CIMOT3D:基于中文引导的单目视角下三维多目标跟踪研究

王 荣[1],胡海祥[1],魏弘凯[1],梁浩翔[2*],钱晓伟[1],李凯飞[1],郭柯宇[1],宋翔宇[3],孙士杰[3]
(1. 长安大学信息工程学院,陕西西安 710064;2. 长安大学电子与控制工程学院,陕西西安 710064;
3. 长安大学数据科学与人工智能研究院,陕西西安 710064)

摘 要: 自然语言描述驱动的目标跟踪通过解析符合人类表达习惯的语言描述,并将其与视觉信息融合,从而实现复杂环境中特定目标的精准识别与持续跟踪. 然而,现有方法主要集中于二维场景或三维单目标跟踪,尚未扩展至三维多目标跟踪,缺乏将文本与三维视觉空间中多个候选目标进行特征对齐与关联建立的能力;此外,现有自然语言描述驱动三维目标跟踪任务在语言层面存在冗余问题,难以模拟人类基于灵活简练的指令对多个特定目标进行跟踪的能力. 针对这些挑战,本文提出基于中文引导的单目视角下三维多目标跟踪新任务(Chinese-Instruction-based monocular 3D Multi-Object Tracking,CIMOT3D),并构建了含有 5 562 个视频序列的数据集 CIMOT3D-5k,且所有序列均标注有符合人类表达习惯的中文描述. 同时,本文设计了一种专用于该任务的神经网络模型 CIMOT3D-SyncTracker(Chinese-Instruction-based monocular 3D Multi-Object tracking Synchronization Tracker),其框架由多模态特征提取器、视觉语言编解码器与检测跟踪模块三部分组成. 相比于基线方法,本文方法在 CIMOT3D-5k 数据集上的跟踪准确率和身份一致性指标上分别提高了 4.1 和 5.0 个百分点,验证了其性能优势. 本文拓展了视觉语言融合在三维多目标跟踪方向的研究深度,并为相关领域的后续探索提供了新的思路.

关键词: 场景理解;三维目标跟踪;多目标跟踪;视觉语言模型;多模态学习;机器视觉

基金项目: 国家重点研发计划(No.2023YFB4301800);国家自然科学基金(No.62576050);国家资助博士后研究人员计划(No.GZC20241447);长安大学中央高校基本科研业务费专项资金(No.300102325101)

中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0102-13

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250826

CIMOT3D: Chinese-Instruction-Based Monocular 3D Multi-Object Tracking

WANG Rong[1], HU Haixiang[1], WEI Hongkai[1], LIANG Haoxiang[2*], QIAN Xiaowei[1], LI Kaifei[1], GUO Keyu[1], SONG Xiangyu[3], SUN Shijie[3]
(1. *School of Information Engineering*, *Chang'an University*, *Xi'an*, *Shaanxi* 710064, *China*;
2. *School of Electronic and Control Engineering*, *Chang'an University*, *Xi'an*, *Shaanxi* 710064, *China*;
3. *School of Data Science and Artificial Intelligence*, *Chang'an University*, *Xi'an*, *Shaanxi* 710064, *China*)

Abstract: Natural language-driven object tracking parses human-like language descriptions and fuses them with visual information to achieve accurate recognition and continuous tracking of specific targets in complex environments. However, existing methods focus on 2D tracking or 3D single-target tracking, and they have not been effectively extended to 3D multi-target tracking. They lack the capability to align text with multiple candidate targets in 3D visual space and to establish associations. In addition, existing natural language-driven 3D object tracking tasks suffer from redundancy in language descriptions, which makes it hard to track multiple specific targets using flexible and concise instructions as humans do. To address these challenges, this paper introduces a new task, chinese-instruction-based monocular 3D multi-object tracking (CIMOT3D). The paper also constructs a new dataset, CIMOT3D-5k, which contains 5 562 video sequences with human-like Chinese descriptions. Furthermore, this paper designs a neural network model chinese-instruction-based monocular 3D multi-object tracking synchronization tracker (CIMOT3D-SyncTracker) for this task, which consists of a multimodal feature extractor, a vision-language encoder-decoder, and a detection-tracking module. Compared with baseline methods, the proposed approach achieves an improvement of 4.1% in tracking accuracy and 5.0% in identity consistency metric on the CIMOT3D-5k dataset, verifying its performance advantage. This paper advances research on vision-language fusion in 3D

收稿日期:2025-09-23;录用日期:2026-01-06;责任编辑:覃怀银
*通信作者:梁浩翔

multi-object tracking and offers new ideas for further exploration in related fields.

Keywords: scene understanding; 3D object tracking; multi-object tracking; vision-language model; multimodal learning; machine vision

Foundation Item(s): National Key Research and Development Program of China (No.2023YFB4301800); National Natural Science Foundation of China (No.62576050); National Postdoctoral Researcher Program (No.GZC20241447); Fundamental Research Funds for the Central Universities, CHD (No.300102325101)

0　引言

目标跟踪[1]旨在在连续的视频序列或传感器数据中,对一个或多个目标(如交通场景中的车辆)进行定位、识别与轨迹估计。然而,传统目标跟踪在复杂空间场景中实现高精度跟踪时,往往依赖昂贵的传感器来获取深度信息[2],这一局限制约了其实际应用与推广。为缓解这一局限对多目标跟踪任务的制约,研究者逐渐将自然语言引入到目标跟踪中,形成了自然语言描述驱动的目标跟踪(Natural Language-driven Object Tracking,NLOT)任务。此任务通过将自然语言描述与视觉特征跨模态互补与融合,实现复杂开放场景中的跟踪任务,同时能够依据自然语言描述跟踪特定目标。

现阶段的NLOT研究主要聚焦于二维或三维单目标场景,对三维多目标场景的研究相对缺乏。其中,自然语言驱动的二维目标跟踪(Natural Language-driven Object Tracking in 2D,NLOT2D)相关方法虽能够实现简单的多目标跟踪和基本的语义理解,但所用文本多为简单提示词形式,缺乏对目标的细粒度刻画与复杂语义建模能力。现有自然语言驱动的三维目标跟踪(Natural Language-driven Object Tracking in 3D,NLOT3D)任务虽然取得了一定进展,但仍存在两方面的局限:一是尚未建立文本与多候选目标之间有效对齐和关联的机制;二是自然语言描述存在冗余问题,不具备人类根据需求使用灵活简练的描述来理解开放世界的能力。

针对上述问题,本文将NLOT3D扩展到一个全新的任务:基于中文引导的单目视角下三维多目标跟踪(Chinese-Instruction-based monocular 3D Multi-Object Tracking,CIMOT3D),见图1。此任务的核心目标为在单目视频序列中,根据简练的中文描述对多个特定目标进行定位、识别与持续跟踪。

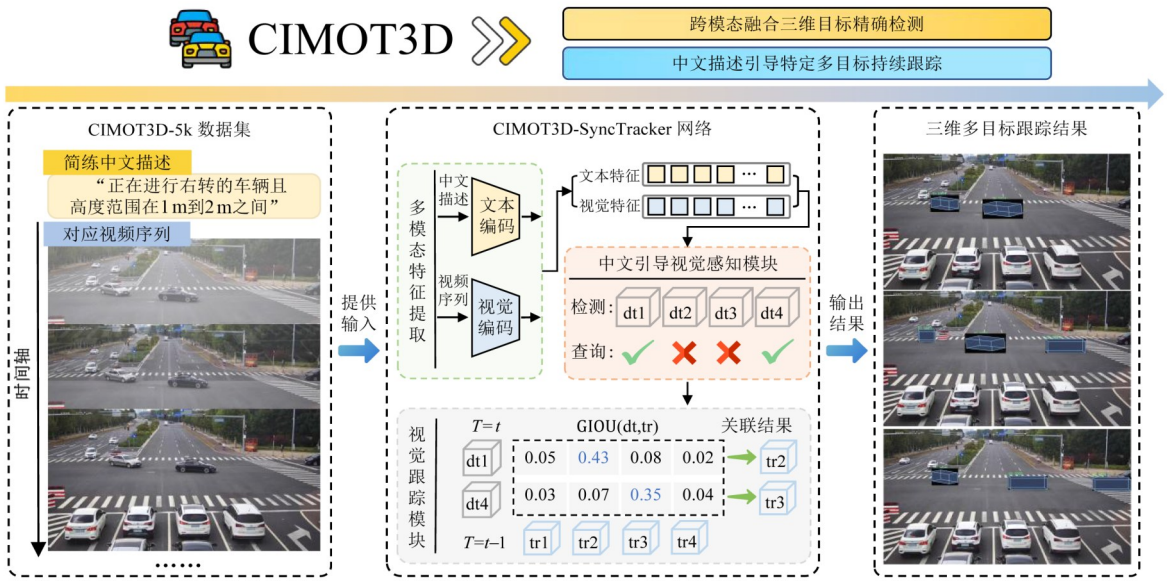


图1　基于中文引导的单目视角下三维多目标跟踪任务总图

Figure 1　Overall framework of the Chinese-instructed monocular 3D multi-object tracking task

