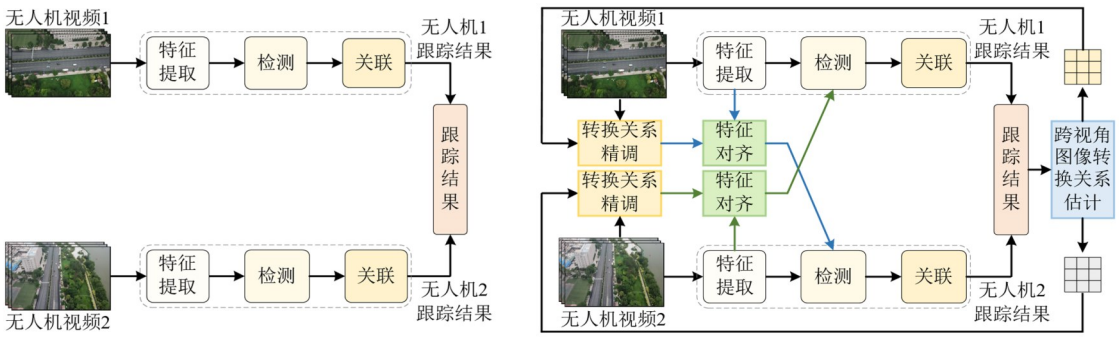
基于以上分析,提出一种利用时序信息引导跨视角特征融合的MDMOT方法——TCFNet(Temporal-guided Cross-view Feature Fusion Network),如图1(b)所示. 针对跨视角图像对齐过程中提取关键特征点和捕获特征点对应关系困难的问题,TCFNet通过挖掘跟踪过程中的目标轨迹先验实现跨视角特征的对齐和融合. 首先,TCFNet利用已有公共跟踪目标的位置作为先验特征点集估计先前时刻跨视角图像间的转换关系. 接着,TCFNet估计相同视角下不同时刻视频帧间的转换关系,并用其精调前一时刻跨视角图像的转换关系,从而得到前后时刻不同视角图像的转换关系. 最后,TCFNet将前一时刻各个视角的图像特征作为线索,基于所预测的转换关系将其与目标视角对齐实现特征融合. 通过上述步骤,TCFNet为复杂的跨视角特征对齐引入了目标跟踪轨迹作为特征点先验,并通过难度更低的单机图像配准辅助预测. 实验结果表明,提出的TCFNet能够有效提升跟踪准确性,改进的MDMOT算法在跟踪评价指标中的跟踪准确率、识别F1值和多机目标关联分数分别超出现有先进算法2.23、1.67和2.15个百分点.

2 相关工作

受益于深度学习技术的快速发展,目标跟踪技术已在多种真实场景中得到广泛应用. 本章节对当前主流的多目标跟踪方法以及专为无人机视频设计的跟踪算法进行全面回顾.

2.1 多目标跟踪

相较于单目标跟踪(Single Object Tracking,SOT),MOT面临更多的不确定性,如目标数量和类别的变化. 因此,大多数MOT方法遵循基于检测的跟踪(Tracking By Detection,TBD)范式,即首先使用一个检测网络定位视频帧中的目标,然后将不同图像中的目标关联成完整的轨迹.



(*a*) 现有方法跟踪方式 (*b*) 本文方法

注:蓝色和绿色箭头分别表示传递来自不同无人机序列的线索.

图1 本文方法与现有方法结构对比

基于TBD范式的跟踪方法主要包括目标检测和目标关联2个阶段. 当前的跟踪器大多将检测和关联视为2个独立步骤,通过独立设计的检测器和关联器更好地优化每个模块,从而提升整体的鲁棒性. 为了提高TBD模型的运算效率,JDE[7]和TrackR-CNN[8]将关联阶段所需的特征提取分支集成到检测网络,为后续关联提供目标的外观表征. 通过使检测网络和关联网络共享骨干网络,这类方法显著减少了计算开销,提升了运算效率. 然而,这种共享特征的方法往往难以在复杂场景中实现准确跟踪,其主要原因在于不同任务之间存在优化矛盾. 具体地,目标检测需要寻找同一类别目标之间的公共特征,而跟踪要求关联算法描述各类别中所有目标之间的特征差异. 这2个子任务间的固有矛盾使得骨干网络提取的特征无法同时满足检测和关联的需求,从而导致了次优的跟踪结果. 因此,CSTrack[9]基于卷积层和注意力机制设计了特征解耦网络用于将骨干网络所提取的共享特征解耦为检测和关联所需的特定特征,实现了准确性和运算效率之间的平衡.

当前MOT算法对目标关联的改进主要集中在提取目标的外观特征或运动特征,从而为后续关联提供更具辨别性的目标表征. 早期的研究中,DeepSORT[10]和DAN[11]使用深度卷积网络从每个目标框中提取目标的外观信息. 随后,TADAM[12]和FineTrack[13]通过学习寻找目标清晰可见的区域来抑制遮挡和杂乱背景引入的冗余信息. QDTrack[14]和MTCL[15]使用对比学习比较相同目标在不同视角下的特征,从而获取更为鲁棒的特征. 为了解决在错误的关联后基于短期的目标表征难以正确找回目标身份标识的问题,BLSTM-MTP[16]和UTM[17]分别基于长短时记忆网络(Long Short-Term Memory,LSTM)和图神经网络(Graph Neural Network,GNN)对各个目标的外观表征进行长期建模. 在运动表征方面,SORT[18]首先使用卡尔曼滤波预测跟踪轨迹的位置,然后计算预测位置和检测目标间的交并比(Intersection-over-Union,IoU),最后通过匈牙利算法输出关联结果. 该方案在大多数场景下能展现出良好的性能,至今仍然是大多MOT模型采取的运动表征建模策略. 为了更好地应对摄像头抖动和目标的不规律运动,OC-SORT[19]将非线性运动模块引入算法从而辅助卡尔曼滤波对目标更为复杂的运动状态进行建模. 随后,MotionTrack[20]通过学习和模拟目标之间的交互关系为关联密集分布的目标提供了更强的运动约束. 在这些关联约束中,当不同目标的位置相近或部分重叠时,外观模型通常相较于运动模型表现出更好的性能;而当目标在复杂背景中被遮挡时,基于运动特征的方法更适合跟踪消失在视野中的目标.

得到目标的外观或运动表征后,将跨帧目标之间的表征相似性送入匈牙利算法计算最终的关联结果是常见的方法. 为进一步提升关联结果的可靠性,MOTDT[21]为待关联的候选目标设计了一个评分机制,从而赋予各个目标不同的关联优先级,减少虚警和冗余轨迹造成的错跟. 随后,为了避免遮挡造成的低置信度检测导致的漏跟,ByteTrack[22]将所有检测目标都送入关联阶段,并利用检测目标和已有跟踪轨迹间的相似性恢复低置信度的目标,同时抑制了背景检测造成的虚警.

2.2 无人机视频目标跟踪

由于具有广泛的应用前景,从无人机视频中跟踪目标已经成为一个备受关注的研究领域. 在无人机视角的SOT领域,基于判别式相关滤波器(Discriminative Correlation Filter,DCF)的方法是早期的主流框架. Huang等人[23]提出了一种异常抑制相关滤波器(Aberrance Repressed Correlation Filters,ARCF),用于缓解背景噪声和目标外观的变化所导致的跟踪困难. Lin等人[24]考虑了跨帧跟踪的可逆性并提出了双向不一致性感知的相关滤波器(Bidirectional Incongruity-aware Correlation Filter,BICF),该算法通过保留目标过去时刻的特征有效地增强了跟踪器的鲁棒性. 在此基础上,Chen等人[25]在训练阶段引入了类高斯函数标签,该方法充

分考虑了目标的宽高比分布,有效增强了模型对不同形状目标的跟踪准确性. 随着深度学习的高速发展,基于孪生网络的跟踪框架因其高参数效率和高准确性成为当前无人机视角SOT的研究主流. 为了应对无人机视频中背景导致的遮挡和目标快速运动等挑战,Cao等人[26]将自注意力和互注意力聚合模块引入孪生网络框架中,提出SiamAPN++,从而增强困难场景下的目标特征. 为了增强无人机在夜间对困难目标的跟踪能力,Ye等人[27]基于Transformer设计了一种空间和通道维度的弱光特征增强器,该网络通过一种鲁棒的非线性曲线投影对夜间拍摄的图像进行去噪和光照增强. 为了进一步提升模型性能,SiamTPN[28]将特征金字塔(Feature Pyramid Network,FPN)和轻量级Transformer整合到孪生网络跟踪器中,从而构建了一个鲁棒的目标感知外观模型.

