中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Wang Jiawei, Tian Jing, Yuan Ziyang, Zhang Xiaodian, Zhao Liangjin. Windowed Attention Driven Multi-View Geometric Perception and Tracking[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260068. (王佳伟, 田璟, 苑子杨, 张晓典, 赵良瑾. 窗口化注意力驱动的多视几何感知与跟踪[J/OL]. 中国图象图形学报,XXXX:1-15. DOI: 10.11834/jig.260068.)[DOI:10.11834/jig.260068]

窗口化注意力驱动的多视几何感知与跟踪

王佳伟[1,2,3,4],田璟[1,4],苑子杨[1,4],张晓典[1,4],赵良瑾[1,4]

1. 中国科学院空天信息创新研究院,北京,100190; 2. 中国科学院大学,北京,100190; 3. 中国科学院大学电子电气;与通信工程学院,北京,100190; 4. 目标认知与应用技术国家级重点实验室,北京,100190

摘　要:**目的**　多摄像头多目标跟踪(Multi-Camera Multi-Object Tracking, MC-MOT)在视频监控与自动驾驶中具有重要应用,但现有方法常因跨视角遮挡导致特征不连续,往往不得不依赖复杂的后端图优化算法进行补偿。**方法**　本文提出一种基于强时空表征的统一感知框架,通过在特征提取阶段深度融合时空信息,得到高质量的目标特征,提升目标跟踪的效果。首先,利用包含相机内参和外参的几何投影模块,将多视角的2D图像特征提升并融合至统一的3D鸟瞰图(Bird's-Eye-View, BEV)空间。其次,设计一种窗口化时空融合模块,利用交叉注意力机制让当前帧BEV特征(Query, $\boldsymbol{Q}$)在局部窗口内动态聚合历史帧特征(Key/Value,$\boldsymbol{K/V}$),实现特征去噪与时序平滑。最后,结合具有鲁棒性的时空表征,采用基于运动模型的简单卡尔曼滤波配合匈牙利算法,完成高精度的跨帧数据关联。**结果**　在Wildtrack多视角数据集上的实验结果表明,该方法取得了极具竞争力的性能,其中*IDF1*为92.23%,相比基线提升2.04%,*MOTA*为89.70%,提升1.8%,同时在MultiviewX上依然取得先进的跟踪结果。消融实验显示了引入窗口化时空融合模块的有效性。**结论**　本文提出的强时空表征统一感知框架有效解决了多摄像头多目标跟踪中的特征对齐与长时遮挡问题。研究证明,端到端时空表征的构建在MC-MOT任务中展现出显著的性能增益,高质量的特征使得基础关联算法足以应对复杂的关联挑战,为该领域提供了一种高效、简洁的新基准。本文相关代码与数据集已在ScienceDB开放共享,DOI:https://doi.org/10.57760/sciencedb.j00240.00082

关键词:多目标跟踪;鸟瞰图;时空特征融合;带窗交叉注意力;卡尔曼滤波

Windowed Attention Driven Multi-View Geometric Perception and Tracking

Wang Jiawei[1,2,3,4], Tian Jing[1,4], Yuan Ziyang[1,4], Zhang Xiaodian[1,4], Zhao Liangjin[1,4]

1. Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 100190, China; 3. School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China; 4. Key Laboratory of Target Cognition and Application Technology (TCAT), Beijing 100190, China

Abstract: **Objective** Multi-Camera Multi-Object Tracking (MC-MOT) has emerged as a cornerstone technology in the evolution of intelligent transportation systems, large-scale video surveillance, and autonomous driving. Unlike single-camera tracking, MC-MOT requires the seamless integration of spatial information from overlapping or non-overlapping fields of view to maintain a consistent global identity for each target. However, contemporary research in this domain faces a significant bottleneck: feature discontinuity caused by severe cross-view occlusions and the inherent perspective distortion of 2D sensors. To mitigate these issues, existing frameworks often rely heavily on computationally expensive backend optimization algorithms, such as complex Graph Neural Networks (GNNs) or multi-frame batch processing, to compensate for the

收稿日期:2026-01-29;**修回日期**:2026-03-24

基金项目:国家自然科学基金(项目编号:62571515)

Supported by: the National Natural Science Foundation of China (62571515)

inaccuracies of frontend detection and re-identification. Such dependencies not only increase the system's inference latency but also limit its deployment in real-time scenarios. Therefore, there is an urgent need for a framework that can fundamentally enhance the robustness of feature representations at the perception stage. **Method** To address these critical challenges, this paper proposes a novel, unified perception framework driven by robust spatio-temporal representation learning. The core philosophy of our approach is to shift the burden of tracking from complex association logic to high-fidelity feature extraction. The proposed method consists of three primary technical contributions: Geometric Projection and Feature Lifting: We first introduce a geometric projection module that leverages camera intrinsic and extrinsic parameters. This module transforms 2D image-plane features into a unified 3D Bird's-Eye-View (BEV) representation. By projecting multi-view features onto a shared ground plane, we effectively eliminate perspective ambiguity and provide a spatially consistent foundation for subsequent multi-sensor fusion. Windowed Spatio-Temporal Fusion (WSTF): To further enhance feature robustness against transient occlusions, we design the Windowed Spatio-Temporal Fusion (WSTF) module. Unlike the generic Window-based Vision Transformer (Swin-Transformer) which focuses on spatial hierarchy for image classification, our WSTF is specifically tailored for spatio-temporal alignment in the BEV domain. It employs a cross-attention mechanism where the BEV features of the current frame serve as Queries ($\boldsymbol{Q}$), while the historical features stored within a local temporal window (T) act as Keys ($\boldsymbol{K}$) and Values ($\boldsymbol{V}$). This localized attention mechanism constrains the search space to a physically plausible neighborhood, effectively realizing feature denoising and temporal smoothing while significantly reducing the quadratic computational complexity typically associated with standard Transformers. Streamlined Association Strategy: Leveraging these high-quality, spatio-temporally aligned representations, we demonstrate that a simplified data association strategy is sufficient. We combine a standard Kalman Filter (KF) for motion state estimation with the Hungarian algorithm for bipartite matching. This design choice explicitly eliminates the need for exhaustive backend graph optimization, thereby significantly improving the Frame Per Second (FPS) performance of the overall system. **Results** The proposed framework was rigorously evaluated on two benchmark datasets: Wildtrack and MultiviewX. On the challenging Wildtrack dataset, which is characterized by dense crowds and frequent occlusions, our method achieved highly competitive results, recording an Identification F1-Score (*IDF1*) of 92.23% (a 2.04% improvement over the current baseline) and a Multi-Object Tracking Accuracy (*MOTA*) of 89.70% (a 1.8% improvement). Similar state-of-the-art (SOTA) performance was maintained on the MultiviewX dataset. Furthermore, comprehensive ablation studies were conducted to analyze the sensitivity of the temporal window length T. The results indicate that a window size of T=5 strikes the optimal balance between tracking precision and computational latency. Quantitative analysis of inference speed shows that our model achieves a real-time processing rate of 18.5 FPS, outperforming many graph-based optimization methods that require batch processing. **Conclusion** In conclusion, this study demonstrates that constructing an end-to-end spatio-temporal representation in the BEV space yields substantial performance gains for MC-MOT. By introducing the WSTF module, we effectively resolve the long-standing issues of feature alignment and long-term occlusion. Crucially, this work proves that high-quality frontend features can render basic association algorithms sufficient for handling complex tracking challenges, providing an efficient and concise new benchmark for the field of multi-sensor fusion and computer vision. **Data and Code Availability** The implementation code and the associated datasets supporting the findings of this study have been openly archived in the Science Data Bank (ScienceDB) athttps://doi.org/10.57760/sciencedb.j00240.00082.