为推动该任务的研究,本文在DAIR-V2X-Seq数据集[3]基础上,构建了CIMOT3D-5k(Chinese-Instructed monocular 3D Multi-Object tracking dataset-5k)数据集,该数据集包含5 562个视频序列,每个视频序列均配有3条符合人类表达习惯的简练中文描述,为单目视频中特定多目标的三维跟踪提供了丰富的语义信息支撑。这些中文描述由千问2.5大语言模型[4]生成,后续经过人工校对与调整,确保描述的精准性与可靠性。

相比传统的三维多目标跟踪(3D Multi-Object

Tracking，MOT3D），CIMOT3D 无需依赖额外的深度信息，而是引入语言模态作为视觉模态的补充。相较于 NLOT2D 和 NLOT3D 任务，CIMOT3D 在既有框架的基础上首次将 NLOT 任务扩展到单目视角下的三维多目标跟踪，且在语言描述层面具备更强的自适应性，进而具备类人的语义理解能力。CIMOT3D 与上述三类任务的具体差异如表 1 所示。

表 1 不同跟踪任务差异对比

Table 1 Comparison of different tracking tasks

任务	输入	任务场景	语言指导
MOT3D	深度信息+视频序列	三维多目标	不支持
NLOT2D	文本描述+视频序列	二维多目标	支持,语言简洁
NLOT3D	文本描述+视频序列	三维单目标	支持,语言冗余
CIMOT3D	文本描述+视频序列	三维多目标	支持,语言简练

针对该任务，本文提出了专用网络模型 CIMOT3D-SyncTracker（Chinese-Instruction-based Monocular 3D Multi-Object Tracking Synchronization Tracker），用于在中文引导下的单目视角下三维多目标检测和跟踪网络。该网络通过多模态特征提取器获取视觉空间特征与中文描述语义特征；随后通过中文描述驱动的视觉编解码器深度挖掘视觉特征，同时构建文本与候选目标特征的相似性度量并进行模态对齐；最后检测跟踪器基于融合特征完成中文引导下的多目标跟踪任务。CIMOT3D-SyncTracker 模型在本文设计的一系列基准实验中性能突出，相较于现有基线方法，其在 CIMOT3D-5k 数据集上的多目标跟踪准确率（Multiple Object TrAcking，MOTA）和身份一致性指标（IDentity F1 score，IDF1）上分别提高了 4.1% 和 5.0%。

本文的主要贡献如下：

（1）提出新任务 CIMOT3D，旨在单目视角下根据中文描述的引导实现特定多目标的三维视觉跟踪。

（2）在 DAIR-V2X-Seq 数据集的基础上构建了 CIMOT3D-5k 数据集，包含了 5 562 个视频序列以及相应的中文描述，用于 CIMOT3D 任务的评估。

（3）设计并实现了专用于 CIMOT3D 任务的网络模型 CIMOT3D-SyncTracker，该模型集成了多模态特征提取、中文与视觉匹配以及三维检测跟踪技术。

（4）构建了 CIMOT3D 任务的全面基准测试体系，通过系统性实验验证了 CIMOT3D-SyncTracker 的性能优势，在 MOTA 和 IDF1 指标上分别提高 4.1 和 5.0 个百分点。

1 相关工作

1.1 目标跟踪

目标跟踪作为计算机视觉领域的核心任务之一，近年取得了快速发展，逐步从单目标扩展到多目标，并从二维场景扩展到三维场景。在二维单目标跟踪（Single Object Tracking，SOT）方面，SiamBAN[5]通过设计目标感知分支解决了分类与回归任务的不一致问题，在 VOT2018[6]、VOT2019[7]、UVA123[8]和 LaSOT[9]等多个基准测试中均取得了具有竞争力的结果。在多目标跟踪方面，由 Maggiolin 等人[10]提出的 Deep-OC-SORT，以运动驱动的 OC-SORT[11]为基础框架，引入自适应外观匹配机制，增强了特征退化与遮挡场景下的跟踪鲁棒性，在 MOT17 以及 MOT20 数据集上展示了较为优越的跟踪性能。在三维场景中，Weng 等人[12]提出的 AB3DMOT，通过激光雷达三维检测结合三维卡尔曼滤波与匈牙利算法，实现了简单高效的实时三维多目标跟踪，在 KITTI[13]数据集上的运行速度可达 207.4 FPS。然而，现有高性能的三维多目标跟踪方法往往依赖精密的传感器数据[14]，不仅成本高昂，更增加了实际应用的复杂度；其次，这些方法只能基于视觉和运动特征进行通用的全目标跟踪，缺乏面向特定目标的选择性关注机制，无法根据需求对特定目标进行跟踪。这些局限制约了目标跟踪技术在高效性与灵活性方面的进一步发展。

1.2 自然语言描述驱动的目标跟踪

自然语言描述驱动的目标跟踪通过引入语言描述为跟踪提供语义信息作为约束和指导，使得系统能够模拟人类对目标外观、类别及行为信息的利用逻辑，以增强跟踪的语义理解能力和跨模态适应性，从而克服传统目标跟踪在复杂场景下的局限。目前，该技术在二维领域已取得了显著进展：TransRMOT[15]是一种基于 Transformer 的在线的自然语言驱动的二维多目标跟踪架构，能够逐帧利用语言描述进行目标检测和关联，在 Refer-KITTI 数据集上取得了优异性能。HFF-Tracker[16]通过层次化细粒度的文本-图像融合、语言引导的解码与预测机制以及回溯式训练策略，在 Refer-KITTI[15]以及 Refer-KITTI2 数据集上取得了多项跟踪性能上的领先表现。

尽管现有自然语言描述驱动的二维跟踪方法取得了一定进展，但它们对于目标的语义信息挖掘有限，局限于二维像素平面，而面向三维目标跟踪的研究仍然稀少。在这一背景下，杨洋等人[17]提出了单目视角下自然语言描述驱动的三维目标跟踪任务 NLOT3D，并构建了对应数据集 NLOT3D-SPD。同时，设计了 NLOT3D-TR 模型，通过跨模态融合视觉与文本特征，实现了优异的跟踪性能。但是，该架构仍局限于单目标跟踪，缺乏将描述文本与三维空间中多个候选目标进行特征对齐的能力，并且所用的中文描述冗长，降低了语言指导下三维目标跟踪的效率，并在一定程度上影响了模型在复杂场景下的实用性。

2　数据集构建

为缓解三维多目标跟踪技术对多传感器架构的依赖,并规避语言冗余导致的跟踪效率降低问题,本文提出基于中文引导的单目视角下三维多目标跟踪新任务 CIMOT3D,并构建了支撑该任务的基准数据集 CIMOT3D-5k。数据集的构建包含视频处理与中文描述生成两个阶段。

在视频处理阶段,本研究基于 DAIR-V2X-Seq 数据集[3]对原始数据进行了系统性的整理与筛选,以完成视频片段的优化构建。首先,将分段式的连续图像帧整合为完整的视频片段;随后依据多目标跟踪任务对目标动态性的要求,剔除运动性不足的样本,最终得到 5 562 个高质量的视频序列。这些视频序列覆盖城市中 15 个十字路口场景,涵盖了光照良好的晴天与光照较弱的阴天和雨天等天气条件。标注目标包括轿车、厢型车、货车和客车四类,并对这四类目标实例进行持续跟踪。

在中文描述生成方面,本文结合千问 2.5 大语言模型[4],构建了自动化语言生成流程,并辅以人工验证以确保中文描述的精确与可靠。图 2 展示了中文描述生成流程,首先通过特征提取模块从原始数据中获得目标的语义类别、几何参数、距离、可见性、朝向,以及基于 DeiT-Ti[18]模型提取的颜色与转向行为特征。其次从目标属性中随机选取 2~6 项进行属性重组,形成描述依据。随后基于千问 2.5 大语言模型[4]构建中文描述生成模板,将重组后的属性转化为简练的中文描述。研究团队对生成的中文描述进行了人工验证,据人为评估,仅不足 5% 的样本需要调整,表明该描述生成方法的高效性与可靠性。