基于无人机视频的MOT技术近年来迅速发展. 为联合多个关联约束增强跟踪性能,IPGAT[29]在网络中设计了多个模块分别建模相机运动、目标运动和目标外观. 随后,Liu等人[30]提出了一种无人机运动自适应的运动滤波器建模目标的线性和非线性运动,从而有效提升了模型的跟踪准确性. 针对无人机视频中小尺寸目标和成像模糊造成的跟踪困难,Wu等人[31]结合多感受野特征和多尺度Transformer块,增强了网络特征对小目标的响应能力. 为了缓解无人机快速运动造成的运动模糊导致的跟踪性能衰退,Cheng等人[32]在模型中引入了超分辨技术重构模糊成像的视频帧用于后续的预测. 针对目标尺寸变化和目标遮挡对鲁棒跟踪的阻碍,Shi等人[33]设计了一种全局和局部感知以及遮挡感知(Global-Local and Occlusion Awareness,GLOA)的跟踪框架,提升了复杂场景下模型对目标尺度变化和遮挡的鲁棒性. 尽管这些方法提升了跟踪准确性,但在无人机视频中关联目标仍然面临挑战. 相较于自然场景,无人机视角下目标间的相似性更高且无人机图像视觉信息较为稀疏,难以提取可靠的外观表征用于区分不同目标. 此外,相比于固定摄像头下的通用MOT任务,机载摄像头的动态特性使得目标在视频中的运动模式更加复杂,为运动建模带来了更大的难度. 更为重要的是,现有方法通常仅依靠单个无人机视频进行跟踪,难以有效应对小目标、严重遮挡、模糊成像和光照变换等复杂问题. 相较之下,多个无人机视频可有效协同多视角的互补信息,从而提升在复杂场景下的感知能力. 因此,多无人机视频协同跟踪近年来逐渐受到研究关注.

在多无人机单目标跟踪领域,Chen等人[34]提出TransMDOT,利用孪生网络和Transformer层对不同无人机视角下的模板和搜索区域特征进行提取、融合和匹配. 在MDMOT方面,MIA-Net[3]是当前的代表算法,其首先在独立视频帧中检测目标,然后通过计算目标框的IoU完成单机内的目标关联,接着通过配准算法对齐不同视角无人机图像的坐标系,从而基于目标的空间位置信息关联跨机的目标. 然而,由于不同无人机所捕获的图像在视角和尺度上存在较大差异,MIA-Net难以在跟踪过程中准确对齐和融合各无人机图像间的互补信息.

相较于通用场景和单个无人机视频的目标跟踪,关于MDMOT的研究尚处于起步阶段,从多个无人机视频中跟踪多个感兴趣的目标仍存在较大的困难. 准确对齐和融合不同视角中的互补信息,发挥不同无人机间相互引导的作用对于提升跟踪性能至关重要. 事实上,在当前仅依赖图像视觉信息进行跨视角特征的对齐与融合的基础上,进一步引入跟踪过程中的目标时序信息能够更好地协同不同无人机捕获的互补信息,从而提升跟踪性能.

3　本文方法

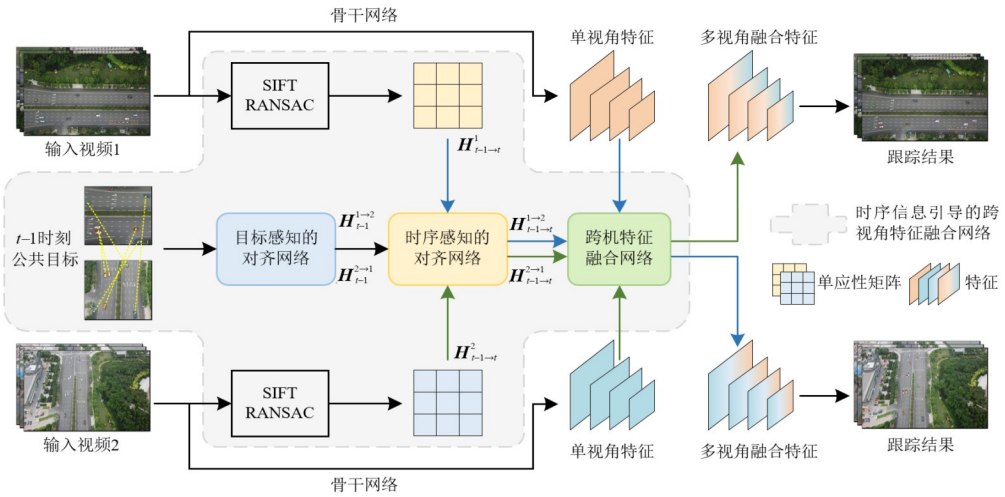
3.1　网络整体框架

本文提出的TCFNet模型旨在在跟踪过程中对齐和融合不同视角无人机的信息,实现不同无人机间的相互引导,提升模型在复杂场景下的跟踪性能. TCFNet的模型结构如图2所示,其主要包括3个关键模块:(1)目标感知的对齐网络(Object-aware Alignment Network,OAN),用于估计先前时刻跨视角图像间的转换关系;(2)时序感知的对齐网络(Temporal-aware Alignment Network,TAN),用于预测从过去到当前时刻的跨视角图像转换关系;(3)跨机特征融合网络(Cross-drone Feature Fusion Network,CFFN),根据所预测的图像间转换关系融合不同无人机捕获的图像特征. 具体而言,首先设计了OAN基于先前时刻不同无人机中相同目标的位置估计视角1和视角2图像间的转换关系. 接着,TCFNet通过图像配准算法估计各无人机序列中从过去到当前时刻的转换关系,同时通过TAN进一步结合和估计跨时空的帧间转换关系. 最后,TCFNet通过CFFN基于交互骨干网络提取的跨视角图像特征,增强特征图对困难样本的表示能力,从而提升模型在多个无人机视频中的跟踪准确性.

3.2　目标感知的对齐网络

针对不同无人机捕获的视频帧难对齐的问题,以各无人机已捕获的公共跟踪轨迹为先验,并提出OAN估计先前时刻无人机1和无人机2所捕获图像之间的转换关系 $\boldsymbol{H}_{t-1}^{i\to j},(i,j)\in\{(1,2),(2,1)\}$. 如图3所示,OAN主要包括目标位置特征提取和目标感知的转换矩阵估计2个步骤.

首先,给定不同无人机中的num个公共目标,OAN将它们的中心点坐标和坐标框面积组成的位置信息



注:蓝色和绿色箭头分别表示传递来自不同无人机的特征.

图2 TCFNet结构图

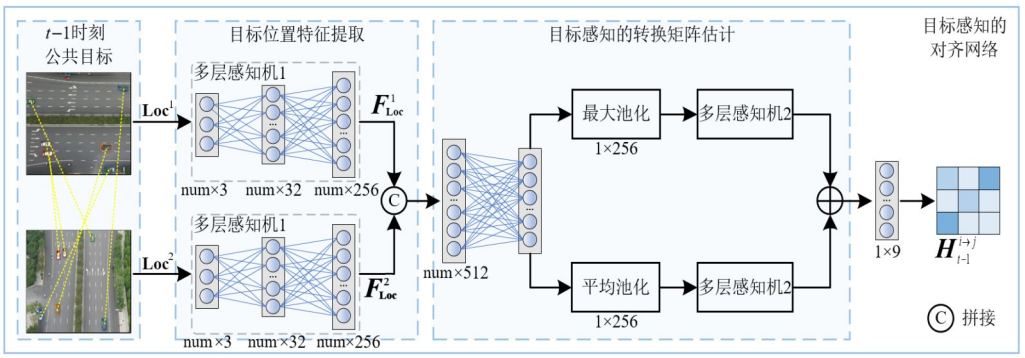


图3 OAN结构图

$\mathbf{Loc}^1, \mathbf{Loc}^2 \in \mathbb{R}^{\text{num}\times 3}$ 输入一个多层感知机(MultiLayer Perceptron, MLP)中提取各目标的位置和尺度特征,并将所得特征 $\boldsymbol{F}^1_{\mathbf{Loc}}, \boldsymbol{F}^2_{\mathbf{Loc}} \in \mathbb{R}^{\text{num}\times 256}$ 拼接为一个包含目标多视角信息的特征 $\boldsymbol{F}_{\mathbf{Loc}} \in \mathbb{R}^{\text{num}\times 512}$,该过程可以表示为

$$\boldsymbol{F}_{\mathbf{Loc}} = \text{Concat}\left(\text{MLP}_1(\mathbf{Loc}^1), \text{MLP}_1(\mathbf{Loc}^2)\right) \tag{1}$$

其中,Concat(·)和 MLP_1 分别表示拼接操作和多层感知机1.