Key words: Multi-Camera Multi-Object Tracking (MC-MOT); Bird's-Eye View (BEV); Spatio-Temporal Feature Fusion; Windowed Cross-Attention; Kalman Filter

1 引言与相关工作

1.1 研究背景与多相机多目标跟踪范式

多摄像头多目标跟踪(Multi-Camera Multi-Object Tracking, MC-MOT)作为计算机视觉领域的关键任务,在广域视频监控、自动驾驶环境感知及智慧城市精细化管理等众多实际应用中发挥着不可替代的作用(Fei and Han, 2023 , Cherdchusakulchai等, 2024 , Tran and Vi, 2024 , Chen 等, 2025)。相比单相机跟踪,MC-MOT的核心优势在于能够利用多视角信息的几何冗余性与互补性(Cheng 等,

2023，Wang 等，2024b，Ortiz 等，2024，Kim and Bras，2022），从根本上缓解了单视角下常见的深度模糊、短暂遮挡等瓶颈（Dendorfer 等，2020）。

根据信息融合阶段的不同，现有方法主要可划分为两类范式。晚期融合（结果级融合）：遵循先检测后关联原则（Wang 等，2024a，Szcs 等，2024，Tran and Vi，2024）。此类方法在各单视角独立检测后，利用相机标定参数将 2D 轨迹投影至统一坐标系，通过聚类（Fleuret 等，2008）或图优化（Zhang 等，2008；Pang 等，2023）匹配。虽模块化程度高，但由于检测阶段未利用视角互补性，当单视图发生严重遮挡时，误差累积会导致频繁的漏检与身份混淆。

早期融合（特征级融合）：主张在特征提取阶段构建统一表征。MVDet（Hou 等，2020）开创性地将多视角 2D 特征投影至鸟瞰图（Bird's-Eye-View，BEV）平面，提升了拥挤场景下的定位精度。随后，SHOT（Song 等，2021）优化了高度估计，EarlyBird（Teepe 等，2024b；Li 等，2024）引入 Transformer 增强跨视角交互。此类方法通过消除透视畸变保持了物理尺度一致性（Yamane 等，2025），实现了高效空间对齐。

1.2 BEV 下的时序建模挑战

尽管早期融合解决了空间一致性，但现有的框架普遍存在空间统一但时序孤立的缺陷。现有的单帧空间融合策略在面对动态目标时，会导致 BEV 特征在时间维度出现抖动。随着视图变换技术的发展（Philion and Fidler，2020），如何高效融合时序上下文成为关键。

早期的时序融合如 BEVDet4D（Huang and Huang，2022）采用朴素的特征拼接，忽略了运动引起的空间错位，导致特征重影。为解决对齐问题，BEVFormer（Li 等，2024）引入时序自注意力机制，利用可变形注意力搜索对应点，但全局搜索的高额计算开销难以满足在线跟踪的实时性要求。而 Sparse4D（Lin 等，2022）等稀疏采样方法在处理高密度人群时易丢失细粒度特征。这种无记忆性或低效的特征提取，迫使下游任务不得不依赖极高复杂度的关联算法来弥补前端表征的缺失。

1.3 本文动机与贡献

针对上述局限性，本文提出了一种窗口化时空注意力（Vaswani 等，2017，Liu 等，2021）驱动的多视几何感知与跟踪框架（Windowed Spatio-Temporal Tracking Framework，简称 WST-Track）。本文的核心观点是：若前端网络能够构建出具有高度时序一致性和抗遮挡能力的 BEV 表征，后端的关联过程可以被极大简化。

不同于以往依赖重后端优化策略的方法，如 MPN-Track（Brasó and Leal-Taixé，2020）利用 GNN（Graph-Neural- Network）推断连接概率，或 TransTrack（Sun 等，2020）通过 Object Query 生成轨迹，本文主张回归轻后端范式。受 SORT（Bewley 等，2016）和 ByteTrack（Zhang 等，2022）启发，我们利用强时空表征生成的平滑轨迹特征，使得基础的线性关联模型即可在复杂场景下取得卓越性能。

本文的主要贡献总结如下：

提出了一种基于强时空表征的统一跟踪框架 WST-track，证明了强化前端特征鲁棒性后，可以使用传统的线性运动模型平衡跟踪精度与效率。

设计了速度引导的窗口化时空融合模块。该模块通过显式的运动对齐与局部注意力机制，有效解决了动态场景下的特征漂移与长时遮挡问题，显著提升了 BEV 表征的判别力。

在 Wildtrack（Chavdarova and Baqu，2018）和 MultiviewX（Hou 等，2020）等多视角数据集上取得了具有竞争力的性能，实验结果验证了所提时空联合建模方法在维持轨迹连续性方面的有效性。

2 方　法

2.1 总体框架

如图 1 所示，本文提出了一种基于强时空表征的端到端多摄多目标跟踪框架 WST-Track。该框架旨在通过构建统一的鸟瞰图表征，解决多视角感知中存在的空间不对齐与长时遮挡问题。系统整体流程遵循重前端表征、轻后端关联的设计原则，主要由三个核心阶段组成：多视几何特征投影、窗口化时空特征融合、以及基于物理运动模型的检测与关联。

首先，在多视几何特征投影阶段，系统接收 N 路时间同步的二维图像序列 $\boldsymbol{I}_t=\{\boldsymbol{I}_t^1,\boldsymbol{I}_t^2,\boldsymbol{I}_t^3,\cdots,\boldsymbol{I}_t^N\}$ 及其对应的相机内外参数作为输入。利用共享权重的骨干网络（Backbone）对单视图进行特征提取，通过视图投影模块引入几何先验，将多视角图像特征从各自的透视坐标系提升（Lifting）并投影至统一的

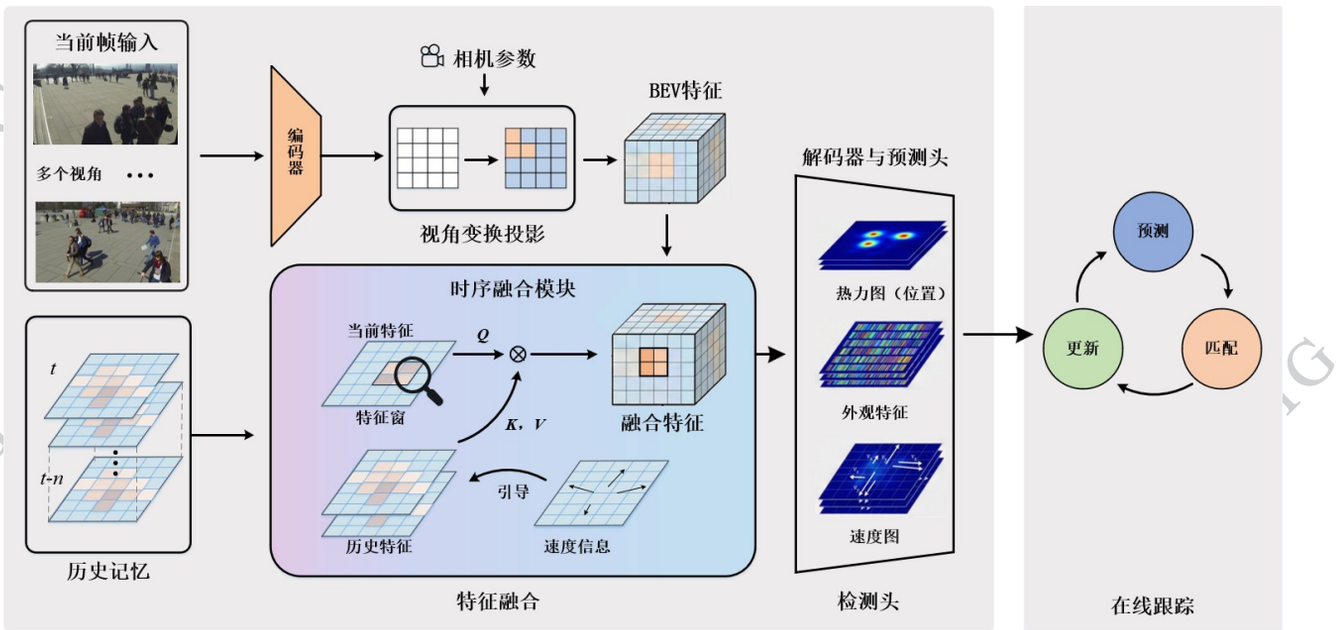

图1　方法框架

Fig. 1　Framework of the Proposed Method

3D BEV空间。这一过程有效地将不同视角的语义信息在几何空间上进行了对齐，生成了初始的空间BEV特征 $\boldsymbol{B}_t^{spatial}$。式中，$\boldsymbol{I}_t^i \in \mathbb{R}^{3\times H\times W}$ 为输入的第 i 路视角图像。