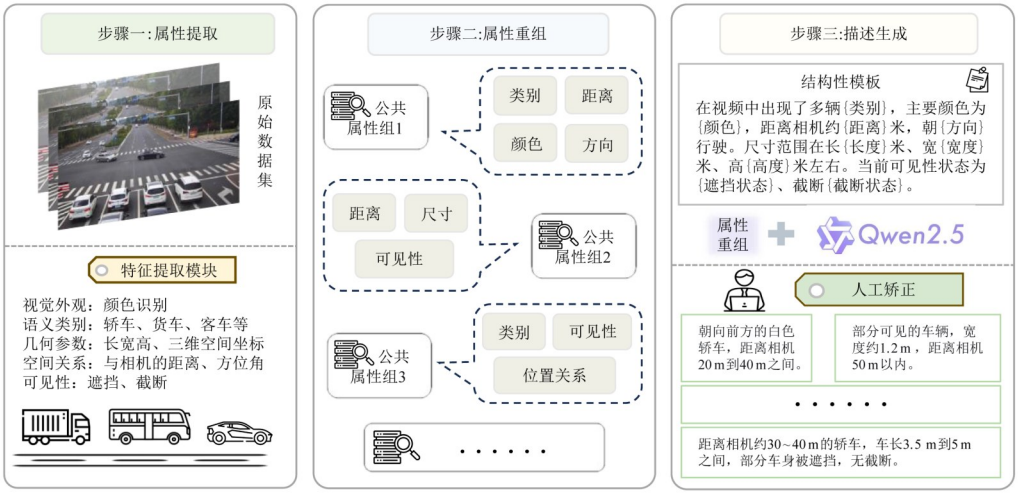


图 2　基于中文引导的单目视角下三维多目标跟踪数据集构建过程图

Figure 2　Construction process of the Chinese-instructed monocular 3D multi-object tracking dataset

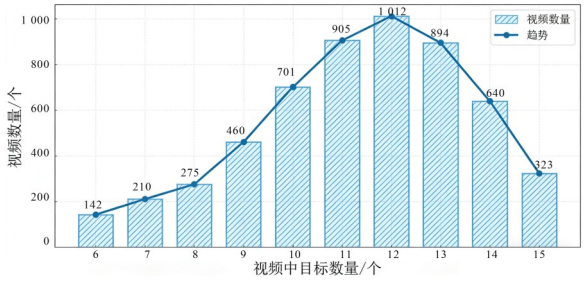
数据集中的每段视频序列均具备精确的二维边界框标注与三维空间坐标标注,每个视频序列还配有三条中文描述以对应不同目标组。为刻画数据集的组成结构,本文统计了不同目标数量条件下的视频样本分布以及各类目标在数据集中的占比,结果如图 3(a)和图 3(b)所示。此外,对全部中文描述进行了词频统计并生成词云,以可视化呈现高频语义要素及描述侧重点,结果如图 3(c)所示。相较于 NLOT3D-SPD 数据集固定使用全部属性生成描述的模式,本方法借助属性重组构建多样化语言表达,在确保描述准确性的基础上增强其灵活性,有效提升了语言-视觉特征的对齐精度与模型训练效率,为三维空间下的 NLOT 任务提供了重要数据支撑。

3　方法

为解决所提出的 CIMOT3D 任务,本文设计了专用的网络模型 CIMOT3D-SyncTracker。如图 4 所示,该模型由多模态特征提取器、中文描述驱动的视觉编码器和解码器以及检测器和跟踪器三部分组成。各组件的具体结构与功能将在后续小节中详细说明。

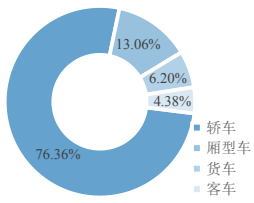
3.1　多模态特征提取器

多模态特征提取器旨在从中文描述和单目 RGB 图像这两种不同模态的数据中提取特征信息:从中文语言模态获取关于几何的表征信息,同时从视觉模态捕获外观特征与深度线索。对于语言模态,采用预训练的 RoBERTa-Chinese[19]模型对输入的中文描述进行编码,经线性变换将其映射到特征空间,从而生成



(a) 不同目标数量条件下的视频样本分布

(a) Distribution of video samples under different numbers of targets



(b) 目标类别占比

(b) Target category distribution

(c) 词云图

(c) Word cloud

图3 数据集统计

Figure 3 Dataset statistics

用于和视觉信息相互匹配、有效对齐的语义特征 $\boldsymbol{F}^l \in \mathbb{R}^{C\times N_l}$，其中 N_l 表示中文描述的长度。对于视觉模态，使用 ResNet-50[20] 作为骨干网络，从其中四个不同层级提取特征，并通过线性层融合为多尺度视觉特征 $\boldsymbol{F}^i \in \mathbb{R}^{C\times N_i}$，以整合不同感受野下的视觉线索，实现对目标更全面的表征，其中 $C=256$，$N_i=HW/8^2+HW/16^2+HW/32^2+HW/64^2$。此外，为补全单目图像缺失的三维深度信息，提取器还引入轻量化深度估计层，对 ResNet-50 提取的视觉特征进行深度预测并生成深度特征 $\boldsymbol{F}^d \in \mathbb{R}^{C\times N_d}$（其中 $N_d=H/16\times W/16$），进而提升最终三维跟踪任务的定位精度。

3.2 中文描述驱动的视觉编码器和解码器

中文描述驱动的视觉编码-解码器由视觉编码器、深度编码器、几何表征解码器（Geometric Representation Decoder，GRD）以及中文与视觉特征对齐模块（Chinese-Visual Feature Alignment，CVFA）四部分组成。视觉和深度编码器分别提取非局部的视觉与深度特征向量，使解码器中的目标查询能够自适应地捕获场景级信息。GRD 负责整合以及抽取关键的三维几何特征，为后续模块提供视觉特征的结构化表示。CVFA 模块将外观特征与蕴含着三维空间信息的中文语义特征对齐和融合在一起，丰富了对三维空间中对象的上下文理解。

3.2.1 视觉和深度编码器

对于给定的视觉和深度特征，本文使用两个基于 Transformer[21-22] 的编码器来生成具有全局感受野的场景级嵌入，表示为 $\boldsymbol{v}_{\text{enc}} \in \mathbb{R}^{HWC/32^2}$ 和 $\boldsymbol{d}_{\text{enc}} \in \mathbb{R}^{HWC/32^2}$。其中，视觉编码器由两个堆叠模块构成，深度编码器仅由一个模块构成，这是因为相较于丰富的视觉外观信息，离散的前景深度信息通常更容易进行编码。深度

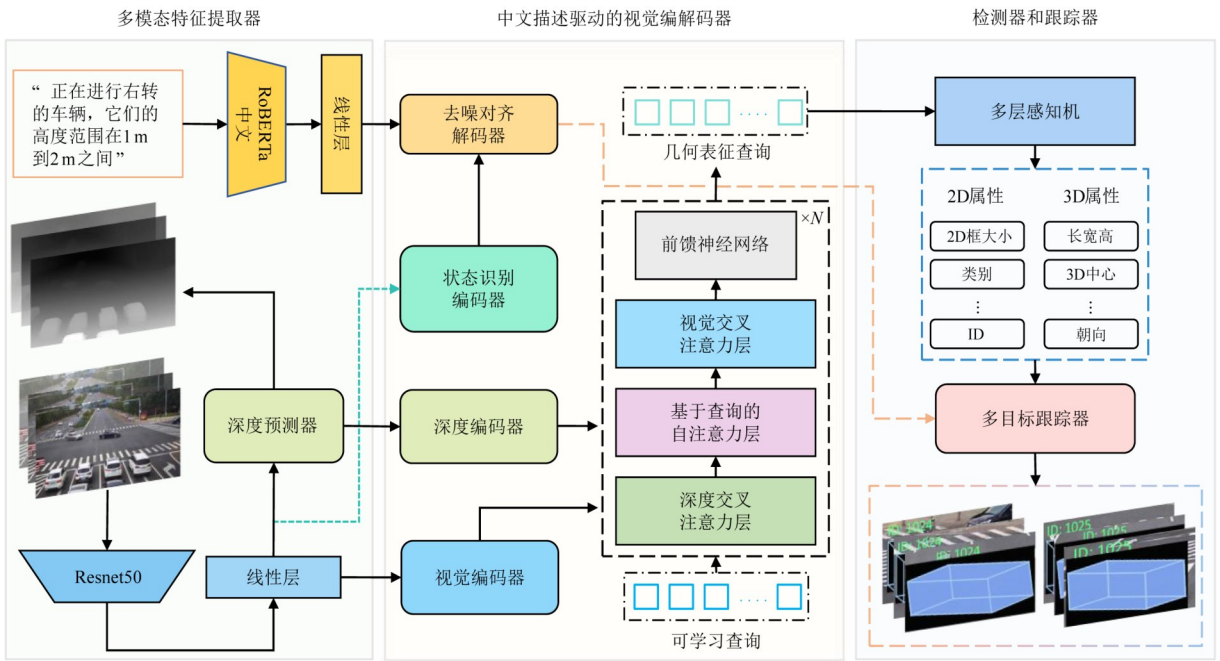


图4 CIMOT3D-SyncTracker网络模型图

Figure 4 CIMOT3D-SyncTracker network architecture

编码器通过全局自注意力机制[23],能够捕获不同前景区域之间深度值的长距离依赖关系,为立体空间提供关键的非局部几何线索。此外,视觉编码器与深度编码器的显式解耦,使二者能够充分挖掘各自模态特征,以互补方式对输入图像进行编码,提取深度几何信息和视觉外观信息。