接着,OAN通过一个全连接层(Fully Connected, FC)交互公共目标的跨机信息,并分别使用最大池化(Max Pooling, MP)和平均池化(Average Pooling, AP)对所得特征进行压缩. 然后,OAN通过 MLP_2 将压缩后的特征向量依次降维至1×128、1×32和1×9. 最后,OAN对2个分支捕获的1×9特征向量求和,并将其重构(Reshape)为用于描述 $t-1$ 时刻跨视角图像的转换矩阵 $\boldsymbol{H}^{i\to j}_{t-1} \in \mathbb{R}^{3\times 3}$. 上述过程可以表述为

$$\begin{aligned}\boldsymbol{H}^{i\to j}_{t-1} = \text{Reshape}\Big(&\text{MLP}_2\Big(\text{MP}\big(\text{FC}(\boldsymbol{F}_{\mathbf{Loc}})\big)\Big)\\ &+\text{MLP}_2\Big(\text{AP}\big(\text{FC}(\boldsymbol{F}_{\mathbf{Loc}})\big)\Big)\Big)\end{aligned} \tag{2}$$

通过以上操作,OAN可估计用于描述先前时刻不同无人机捕获视频帧间的转换矩阵 $\boldsymbol{H}^{i\to j}_{t-1}$,其被送入后续的TAN中用于增强当前时刻各无人机捕获图像的特征.

3.3 时序感知的对齐网络

TAN旨在融合同一无人机第 $t-1$ 帧到第 t 帧图像的转换关系 $\boldsymbol{H}^{i}_{t-1\to t}$ 以及 $t-1$ 时刻不同无人机所捕获图像的转换关系 $\boldsymbol{H}^{i\to j}_{t-1}$,从而预测 $t-1$ 时刻到 t 时刻不同无人机捕获图像间的转换关系 $\boldsymbol{H}^{i\to j}_{t-1\to t}, (i,j)\in\{(1,2),(2,1)\}$,以引导后续的跨视角特征融合.

由于同一无人机前后时刻捕获图像的偏差相对小,TCFNet直接使用尺度不变特征变换(Scale Invariant Feature Transform, SIFT)[35]和随机样本一致(RANdom SAmple Consensus, RANSAC)[36]算法提取 $\boldsymbol{H}^{i}_{t-1\to t}$,并将其送入TAN中联合 $\boldsymbol{H}^{i\to j}_{t-1}$ 估计 $\boldsymbol{H}^{i\to j}_{t-1\to t}$. 如图4所示,TAN首先使用线形层和1×1卷积提取 $\boldsymbol{H}^{i\to j}_{t-1}$ 和 $\boldsymbol{H}^{i}_{t-1\to t}$ 的特征并将其重构为 $\boldsymbol{h}^x, \boldsymbol{h}^y \in \mathbb{R}^{8\times 256}$. 然后,TAN通过由2个并行的交叉注意力(Cross Attention, CA)模块组成的时序感知的转换矩阵估计交互 $\boldsymbol{h}^x$ 和 $\boldsymbol{h}^y$. 具体来说,TAN为 $\boldsymbol{h}^x$ 和 $\boldsymbol{h}^y$ 添加一个位置编码得到相应的特征表示 $\boldsymbol{h}^x_p$ 和 $\boldsymbol{h}^y_p$. 接

着,TAN将$\boldsymbol{h}_p^x$和$\boldsymbol{h}_p^y$分别作为2个交叉注意力的键(Key)和值(Value),并将它们的和作为公共的查询(Query):

$$\boldsymbol{h}^1=\mathrm{softmax}\left[\frac{(\boldsymbol{h}_p^x+\boldsymbol{h}_p^y)\otimes(\boldsymbol{h}_p^x)^{\mathrm{T}}}{\sqrt{d_k}}\right]\odot\boldsymbol{h}_p^x \tag{3}$$

$$\boldsymbol{h}^2=\mathrm{softmax}\left[\frac{(\boldsymbol{h}_p^x+\boldsymbol{h}_p^y)\otimes(\boldsymbol{h}_p^y)^{\mathrm{T}}}{\sqrt{d_k}}\right]\odot\boldsymbol{h}_p^y \tag{4}$$

其中,$\otimes$和$\odot$分别代表矩阵乘法和点积,d_k表示键的维度,T为转置.

随后,TAN将CA模块输出的$\boldsymbol{h}^1$和$\boldsymbol{h}^2$相加后进行层正则化(LayerNorm),并通过一个由1×1卷积和全连接层组成的前馈网络计算最终的转换矩阵$\boldsymbol{H}_{t-1\to t}^{i\to j}\in\mathbb{R}^{3\times3}$.该矩阵作为描述无人机$i$第$t$–1帧图像到无人机$j$第$t$帧图像转换关系的线索被送入后续的CFFN中.

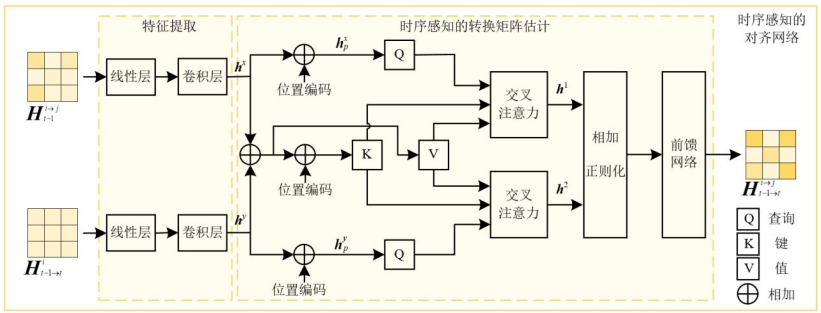


图4　TAN结构图

3.4　跨机特征融合网络

CFFN旨在融合不同无人机所捕获视频帧的图像特征从而实现多无人机协同跟踪,其网络结构如图5所示.在进行无人机j的跟踪时,从无人机i中寻求线索.

给定t时刻无人机j捕获图像的特征$\boldsymbol{F}_t^j$,CFFN通过t–1时刻无人机i捕获图像的特征$\boldsymbol{F}_{t-1}^i$增强$\boldsymbol{F}_t^j$.具体地,CFFN首先通过由核为1×1和3×3的卷积组成的瓶颈网络对$\boldsymbol{F}_t^j$和$\boldsymbol{F}_{t-1}^i$进行编码.然后,CFFN基于TAN预测的$\boldsymbol{H}_{t-1\to t}^{i\to j}$对$\boldsymbol{F}_{t-1}^i$编码得到的特征进行仿射变换[37]后,通过像素级相加融合跨视角特征.最终,一个瓶颈网络被用于输出最终的特征$\tilde{\boldsymbol{F}}_t^j$.