其次，为了弥补单帧感知在遮挡恢复与深度估计上的不足，引入了窗口化时空特征融合阶段。该模块维护一个历史特征记忆库，将当前帧的BEV特征作为查询向量（Query，$\boldsymbol{Q}$），与历史帧缓存的特征（Key/Value，$\boldsymbol{K}/\boldsymbol{V}$）进行交互。通过设计的窗口化交叉注意力机制（Windowed Cross-Attention），模型能够自适应地聚合局部时空邻域内的上下文信息。这一过程不仅平滑了特征的抖动，还能够利用时序一致性补全被遮挡目标的特征表示，从而输出高质量的时空融合特征 $\boldsymbol{B}_t^{fused}$。

最后，在检测与关联阶段，增强后的时空特征被送入多任务检测头，以全卷积的方式并行回归出目标的热力图置信度、3D位置偏移量及尺寸信息。得益于前端生成的鲁棒时空表征，目标的运动轨迹呈现出高度的线性与连续性。因此，本框架摒弃了复杂的图优化后处理，转而采用基于线性匀速运动模型的卡尔曼滤波进行状态估计，并结合匈牙利算法完成最终的跨帧数据关联，实现了高效且精准的多目标跟踪。

2.2　多视几何投影与BEV特征构建

为了实现跨视角的特征融合，本文首先需要将多路输入的二维图像特征转换至统一的三维空间。本节详细阐述特征提取主干网络的设计，以及基于深度估计的几何投影机制。

2.2.1　二维特征提取

给定 N 路时间同步的输入图像序列 $\boldsymbol{I}_t=\{\boldsymbol{I}_t^1,\boldsymbol{I}_t^2,\boldsymbol{I}_t^3,\cdots,\boldsymbol{I}_t^N\}$，式中 $\boldsymbol{I}_t\in\mathbb{R}^{H\times W\times 3}$。为了保证特征提取的高效性与一致性，本文采用权重共享的ResNet-18网络作为Backone。每张输入图像经过主干网络的层级处理后，输出高维语义特征图 $\boldsymbol{F}_{img}\in\mathbb{R}^{H_f\times W_f\times C}$，式中 H_f，W_f 为下采样后的特征图尺寸，C 为特征通道数。为了更好地捕捉多尺度的上下文信息，在主干网络末端引入了特征金字塔结构（Feature Pyramid Network，FPN），以增强对不同大小目标的表征能力。

2.2.2　基于深度分布的几何视图变换

传统的逆透视映射（Inverse Perspective Mapping，IPM）依赖于平坦地面的假设，在复杂场景下容易产生畸变。为了解决这一问题，本文采用了基于隐式深度估计的视图变换策略，通过联合学习图像特征与深度分布，将二维特征“提升”为三维视锥点云。具体而言，视图投影模块包含两个并行的预测头：一个用于优化图像语义特征 $\boldsymbol{F}_{img}$，另一个用于预

测每个像素的离散深度分布 $\boldsymbol{D} \in \mathbb{R}^{H_f \times W_f \times D_{bins}}$，式中 D_{bins} 为深度预测位数。深度分布预测头通过分类的方式，计算每个特征点落在预定义深度区间内的概率。基于预测的深度分布 $\boldsymbol{D}$ 和语义特征 $\boldsymbol{F}_{img}$，利用相机的内参矩阵 $\boldsymbol{K}$ 和外参矩 $\boldsymbol{E}$，执行从图像坐标系到世界坐标系的几何投影。该过程可形式化地表示为：

$$\boldsymbol{P}_{bev} = \mathrm{P}(\boldsymbol{F}_{img}, \boldsymbol{D}, \boldsymbol{K}, \boldsymbol{E}) \tag{1}$$

式中，$\mathcal{P}$ 表示几何投影函数，$\boldsymbol{P}_{bev}$ 为生成的三维特征点云。在此过程中，每个二维像素点 (u, v) 根据其预测的深度 d，结合相机参数被映射到三维空间坐标 (x, y, z)：

$$d \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \boldsymbol{K} \cdot \boldsymbol{E} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{2}$$

通过上述变换，每个相机的特征图被转化为一个三维视锥特征体。随后，为了构建规则的鸟瞰图表征，采用体素池化操作，将所有视锥特征点拍平到预定义的 BEV 网格上。落入同一网格内的特征通过求和或平均池化进行聚合，最终生成统一的 BEV 空间特征 $\boldsymbol{B} \in \mathbb{R}^{X \times Y \times C'}$，式中 X, Y 为 BEV 网格在空间维度上的宽度与高度，C' 表示特征通道数。为后续的窗口化时空融合模块提供了物理对齐的特征基础。

2.3 窗口化时空特征融合

尽管多视几何投影实现了空间上的特征对齐，但在复杂动态场景中，单帧 BEV 表征仍面临严重的挑战：一方面，物体间的相互遮挡会导致瞬时特征缺失；另一方面，深度估计的抖动会引入位置噪声。为了解决上述问题，本节提出了一种速度引导的窗口化时空特征融合模块。该模块旨在通过运动补偿和局部注意力机制，从历史记忆中检索并聚合有效信息，以增强当前帧表征的时序一致性与鲁棒性。需说明的是，本文窗口化注意力与 Swin-Transformer 仅形式相似，核心差异在于前者针对多摄跟踪动态场景设计，后者聚焦单视图静态特征建模。

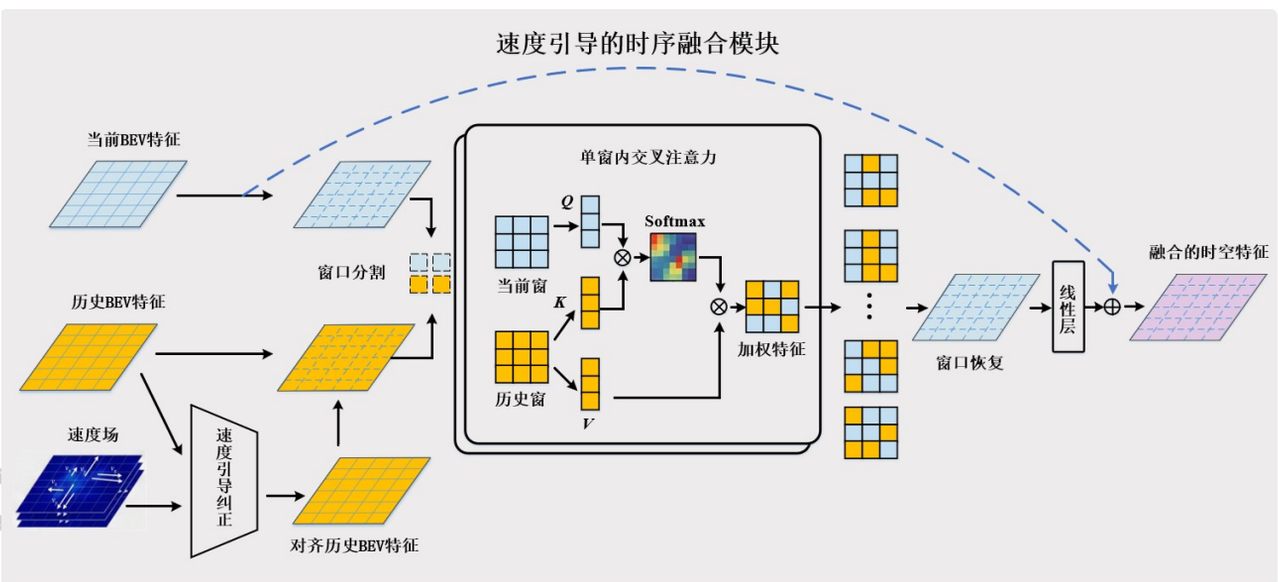

图2 速度引导的时序融合模块示意图

Fig. 2 Schematic of Velocity-Guided Temporal Fusion Module

2.3.1 速度引导的运动对齐与历史记忆构建

为了利用时序信息，构建一个历史特征记忆库，用于缓存上一帧的融合特征 $\boldsymbol{F}_{(t-1)} \in \mathbb{R}^{(C \times H \times W)}$。然而，直接将历史特征 $\boldsymbol{F}_{(t-1)}$ 与当前帧特征 $\boldsymbol{F}_t$ 进行融合是不可行的，因为目标在两帧之间发生了位移。