3.2.2　几何表征解码器

GRD用于将编码阶段提取的视觉和深度特征进行有效解码,并融合为结构化的几何表征。基于非局部的视觉嵌入 $\boldsymbol{v}_{\text{enc}}$ 和深度嵌入 $\boldsymbol{d}_{\text{enc}}$,GRD引入一组可学习的查询向量 $\boldsymbol{Q}_t \in \mathbb{R}^{N\times C}$(其中 N 表示输入图像中对象的预设最大数量),通过级联的交互过程,逐步将这些抽象的查询向量解码为包含丰富三维信息的几何表征查询,用于后续3D对象检测和跟踪。在此过程中,查询向量依次通过一个深度交叉注意力层,一个查询自注意力层和一个视觉交叉注意力层完成解码。嵌入特征 $\boldsymbol{d}_{\text{enc}}$ 和 $\boldsymbol{v}_{\text{enc}}$ 在深度和视觉交叉注意力层中,分别通过各自的权重矩阵生成对应的键和值,与查询向量进行交互。最终GRD输出经充分交互和增强后的几何表征查询 $\widetilde{\boldsymbol{Q}}^l$。这一机制能够使每个可学习查询可动态地将注意力聚焦于图像中几何信息最显著的区域,以高效提取空间线索,最终实现对场景更为细致的全局空间认知。

3.2.3　中文与视觉特征对齐模块

中文与视觉特征对齐模块由一个状态识别编码器(State-Aware-Encoder,SAE)和一个去噪对齐解码器(Denoising Alignment Decoder,DAD)构成,用于在视觉与语言模态之间建立鲁棒且高效的对齐关系并生成联合嵌入特征。如图5①所示,多模态特征提取器得到的多尺度视觉特征 $\boldsymbol{F}^i$ 输入到SAE中,SAE通过状态识别与状态编码模块,从输入特征中提取语义类别与局部结构属性,并将其编码为状态提示向量。同时,利用预先训练的ViT(Vision Transformer)[24]模型对输入特征中的局部区域提取二维外观特征,经过2D视觉编码器进一步处理后,与状态提示向量在特征融合层进行交互,使视觉特征在保持原始感知信息的同时,还融入了状态相关的区分性线索。随后,经由多层感知机的非线性映射,得到同时富含视觉、几何与状态信息的嵌入特征 $\boldsymbol{F}^v \in \mathbb{R}^{C\times N}$,其中 C 为特征维度,N 是图像块的数量,用于在DAD中与语义特征进行融合。

图5②展示了去噪对齐解码器(DAD),用于缓解对齐过程中可能产生的噪声和冗余问题,并进行跨模态对齐得到联合特征。考虑到视觉与语言模态本身固有的信息冗余,在进行多模态对齐时,这些冗余或不相关特征很容易被错误地关联并放大为跨模态噪声,进而影响融合表示的有效性。为抑制此类噪声并提升特征质量,本文设计了特征平滑去噪(Smoothed Feature Denoising,SFD)模块,其核心思想为,在进行多模态融合前,先将各单模态特征映射到一个共享的潜在空间中进行提纯与去噪。具体而言,对于给定的任一单模态嵌入 $\boldsymbol{f}^v$ 或 $\boldsymbol{f}^l$,SFD都为其去噪并生成潜在表示 $\boldsymbol{h}^i$,其中 i 代表视觉 v 或语言模态 l。该映射由一个学习到的条件分布所控制,其建模如下:

$$\begin{aligned} &p\left(\hat{\boldsymbol{h}}_i \middle| \hat{\boldsymbol{f}}_i\right) \sim N\left(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i{}^2 \boldsymbol{I}\right), \\ &\hat{\boldsymbol{\mu}}_i = \phi_{\hat{\theta}^{(1)}}\left(\hat{\boldsymbol{f}}_i\right), \hat{\boldsymbol{\sigma}}_i = \phi_{\hat{\theta}^{(2)}}\left(\hat{\boldsymbol{f}}_i\right) \end{aligned} \tag{1}$$

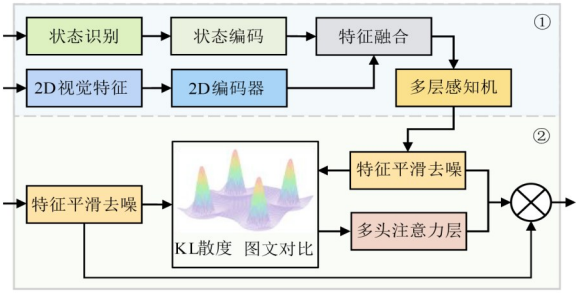
其中,$\phi_{\hat{\theta}^{(1)}}$ 和 $\phi_{\hat{\theta}^{(2)}}$ 是两组可学习的参数函数。随后,使用重新参数化策略[25]得到去噪化的潜在表示:

$$\hat{\boldsymbol{h}}_i = \hat{\boldsymbol{\mu}}_i + \boldsymbol{\varepsilon}\hat{\boldsymbol{\sigma}}_i, \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}) \tag{2}$$

为构建兼具鲁棒性和动态自适应能力的联合多模态表示,本文在去噪对齐解码器中进一步设计了对齐与融合机制。该机制的核心在于,引入可学习的查询,并借鉴对齐引导融合[26]的设计范式,通过查询与多模态特征之间的动态交互实现特征对齐与融合。具体而言,给定视觉嵌入 $\boldsymbol{f}^v \in \mathbb{R}^N$ 和语言嵌入 $\boldsymbol{f}^l \in \mathbb{R}^L$,经SFD模块后到维度 $D=512$ 的去噪潜在表示 $\hat{\boldsymbol{h}}_v$ 和 $\hat{\boldsymbol{h}}_l$。随后对其进行L2归一化,并采用带负样本挖掘策略的对比学习损失[27]完成对齐,由此得到对齐良好的分类标记 $\hat{\boldsymbol{h}}'_v$ 和 $\hat{\boldsymbol{h}}'_l$。随后的融合过程在一个多头交叉注意力(Multi-Head Cross-Attention,MHCA)模块[28]中完成:该模块引入一组可学习的融合查询,并将前一阶段得到的对齐标记 $\hat{\boldsymbol{h}}'_v$ 和 $\hat{\boldsymbol{h}}'_l$ 进行拼接,作为注意力机制中的键和值,通过让融合查询与这个高度对齐的序列进行交互,生成包含跨模态交互信息的联合嵌入 $\boldsymbol{f}_{vl} \in \mathbb{R}^C$。为适配下游任务的需求,初步的联合嵌入 $\boldsymbol{f}_{vl}$ 会进一步与代表提纯后单模态信息的对齐分类标记 $\hat{\boldsymbol{h}}'_v$ 和 $\hat{\boldsymbol{h}}'_l$ 进行拼接,构成包含三部分信息(视觉、语言和交互)的复合序列,作为分层且全面的多模态联合嵌入 $\boldsymbol{f}_{vl} \in \mathbb{R}^C$。同时,为减弱融合过程中可能放大的不确定性驱动噪声,再次使用SFD模块对该复合序列进行去噪,得到多模态联合潜在表示 $\hat{\boldsymbol{h}}_{vl}$,用于下游任务。