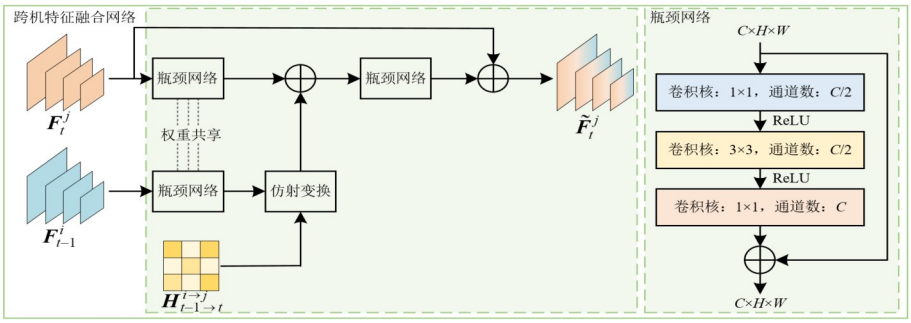


图5　CFFN结构图

3.5　损失函数

TCFNet旨在对齐和融合跨视角图像的特征,提升模型跟踪目标的能力.因此,TCFNet首先采用对齐损失优化OAN和TAN预测的跨视角图像转换关系,确保模型对齐跨视角特征的可靠性,然后通过检测损失增强模型利用CFFN融合的特征定位目标的能力.具体而言,首先通过对齐损失L_{match}^1和L_{match}^2优化OAN和TAN估计的$\boldsymbol{H}_{t-1}^{i\to j}$和$\boldsymbol{H}_{t-1\to t}^{i\to j}$.假定无人机$i$和无人机$j$捕获了num个公共跟踪目标,$L_{\mathrm{match}}^1$和$L_{\mathrm{match}}^2$详细的计算过程如下:

$$L_{\mathrm{match}}^1=\frac{1}{\mathrm{num}}\sum_{n=1}^{\mathrm{num}}\left(\left|x_{j\_(t-1)}^n-\bar{x}_{j\_(t-1)}^n\right|+\left|y_{j\_(t-1)}^n-\bar{y}_{j\_(t-1)}^n\right|\right) \tag{5}$$

$$L_{\mathrm{match}}^2=\frac{1}{\mathrm{num}}\sum_{n=1}^{\mathrm{num}}\left(\left|x_{j\_t}^n-\bar{x}_{j\_t}^n\right|+\left|y_{j\_t}^n-\bar{y}_{j\_t}^n\right|\right) \tag{6}$$

其中,$(x_{j_(t-1)}^n,y_{j_(t-1)}^n)$和$(x_{j_t}^n,y_{j_t}^n)$分别表示$t$–1时刻和$t$时刻各公共目标在无人机$j$图像中的中心点坐标,$(\bar{x}_{j_(t-1)}^n,\bar{y}_{j_(t-1)}^n)$和$(\bar{x}_{j_t}^n,\bar{y}_{j_t}^n)$分别表示TCFNet基于$\boldsymbol{H}_{t-1}^{i\to j}$和$\boldsymbol{H}_{t-1\to t}^{i\to j}$对无人机$i$中的公共目标进行仿射变换后的中心点坐标,其计算公式为

$$\left(\bar{x}^n_{j\_(t-1)},\bar{y}^n_{j\_(t-1)}\right)=\left(\frac{\alpha_{11}x^n_{i\_(t-1)}+\alpha_{12}y^n_{i\_(t-1)}+\alpha_{13}}{\alpha_{31}x^n_{i\_(t-1)}+\alpha_{32}y^n_{i\_(t-1)}+\alpha_{33}},\frac{\alpha_{21}x^n_{i\_(t-1)}+\alpha_{22}y^n_{i\_(t-1)}+\alpha_{23}}{\alpha_{31}x^n_{i\_(t-1)}+\alpha_{32}y^n_{i\_(t-1)}+\alpha_{33}}\right) \tag{7}$$

$$\left(\bar{x}^n_{j\_t},\bar{y}^n_{j\_t}\right)=\left(\frac{\beta_{11}x^n_{i\_(t-1)}+\beta_{12}y^n_{i\_(t-1)}+\beta_{13}}{\beta_{31}x^n_{i\_(t-1)}+\beta_{32}y^n_{i\_(t-1)}+\beta_{33}},\frac{\beta_{21}x^n_{i\_(t-1)}+\beta_{22}y^n_{i\_(t-1)}+\beta_{23}}{\beta_{31}x^n_{i\_(t-1)}+\beta_{32}y^n_{i\_(t-1)}+\beta_{33}}\right) \tag{8}$$

其中，$(x^n_{i_(t-1)},y^n_{i_(t-1)})$是$t$–1时刻无人机$i$中各公共目标的中心点坐标，$\{\alpha_{11},\alpha_{22},\cdots,\alpha_{33}\}$和$\{\beta_{11},\beta_{22},\cdots,\beta_{33}\}$是组成转换矩阵$\boldsymbol{H}^{i\to j}_{t-1}$和$\boldsymbol{H}^{i\to j}_{t-1\to t}$的元素.

在计算对齐损失后，TCFNet通过检测损失提升模型利用融合特征定位目标的能力，具体使用L1损失作为回归损失$L^{\mathrm{reg}}_{\mathrm{det}}$，并使用交叉熵损失作为分类损失$L^{\mathrm{cls}}_{\mathrm{det}}$. 基于上述多个损失函数，通过超参数$\lambda_1$和$\lambda_2$计算总体损失$L$，从而对TCFNet进行全局优化：

$$L=L^{\mathrm{reg}}_{\mathrm{det}}+L^{\mathrm{cls}}_{\mathrm{det}}+\lambda_1L^1_{\mathrm{match}}+\lambda_2L^2_{\mathrm{match}} \tag{9}$$

4 本文方法

4.1 数据集和评价指标

在公开数据集MDMT[3]上验证所提模型的性能，该数据集提供了44对由2架不同位置、不同视角的无人机同时拍摄的视频序列，其中14对用于测试，30对用于训练和验证. MDMT数据集标注了车辆、行人和自行车3类目标且涵盖了多样具有挑战性的真实场景，包括广场、乡村道路、十字路口和城市道路等. 同时，该数据集包含了晴天、多云和夜晚等多种天气和光照条件. MDMT中所有图像的分辨率均为1 920 × 1 080，平均每帧图像包含超过55个目标. 其中，约有11.0%的目标所占像素小于图像总像素的0.1%，约56.3%的目标尺寸为整张图像面积的0.1%~1.0%，大量的小目标给鲁棒跟踪带来了相当大的困难. MDMT数据集中2架无人机拍摄的所有视频中分别捕获了90 829个和1 296 361个目标实例，这种数量上的明显差异表明每架无人机不仅捕获了不同视角下的共同目标，还观测了大量特有的目标.

为了全面比较所提出的TCFNet和现有先进方法的跟踪性能，结合多个评价指标构建实验并展开分析，包括多目标跟踪准确率(Multiple Object Tracking Accuracy，MOTA)，识别F1值(ID F1 score，IDF1)和多机目标关联分数(Multi-Device target Association score，MDA). 其中，MOTA被广泛视作体现算法跟踪性能的关键指标，其计算公式如下：

$$\mathrm{MOTA}=1-\frac{\mathrm{FP}+\mathrm{FN}+\mathrm{IDS}}{\mathrm{GT}} \tag{10}$$

其中，GT表示真实目标框的总数，FP、FN和IDS分别表示虚警数(False Positives)、漏跟数(False Negatives)和目标身份交换次数(ID Switches). IDF1评价算法对目标身份的识别能力是衡量模型跟踪性能鲁棒性的重要指标. MDA评价多个设备同时跟踪多个目标时，不同视角下重叠区域内公共目标的关联准确性. 此外，通过参数量(Parameters)和浮点运算次数(FLoating-point Operations Per second，FLOPs)测试模型的复杂度.

4.2 实验设置

TCFNet选用ResNet50作为骨干网络，并以集成FPN的Faster R-CNN[38]为基准检测网络. TCFNet的训练采用SGD优化器，训练周期为12轮，初始学习率设置为0.001并在第8轮训练衰减至0.000 1，批次大小为8. 在训练和测试过程中，图像输入TCFNet网络的图像分辨率为1 344 × 768. 式(9)中的超参数λ_1和λ_2均被设置为1.0. 推理过程中，TCFNet使用MIA-Net作为多机目标关联算法. 对于连续30帧关联失败的跟踪目标，不再将其作为后续关联的候选轨迹. 为确保公平的对比，TCFNet中未特别说明的超参数均沿用了MIA-Net的设置.

所有实验基于Python 3.8、PyTorch 1.10和MMDetection 2.25[39]构建，实验的硬件平台为一台搭载24 GB显存的RTX 4090 GPU，Inter i9-13900K CPU和64 GB内存的工作站.