为此，本文引入了速度引导的显式对齐机制。利用检测分支预测的目标速度场 $\boldsymbol{V} \in \mathbb{R}^{2 \times H \times W}$，对相邻帧间各空间位置的位移偏移量 $(\Delta x, \Delta y)$ 进行编码。基于该速度场，采用双线性插值操作 $\mathcal{W}(\cdot)$ 对历史特征进行空间重采样，使其在空间分布上与当前帧对齐：

$$\hat{\boldsymbol{F}}_{(t-1)}(\boldsymbol{p}) = \mathcal{W}(\boldsymbol{F}_{(t-1)}, \boldsymbol{V}) = \boldsymbol{F}_{(t-1)}(\boldsymbol{p} - \Delta \boldsymbol{p}(\boldsymbol{V})) \tag{3}$$

式中，$\boldsymbol{p}=(x,y)$表示网格坐标，$\Delta\boldsymbol{p}$为归一化后的位移矢量。经过对齐后的历史特征$\hat{\boldsymbol{F}}_{t-1}$消除了运动引起的空间错位，为后续的细粒度特征交互提供了像素级对应的先验基础。Swin-Transformer仅针对单张静态图像进行空间特征建模，无需考虑时序帧间的目标位移，更无需引入速度引导的对齐机制，无法适配多摄跟踪中目标跨帧、跨视图的动态移动特性。

2.3.2　窗口化交叉注意力机制

虽然全局注意力机制能够捕获长距离依赖，但在高分辨率的BEV特征图上，其计算复杂度高达$O((HW)^2)$，且容易受到背景噪声的干扰。考虑到经过运动对齐后，目标特征主要集中在局部邻域内，本文设计了窗口化交叉注意力机制。该机制引入了局部性的归纳偏置，仅在限制的$w\times w$窗口内进行特征交互，从而在大幅降低计算开销的同时，聚焦于目标核心区域的特征增强，此外，本文所设置的8×8空间窗口在实际物理空间中对应0.2 m×0.2 m的物理感受野范围。该尺度能够精准聚焦目标局部区域，恰好覆盖行人主体的关键特征区域，使空间特征聚合更贴合真实监控场景下的目标分布特性，保证了特征提取的稳定性与合理性。

具体而言，将当前帧特征$\boldsymbol{F}_t$和对齐后的历史特征$\hat{\boldsymbol{F}}_{t-1}$划分为$M$个互不重叠的局部窗口$\{\boldsymbol{F}_t^i\}_{i=1}^M$和$\{\hat{\boldsymbol{F}}_{t-1}^i\}_{i=1}^M$。在每个窗口内，构建Query、Key和Value向量以执行交叉注意力运算：

$$\boldsymbol{Q}_i=\boldsymbol{F}_t^i\boldsymbol{W}_Q,\boldsymbol{K}_i=\hat{\boldsymbol{F}}_{(t-1)}^i\boldsymbol{W}_K,\boldsymbol{V}_i=\hat{\boldsymbol{F}}_{(t-1)}^i\boldsymbol{W}_V \tag{4}$$

式中，$\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V$为可学习的线性投影矩阵。这里，查询向量$Q$来自当前帧，而键值对$K$，$V$来自历史帧，这种设计使得模型能够以当前观测为基准，自适应地从历史记忆中检索互补的特征线索（如被遮挡前的外观纹理）。窗口内的注意力输出$\boldsymbol{Z}_i$计算如下：

$$\boldsymbol{Z}_i=Attention(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i)=Softmax\left(\frac{(\boldsymbol{Q}_i\boldsymbol{K}_i^{\top})}{\sqrt{(d_k)}}\right)\boldsymbol{V}_i$$

(5)式中$\sqrt{(d_k)}$为缩放因子，用于防止梯度消失。通过这种机制，当当前帧某位置因遮挡导致特征响应微弱时（$\boldsymbol{Q}_i$较弱），如果历史帧中该位置存在高置信度特征（$\boldsymbol{V}_i$较强）且两者语义相似度高（Attention权重与$Q_iK_i^{\top}$相关），模型便能自动将历史信息注入当前帧，实现特征修复

2.3.3　融合输出

最后，将所有窗口的输出特征$\{\boldsymbol{Z}_i\}_{i=1}^M$逆操作还原为原始的空间布局，并通过一个线性投影层$\boldsymbol{W}_O$进行特征聚合。为了保持训练的稳定性并保留当前帧的原始语义，引入了残差连接：

$$\boldsymbol{F}_{fuse}=Norm(\boldsymbol{F}_t+\boldsymbol{W}_O(\mathop{\oplus}_{(i=1)}^{M}\boldsymbol{Z}_i)) \tag{6}$$

式中⊕表示窗口拼接与还原操作，$Norm$表示层归一化。最终生成的融合特征$\boldsymbol{F}_{fuse}$兼具了当前帧的空间精确性与历史帧的时间完整性，显著缓解了运动模糊与遮挡带来的特征漂移问题，为后端的简单关联策略提供了鲁棒的输入表征。Swin-Transformer的特征输出仅服务于静态任务，未考虑多摄跟踪所需的时序一致性，无法为后端关联提供有效支撑，本文融合输出设计正是针对该痛点优化。

2.4　目标检测头设计

解码器采用与编码器对称的多级上采样架构，并引入跳跃连接以恢复空间分辨率。经过时空融合模块处理后，获得了一个具备高语义响应且时序一致的BEV特征图$\boldsymbol{F}_{fuse}\in\mathbb{R}^{(C\times H\times W)}$。为了从该特征中解码出具体的目标状态，本文设计了一个轻量级的多任务全卷积检测头。该模块遵循无锚框的设计范式，避免了复杂的锚框计算，从而保证了系统的高效运行。

2.4.1　网络结构与输出定义

检测头由三个并行的卷积分支组成，分别用于预测目标的中心热力图、速度估计以及重识别（Person Re-identification，ReID）特征嵌入。所有分支均共享$\boldsymbol{F}_{fuse}$作为输入，并通过3×3卷积层进行特征适配，最后由1×1卷积层输出最终的预测结果。

中心热力图分支：输出张量为$\hat{\boldsymbol{Y}}\in[0,1]^{(K\times H\times W)}$，式中$K$表示目标类别数，在本任务中$K$=1。热力图中坐标$(x,y)$处的数值$\hat{Y}_{xy}$表示该位置存在目标中心的置信度。为了生成训练过程中的真值，将每个真实目标中心映射到BEV网格上，并利用二维高斯核函数$Y_{xy}=\mathrm{e}^{(-((x-x_i)^2+(y-y_i)^2)/(2\sigma^2))}$生成软标签，以缓解正负样本的不平衡问题。

速度估计分支：为了辅助运动模型，该分支直接回归目标的二维速度矢量。输出张量为$\hat{\boldsymbol{V}}\in\mathbb{R}^{(2\times H\times W)}$，分别对应BEV平面上的$x$轴和$y$轴速度分量$(v_x,v_y)$。与传统方法仅依赖位置差分不

同，显式的速度预测使得系统在遮挡发生时仍能保持对目标运动趋势的鲁棒估计。其中速度真值的获取并非直接来自单帧标注，而是通过多帧计算所得，具体为在目标对齐全局坐标系后，对位移差分计算并进行平滑处理得到。

重识特征分支：为了解决长时遮挡后的身份恢复及跨视角关联问题，该分支旨在提取具有判别力的外观特征。输出张量为$\hat{\boldsymbol{E}} \in \mathbb{R}^{(C_{id} \times H \times W)}$，式中$C_{id}$为特征嵌入维度。在推理阶段，提取热力图峰值位置对应的特征向量$\boldsymbol{e}_i$，并对其进行 L2 归一化，使其作为该目标的身份描述符。

2.4.2　损失函数

在训练阶段，采用多任务损失函数对上述分支进行联合优化，确保定位精度、运动估计与身份识别能力的同步提升。

（1）分类损失（定位），对于中心热力图，采用改进的焦点损失（Focal Loss）来平衡正负样本：$L_{xy} = \frac{-1}{N}\sum_{xy}\begin{cases}(1-\hat{Y}_{xy})^{\alpha}\log(\hat{Y}_{xy}) & Y_{xy}=1\\ (1-Y_{xy})^{\beta}(\hat{Y}_{xy})^{\alpha}\log(1-\hat{Y}_{xy}) & Y_{xy}\neq 1\end{cases}$　（7）