3.3　多目标检测跟踪器与损失函数设计

模型的最终输出由一组基于MLP的三维属性预测头生成。这些预测头以结构化的几何表征查询作为输入,对目标的多项属性进行预测,涵盖类别、颜色到三维尺寸和朝向等多个维度。为训练整个网络,模型的总体损失由两部分构成:一部分是上述检测头产生的检测损失,另一部分则是来自视觉-语言对齐模块的KL散度损失与图文对比损失。



注：①状态提示编码器；②去噪对齐解码器。

图5 中文与视觉特征对齐模块

Figure 5 CVFA module

3.3.1 双向匹配

为实现查询与目标之间的精确匹配，首先针对每一对查询-标签计算匹配损失，然后利用匈牙利算法[29]寻找全局最优分配。对于任一匹配对，其损失可分为两类：第一类包括目标类别、目标颜色、二维尺寸以及二维中心点，这些属性主要描述目标的二维视觉外观；第二类则由深度、三维尺寸、投影三维中心以及朝向构成，这些属于目标的三维空间属性。分别对两类损失求和，记为 $L_{2\mathrm{d}}$ 和 $L_{3\mathrm{d}}$。由于在训练初期，网络对三维属性的预测精度低于二维属性，$L_{3\mathrm{d}}$ 的数值不稳定且可能干扰匹配过程，因此在计算每一对查询-标签的匹配代价时，仅采用 $L_{2\mathrm{d}}$ 进行计算。

3.3.2 视觉与语言对比损失

为在多模态融合之前学习更优的单模态去噪潜在表示，本文引入具有双向约束的三元组排序损失[30]用于特征对齐。该损失函数通过约束匹配样本对的相似度得分显著高于非匹配样本对，以显式强化跨模态判别能力，其定义如下：

$$\begin{aligned}L_{\mathrm{vlc}}=&\mathbb{E}_{(v,l)\sim D}\left[d-s\left(\hat{\boldsymbol{h}}_v,\hat{\boldsymbol{h}}_l\right)+s\left(\hat{\boldsymbol{h}}_v,\hat{\boldsymbol{h}}_l^-\right)\right]_+\\&+\left[d-s\left(\hat{\boldsymbol{h}}_v,\hat{\boldsymbol{h}}_l\right)+s\left(\hat{\boldsymbol{h}}_v^-,\hat{\boldsymbol{h}}_l\right)\right]_+\end{aligned}\tag{3}$$

$$\hat{\boldsymbol{h}}_l^-=\arg\max\left[s\left(\boldsymbol{z}_v,\hat{\boldsymbol{h}}_l\right)\right],\boldsymbol{z}_v\neq\hat{\boldsymbol{h}}_v\tag{4}$$

$$\hat{\boldsymbol{h}}_v^-=\arg\max\left[s\left(\boldsymbol{z}_l,\hat{\boldsymbol{h}}_v\right)\right],\boldsymbol{z}_l\neq\hat{\boldsymbol{h}}_l\tag{5}$$

其中，d 表示间隔超参数；$[\cdot]_+$ 表示 $\max(0,\cdot)$ 函数；$\hat{\boldsymbol{h}}_l^-$ 和 $\hat{\boldsymbol{h}}_v^-$ 分别表示 mini-batch 内挖掘得到的困难负样本，其特征表示在优化过程中被显式拉远，以增强其与真实匹配样本的可分性。$s(\cdot)$ 表示视觉嵌入与语言嵌入之间的相似度，通常采用余弦相似度进行计算。

3.3.3 Kullback-Leibler 散度损失

为进一步降低模态对齐与融合过程中由不确定性引入的噪声，本文采用 KL 散度（Kullback-Leibler Divergence，KLD）[31]作为监督信号，通过逼近单模态与多模态高斯分布实现去噪。具体而言，根据式（1）中获得的视觉、语言及其联合嵌入的均值与方差，分别构建三个高斯分布：$N(\hat{\boldsymbol{\mu}}_v,\hat{\boldsymbol{\sigma}}_v^2\boldsymbol{I})$、$N(\hat{\boldsymbol{\mu}}_l,\hat{\boldsymbol{\sigma}}_l^2\boldsymbol{I})$ 和 $N(\hat{\boldsymbol{\mu}}_{vl},\hat{\boldsymbol{\sigma}}_{vl}^2\boldsymbol{I})$。随后引入三项 KL 散度损失，通过促使这三个高斯分布在两两之间逐步收敛，从而共同推动特征的平滑与去噪。三项损失的具体形式定义如下：

$$L_{\mathrm{KLD}}=\frac{1}{3}\mathbb{E}_{(v,l)\sim D}\begin{bmatrix}\mathrm{KL}\left(N(\hat{\boldsymbol{\mu}}_v,\hat{\boldsymbol{\sigma}}_v^2\boldsymbol{I})\middle\|N(\hat{\boldsymbol{\mu}}_l,\hat{\boldsymbol{\sigma}}_l^2\boldsymbol{I})\right)\\+\mathrm{KL}\left(N(\hat{\boldsymbol{\mu}}_v,\hat{\boldsymbol{\sigma}}_v^2\boldsymbol{I})\middle\|N(\hat{\boldsymbol{\mu}}_{vl},\hat{\boldsymbol{\sigma}}_{vl}^2\boldsymbol{I})\right)\\+\mathrm{KL}\left(N(\hat{\boldsymbol{\mu}}_l,\hat{\boldsymbol{\sigma}}_l^2\boldsymbol{I})\middle\|N(\hat{\boldsymbol{\mu}}_{vl},\hat{\boldsymbol{\sigma}}_{vl}^2\boldsymbol{I})\right)\end{bmatrix}\tag{6}$$

3.3.4 整体损失函数

类似于视觉-语言匹配任务，检测头的最终任务是判定一个“物体-描述”对是否关联。将多模态联合潜在表示 $\hat{\boldsymbol{h}}_{vl}$ 输入到一个由两层多层感知机（Multi-Layer Perceptron，MLP）构成的二分类器中，得到物体-描述对的关联概率 P_{ODA}，据此在同一条中文描述下完成多目标关联。相应的物体-描述关联度（Object-Description Association，ODA）损失记为 L_{ODA}，其定义采用标准的交叉熵形式，表示如下：

$$L_{\mathrm{ODA}}=\mathbb{E}_{(v,l)\sim D}\mathbb{B}\left(Y^{\mathrm{ODA}},P^{\mathrm{ODA}}\right)\tag{7}$$

其中，$\mathbb{B}$ 表示二元交叉熵损失函数；Y^{ODA} 表对应独热编码的真值标签。

此外，为对模型生成的深度图进行监督，引入基于 Focal Loss 计算的深度图预测损失 L_{dm}，并纳入总体损失的计算中。综合上述各损失项，模型的总体损失 L_{total} 表示如下：

$$L_{\mathrm{total}}=\frac{1}{N_{\mathrm{gt}}}\left(L_{2\mathrm{d}}+L_{3\mathrm{d}}\right)+L_{\mathrm{dm}}+L_{\mathrm{vlc}}+L_{\mathrm{KLD}}+L_{\mathrm{ODA}}\tag{8}$$

其中，$L_{2\mathrm{d}}$ 和 $L_{3\mathrm{d}}$ 分别表示二维检测损失和三维检测损失。

3.3.5 三维多目标跟踪器

本文采用了卡尔曼滤波[32]作为三维多目标跟踪任务的基础运动模型。借鉴 AB3DMOT[12]的工作，本研究构建了十维状态向量 $(x,y,z,\theta,l,w,h,x',y',z')$ 用于表征目标，其中 (x,y,z) 表示三维包围框的中心坐标，(l,w,h) 为其尺寸，θ 为该目标的朝向角，而 (x',y',z') 则对应其三维空间的速度。在该状态空间基础上，采用配置为匀速运动模型和线性观测模型的卡尔曼滤波器，将轨迹状态更新为预测状态和匹配观测状态的加权平均，权重由各自的不确定性大小确定。

借鉴 ByteTrackv2[33]中的思想，本文在 CIMOT3D-SyncTracker 框架中引入一种互补的三维运动预测策略，并在实验中验证了其有效性。该策略的设计基于以下观察：最新的三维检测器[34]能够通过时序建模，提供精确的短期瞬时速度；而卡尔曼滤波器则更擅长利用历史轨迹数据建模平滑的长期平均速度。为充

分利用这两种速度信息的互补优势,本文采用了一种双向预测的方法:前向预测利用卡尔曼滤波器完成,依托平滑的长期速度,对丢失轨迹进行稳定的状态预测;后向预测则基于检测器得到的速度完成,凭借较高的速度精度,能够高效地对活跃轨迹执行短期运动预测,尤其适用于物体发生剧烈或非线性运动的场景。