4.3 消融实验

4.3.1 各模块有效性分析

为了验证TCFNet中各模块的有效性，本节对TCFNet中的OAN、TAN和CFFN进行了消融实验，实验结果如表1，其中粗体标识该指标的最优得分.

由表1可见，不使用时序信息和轨迹先验对齐特征而直接通过CFFN基于SIFT和RANSAC算法估计的跨视角图像转换矩阵进行特征融合时(②对比①)，模型的性能在MOTA和IDF1指标上分别衰退了0.25个百分点和0.39个百分点. 这种性能衰退说明仅依靠图像的视觉信息估计不同无人机捕获图像间的转换关系十分困难，有必要寻找额外的线索引导跨视角特征融合. 而在此基础上添加OAN后(③对比②)，模型的3项指标分别提升了1.68、1.20和1.58个百分点. 这种显著的性能提升说明了通过公共目标的轨迹先验引导不同无人机间信息交互的有效性和优越性. 添加TAN后(④对比③)，模型在3项指标上进一步取得了0. 80、0. 86和0. 43个百分点的性能增益. 整体上(④对比①)，所提出的TCFNet相比基准算法将跟踪性能分别提升了2.23、1.67和2.15个百分点.

4.3.2 有效性分析

为了进一步验证额外引入的损失函数L^1_{match}和L^2_{match}对提升TCFNet跟踪性能的有效性，该节在MDMT

表1　提出模块对跟踪性能的影响

模型	方法				整体性能		
	基准	OAN	TAN	CFFN	MOTA/%	IDF1/%	MDA/%
①	√				51.58	66.97	38.47
②	√			√	51.33	66.58	38.61
③	√	√		√	53.01	67.78	40.19
④	√	√	√	√	**53.81**	**68.64**	**40.62**

测试集上对L^1_{match}和L^2_{match}进行了消融实验,实验结果总结在表2中. 对比②和①可以得出,引入L^1_{match}后MOTA和MDA指标分别提升了0.39个百分点和1.48个百分点. 同时,从③和②中可以发现使用L^2_{match}进一步将MOTA和IDF1提升0.46个百分点和0.67个百分点. 在多个指标上的性能提升证明了为TCFNet设计的损失函数对提升模型的跟踪准确性具有重要作用.

表2　损失函数对跟踪性能的影响

模型	损失函数		整体性能		
	L^1_{match}	L^2_{match}	MOTA/%	IDF1/%	MDA/%
①			52.96	67.63	39.04
②	√		53.35	67.97	40.52
③	√	√	**53.81**	**68.64**	**40.62**

4.3.3　超参数选择

在计算损失函数的过程中,式(9)中的超参数λ_1和λ_2控制了L^1_{match}和L^2_{match}在整体损失中所占的权重,对网络的优化方向具有重要的影响. 为了研究λ_1和λ_2取值对跟踪性能的影响,该节将λ_1和λ_2不同取值的TCFNet在MDMT数据集上进行测试,对比结果如表3所示.

表3　λ_1和λ_2取值对跟踪性能的影响

超参数		整体性能		
λ_1	λ_2	MOTA/%	IDF1/%	MDA/%
0.50	0.50	53.16	67.99	40.51
0.75	0.75	53.57	68.35	40.60
1.00	1.00	**53.81**	**68.64**	**40.62**
1.25	1.25	53.68	68.58	40.54
1.50	1.50	53.31	68.52	40.21

可以得出,随着λ_1和λ_2的取值从0.50增大到1.00,TCFNet的跟踪性能逐步提升. 进一步加大λ_1和λ_2的值后,TCFNet的各项性能指标均开始衰退. 根据实验结果,为TCFNet额外引入的2个损失函数L^1_{match}和L^2_{match}对于跟踪性能的提升具有和目标分类和回归相同或相似的重要性. 因此,将λ_1和λ_2的取值都设置为1.00,使模型在训练过程中更好地平衡网络内的多个子任务,从而提升跟踪性能.

4.3.4　不同特征融合方法对比

为了进一步验证本文方法的有效性,该节在表4中对比了TCFNet与其他特征融合方法的性能. 其中,LightGlue[40]对跨视角图像进行稀疏局部特征匹配后进行特征融合,RoMa[41]对图像特征进行密集匹配再融合,MFPT[42]通过多尺度语义提取来捕获跨视角的互补信息,而CDDFuse[43]则通过建模不同视角的共享特征和专属特征完成跨视角特征融合. 结果显示,TCFNet在各项指标上均显著优于MFPT和CDDFuse这2种从特征语义中直接捕获特征相关性的融合方法. 这主要是因为不同无人机捕获的图像重叠区域较小且特征相差较大,在未进行空间对齐的情况下直接融合特征的可靠性较低. 此外,相比于基于对齐特征后再进行融合的LightGlue和RoMa,TCFNet在3项指标上分别取得了2.39~2.75个百分点、3.38~4.46个百分点和11.84~13.28个百分点的性能提升. 这种性能优势的关键在于相比于依靠图像视觉信息的特征对齐,TCFNet利用跟踪过程中的时序信息和目标轨迹先验能更准确地对齐跨视角的图像特征.

表4　不同特征融合方法的对比结果　　单位:%

融合方法	整体性能		
	MOTA	IDF1	MDA
LightGlue [40]	51.06	64.18	37.34
RoMa [41]	51.42	65.26	38.78
MFPT [42]	45.77	62.39	24.16
CDDFuse [43]	49.98	59.27	26.85
TCFNet	**53.81**	**68.64**	**40.62**

4.4　定量对比

4.4.1　在MDMT数据集上的实验结果

为了验证所提TCFNet的性能,该节在表5将其与现有先进MDMOT算法进行比较,这些方法使用的检测网络包括Faster R-CNN[38]、TooD[44]、AutoAssign[45]、Carafe[46]和Cascade RPN[47],关联算法包括ByteTrack[22]、MIA-Net[3]、DeepSORT[10]、SBS-50[48]和QDTrack[14]. 为表述方便,表5中用检测和关联算法的首字母给了每种方法一个缩写名称,同时使用红色和蓝色字体表示每个指标的最优和次优分数.

TCFNet以Faster R-CNN为基准检测网络,与使用相同Faster R-CNN为检测器的FB、FD和FQ相比,TCFNet在多个无人机视频整体的MOTA和IDF1指标上超出它们2.85~6.24个百分点和2.71~12.18个百分点.

在表5对比算法中,FD、FQ、CQ、CRQ、ABS和CBS使用了目标的外观特征作为关联约束,与这些方法相比,TCFNet在MOTA指标上优于它们4.95~11.24个百分点,同时在IDF1上超出了6.68~13.35个百分点. 这种明显的

表5 不同算法在MDMT测试集上的对比结果 单位:%

方法	缩写名称	无人机1		无人机2		整体性能		
		MOTA	IDF1	MOTA	IDF1	MOTA	IDF1	MDA
Faster R-CNN[38] + ByteTrack[22]	FB	53.88	67.71	47.98	64.14	50.92	65.93	—
TooD[44] + ByteTrack[22]	TB	50.95	66.42	48.02	63.92	49.49	65.18	—
AutoAssign[45] + ByteTrack[22]	AB	49.52	67.38	44.52	63.75	47.01	65.56	—
Carafe[46] + ByteTrack[22]	CB	54.13	68.22	48.42	64.93	51.38	66.58	—
AutoAssign[45] + MIA-Net[3]	AM	51.90	69.67	47.46	66.81	49.68	68.24	41.72
Carafe[46] +MIA-Net[3]	CM	54.92	68.82	48.23	65.12	51.58	66.97	38.47
Faster R-CNN[38] + DeepSORT[10]	FD	50.20	60.48	41.52	52.44	48.86	56.46	—
Faster R-CNN[38] + QDTrack[14]	FQ	52.12	66.23	43.02	57.68	47.57	61.96	—
Carafe[46] + QDTrack[14]	CQ	53.20	66.28	43.06	57.46	48.13	61.87	—
Cascade RPN[47] + QDTrack[14]	CRQ	51.05	65.00	45.75	58.66	48.40	61.82	—
AutoAssign[45] + ByteTrack[22] + SBS-50[48]	ABS	44.49	55.89	40.65	54.69	42.57	55.29	18.19
Carafe[46] + ByteTrack[22] + SBS-50[48]	CBS	50.46	57.04	45.33	55.86	47.89	56.44	18.30
TCFNet	—	56.12	70.27	51.62	67.01	53.81	68.64	40.62

性能优势证明了即使不依靠目标的外观特征作为目标关联的线索,提出的TCFNet依然具有强大的跟踪性能.