（2）回归损失（运动）对于速度估计分支，仅在目标中心处计算 L1 损失：

$$L_{vel} = 1/N\sum_{(k=1)}^{N}|\hat{\boldsymbol{V}}_k - \boldsymbol{V}_k^{gt}| \tag{8}$$

式中，$\hat{\boldsymbol{V}}_k$和$\hat{\boldsymbol{V}}_k^{gt}$分别为第$k$个目标的预测速度与真实速度。

（3）身份识别损失，将 ReID 训练视为一个分类任务。将提取的特征嵌入$\boldsymbol{e}_k$映射到类别空间，并使用 Softmax 交叉熵损失进行监督，以最大化类间差异并最小化类内方差：

$$L_{\mathrm{id}} = -\sum_{(k=1)}^{N}\log((e^{(\boldsymbol{w}_c^{T}\boldsymbol{e}_k)})/(\sum_{(j)}\mathrm{e}^{(\boldsymbol{w}_j^{T}\boldsymbol{e}_k)})) \tag{9}$$

式中c为目标k的真实身份标签ID，$\boldsymbol{w}$为分类器权重。为了进一步提升特征的鲁棒性，该部分三元组损失（Triplet Loss）进行辅助训练。

（4）总损失 最终的总损失函数定义为：$L_{total} = \lambda_{heat}L_{heat} + \lambda_{vel}L_{vel} + \lambda_{id}L_{id}$通过联合优化，网络能够在特征提取阶段自适应地平衡定位对齐、运动平滑与身份鉴别之间的关系。

2.5　基于卡尔曼滤波的在线关联

鉴于时空融合模块已提供了高精度的目标定位与运动特征，本文采用经典的卡尔曼滤波（Kalman，1960，Wojke 等，2017）进行高效的在线轨迹关联。

将目标状态建模为遵循线性匀速运动模型，其状态向量定义为$\boldsymbol{X} = [x, y, \dot{x}, \dot{y}, w, h, l]^T$，分别对应目标的中心坐标、速度分量、物理尺寸。在跟踪流程中，首先利用卡尔曼滤波器执行预测步骤，不同于传统 DeepSORT（Wojke 等，2017）仅依赖历史位置差分推算运动状态，本文直接利用前端分支回归的瞬时二维速度矢量引导状态转移过程，使得系统在目标快速运动或变速工况下的先验估计更贴近物理实际。

在关联阶段，系统采用多级匹配策略：首先执行关联匹配，优先对活跃度高的轨迹进行处理，通过融合空间位置约束与运动特征一致性构建代价矩阵，并在计算马氏距离时引入显式速度校验，利用速度方向的一致性排除 BEV 空间内位置相近但运动趋势相左的误匹配目标。随后，系统采用匈牙利算法求解全局最优匹配方案，并利用包含位置与速度的双重观测向量更新卡尔曼滤波器，使轨迹状态在检测波动时仍能快速收敛。最后，通过基于计数的生命周期管理机制，系统能自动初始化高置信度新目标并剔除长时间失配的冗余轨迹，从而在目标发生短时遮挡或检测波动时，仍能凭借显式速度引导的惯性预测维持鲁棒的长时跟踪。

3　实　验

3.1　数据集介绍

为了全面验证所提方法在不同环境下的感知与跟踪性能，本文选用了两个广泛使用的多视角多目标跟踪基准数据集：Wildtrack 和 MultiviewX。

Wildtrack 数据集：这是一个极具挑战性的真实场景数据集，由 7 台经过时间同步与参数标定的静态相机在拥挤的公共广场采集而成。其有效视场覆盖范围为 12m × 36m，存在严重的行人间相互遮挡以及背景干扰。为了适配鸟瞰图空间的特征投影，场景地面被离散化为 480 × 1440 的高分辨率网格，单个网格单元尺寸为 2.5cm × 2.5cm。该数据集记录了行人无预设的自然运动轨迹，平均每帧包含 20 名行人，每个空间位置平均被 3.74 台相机同时覆盖，对算法的抗遮挡能力提出了极高要求。原始视频分辨率为 1080 × 1920，采样帧率为 2 *FPS*，数据总时长约为 35 分钟。

MultiviewX 数据集：这是一个基于 Unity 引擎构建的合成数据集，旨在提供具有像素级精确真值的受控实验环境。该数据集由 6 台虚拟相机采集，其场景布局与相机配置高度模仿了 Wildtrack 的设计风格。场景覆盖范围为 16m × 25m，地面平面同样被量化为 2.5cm × 2.5cm 的精细网格（网格总数为 640 × 1000）。相比于 Wildtrack，MultiviewX 具有更高的人群密度，平均每帧包含 40 名行人，且每个空间位置的平均相机覆盖数高达 4.41 台，提供了更丰富的多视角几何冗余信息。该数据集的图像分辨率（1080 × 1920）、帧率（2 *FPS*）以及标准测试序列长度（400 帧）均与主流实验设置保持一致。

3.2 实验细节

在时序特征提取模块中，为了平衡计算效率与特征表达能力，我们配置了 8 头多头自注意力机制，并将 Dropout 比率设定为 0.1 以防止过拟合。此外，为了强化模型对局部运动模式的捕捉能力，引入了窗口大小为 8 的窗口化注意力机制，以实现局部时空邻域内的特征交互。数据预处理与增强为了提升模型的泛化性能并抑制训练过拟合，本文采用了一套混合数据增强策略（Shorten and Khoshgoftaar，2019），包括随机水平翻转、随机缩放裁剪（Zhong 等，2020）以及颜色抖动。输入图像的分辨率统一调整为 1080 × 1920，在推理阶段，检测置信度的过滤阈值设定为 0.5。优化策略与训练设置模型训练采用 Adam 优化器（Kingma，2014）动量参数设置为 $\beta_1 = 0.9, \beta_2 = 0.99$。初始学习率设为 1×10^{-3}，并采用 OneCycle 学习率策略 动态调整。模型共训练 50 个 Epoch，Batch Size 设置为 16。在损失函数配置中，分类、回归、重识别的权重为：10、10、1。其中，三元组损失的边界值（Margin）α 设定为 0.2，权重系数设为 0.5。针对高分辨率输入带来的显存资源约束，本文实施了步长为 8 的梯度累积策略，并配合阈值为 0.5 的梯度裁剪技术，在有效降低显存占用的同时保证了训练梯度的稳定性。实验环境算法基于 PyTorch 深度学习框架构建，并利用 PyTorch Lightning 库进行标准化的训练流程管理。所有实验均在单块 NVIDIA RTX 4090 GPU 上部署并完成。

3.3 评估指标

为了全面、客观地衡量跟踪系统的综合性能，本文遵循标准的 CLEAR-MOT 评测协议，选取了六个关键指标进行量化分为了全面衡量跟踪系统的综合性能，本文采用标准的 CLEAR-MOT 评测协议。其中，重点关注以下两个核心指标：多目标跟踪准确度（*MOTA*）：作为衡量系统整体性能的首要指标，它综合计算了误检（*FP*）、漏检（*FN*）及身份切换（*IDs*）带来的误差，直观反映了算法在目标检测与轨迹维持上的综合能力。计算公式如式 10：

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDs_t)}{\sum_t GT_t} \tag{10}$$

式中 GT_t 为第 t 帧的真值目标数。该指标直观反映了算法在目标检测与轨迹维持上的综合能力。

IDF1 得分（*IDF1*）：该指标侧重于评估跟踪器在长时段内维持身份一致性的能力。

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{11}$$

相比于 *MOTA*，*IDF1* 对频繁的身份跳变更为敏感，能更准确地反映模型在严重遮挡场景下的抗轨迹断裂性能，是验证本文时空融合模块有效性的关键依据。

此外，本文辅以以下指标进行综合考量：身份切换次数（*IDs*）用于统计轨迹中断后的错误关联次数，直接反映特征的鲁棒性；多目标跟踪精度（*MOTP*）用于衡量定位的几何偏差；主要跟踪（*MT*）与主要丢失（*ML*）则分别表征了被成功跟踪时长超过 80% 和不足 20% 的轨迹比例，反映算法对目标轨迹完整周期的跟踪水平。

3.4 与其他方法的对比

为了全面验证本文提出的方法的有效性，将本方法与当前主流的多相机多目标跟踪算法进行了定量对比。对比方法涵盖了经典的晚期融合方法（如 KSP-DO（Chavdarova and Baqu，2018））、基于图优化的方法（如 GLMB（Ong 等，2020））以及最新的基于早期融合与 Transformer 的 SOTA 方法（如 ReST（Cheng 等，2023），EarlyBird（Teepe 等，2024））。

