NLOT3D:单目视角下自然语言描述驱动的三维目标跟踪研究

杨 洋[1],魏弘凯[1],孙士杰[2*],宋翔宇[2],胡红利[1],郭柯宇[1],宋焕生[1]
(1. 长安大学信息工程学院,陕西西安 710064;2. 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是指通过自然语言描述引导视觉目标跟踪,通过融合文本描述和图像视觉信息,使机器能够"像人类一样"感知和理解真实的三维世界.随着深度学习的发展,自然语言描述驱动的视觉目标跟踪领域不断涌现新的方法.但现有方法大多局限于二维空间,未能充分利用三维空间的位姿信息,因此无法像人类一样自然地进行三维感知;而传统三维目标跟踪任务又依赖于昂贵的传感器,并且数据采集和处理存在局限性,这使得三维目标跟踪变得更加复杂.针对上述挑战,本文提出了单目视角下自然语言描述驱动的三维目标跟踪(Natural Language-driven Object Tracking in 3D,NLOT3D)新任务,并构建了对应的数据集NLOT3D-SPD.此外,本文还设计了一个端到端的NLOT3D-TR(Natural Language-driven Object Tracking in 3D based on Transformer)模型,该模型融合了视觉与文本的跨模态特征,在NLOT3D-SPD数据集上取得了优异的实验结果.本文为NLOT3D任务提供了全面的基准测试,并进行了对比实验与消融研究,为三维目标跟踪领域的进一步发展提供了支持.

关键词: 场景理解;三维目标跟踪;单目标跟踪;多模态学习;机器视觉
基金项目: 江西省青年基金(No. S2024QNJJL0062);长安大学中央高校基本科研业务费专项资金(No.300102244202)
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2025)06-2038-12
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20241160

NLOT3D: Natural-Language-Driven 3D Object Tracking in Monocular View

YANG Yang[1], WEI Hong-kai[1], SUN Shi-jie[2*], SONG Xiang-yu[2], HU Hong-li[1], GUO Ke-yu[1], SONG Huan-sheng[1]
(1. *School of Information Engineering*, *Chang'an University*, *Xi'an*, *Shaanxi* 710064, *China*;
2. *School of Data Science and Artificial Intelligence*,*Chang'an University*, *Xi'an*, *Shaanxi* 710064, *China*)

Abstract: Natural language description-driven object tracking refers to guiding the visual tracking task through natural language descriptions, and fusing textual descriptions and image visual information to realize the model's perception and understanding of the world "like a human". With the development of deep learning, new methods in the field of natural language description-driven visual tracking are emerging. However, most of the existing methods are limited to two-dimensional space and fail to fully utilize the position information in three-dimensional space, and thus are unable to naturally perceive the world in three dimensions as humans do. Most of the existing 3D object tracking tasks rely on expensive sensors and have limitations in data acquisition, which makes 3D object tracking even more complicated. To address the above challenges, this paper proposes a new task of natural language-driven object tracking in 3D(NLOT3D) in monocular view and constructs the corresponding dataset, NLOT3D-SPD. In addition, this paper designs an end-to-end NLOT3D-TR(Natural Language-driven Object Tracking in 3D based on Transformer) model, which fuses visual and textual cross-modal features and achieves excellent experimental results. This paper provides a comprehensive benchmarking of the NLOT3D task with several comparative experiments and ablation studies, providing strong support for further development in the field of 3D object tracking.

收稿日期:2024-12-25;修回日期:2025-05-13;责任编辑:覃怀银
*通讯作者:孙士杰

Key words:　scene understanding; 3D object tracking; single object tracking; multimodal learning; machine vision
Foundation Item(s):　Jiangxi Provincial Youth Fund Project (No.S2024QNJJL0062); Fundamental Research Funds for the Central Universities, Chang'an University (No.300102244202)

1　引言

目标跟踪是人类通过视觉系统识别并持续追踪物体的基本能力. 近年来,越来越多的研究致力于开发能够模拟人类行为的跟踪模型,其中单目标跟踪作为机器视觉的基础任务之一,旨在通过连续的视频帧自动识别并追踪特定物体. 然而,单模态数据的局限性使得传统单目标跟踪方法在场景感知上面临挑战,尤其在复杂和动态的环境中,难以应对目标外观变化、遮挡及运动模糊等问题. 为克服这些挑战,自然语言描述驱动的目标跟踪(Natural Language-driven Object Tracking,NLOT)应运而生. 通过引入自然语言文本模态,NLOT方法能够为模型提供额外的自然语言语义信息,帮助模型实现更加自然的感知,使其更接近人类跟踪物体的方式.