ByteTrack是MOT领域的先进算法,已在多个领域中展现出卓越的性能. 与使用ByteTrack作为关联算法的FB、TB、AB和CB相比,TCFNet在MOTA和IDF1指标上分别取得了2.43~6.54个百分点和2.06~3.46个百分点的优势.

MIA-Net是当前MDMOT领域最具代表性的方法,与同样采用MIA-Net作为关联算法的AM和CM相比,TCFNet在MOTA指标上超出它们4.13个百分点和2.23个百分点,在IDF1指标上超过它们0.40个百分点和1.67个百分点. 同时,TCFNet相比CM在MDA指标上取得了2.15个百分点的优势.这些显著的性能差异证明了TCFNet不仅能更准确地跟踪多个无人机视频中的目标,还能更好地对齐不同视频中公共目标的身份标识.

相比于其他先进的MDMOT方法,TCFNet在MOTA和IDF1上均取得最优的性能,且分别高出次优的方法2.23个百分点和0.40个百分点. 这些显著的性能优势证明了所提的TCFNet通过对齐和融合跨视角特征的策略能有效地提升模型的跟踪性能.

4.4.2 不同光照条件下的实验结果

为了验证本文方法在不同光照条件下的有效性,该节从MDMT数据集中选取了晴天、阴天和夜晚条件下的序列,并在所选取的视频序列中测试了TCFNet的性能. 同时,该节从表5中选择了使用目标外观特征作为关联约束的算法CQ、CQR、ABS和CBS,以及关联过程中不使用目标外观特征的算法AB、CB、AM和CM与提出的TCFNet进行对比,实验结果列于表6,红色和蓝色表示每个指标的最优和次优分数.

可以观察到,所提的TCFNet在各种光照条件下都展现出优于其他方法的性能. 在MOTA和IDF1指标上,本文方法在各个环境下均领先其他算法,尤其在夜晚,TCFNet在MOTA和IDF1指标上分别超出第2名的方法5.09个百分点和1.51个百分点. 同时,在晴天和阴天场景中,TCFNet在MOTA指标上分别超出次优算法3.12个百分点和0.49个百分点,同时在IDF1指标上领先第2名的算法0.22个百分点和0.88个百分点. 在阴天和夜晚条件下,TCFNet虽然在MDA指标上未取得最佳性能,但仅落后0.19个百分点和0.05个百分点,说明TCFNet同样能准确地对齐跨机目标的身份标识. 在多个指标上优越的性能证明了提出的TCFNet对光照条件的变化具有良好的鲁棒性,具有在复杂场景下进行准确跟踪的能力.

4.4.3 自然场景下的实验结果

在无人机视频之外,本节在2个自然场景数据集Wildtrack[49]和MvMHAT[50]上验证了TCFNet的泛化能力. 这2个数据集均为真实世界采集的多摄像头行人跟踪数据集. 其中,Wildtrack包含60 min的视频和约66 000个标注实例,MvMHAT则包括98段视频和超过90 000帧图像. Wildtrack和MvMHAT数据集上的性能对比展示在表7中,红色和蓝色字体表示每个指标的最优和次优分数.

如表7所示,本文方法在Wildtrack和MvMHAT上均取得了最佳性能. 在Wildtrack数据集上,TCFNet在MOTA和IDF1指标上分别以1.1个百分点和1.2个百分点的优势超出现有的先进方法ReST. 相比TrackFormer,本文方法在MvMHAT数据集上取得了0.9个百分点和1.6个百分点的优势. 可以得出,相较于自然场景中的跟踪,TCFNet在无人机视频中的优势更为显著. 造成这个现象的原因是Wildtrack和MvMHAT是由固定的监控摄像头捕获,避免了无人机视频中摄像头运

表6　不同光照条件下的对比结果　　单位:%

光照条件	方法	整体性能		
		MOTA	IDF1	MDA
晴天	CQ[14,46]	46.99	59.54	—
	CRQ[14,47]	47.18	59.20	—
	ABS[22,45,48]	43.51	55.98	14.21
	CBS[22,46,48]	46.81	57.25	14.46
	AB[22,45]	45.99	64.27	—
	CB[22,46]	48.07	65.18	—
	AM[3,45]	47.11	68.01	34.37
	CM[3,46]	48.86	65.52	27.26
	TCFNet	51.98	68.23	34.39
阴天	CQ[14,46]	45.01	57.28	—
	CRQ[14,47]	44.97	56.83	—
	ABS[22,45,48]	44.06	52.02	19.83
	CBS[22,46,48]	45.24	51.23	19.40
	AB[22,45]	44.19	61.09	—
	CB[22,46]	46.11	62.24	—
	AM[3,45]	44.32	62.97	39.98
	CM[3,46]	46.52	62.01	37.91
	TCFNet	47.01	63.85	39.79
夜晚	CQ[14,46]	66.98	73.98	—
	CRQ[14,47]	67.24	73.71	—
	ABS[22,45,48]	60.64	68.88	33.74
	CBS[22,46,48]	66.82	70.22	34.14
	AB[22,45]	63.61	79.07	—
	CB[22,46]	69.85	80.26	—
	AM[3,45]	68.26	84.25	74.94
	CM[3,46]	70.17	80.90	69.17
	TCFNet	75.26	85.76	74.89

表7　Wildtrack 和 MvMHAT 数据集上的对比结果　　单位:%

数据集	方法	MOTA	IDF1
Wildtrack	KSP-DO[49]	69.6	73.2
	KSP-DO-ptrack[49]	72.2	78.4
	GLMB-DO[51]	70.1	72.5
	DMCT[52]	72.8	77.8
	ReST[53]	84.9	86.7
	TCFNet	86.0	87.9
MvMHAT	Tracktor++[54]	66.5	46.1
	CenterTrack[55]	63.5	38.1
	TraDes[56]	69.5	44.9
	TrackFormer[57]	70.4	49.6
	DeepCC[58]	63.9	44.4
	TCFNet	71.3	51.2

动问题,且自然场景下不同视角的成像尺度差异相对无人机视频更小. 因此,在自然场景中协同跨视角图像信息的难度更低,弱化了本文方法的优势. 尽管如此,TCFNet 仍然在 Wildtrack 和 MvMHAT 上超出了现有的方法,证明了 TCFNet 的有效性.

4.4.4　模型复杂度分析

为了进一步验证本文方法的实际应用价值,表8中对比了不同模型在相同设置下的复杂度. 结果显示,TCFNet 的参数量和运算量均处于所有跟踪算法中的中间水平. 值得注意的是,计算量低于 TCFNet 的方法均未协同多视角的图像特征跟踪目标,因此其跟踪准确性相对逊色. 这一对比结果证明,本文方法在提升跟踪准确性的前提下,实现了准确性与模型复杂度的良好平衡.

表8　模型复杂度对比

方法	Parameters/M	FLOPs/G
FB[22,38]	41.13	394.24
TB[22,44]	**31.98**	**376.04**
AB[22,45]	35.97	403.43
CB[22,46]	46.74	400.16
AM[3,45]	67.73	403.44
CM[3,46]	46.74	400.17
FD[10,38]	67.73	570.24
FQ[14,38]	47.50	435.68
CQ[14,46]	53.24	441.60
CRQ[14,47]	59.70	474.08
ABS[22,45,48]	59.47	573.33
CBS[22,46,48]	70.24	570.06
TCFNet	48.45	414.39

为了进一步降低模型的运算量并增强其实际应用价值,该节对 TCFNet 进行了轻量化处理并提出 TCFNet-tiny. 该版本通过减少 OAN 和 TAN 的特征维度来降低特征复杂度. 具体而言,OAN 中减少了多层感知机和转换矩阵估计中全连接层的数量,轻量化处理后的 OAN 如图6所示. 同时,TCFNet-tiny 减小了 TAN 中的线性层提取特征的维度,将 $\boldsymbol{h}^x$ 和 $\boldsymbol{h}^y$ 的维度从 8×256 降至 8×32. TCFNet-tiny 和 TCFNet 的性能对比如表9所示. 可以得出,通过降低 OAN 和 TAN 中特征的维度,TCFNet-tiny 的运算量可以降至与 AB 和 AM 相当的水平. 同时,TCFNet-tiny 的跟踪准确率仅轻微下降,仍在 MOTA 和 IDF1 指标上超越了当前的跟踪算法,表明了本文方法具有良好的实际应用前景.