如表 1 所示，本文方法在 Wildtrack 数据集上表现出显著的性能优势。在核心指标 *MOTA* 上，本文方法达到了 89.70%，优于现有的最佳方法 EarlyBird 约 1.78%。更为显著的优势体现在反映轨迹稳定性的 *IDF1* 和 *IDs* 指标上。与 ReST 和 EarlyBird 相比，*IDF1* 分别提升了 5.53% 和 2.04%，这表明本文模型在处理真实场景下的拥挤与遮挡问题时，能够维持更长时间的身份一致性。

表1　Wildtrack与MultiviewX数据集上的多目标跟踪性能对比

Table 1　Performance comparison of multi-object tracking on Wildtrack and MultiviewX datasets

数据集	方法	来源	*MOTA* ↑	*MOTP* ↑	*IDF1* ↑	*MT* ↑	*IDs* ↓
Wildtrack	KSP-DO(Chavdarova and Baqu, 2018)	CVPR2018	69.6	61.5	73.2	28.7	–
	KSP-DO-ptrack (Chavdarova and Baqu, 2018)	CVPR2018	72.2	60.3	78.4	42.1	–
	GLMB-YOLOv3(Ong 等, 2020)	TPAMI2022	69.7	73.2	74.3	79.5	–
	GLMB-DO (Ong 等, 2020)	TPAMI2022	70.1	63.1	72.5	93.6	–
	DMCT (You and Jiang, 2020)	Arxiv2020	72.8	79.1	77.8	61.0	–
	DMCT-Stack(You and Jiang, 2020)	Arxiv2020	74.6	78.9	81.9	65.9	–
	ReST(Cheng 等, 2023)	ICCV2023	84.9	84.1	86.7	87.8	–
	EarlyBird(Teepe 等, 2024)	WACV2024	87.92	**87.81**	90.19	78.0	**10**
	WST-track	–	**89.70**	77.74	**92.23**	**87.80**	**7**
MultiviewX	Earlybird(Teepe 等, 2024)	WACV2024	88.42	85.77	81.52	84.2	45
	MVTr(Yang 等, 2024)	CVIU2024	**91.4**	**95.0**	82.9	**96.1**	–
	REMP(Dakic 等, 2025)	PerCom2025	81.0	85.8	–	–	–
	WST-track	–	89.42	86.7	**84.42**	88.13	**8**

注:加粗字体为每行最优值。

在MultiviewX数据集上,本文方法同样保持了极具竞争力的水平。*MOTA*与*MT*均优于基准方法EarlyBird,说明模型对目标的召回与覆盖能力更强。尽管*IDF1*指标略有波动,但*IDs*从EarlyBird的45次大幅下降至8次,进一步印证了本文方法在跨视角关联上的极高稳定性。极低的身份切换次数是本文实验结果的最大亮点。得益于窗口化时空融合模块对历史特征的有效聚合,模型能够利用时序上下文信息在BEV空间中补全被遮挡目标的特征表示。这种特征层面的连续性使得绝大多数目标在整个生命周期内未发生身份跳变,有效解决了传统方法在遮挡区域易丢失*ID*的痛点。值得注意的是,与对比方法普遍采用复杂的图优化或Transformer关联不同,本文仅采用了基于线性运动模型的在线跟踪进行后端关联。*IDF1*和*IDs*数据有力地证明了当前端网络能够构建出强鲁棒性的时空表征时,复杂的后端优化并非多目标跟踪的必要条件。 高质量的BEV特征显著降低了数据关联的难度,使得简单的线性预测即可实现良好的跟踪效果。

为了验证本文提出的方法的有效性,并探究优异跟踪性能的来源,评估了模型在纯检测任务上的表现。表2展示了本文方法与Deep-Occ(Baqu, 2017),MVDet(Hou 等, 2020),MVTT(Lee 等, 2023)及EarlyBird(Teepe 等, 2024)等主流方法在Wildtrack和MultiviewX数据集上的检测指标对比。

表3数据显示,本文方法在两个数据集上均展现出优异的检测召回率。特别是在MultiviewX数据集上,Recall高达99.10%,MODA达95.70%,均刷新了该数据集的SOTA记录;Wildtrack数据集的Recall亦保持在95.79%的高位。极高的召回率显著降低了漏检率,有力验证了强前端策略的有效性:即通过提供连续、稳定的观测输入,确保后端卡尔曼滤波不易发生轨迹中断,从而保障了跟踪阶段优异的*IDF1*表现。

为了进一步佐证本方法的实用高效性,表2展示了本模型与主流多视角跟踪算法在推理速度(Frames-Per-Second,*FPS*)上的定量对比。本方法在单张NVIDIA RTX 4090 GPU上实现了18.5 *FPS*的处理速率。尽管在绝对数值上略低于纯空间投影方案MVDet(Hou 等, 2020)和EarlyBird(Teepe 等, 2024),但本方法在感知精度与计算开销之间取得了更优的平衡。这种性能差异源于本框架引入了显式的速度引导窗口化时空融合模块,虽然该设计增加了时序对齐延迟,但其有效解决了多视角下的长时

表2　Wildtrack 与 MultiviewX 数据集上检测性能对比

Table 2　Performance comparison of object detect on Wildtrack and MultiviewX datasets

方法	Wildtrack				MultiviewX				
	MODA	*MODP*	*Precision*	*Recall*	*MODA*	*MODP*	*Precision*	*Recall*	*FPS*
Deep-Occlusion(Baqu, 2017)	74.1	53.8	95.0	80.0	75.2	54.7	97.8	80.2	~1.5
MVDet (Hou 等, 2020)	88.2	75.7	94.7	93.6	83.9	79.6	96.8	86.7	**35.2**
3DROM†(Qiu 等, 2022)	91.2	76.9	95.9	95.3	90.0	83.7	97.5	92.4	–
MVDeTr (Hou and Zheng, 2021)	91.5	**82.1**	97.4	94.0	93.7	91.3	**99.5**	94.2	12.5
MVTT(Lee 等, 2023)	**94.1**	81.3	**97.6**	96.5	95.0	**92.8**	99.4	95.6	15.6
EarlyBird(Teepe 等, 2024)	90.4	81.2	93.7	**96.6**	94.4	90.5	98.7	95.6	22.8
WST-track	91.28	70.14	95.49	95.79	95.7	87.1	96.5	**99.1**	18.5

注:加粗字体为每行最优值,~所表数据为根据论文内容的预估值。

遮挡问题,使得 MODA 指标提升至 91.28%。相比于同样具备注意力机制的 MVDeTr (12.5 *FPS*) 和 MVTT (15.6 *FPS*),本方法利用局部窗口限制计算范围,规避了全局注意力带来的二次方级计算负担。综上所述,本方法实现了感知精度与运行速度的高效统一,一定程度上的部署需求。

3.5　消融实验

为了深入探究本文所提框架中各核心模块的贡献及其对整体性能的影响,在 Wildtrack 数据集上开展了系统的消融实验。实验主要围绕核心组件的有效性以及不同时空融合机制的差异性展开定量分析。

表3　核心组件消融实验

Table 3　Ablation of core components

模型	集合投影	时空融合	速度引导	*MOTA*	*IDF1*	*IDs*
Baseline	√	–	–	86.50	82.15	45
单融合	√	√	–	**90.12**	90.40	18
本方法	√	√	√	89.70	**92.23**	**7**

注:加粗字体为每行最优值。

表 3 详细展示了逐步引入窗口化时空融合 与速度引导分支后,模型跟踪性能的演变过程。首先,在仅使用单帧几何投影的基准模型中,系统缺乏对时序上下文的建模能力。尽管 *MOTA* 达到了 86.50%,但由于无法有效应对遮挡导致的特征丢失,反映轨迹连续性的 *IDF1* 指标仅为 82.15%,且身份切换次数高达 45 次。这表明单纯的空间感知在动态复杂场景下难以维持长时稳定的特征表征。

随后,在引入窗口化时空融合模块后,模型性能取得了显著突破。*IDF1* 大幅跃升至 90.40%(提升 8.25%),*IDs* 锐减至 18 次。这一结果有力地证明了利用历史帧信息对当前帧进行特征补全的必要性,时空融合机制成功修复了因遮挡造成的检测盲区,显著增强了 BEV 表征的鲁棒性。