当前的NLOT技术大多局限于二维空间,尚未充分实现三维物体的感知与追踪,也未能有效模拟人类在复杂三维环境中的自然感知能力. 虽然三维目标跟踪技术已有一定的进展,但大多数方法仍依赖于激光雷达、毫米波雷达和深度摄像头等多种传感器[1],这与人类主要通过视觉和自然语言描述来感知与追踪物体的自然方式存在较大差距.

为了解决上述问题,本文提出了一个全新的任务:单目视角下自然语言描述驱动的三维目标跟踪(Natural Language-driven Object Tracking in 3D,NLOT3D). 该任务旨在通过自然语言文本提示对单目视频中的目标进行三维跟踪. 如图1(*a*)所示,NLOT3D通过自然语言描述对视频中目标车辆进行三维跟踪,使用一段文本描述对目标车辆进行指导跟踪,同时摆脱对深度相机、雷达等额外传感器的依赖.

单目标三维跟踪(Single Object Tracking in 3D,SOT3D)任务需要通过额外的深度图或点云信息进行三维跟踪,如图1(*b*)所示,这些模型需要额外输入深度信息帮助模型进行3D定位,且无法实现基于语义指导的3D跟踪. 自然语言引导的二维目标跟踪(Natural Language-driven Object Tracking in 2D,NLOT2D)通过自然语言描述对物体进行二维跟踪,如图1(*c*)所示,只能通过简单语句指导进行二维跟踪,无法实现三维跟踪. 3种跟踪任务的区别如表1所示.

表1　不同跟踪任务差异对比

任务	模型输入	跟踪类型	文本描述情况
SOT3D	深度信息+视频序列	3D	无文本描述
NLOT2D	文本描述+视频序列	2D	初始简单描述
NLOT3D	文本描述+视频序列	3D	完整过程描述