设在 t 时刻共得到 M 个检测结果,其集合记为 $\boldsymbol{O}_t \in \mathbb{R}^{M\times 7}$,包含目标的中心位置、转向角与尺寸。同时记检测器在 x 与 z 方向上估计的速度为 $\boldsymbol{S}_t \in \mathbb{R}^{M\times 2}$,其反向预测计算如下:

$$\hat{\boldsymbol{O}}_{t-1}^{(x)} = \boldsymbol{O}_t^{(x)} - \boldsymbol{S}_t^{(x)}, \hat{\boldsymbol{O}}_{t-1}^{(z)} = \boldsymbol{O}_t^{(z)} - \boldsymbol{S}_t^{(z)} \tag{9}$$

给定在 $t-1$ 时刻的 N 个轨迹 $\boldsymbol{T}_{t-1} \in \mathbb{R}^{N\times 7}$,使用卡尔曼滤波器执行前向预测的公式为

$$\hat{\boldsymbol{T}}_t^{(x,z)} = \boldsymbol{T}_{t-1}^{(x,z)} + \dot{\boldsymbol{S}}_{t-1}^{(x,z)} \tag{10}$$

其中,$\dot{\boldsymbol{S}}_{t-1}^{(x,z)}$ 表示卡尔曼滤波器计算得到的平滑速度。在完成双向预测后,本文采用统一的数据关联策略。具体而言,对每一个经过后向预测得到的检测结果 $\hat{\boldsymbol{O}}_{t-1}$ 与每条已有轨迹 $\boldsymbol{T}_{t-1}$ 通过三维广义交并比(3D Generalized Intersection Over Union, 3D GIOU)[35]计算得到关联矩阵 $\boldsymbol{A}_{t-1} \in \mathbb{R}^{M\times N}$,以刻画包围框无重叠情况下的空间匹配度。随后,基于该关联矩阵,使用匈牙利算法完成轨迹与检测结果之间的身份匹配。最后,成功匹配的检测结果 $\boldsymbol{O}_t^{\text{assoc}}$ 通过标准的卡尔曼滤波器更新规则来更新对应的轨迹 $\boldsymbol{O}_t^{\text{assoc}}$ 以完成第一次的关联。

前述式(10)中定义的前向预测机制,在处理轨迹找回的场景中扮演着至关重要的角色。更具体地说,对于那些暂时丢失的轨迹,该机制能够为其提供一个预测状态。一旦这些物体在后续帧中重新出现,模型便能够使用这个预测状态,成功地将其与原有的轨迹重新关联起来。为实现这种重关联,模型会进行第二次关联,其流程与第一次关联保持一致。

4　实验

4.1　实验配置

所构建的CIMOT3D-5k数据集包含5 562个视频,将这些视频按照7∶2∶1的比例,划分为训练集、测试集和验证集。每个视频序列均配备了三条准确且贴近人类自然表达方式的中文描述,为模型训练与推理提供支撑。本文提出的CIMOT3D-SyncTracker模型基于Pytorch框架实现,训练过程中采用AdamW优化器[36],初始学习率和权重衰减率分别设置为 2×10^{-4} 和 1×10^{-4},训练周期设置为40轮,所有实验均在搭载2张Tesla V100 32 GB GPU的硬件环境下开展。

4.2　评估指标

本实验采用了八项评估指标来衡量最终的实验结果:多目标跟踪准确度(MOTA)、多目标跟踪精度(Multiple Object Tracking Precision, MOTP)、身份判别精度(IDentity Precision, IDP)、身份保持召回率(IDentity Recall, IDR)、身份一致性指标(IDF1)、身份标识切换次数(ID Switches, IDS)、跟踪帧数占比20%~80%(Partially Tracked, PT)以及跟踪帧数占比大于80%的目标比例(Mostly Tracked, MT)。其中,MOTA和MOTP反映跟踪的整体准确性和位置精度;IDP、IDR、IDF1评估身份保持能力,衡量目标身份的连续性;IDS记录身份切换次数,反映轨迹稳定性,为增强指标直观性,采用比率形式呈现;PT和MT则通过衡量目标被成功跟踪的时间占比,体现算法的长期跟踪能力。上述指标体系可全面覆盖算法在检测精度、身份保持以及轨迹稳定性等核心维度的性能表现。

4.3　实验基准

为全面验证不同方法在CIMOT3D任务中的有效性,本文基于CIMOT3D-5k数据集设计了一系列基准实验。实验主要囊括两类基线方法:第一类为自然语言描述驱动的二维多目标跟踪集合逆投影方法,包括了TransRMOT[15]、HFF-Tracker[16]、DeepSORT[37]、CSTracker[38]以及EchoTrack[39]。第二类为单目三维多目标跟踪方法与本研究的中文指导机制(Chinese-Instructed Mechanism, CIM)相结合来完成匹配,包括QD-3DT[40]和MoMA-M3T[41]。此外,为适配CIMOT3D任务,本文还在NLOT3D-TR[17]的基础上构建了三维多目标跟踪变体作为额外对比基线。通过将上述基线方法与所提出的CIMOT3D-SyncTracker模型进行系统对比,能够系统性地评估各模型的性能表现,进而验证本文方法的有效性。

4.4　结果分析与可视化

在基线实验对比中,CIMOT3D-SyncTracker模型在多项关键指标上均优于现有方法。由表2可知,该模型的整体跟踪准确度(MOTA)达到31.5%,相较于其他基线方法中表现最佳的基线模型HFF-Tracker+backproj提升4.1%,表明所提模型在跟踪准确性方面具有明显优势。在边界框精度(MOTP)方面,CIMOT3D-SyncTracker的结果为19.8%,为所有方法中最优,说明其在目标三维空间位置与尺寸的预测上更加精确。此外,其身份一致性指标(IDF1)达到52.3%,相较于次优的CSTracker+backproj提升5.0%,进一步体现了模型在身份保持上的稳定性与可靠性。这些核心指标的提升共同表明CIMOT3D-SyncTracker在CIMOT3D任务中具有较为优越的跟踪性能。

基于二维视觉感知与逆投影的基线方法,其三维

表2 基线实验对比结果 单位:%

Table 2 The benchmark comparison results unit: %

方法	MOTA(↑)	MOTP(↓)	IDR(↑)	IDP(↑)	IDF1(↑)	IDS(↓)	PT(↑)	MT(↑)
DeepSORT + backproj[37]	25.7	27.9	36.1	51.0	42.3	13.7	54.9	28.8
TransRMOT + backproj[15]	26.8	26.7	36.5	49.8	42.1	12.7	56.8	29.1
CSTracker + backproj[38]	25.1	24.9	39.2	53.2	45.3	11.9	56.1	31.3
QD-3DT + CIM[40]	26.7	25.0	38.2	53.1	44.7	11.4	56.9	31.4
EchoTrack + backproj[39]	26.9	25.6	37.1	53.8	44.0	15.4	58.0	33.8
NLOT3D-TR(MOT)[17]	26.9	24.5	38.5	53.3	44.9	11.2	57.1	31.7
MoMA-M3T + CIM[41]	27.1	23.2	39.4	53.6	45.1	10.9	57.3	31.9
HFF-Tracker + backproj[16]	27.4	24.8	39.0	54.0	45.2	10.2	57.6	32.2
CIMOT3D-SyncTracker	**31.5**(+4.1)	**19.8**(−3.4)	**41.5**(+2.1)	**61.0**(+7.0)	**50.3**(+5.0)	**9.4**(−0.8)	**62.0**(+4.0)	**36.2**(+2.4)

注:表格中↑表示该指标的值越高性能越好,↓则表示该指标的值越低性能越好,加粗表示每个指标的最佳性能值,下划线表示次优,括号内表示本文提出的模型相比次优方法指标提升的幅度值。

结果完全依赖前一阶段的二维检测,对二维预测误差极为敏感。对于将单目三维多目标跟踪方法与中文指导机制结合的基线,由于缺乏体系化的中文-视觉联合建模,中文提示仅在匹配阶段提供有限约束。NLOT3D-TR多目标跟踪变体虽然能够支撑多目标跟踪,但其架构仍依赖冗长、属性齐全的文本模板,难以适应简练自然的中文描述。这些结构性局限共同制约了上述基线方法在三维定位精度及身份保持性能上的进一步提升。