表9　TCFNet-tiny 和 TCFNet 性能对比

方法	Parameters / M	FLOPs / G	MOTA/%	IDF1/%	MDA/%
TCFNet	48.45	414.39	**53.81**	**68.64**	**40.62**
TCFNet-tiny	**47.11**	**403.81**	52.86	68.27	40.09

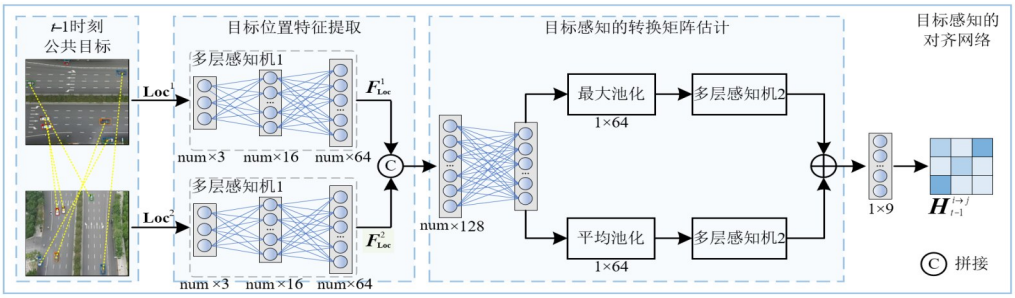


图6 轻量化OAN结构图

4.5 定性分析

为了直观展现所提TCFNet的有效性，在图7中将TCFNet和以Carafe为检测网络的MIA-Net（简称CM）在MDMT测试集中部分场景下的跟踪结果进行可视化，并使用红色方框标记了算法漏跟的目标.

图7（a）中存在大量的小目标受到不同程度的背景遮挡. 在这种场景下，CM在无人机1视角下第57帧和第234帧丢失了大量被遮挡的目标. 作为对比，由于这些被背景遮挡的目标在无人机2的视角下清晰可见，TCFNet通过从无人机2视角中挖掘了多机的互补信息，有效提升了对无人机1中被遮挡目标的跟踪性能.

图7（b）中的杂乱背景为鲁棒跟踪带来了挑战. 第115帧时，CM在2个无人机视频中均丢失了同一个目标，而TCFNet在2个视频中均对其进行了成功跟踪. 在第136帧中，CM因关联失败为相同的目标分配了不同的身份标识，即无人机1中的目标385和588在无人机2中被识别为目标733和385. 作为对比，TCFNet在跨视角图像中成功对齐了这些公共目标的身份标识，表明本文方法对齐和融合跨摄像头特征的策略能在复杂场景下提升模型识别目标的能力.

（a）背景遮挡

（b）背景杂乱

（c）光照变换

图7 跟踪结果对比可视化

图7(*c*)中的光照条件复杂,大量的阴影对模型造成了干扰. 在这种复杂的光照条件下,CM在第127帧时因阴暗的光照丢失了无人机1中的部分目标,且受遮挡的干扰在无人机2的视角下中丢失了许多目标. 在这些场景下,提出的TCFNet对这些目标进行了准确跟踪,证明本文算法对遮挡和光照条件变换具有较好的鲁棒性.

为了进一步说明TCFNet的跟踪性能,在图8中展示了其他涵盖各种真实世界跟踪场景的案例,包括杂乱背景、昼夜时段、光照变换和不同目标尺寸等挑战. 从图8中可以得出,在这些具有挑战性的真实场景下,TCFNet仍在不同无人机视频中对多个目标实现了准确的定位和识别,证明所提方法在多种复杂场景下都具有良好的性能.

图8　TCFNet跟踪结果可视化

4.6　局限性分析

尽管提出的TCFNet在许多复杂场景下取得了领先的跟踪性能,但在某些情况下仍然会出现一定的跟踪错误. 图9展示了TCFNet跟踪中2个典型的错误情形,描述了目标丢失的具体示例. 如图9所示,场景1中的209号目标和场景2的277号目标均在无人机1的视角中被成功跟踪,但在无人机2的视角却出现了目标丢失,造成这一错误的主要原因在于目标的尺寸太小. 对于小尺寸目标,进行跨视角图像特征对齐的过程中空间维度上微小的误差就可能造成目标关键信息被映射到另一视角的非目标区域. 这种现象会限制模型协同多视角图像互补信息提升跟踪性能的能力,从

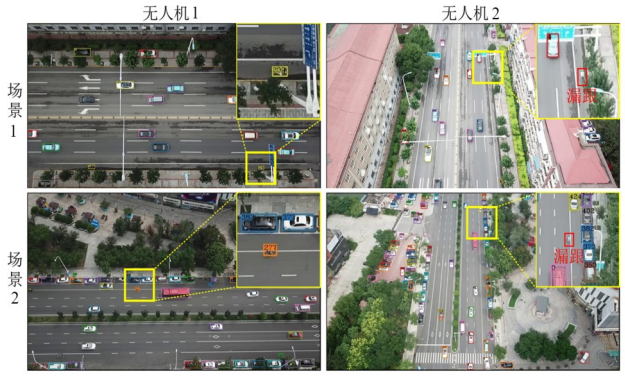


图9 TCFNet错误结果示例

而对网络对齐跨视角特征的准确性提出更高的要求.在未来的研究中,如何进一步提升不同无人机所捕获图像特征的对齐和融合准确性以提升模型对各尺度目标的跟踪性能仍是一个研究重点.

5 结论

针对不同无人机捕获图像间视角和尺度差异大造成跨机特征对齐和融合困难的问题,为多无人机多目标跟踪设计了一种TCFNet算法.与现有仅在决策级融合多个无人机跟踪结果的方法不同,TCFNet使用跟踪过程中跨视角的时序先验估计不同无人机图像间的转换关系,从而实现不同无人机图像特征间的对齐和融合,缓解复杂场景下的跟踪性能下降.在MDMT数据集上的对比结果表明,相较于当前先进的跟踪方法,TCFNet将跟踪准确率、识别F1值和多机目标关联分数分别提升了2.23、1.67和2.15个百分点.后续的工作将针对目标间的交互关系展开进一步研究,提升复杂场景下关联不同无人机视频中公共目标身份标识的能力.

参考文献

[1] 陈阳, 皮德常, 代成龙, 等. 多无人机协同陆地设施辅助移动边缘计算的系统能耗最小化方法[J]. 电子学报, 2023, 51(4): 984-992.
CHEN Y, PI D C, DAI C L, et al. Energy minimization for multi-UAVs cooperative ground access points assisted mobile edge computing[J]. Acta Electronica Sinica, 2023, 51(4): 984-992. (in Chinese)

[2] 任双, 周洁, 高嵩, 等. 基于注意力机制的无人机集群协同分群控制算法[J]. 电子学报, 2023, 51(7): 1898-1905.
REN S, ZHOU J, GAO S, et al. Cooperative fission control algorithm of UAV swarm based on attention mechanism[J]. Acta Electronica Sinica, 2023, 51(7): 1898-1905. (in Chinese)

[3] LIU Z H, SHANG Y Y, LI T M, et al. Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark[J]. IEEE Transactions on Multimedia, 2023, 25: 1462-1476.

[4] LIU R S, LIU Z, LIU J Y, et al. A task-guided, implicitly-searched and meta-initialized deep model for image fusion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(10): 6594-6609.

[5] 乔通, 陈彧星, 谢世闯, 等. 多色彩通道特征融合的GAN合成图像检测方法[J]. 电子学报, 2024, 52(3): 924-936.
QIAO T, CHEN Y X, XIE S C, et al. GAN synthetic image detection using fused features in the multi-color channels[J]. Acta Electronica Sinica, 2024, 52(3): 924-936. (in Chinese)

[6] ZHAO Z X, BAI H W, ZHANG J S, et al. Equivariant multi-modality image fusion[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 25912-25921.