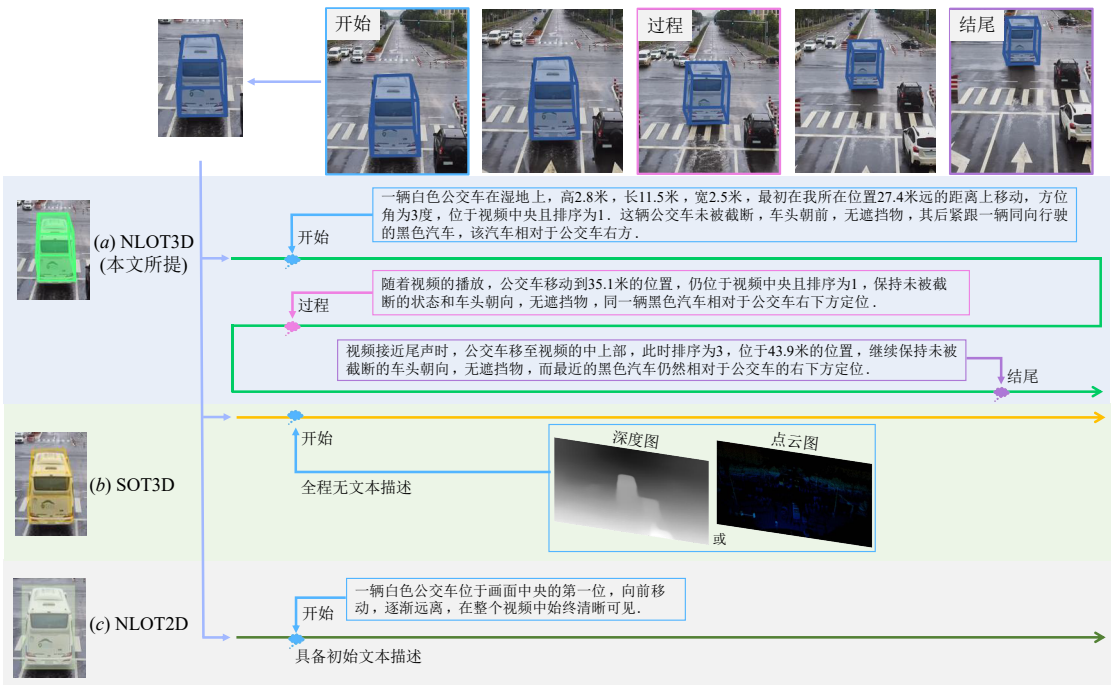


图1　NLOT3D任务与其他目标跟踪任务对比图

为此,本文基于V2X-SPD数据集,构建了一个大规模的NLOT3D-SPD数据集,包含36 956条自然语言描述,用于对视频中的单目标进行三维跟踪.这些自然语言描述由ChatGLM生成,经过人工精细调整以保证准确性.该任务结合了视觉信息和自然语言描述,促使机器能够在单目视频中理解三维物体的空间信息,使机器在三维目标跟踪领域实现了更好的人机交互.

此外,本文针对该任务提出首个端到端的网络模型NLOT3D-TR.模型构架包括多模态特征提取器、自然语言描述驱动的编码器、解码器及跟踪头,模型通过像素级注意力机制进行自然语言描述引导的特征学习,然后由视觉、文本和深度信息融合注意力机制进行特征融合,实现不同模态特征高效整合,最终通过跟踪头实现三维跟踪任务.模型NLOT3D-TR在NLOT3D-SPD数据集上实现了全面基准测试和对比实验,验证了模型在该任务中的有效性,进一步推动了复杂场景下三维目标跟踪技术的发展.

本文的主要贡献如下:

(1)提出了新任务NLOT3D,用于在单目视频中通过自然语言描述实现单目标的三维视觉语言跟踪.

(2)基于V2X-SPD数据集构建了大规模NLOT3D-SPD数据集,包含36 956条单目标跟踪的自然语言描述和视频序列.

(3)设计并实现了端到端的网络模型NLOT3D-TR,该模型在多模态框架中有效地融合了语言、视觉和深度几何特征.

(4)提供了NLOT3D任务的系统基准测试,并通过详细实验验证了NLOT-TR模型在各项基准测试中的显著优势.

2 相关工作

2.1 目标跟踪方法及发展

目标跟踪[2]作为机器视觉领域重要的研究方向,近年来不断取得发展[3].其中,二维目标跟踪主要关注目标在二维平面中的运动和位置变化[4,5].Nam等人[6]提出的MDNet(Multi-Domain Network)通过在多个数据集上进行预训练,利用深度卷积神经网络(Convolutional Neural Network,CNN)提取特征,并通过多个任务(例如分类和回归)来实现高效的目标跟踪.SiamRPN[7]结合了孪生网络(siamese network)和区域提议网络(Region Proposal Network,RPN),通过共享卷积特征提取器计算模板和搜索区域的相似性,并利用回归方法预测目标的位置.Li等人[8]对其进行改进得到SiamRPN++,通过引入优化的特征提取模块和网络结构,解决了目标尺度变化和遮挡问题,并加入了注意力机制来增强目标的匹配能力使得模型整体表现更加稳定.

尽管二维目标跟踪取得了很大的进展,但其依然存在一定的局限性,在处理复杂场景时,二维跟踪方法无法准确捕捉目标的深度信息和三维姿态.

2.2 自然语言描述引导的多模态学习

多模态技术近年来发展迅速[9],其中语义指导任务[10,11]作为其重要应用之一,已成为广泛研究的热点[12].DALL-E模型通过文本语义指导生成图像[13],利用多模态融合技术,根据文本描述生成相应图像,展示了文本与视觉信息的深度结合.CLIP(Contrastive Language-Image Pre-training)模型[14]则通过联合训练图像和文本,使得模型能够理解并关联视觉内容与文本描述,推动了多模态技术在跨模态检索、图像生成以及图像描述等任务中的广泛应用.在医学图像分析领域,樊琳等人[15]提出的基于文本引导的多模态医学图像分析框架,通过多阶段诊断文本引导模型学习,有效提取图像中的关键诊断信息特征.该框架利用交叉模态注意力机制,促进了图像与文本特征之间的深度交互,实现了文本指导的医学图像检测.在图像生成领域,Xia等人[16,17]提出的TediGAN模型引入视觉语言相似性,将图像和文本映射到共同的嵌入空间,学习图像与文本之间的匹配关系,最终生成高质量图像.该模型不仅提升了图像生成的质量,还增强了图像与文本之间的互联性与一致性,为生成式任务带来了新的突破.