相比之下,CIMOT3D-SyncTracker能够充分理解和挖掘这种简练而语义丰富的中文描述,并以中文描述作为全流程的结构化引导,在三维视觉空间中完成目标定位与跟踪。这种核心优势不仅让模型在整体性能上表现突出,还使其在跟踪精度与鲁棒性方面均优于其他方法。

为了使实验结果更加直观,本文采用OpenCV框架对部分方法的跟踪过程进行可视化。具体做法是在同一个场景中连续的三帧中,将具有代表性的TransRMOT和CSTracker两种跟踪法结合逆投影得到的三维轨迹框、本文方法得到的轨迹框与目标的真实边界框进行对比,以清晰展示不同方法在目标空间位置、朝向和尺度预测上的差异。得到的可视化结果如图6所示。从可视化结果可以看出,TransRMOT结合逆投影的方法存在较为明显的不足:其预测轨迹的边界框与车辆真实边界框的重合度较低,位置出现较大偏移,且三维框的尺寸也不稳定,导致轨迹整体连续性较差。相比之下,CSTracker结合逆投影的方法在一定程度上减弱了边界框尺寸波动并改善了轨迹连续性,但是在车辆转弯时仍然存在较为显著的朝向以及位置的偏差,未能够保持与真实目标的稳定一致。

相比之下,本文提出的CIMOT3D-SyncTracker在轨迹预测方面表现较为稳健。可视化结果显示,其预测的轨迹边界框在连续帧中保持着良好的空间一致性,预测框尺寸与目标真实尺寸匹配度更高,且朝向能够紧密跟随目标运动变化,表明其在跟踪精度方面具备较为优越的性能。

4.5 消融实验

为深入评估各模块对整体模型性能的贡献,本文针对视觉-深度编码器与视觉-语言对齐模块开展了系统性的消融实验。通过对不同模块配置组合的对比,可以定量分析这些组件在信息表征、特征融合以及跨模态语义对齐过程中的作用。实验结果如表3和表4所示,数据表明,合理的模块设计不仅能够提升跟踪准确度和身份保持能力,还能有效降低错误率与轨迹不稳定性,从而验证整体模型结构的有效性。

4.5.1 编码器模态堆叠分析

在中文引导的视觉-深度编码器中,视觉特征与深度几何信息通过语义提示进行交互和增强,从而更充分地建模多模态语义与空间几何信息的互补性。为分析编码模块层数对性能的影响,本文针对视觉编码器与深度编码模块设计了多种层数组合,并开展对比实验。实验结果如表3所示,当视觉编码模块设置为2层、深度编码模块设置为1层时,模型在MOTA、MOTP和IDF1三个关键指标上取得最优表现。这一配置能够在充分提取有效信息和避免冗余之间实现平衡,使模型能够准确捕获与语言描述紧密相关的特征,提升跨帧目标身份的一致性与跟踪精度。相比之下,过浅的结构难以捕获复杂语义,过深的结构则易引入冗余与噪声,二者均会导致模型性能下降。

4.5.2 中文与视觉特征对齐模块配置分析

在视觉语言对齐模块的实验中,本文重点考察了状态识别编码器(SAE)与视觉-语言对比学习(Vision-Language Contrastive learning, VLC)两部分的性能贡献。其中,SAE用于生成状态提示信息,动态引导模

注:橙色线框为对应目标实际边界框,蓝色框为对应方法预测的轨迹框。

图6　实验结果可视化图

Figure 6　Visualization of experimental results

表3　不同视觉深度编码器层数配置性能对比结果　单位:%

Table 3　comparison of different visual depth encoder layer　unit: %

模块	层数	MOTA(↑)	MOTP(↓)	IDF1(↑)
中文描述引导的视觉编码模块	$N=1$	30.1	23.0	46.7
	$N=2$	**31.5**	**19.8**	**50.3**
	$N=3$	27.2	22.1	48.1
中文描述引导的深度编码模块	**$N=1$**	**31.5**	**19.8**	**50.3**
	$N=2$	29.8	21.0	49.0
	$N=3$	28.7	24.0	44.9

注:表格中↑表示该指标的值越高性能越好,↓则表示该指标的值越低性能越好,加粗表示每个指标的最佳性能值。

型在时空维度上关注与当前目标相关的区域,从而提升状态感知的准确性及提示信息的利用效率;而VLC则通过跨模态一致性约束与对比学习损失,有效增强视觉与语言表征之间的关联性与判别性。实验结果(如表4所示)表明,二者均能显著提升模型性能:SAE改善了目标在连续帧中的定位精度,VLC强化了跨模态语义对齐与判别特征学习。当二者协同作用时,模型在身份保持能力、轨迹稳定性以及跨模态理解能力上均实现了更优表现,能够进一步提升复杂场景下的三维多目标跟踪性能。

表4　SAE和VLC对系统性能影响的对比结果　单位:%

Table 4　Impact of SAE and VLC on system performance　unit: %

SAE	VLC	MOTA(↑)	MOTP(↓)	IDF1(↑)
		21.6	28.0	38.9
√		<u>26.2</u>	27.1	<u>46.7</u>
	√	23.9	<u>21.5</u>	43.6
√	√	**31.5**(+5.3)	**19.8**(−1.7)	**50.3**(+3.6)

注:表格中↑表示该指标的值越高性能越好,↓则表示该指标的值越低性能越好,加粗表示每个指标的最佳性能值,下划线表示次优,括号内表示本文提出的模型相比次优方法指标提升的幅度值。

5　结束语

当前自然语言描述驱动的目标跟踪方法集中于二维目标跟踪,或依赖冗长语言描述的三维单目标跟踪。对此,本文提出了CIMOT3D这一新任务,旨在在单目视频中通过简练的中文描述引导来完成特定多目标的三维跟踪,从而增强机器对复杂开放交通场景的语义理解能力。为支撑该任务,本文构建了CIMOT3D-5k数据集,其中包含多样化的单目视频序列,并配有符合人类表达习惯的中文描述。这些描述由千问2.5大语言模型生成,并经过人工校验,以提高训练与评估的可靠性与有效性。在此基础上,本文设

计了首个面向CIMOT3D任务的神经网络，CIMOT3D-SyncTracker，该方法整合了多模态特征提取器、视觉语言编码-解码器以及几何表征解码器，能够建立文本与多目标之间的关联，进而实现在中文指导下的单目视角三维多目标跟踪任务。本文设计的系统性实验表明，所提出模型在多个评估指标上均显著优于现有基线方法。

本研究推动了视觉语言模型在三维多目标跟踪任务中的发展，并为相关方向的进一步研究提供了新的思路和参考。未来工作将进一步探索更丰富的跨模态交互和更具挑战性的开放世界应用场景，以提升模型在复杂环境中的鲁棒性与实用性。

致谢 感谢张朝阳老师、宋焕生教授，以及雷琪博士给本文提出的参考性意见。

参考文献

[1] 伍瀚, 孙浩, 计科峰, 等. 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[J]. 电子学报, 2025, 53(3): 728-743.
Wu Han, Sun Hao, Ji Kefeng, et al. Temporal-guided cross-view feature fusion network for multi-drone multi-object tracking[J]. Acta Electronica Sinica, 2025, 53(3): 728-743. (in Chinese)

[2] 郑锦, 蒋博韬, 彭微, 等. LiDar点云指导下特征分布趋同与语义关联的3D目标检测[J]. 电子学报, 2024, 52(5): 1700-1715.
Zheng Jin, Jiang Botao, Peng Wei, et al. 3D object detection based on feature distribution convergence guided by LiDar point cloud and semantic association[J]. Acta Electronica Sinica, 2024, 52(5): 1700-1715. (in Chinese)

[3] Yu Haibao, Yang Wenxian, Ruan Hongzhi, et al. V2X-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 5486-5495.

[4] Bai Shuai, Chen Keqin, Liu Xuejing, et al. Qwen2.5-vl technical report[PP/OL]. V1.arXiv(2025-02-19)[2025-09-10]. https://arxiv.org/abs/2502.13923.

[5] Chen Zedu, Zhong Bineng, Li Guorong, et al. SiamBAN: Target-aware tracking with Siamese box adaptive network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 5158-5173.

[6] Kristan M, Leonardis A, Matas J, et al. The sixth visual object tracking VOT2018 challenge results[C]//International Conference on Computer Vision. Munich: Springer, 2019: 3-53.

[7] Kristan M, Matas J, Leonardis A, et al. The seventh visual object tracking VOT2019 challenge results[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop. Piscataway: IEEE, 2019: 2206-2241.

[8] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[C]//14th European Conference on Computer Vision. Cham: Springer, 2016: 445-461.