[7] WANG Z D, ZHENG L, LIU Y X, et al. Towards real-time multi-object tracking[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 107-122.

[8] VOIGTLAENDER P, KRAUSE M, OSEP A, et al. MOTS: Multi-object tracking and segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 7934-7943.

[9] LIANG C, ZHANG Z P, ZHOU X, et al. Rethinking the competition between detection and ReID in multiobject tracking[J]. IEEE Transactions on Image Processing, 2022, 31: 3182-3196.

[10] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2017: 3645-3649.

[11] SUN S J, AKHTAR N, SONG H S, et al. Deep affinity network for multiple object tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 104-119.

[12] GUO S, WANG J Y, WANG X C, et al. Online multiple object tracking with cross-task synergy[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 8132-8141.

[13] REN H, HAN S D, DING H L, et al. Focus on details: Online multi-object tracking with diverse fine-grained representation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 11289-11298.

[14] PANG J M, QIU L L, LI X, et al. Quasi-dense similari-

ty learning for multiple object tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 164-173.

[15] YU E, LI Z L, HAN S D. Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8824-8833.

[16] KIM C, LI F X, ALOTAIBI M, et al. Discriminative appearance modeling with multi-track pooling for real-time multi-object tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 9548-9557.

[17] YOU S S, YAO H T, BAO B K, et al. UTM: A unified multiple object tracking model with identity-aware feature enhancement[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 21876-21886.

[18] BEWLEY A, GE Z Y, OTT L, et al. Simple online and realtime tracking[C]//2016 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2016: 3464-3468.

[19] CAO J K, PANG J M, WENG X S, et al. Observation-centric sort: Rethinking sort for robust multi-object tracking[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 9686-9696.

[20] QIN Z, ZHOU S P, WANG L, et al. MotionTrack: Learning robust short-term and long-term motions for multi-object tracking[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 17939-17948.

[21] CHEN L, AI H Z, ZHUANG Z J, et al. Real-time multiple people tracking with deeply learned candidate selection and person re-identification[C]//2018 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2018: 1-6.

[22] ZHANG Y F, SUN P Z, JIANG Y, et al. Bytetrack: Multi-object tracking by associating every detection box[M]// Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 1-21.

[23] HUANG Z Y, FU C H, LI Y M, et al. Learning aberrance repressed correlation filters for real-time UAV tracking[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 2891-2900.

[24] LIN F L, FU C H, HE Y J, et al. BiCF: Learning bidirectional incongruity-aware correlation filter for efficient UAV object tracking[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2020: 2365-2371.

[25] CHEN J J, XU T F, HUANG B, et al. ARTracker: Compute a more accurate and robust correlation filter for UAV tracking[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 6514605.

[26] CAO Z A, FU C H, YE J J, et al. SiamAPN: Siamese attentional aggregation network for real-time UAV tracking[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2021: 3086-3092.

[27] YE J J, FU C H, CAO Z A, et al. Tracker meets night: A transformer enhancer for UAV tracking[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 3866-3873.

[28] XING D T, EVANGELIOU N, TSOUKALAS A, et al. Siamese transformer pyramid networks for real-time UAV tracking[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2022: 1898-1907.

[29] YU H Y, LI G R, SU L, et al. Conditional GAN based individual and global motion fusion for multiple object tracking in UAV videos[J]. Pattern Recognition Letters, 2020, 131: 219-226.

[30] LIU S, LI X, LU H C, et al. Multi-object tracking meets moving UAV[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8866-8875.

[31] WU H, NIE J H, HE Z W, et al. One-shot multiple object tracking in UAV videos using task-specific fine-grained features[J]. Remote Sensing, 2022, 14(16): 3853-3872.

[32] CHENG S, YAO M B, XIAO X M. DC-MOT: Motion deblurring and compensation for multi-object tracking in UAV videos[C]//2023 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2023: 789-795.

[33] SHI L K, ZHANG Q R, PAN B, et al. Global-local and occlusion awareness network for object tracking in UAVs[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 8834-8844.

[34] CHEN G L, ZHU P F, CAO B, et al. Cross-drone transformer network for robust single object tracking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9): 4552-4563.

[35] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[36] FISCHLER M A, BOLLES R C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography[M]//Readings in Computer Vision. Amsterdam: Elsevier, 1987: 726-740.

[37] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer network[C]//NIPS´15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. Massachusetts, MIT Press, 2015: 2017-2025.

[38] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[39] CHEN K, WANG J Q, PANG J M, et al. MMDetection: Open MMLab detection toolbox and benchmark[EB/OL]. (2019-06-17)[2024-03-25].https://arxiv.org/abs/1906.07155v1.

[40] LINDENBERGER P, SARLIN P E, POLLEFEYS M. LightGlue: Local feature matching at light speed[C]// 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 17581-17592.

[41] EDSTEDT J, SUN Q Y, BÖKMAN G, et al. RoMa: Robust dense feature matching[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 19790-19800.

[42] ZHU Y H, SUN X Y, WANG M, et al. Multi-modal feature pyramid transformer for RGB-infrared object detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(9): 9984-9995.

[43] ZHAO Z X, BAI H W, ZHANG J S, et al. CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 5906-5916.

[44] FENG C J, ZHONG Y J, GAO Y, et al. TOOD: Task-aligned one-stage object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 3490-3499.

[45] ZHU B J, WANG J F, JIANG Z K, et al. AutoAssign: Differentiable label assignment for dense object detection[EB/OL]. (2020-07-07)[2024-03-25]. https://arxiv.org/abs/2007.03496v3.

[46] WANG J Q, CHEN K, XU R, et al. CARAFE: Content-aware reassembly of features[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 3007-3016.

[47] VU T, JANG H, PHAM T X, et al. Cascade RPN: Delving into high-quality region proposal network with adaptive convolution[EB/OL]. (2019-09-15) [2024-03-25]. https://arxiv.org/abs/1909.06720v2.

[48] HE L X, LIAO X Y, LIU W, et al. FastReID: A pytorch toolbox for general instance re-identification[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 9664-9667.

[49] CHAVDAROVA T, BAQUÉ P, BOUQUET S, et al. WILDTRACK: A multi-camera HD dataset for dense unscripted pedestrian detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 5030-5039.

[50] FENG W, WANG F F, HAN R Z, et al. Unveiling the power of self-supervision for multi-view multi-human association and tracking[EB/OL]. (2024-01-31) [2024-03-25]. https://arxiv.org/abs/2401.17617v1.

[51] ONG J, VO B T, VO B N, et al. A bayesian filter for multi-view 3D multi-object tracking with occlusion handling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(5): 2246-2263.

[52] YOU Q Z, JIANG H. Real-time 3D deep multi-camera tracking[EB/OL]. (2020-03-26) [2024-03-25]. https://arxiv.org/abs/2003.11753v1.

[53] CHENG C C, QIU M X, CHIANG C K, et al. ReST: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 10017-10026.

[54] BERGMANN P, MEINHARDT T, LEAL-TAIXÉ L. Tracking without bells and whistles[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 941-951.

[55] ZHOU X Y, KOLTUN V, KRÄHENBÜHL P. Tracking objects as points[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 474-490.

[56] WU J L, CAO J L, SONG L C, et al. Track to detect and segment: An online multi-object tracker[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 12347-12356.

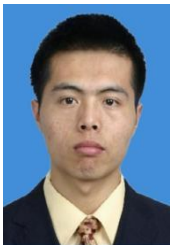
[57] MEINHARDT T, KIRILLOV A, LEAL-TAIXÉ L, et al. Trackformer: Multi-object tracking with transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8834-8844.

[58] RISTANI E, TOMASI C. Features for multi-target multi-camera tracking and re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6036-6046.

作者简介

伍　瀚　男,1999年生,湖南邵阳人. 国防科技大学电子科学学院CEMEE国家重点实验室在读博士研究生. 主要研究方向为多源影像智能处理、视频目标检测与跟踪、机器学习与神经网络.

E-mail: wuhan0326@nudt.edu.cn



孙　浩　男,1984年生,陕西三原人. 博士,国防科技大学电子科学学院CEMEE国家重点实验室副教授. 主要研究方向为多源遥感图像协同解译与语义挖掘.