语义指导的多模态任务在多个领域取得了显著进展[18,19],但在交通视觉领域的应用研究相对较少.将语义指导引入交通视觉任务,能够显著提升目标检测、跟踪与行为分析的准确性和鲁棒性.通过结合文本描述与交通场景中的视觉信息,模型能够更好地理解复杂交通环境中目标的行为和交互.

2.3 自然语言描述引导的目标检测

将多模态技术与目标跟踪技术相结合,实现语义描述指导的视觉跟踪任务.目标检测作为目标跟踪的基础,为跟踪任务提供了重要支持.Liu等人[20]提出的Grounding DINO模型,使用特征提取网络提取图像特征,并将其与文本特征进行跨模态融合[21];通过采用语义指导的查询选择模块[22],网络能够从图像特征中选择相关的跨模态查询,最终通过解码器输出进行边界框预测[23,24],实现基于文本描述的开放集目标检测.该模型通过图像区域与文本描述对齐,显著提升了目标识别与定位的准确性.Yang等人[25]提出的Mono3DVG模型,融合文本与视觉多模态特征[26],提出双文本引导适配器,执行像素级文本引导特征学习;该模型采用深度文本视觉堆叠注意力机制,将对象级深度几何特征与视觉外观特征融合,最终实现基于文本描述的物体三维检测[27,28].

3　数据集构建

传统的视觉语言跟踪任务局限于二维空间,未能充分利用图像中的空间位置信息,且二维跟踪任务获得二维边界框所提供的信息有限. 对于自动驾驶、机器人操作和军事应用等实际场景,通常需要更高精度的三维物体边界框. 而现有的三维目标跟踪方法往往依赖昂贵的专业设备来采集数据,这大大增加了三维跟踪任务的成本和复杂性.

针对上述挑战,本文提出了单目视角下自然语言描述驱动的三维目标跟踪(NLOT3D)任务,这是一种异于传统思路的三维视觉跟踪任务. 为推进该任务,本文构建了单目视角下自然语言描述驱动的三维目标跟踪数据集(NLOT3D-SPD),并提出了相应的基准评估体系. 所提出的数据集基于车路协同时序感知数据集(DAIR-V2X-SPD)[29,30],该数据集来源于真实场景中的车路协同感知数据,包含了完整的三维目标框和跟踪ID标注,主要用于车路协同中的三维目标检测和跟踪任务. 这为单目视角下自然语言描述驱动的三维目标跟踪任务研究提供了宝贵的资源,推动了该领域的进一步发展.

本文对数据集进行了筛选和处理,为每个不同场景和车辆提供了相应的视频描述. 对于每个视频序列,从中抽取开始帧、中间帧和结束帧,以获取车辆的综合状态,包括是否行驶、是否停车、拐弯方向以及距离变化等信息. 通过这种方式,确保描述覆盖视频片段中的全部信息,全面反映车辆在视频序列中的运动状态和特征. 对于长时间被遮挡的目标车辆,由于生成的视频描述无法覆盖整个视频序列,相关样本被过滤. 对于短时间被遮挡的车辆,保留这些样本,但确保其在数据集中的占比不超过5%,以此保障数据集的多样性和完整性,为三维视觉语言驱动的目标跟踪任务提供了坚实的基础.

以往自然语言描述驱动的视觉目标跟踪数据集通常依赖人工标注和描述,这种方式成本高昂且不适用于大规模数据集的制作. 为此,本文采用了一种新的方法生成数据集描述. 受到大语言模型的启发,本文采用ChatGLM[31]按照模板生成对应场景的准确描述. 数据集的生成过程如图2所示.