最后,通过进一步加载速度引导分支,完整模型的 *IDF1* 最终达到92.23%,*IDs* 进一步降至最低的7次。尽管 *MOTA* 在此阶段略有波动(89.70%),但身份一致性的显著提升表明,显式的速度回归为后端的卡尔曼滤波提供了更为精准的运动先验,有效平滑了轨迹预测,从而实现了跟踪稳定性的最大化。

为了验证本文所选融合策略的优越性,表 4 进一步对比了不同时序融合机制的性能差异。

表4　融合机制对比

Table 4　Fusion Mechanisms

融合机制	计算复杂度	*MOTA* ↑	*IDF1* ↑	*IDs* ↓
直接拼接	低	88.45	89.10	24
全局注意力	高	89.50	91.80	12
窗口化注意力	**中**	**89.70**	**92.23**	**7**

注:加粗字体为每行最优值。

实验结果显示,朴素的直接拼接策略虽然计算成本最低,但由于缺乏特征对齐与交互,*IDF1* 仅为 89.10%,*IDs* 较多(24 次)。相比之下,全局注意力机制虽然提升了感受野,将 *IDF1* 提升至 91.80%,

但其全图计算引入了巨大的计算开销，且易受远距离背景噪声的干扰。

与之相比，本文提出的窗口化注意力机制在计算复杂度与性能之间取得了最佳平衡。其 *IDF1* 高达92.23%，且 *IDs* 最少。这表明，将注意力限制在局部时空邻域内，不仅足以捕捉行人运动的时空关联性，还能通过引入局部归纳偏置有效滤除无关的背景噪声，从而实现了最高效的特征增强。

表5　参数敏感性分析

Table 5　Parameter Sensitivity Analysis

时序窗口大小（T）	对应时长	*MOTA* ↑	*IDF1* ↑
1（无时序）	–	86.50	82.15
3	1.5s	89.50	91.80
5	2.5s	**89.70**	**92.23**
7	3.5s	89.65	89.65
9	4.5s	89.20	89.20

注：加粗字体为每行最优值。

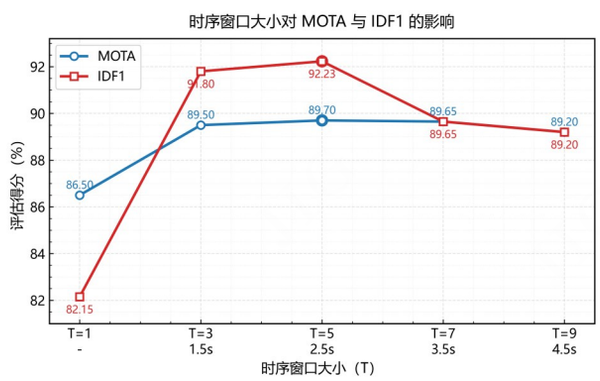

图3　时序窗口大小对 *MOTA* 与 *IDF1* 的影响

Fig. 3　The influence of the temporal window size on *MOTA* and *IDF1*

为了探究时序记忆长度对特征融合质量的影响，我们对时序窗口大小 *T* 进行了敏感性测试。实验结果如表5和图3所示。当 *T* 从1增加到5时，*MOTA* 和 *IDF1* 指标呈现显著上升趋势。这表明，适当扩充时序感受野使得模型能够有效利用历史观测来填补当前的视觉盲区，显著增强了抗遮挡能力。在 *T*=5时观察到了性能峰值。随着 *T* 进一步增加至7或9，模型性能并未持续提升，反而出现轻微下降。这主要是因为：在低帧率场景下，过长的时序跨度会导致行人空间位移过大，即使引入速度引导，特征对齐的累积误差仍会引入背景噪声。综合考量跟踪精度与系统实时性，本文最终选取 *T* 为5作为默认的时序窗口设置，此时系统在保持 推理速度的同时，实现了最优的身份一致性。

表6　损失函数分析

Table 6　Loss Function Analysis

重识别损失	*MOTA* ↑	*IDF1* ↑
交叉熵损失	88.50	92.65
交叉熵损失+三元组损失	89.70	93.23

为了探究为了验证三元组损失的有效性，我们设计了如表6所示消融实验：实验结果显示，引入三元组损失后，多目标跟踪指标 *MOTA* 提升了1.2%，*IDF1* 提升了1.5%。这证明三元组损失通过强化难样本挖掘，有效改善了复杂场景下的关联性能。

3.6　可视化实验

为了直观验证本文所提方法在严重遮挡场景下的鲁棒性，图3展示了在 Wildtrack 数据集典型拥挤帧序列 *t1* 到 *t4* 上的定性跟踪结果对比。如图4所示，被跟踪目标（红框标注）在 *t2* 至 *t3* 时刻遭遇了前方行人的严重动态遮挡。图4底部一行的多视角快照进一步证实，该目标在关键时刻 t3 的其余辅助视角中同样处于部分或完全遮挡状态，视觉信息极度匮乏，这为跨时空关联带来了极大挑战。如第一行所示，EarlyBird 在 *t1* 时刻能够正常跟踪目标。然而，当 *t2* 时刻发生严重遮挡时，由于其特征提取主要依赖单帧空间投影，缺乏对长时序上下文的有效利用，导致目标特征在遮挡期间退化或丢失。因此，当目标在 *t4* 时刻重新完全出现时，系统未能将其与历史轨迹关联，错误地分配了新的身份 *ID*，发生了身份切换。相比之下，本文方法得益于窗口化时空注意力模块对历史记忆的有效聚合，能够 *t2* 时刻的“盲区”期间，自适应地检索并利用 *t1* 时刻的高质量历史特征来补全当前表征。因此，即便经历了严重的视觉遮挡，模型依然成功维持了特征的连续性，在 *t4* 时刻准确保持了目标的原始身份 *ID*。这一可视化结果直观地解释了定量实验中本文方法为何能取得低 *IDs* 与高 *IDF1* 的原因，充分验证了时空联合建模策略在解决复杂遮挡与轨迹断裂问题上的卓越性能。

图4　Wildtrack 数据集严重遮挡场景下的跟踪效果对比

Fig. 4　Qualitative comparison of tracking results under severe occlusion on Wildtrack dataset

如表的定量验证所示，本文模型所提供的改进的连续性直接促进了高阶跟踪指标的提升，例如更高的主要跟踪轨迹比率和更少的身份切换。这一定性分析证实，所提出的架构组件有效地增强了多目标跟踪的时序一致性和可靠性。

图5进一步展示了本文方法在四个具有显著视差的相机视角下的多目标轨迹生成效果。可以看出，尽管各视角的视场范围（Field of View，FOV）、拍摄角度及局部遮挡情况存在显著差异，本文方法生成的轨迹（图中彩色线条）在所有视图中均表现出高度的空间一致性与几何对应关系。特别是在人群密集的交互区域，投影回各个图像平面的二维轨迹依然能够保持精准的同步，未出现因视角变换导致的轨迹漂移或错位。此外，轨迹线条整体呈现出自然的平滑度，未见明显的锯齿或抖动，这得益于前端网络对目标运动状态的稳定预测。上述结果表明，通过将多视角特征统一映射至3D BEV空间并进行时空联合建模，即使在部分视角受限的情况下，也能生成鲁棒且全局一致的跟踪轨迹。

图5　不同视角下的多目标跟踪轨迹可视化

Fig. 5　Visualization of multi-object trajectories across different camera views

4　结　论

本文提出一种新颖的面向多相机多目标跟踪的窗口化时空注意力驱动框架，使用速度引导的窗口化时空融合模块来集中捕获局部时空关联，并在特征提取阶段采取显式的运动补偿策略来增强表征，有效提升了跟踪性能。通过窗口化注意力机制自适应聚合历史特征，为当前视域补充被遮挡的目标信

息，同时引入轻量级的卡尔曼滤波后端，进一步促进了感知与关联的联合优化。在 Wildtrack 和 MultiviewX 上进行的实验结果表明本文方法已取得较好的跟踪性能，特别是在身份一致性指标上表现优异。消融实验表明该框架能够有效利用历史记忆修复特征缺失，并证明了在高质量时空表征下，简单的线性运动模型足以实现高精度的跨帧关联。