[9] Fan Heng, Lin Liting, Yang Fan, et al. LaSOT: A high-quality benchmark for large-scale single object tracking[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 5369-5378.

[10] Maggiolino G, Ahmad A, Cao J K, et al. Deep OC-sort: Multi-pedestrian tracking by adaptive re-identification[C]//2023 IEEE International Conference on Image Processing. Piscataway: IEEE, 2023: 3025-3029.

[11] Cao Jinkun, Pang Jiangmiao, Weng Xinshuo, et al. Observation-centric SORT: Rethinking SORT for robust multi-object tracking[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 9686-9696.

[12] Weng Xinshuo, Wang Jianren, Held D, et al. AB3DMOT: A baseline for 3D multi-object tracking and new evaluation metrics[PP/OL]. V1.arXiv (2020-08-18)[2025-09-23]. https://arXiv.org/abs/2008.08063.

[13] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The KITTI dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[14] Yin Junbo, Shen Jianbing, Chen Runnan, et al. IS-fusion: Instance-scene collaborative fusion for multimodal 3D object detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 14905-14915.

[15] Wu Dongming, Han Wencheng, Wang Tiancai, et al. Referring multi-object tracking[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 14633-14642.

[16] Zhao Zeyong, Hao Yanchao, Zhang Minghao, et al. HFF-tracker: A hierarchical fine-grained fusion tracker for referring multi-object tracking[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(10): 10528-10536.

[17] 杨洋, 魏弘凯, 孙士杰, 等. NLOT3D: 单目视角下自然语言描述驱动的三维目标跟踪研究[J]. 电子学报, 2025, 53(6): 2038-2049.
Yang Yang, Wei Hongkai, Sun Shijie, et al. NLOT3D:

Natural-language-driven 3D object tracking in monocular view[J]. Acta Electronica Sinica, 2025, 53(6): 2038-2049. (in Chinese)

[18] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]//Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021: 10347-10357.

[19] Cui Yiming, Che Wanxiang, Liu Ting, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[20] Koonce B. ResNet 50[M]//Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization. Berkeley: Apress, 2021: 63-72.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017: 6000-6010.

[22] 袁丁, 李源, 孟羽倩, 等. 基于时空注意力Transformer的自动驾驶运动规划方法[J]. 电子学报, 2025, 53(7): 2418-2427.
Yuan Ding, Li Yuan, Meng Yuqian, et al. A motion planning method for autonomous driving based on spatiotemporal attention Transformer[J]. Acta Electronica Sinica, 2025, 53(7): 2418-2427. (in Chinese)

[23] 钟芯, 唐春明, 彭凌西. 基于注意力融合多尺度特征的解压缩点云质量增强方法[J]. 电子学报, 2025, 53(8): 2794-2804.
Zhong Xin, Tang Chunming, Peng Lingxi. A method for enhancing the quality of decompressed point clouds based on attention-fused multi-scale features[J]. Acta Electronica Sinica, 2025, 53(8): 2794-2804. (in Chinese)

[24] Han Kai, Wang Yunhe, Chen Hanting, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 87-110.

[25] Lopez R, Regier J, Jordan M I, et al. Information constraints on auto-encoding variational Bayes[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 6117-6128.

[26] Li Junnan, Selvaraju R R, Gotmare A D, et al. Align before fuse: Vision and language representation learning with momentum distillation[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021: 742.

[27] Ridnik T, Ben-Baruch E, Zamir N, et al. Asymmetric loss for multi-label classification[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 82-91.

[28] 王炼红, 罗志辉, 林飞鹏, 等. 采用多头注意力机制的C&RM-MAKT预测算法[J]. 电子学报, 2023, 51(5): 1215-1222.
Wang Lianhong, Luo Zhihui, Lin Feipeng, et al. C&RM-MAKT prediction algorithm using multi-head attention mechanism[J]. Acta Electronica Sinica, 2023, 51(5): 1215-1222. (in Chinese)

[29] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//16th European Conference on Computer Vision. Cham: Springer, 2020: 213-229.

[30] Gong Yan, Cosma G. Improving visual-semantic embeddings by learning semantically-enhanced hard negatives for cross-modal information retrieval[J]. Pattern Recognition, 2023, 137: 109272.

[31] Rahad M, Shabab R, Ahammad M S, et al. KL-FedDis: A federated learning approach with distribution information sharing using Kullback-Leibler divergence for non-IID data[J]. Neuroscience Informatics, 2025, 5(1): 100182.

[32] Kim S, Petrunin I, Shin H S. A review of Kalman filter with artificial intelligence techniques[C]//2022 Integrated Communication, Navigation and Surveillance Conference. Piscataway: IEEE, 2022: 1-12.

[33] Stadler D, Beyerer J. BYTEv2: Associating more detection boxes under occlusion for improved multi-person tracking[C]//Proceedings of the International Conference on Pattern Recognition, Computer Vision, and Image Processing. Cham: Springer, 2023: 79-94.

[34] Liu Yingfei, Yan Junjie, Jia Fan, et al. PETRv2: A unified framework for 3D perception from multi-camera images[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 3239-3249.

[35] Ali Kamal Mohammed S, Ab Razak M Z, Abd Rahman A H. 3D-DIoU: 3D distance intersection over union for multi-object tracking in point cloud[J]. Sensors, 2023, 23(7): 3390.

[36] Zhou Pan, Xie Xingyu, Lin Zhouchen, et al. Towards understanding convergence and generalization of AdamW[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(9): 6486-6493.

[37] Sim H S, Cho H C. Enhanced DeepSORT and StrongSORT for multicattle tracking with optimized detection and re-identification[J]. IEEE Access, 2025, 13: 19353-19364.

[38] Chen Yao, Ding Shuyan, Guo Jianhui, et al. CSTrack: A comprehensive and concise vision transformer tracker[C]// 6th Chinese Conference on Pattern Recognition and Computer Vision. Singapore: Springer, 2024: 120-132.

[39] Lin Jiacheng, Chen Jiajun, Peng Kunyu, et al. EchoTrack: Auditory referring multi-object tracking for autonomous driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(11): 18964-18977.

[40] Hu Houning, Yang Y H, Fischer T, et al. Monocular quasi-dense 3D object tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1992-2008.

[41] Huang K C, Yang M H, Tsai Y H. Delving into motion-aware matching for monocular 3D object tracking[C]// 2023 IEEE/CVF international conference on computer vision. Piscataway: IEEE, 2023: 6886-6895.

作者简介

王 荣 男,2002年3月出生于山西省晋中市。现为长安大学信息工程学院硕士研究生。主要研究方向为计算机视觉、多模态融合与目标跟踪。

E-mail: wr123@chd.edu.cn

李凯飞 男,2002年12月出生于山西省吕梁市。现为长安大学信息工程学院硕士研究生。主要研究方向为计算机视觉。

E-mail: fei122813@chd.edu.cn

胡海祥 男,1998年2月出生于山东省济宁市。现为长安大学信息工程学院硕士研究生。主要研究方向为计算机视觉、目标分割与目标跟踪。

E-mail: 2024224045@chd.edu.cn

郭柯宇 男,1999年9月出生于贵州省黔南州。现为长安大学信息工程学院博士研究生。主要研究方向为计算机视觉与场景理解。

E-mail: keyuguo@chd.edu.cn

魏弘凯 男,2001年5月出生于福建省南平市。现为长安大学信息工程学院博士研究生。主要研究方向为三维计算机视觉、交通场景理解与医学图像处理。

E-mail: hongkaiwei@chd.edu.cn

宋翔宇 男,1991年3月出生于陕西省西安市。现为长安大学数据科学与人工智能研究院副教授、博士生导师。主要研究方向为交通异常事件视频语言大模型与多模态学习应用。

E-mail: xiangyu.song@chd.edu.cn

梁浩翔 男,1995年3月出生于陕西省西安市。现为长安大学电子与控制工程学院讲师、硕士生导师。主要研究方向为计算机视觉与智能交通、人工智能。

E-mail: lhx@chd.edu.cn

孙士杰 男,1989年10月出生于河南省商丘市。现为长安大学数据科学与人工智能研究院副教授、国际生博士生导师。主要研究方向为多目标检测跟踪、交通三维重建与多目标位姿估计。

E-mail: shijieSun@chd.edu.cn

钱晓伟 男,2002年4月出生于浙江省义乌市。现为长安大学信息工程学院硕士研究生。主要研究方向为三维计算机视觉。

E-mail: xiaoweiqian@chd.edu.cn