本研究尚存在一些局限性，未来将探索新的模型轻量化与剪枝技术，增强边缘设备的部署能力，提高推理速度。本研究尚存在一些局限性。首先，本文实验基于 Wildtrack 和 MultiviewX 标准数据集开展，这些数据集的视频帧率均为 2 *FPS*，因此本文在消融实验中确定的最佳时序窗口长度 T=5 对应时间跨度为 2.5 秒。在实际 30 *FPS* 监控视频场景中，若不进行抽帧直接应用，单位时间内的帧数会显著增多，可能带来一定的计算量上升；而在高密度帧序列下，速度引导的特征对齐策略仍能保持稳定的对齐效果。实际部署时可通过自适应帧采样等方式，使输入帧率与模型训练时序分布保持一致，从而在保证精度的同时维持合理计算开销。

其次，本研究使用网格化的 BEV 特征作为空间信息的来源，尽管并未影响目标检测的召回率，但离散化的网格导致定位精度存在一定的量化误差，未来将尝试引入坐标细化子网络或连续值回归的方法以保留更多几何细节信息。

参考文献(References)

Baqué P, Fleuret F and Fua P. 2017. Deep occlusion reasoning for multi-camera multi-target detection//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 271-279 [DOI: 10.1109/ICCV.2017.38]

Berclaz J, Fleuret F, Turetken E and Fua P. 2011. Multiple object tracking using k-shortest paths optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(9): 1806-1819

Chavdarova T, Baqué P, Bouquet A, et al. 2018. Wildtrack: A multi-camera HD dataset for dense unscripted pedestrian detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5030-5039

Chen F, Zhu G, Lv Z, et al. 2025. An enhanced multiple object tracking method with improved feature fusion and cross-camera trajectory matching for closed scenes. IEEE Access, 13: 1-1

Cheng C C, Qiu M X, Chiang C K, et al. 2023. Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 10051-10060

Cherdchusakulchai R, Phimsiri S, Trairattanapa V, et al. 2024. Online multi-camera people tracking with spatial-temporal mechanism and anchor-feature hierarchical clustering//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 7198-7207

Dakic K, Thilakarathna K, Calheiros R N, et al. 2025. Resource-efficient multiview perception: Integrating semantic masking with masked autoencoders//Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom). Washington D. C., USA: IEEE: 145-151 [DOI: 10.1109/PerCom61555.2025.00032]

Dendorfer P, Rezatofighi H, Milan A, et al. 2020. Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003

Fei L and Han B. 2023. Multi-object multi-camera tracking based on deep learning for intelligent transportation: A review. Sensors, 23(8): 3852

Fleuret F, Berclaz J, Lengagne R, et al. 2008. Multicamera people tracking with a probabilistic occupancy map. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2): 267-282

Hou Y and Zheng L. 2021. Multiview detection with shadow transformer (and view-coherent data augmentation)//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1673-1682

Hou Y, Zheng L and Gould S. 2020. Multiview detection with feature perspective transformation//Proceedings of the European Conference on Computer Vision. Glasgow, UK: Springer: 1-18

Huang J and Huang G. 2022. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054

Kalman R E. 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82(1): 35-45

Kang B, Chen X, Zhao J and Wang D. 2025. TMamba: a visual state-space model for efficient object tracking. Journal of Image and Graphics, 30(10): 3199-3214 (康奔, 陈鑫, 赵洁, 王栋. 2025. TMamba: 面向高效目标跟踪的视觉状态空间模型. 中国图象图形学报, 30(10): 3199-3214) [DOI: 10.11834/jig.240587]

Kim A and Bras P. 2022. PolarMOT: How far can geometric relations take us in 3D multi-object tracking? //Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel: Springer: 41-58

Kingma D P and Ba J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Lee W Y, Jovanov L and Philips W. 2023. Multi-view target transformation for pedestrian detection//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 90-99

Li Z, Wang W, Li H, et al. 2022. Bevformer: Learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers//

Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel: Springer: 1-18

Lin X, Lin T, Pei Z, et al. 2022. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. arXiv preprint arXiv:2211.10581

Liu Z, Lin Y, Cao Y, et al. 2021. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10012-10022

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

Ortiz M M, Iriarte F J, Rodrguez H D, et al. 2024. Context-aware model training for attention-based multicamera multiobject tracking//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 189-197

Pang Z, Li J, Tokmakov P, et al. 2023. Standing between past and future: Spatio-temporal modeling for multi-camera 3d multi-object tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 17928-17938

Philion J and Fidler S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d//Proceedings of the European Conference on Computer Vision. Glasgow, UK: Springer: 194-210

Qiu R, Xu M, Yan Y, et al. 2022. 3d random occlusion and multi-layer projection for deep multi-camera pedestrian localization//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel: Springer: 695-710

Shorten C and Khoshgoftaar T M. 2019. A survey on image data augmentation for deep learning. Journal of Big Data, 6(1): 1-48

Song L, Wu J, Yang M, et al. 2021. Stacked homography transformations for multi-view pedestrian detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6049-6057

Szűcs G, Borsodi R and Papp D. 2024. Multi-camera trajectory matching based on hierarchical clustering and constraints. Multimedia Tools and Applications, 83: 44879-44902

Teepe T, Wolters P, Gilg J, et al. 2023. Lifting multi-view detection and tracking to the bird's eye view//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 667-676

Teepe T, Wolters P, Gilg J, et al. 2024. EarlyBird: early-fusion for multi-view tracking in the bird's eye view//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 102-111

Tran L Q and Vi H D. 2024. Efficient online multi-camera tracking with memory-efficient accumulated appearance features and trajectory validation//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 7217-7226

Vaswani A, Shazeer N, Parmar N, et al. 2017. Attention is all you need//Advances in Neural Information Processing Systems. Long Beach, USA: Curran Associates, Inc.: 5998-6008

Wang Y, Meinhardt T, Cetintas O, et al. 2024a. MCBLT: Multi-Camera Multi-Object 3D Tracking in Long Videos. arXiv preprint arXiv:2412.00692

Wang Y, Meinhardt T, Cetintas O, et al. 2024b. BEV-SUSHI: Multi-Target Multi-Camera 3D Detection and Tracking in Bird's-Eye View. arXiv preprint arXiv:2412.00000

Wojke N, Bewley A and Paulus D. 2017. Simple online and realtime tracking with a deep association metric//Proceedings of the IEEE International Conference on Image Processing. Beijing, China: IEEE: 3645-3649

Yamane T, Masumura R, Suzuki S, et al. 2025. MVTrajecter: Multi-View Pedestrian Tracking with Trajectory Motion Cost and Trajectory Appearance Cost//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE: 13270-13280

Yang Y, Xu M, Ralph J F, et al. 2024. An end-to-end tracking framework via multi-view and temporal feature aggregation. Computer Vision and Image Understanding, 249: 104203 [DOI: 10.1016/j.cviu.2024.104203]

You Q and Jiang H. 2020. Real-time 3d deep multi-camera tracking. arXiv preprint arXiv:2003.11753

Zhang L, Li Y and Nevatia R. 2008. Global data association for multi-object tracking using network flows//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA: IEEE: 1-8

Zhong L, Ou Q F, Shen M J and Xiong B S. 2025. STNet: a small moving target localization and tracking network for badminton. Journal of Image and Graphics, 30(1): 148-160 (钟亮, 欧巧凤, 沈敏杰, 熊邦书. 2025. STNet: 羽毛球运动小目标定位跟踪网络. 中国图象图形学报, 30(1): 148-160) [DOI: 10.11834/jig.230800]

Zhong Z, Zheng L, Kang G, et al. 2020. Random erasing data augmentation//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press: 13001-13008

作者简介

王佳伟,男,硕士研究生,主研究方向为多视角多目标跟踪和计算机视觉。 E-mail: wangjiawei235@mails.ucas.ac.cn

田璟,通信作者,女,副研究员,主要研究方向计算机视觉与多源数据智能融合处理。 E-mail:tianjing@aircas.ac.cn

苑子杨,男,助理研究员,主研究方向为遥感图像的智能解译与理解。 E-mail: yzy30997@gmail.com

张晓典,男,助理研究员,主研究方向为为光学成像、计算机视觉、多模态智能模型。 E-mail:zhangxd@aircas.ac.cn

赵良瑾，男，助理研究员，主要研究方向为自主无人机的遥感图像理解和定位。 E-mail：zhaolj0489@aircas.ac.cn