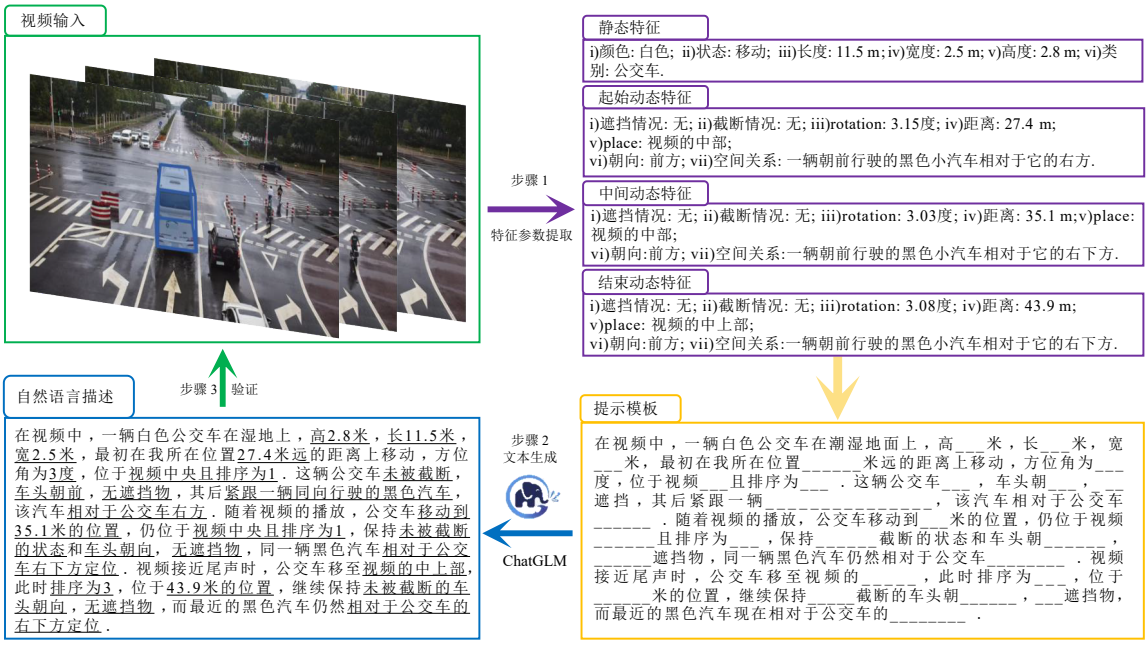


图2　采用ChatGLM进行数据集生成过程

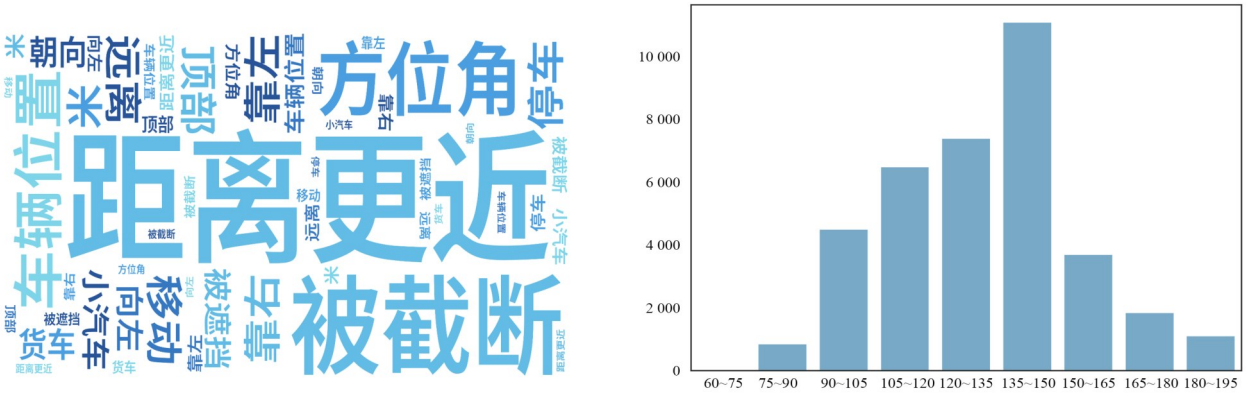
首先,对数据进行筛选和清理,提取所有车辆目标物体的属性信息,包括动态属性与静态属性. 静态属性包括车辆颜色、状态、长宽高和车辆类型,而动态属性则包括车辆的截断、遮挡、方位角、和空间位置关系等信息. 其中截断、遮挡、方位角和长宽高均来源于V2X-SPD原有的标签,其他属性则通过Python代码程序进行数学运算得到对应的属性值,如物体状态属性通过计算车辆位置是否随时间变化而得到,方向和空间位置关系通过计算距离和方位角等信息获得.

对于具体的文本描述,本文采用ChatGLM来生成模板对应的描述,将每个目标物体的属性填入模板中并生成完整的中文自然语言描述. 随后对每个描述信息进行核查以及纠正,经过筛查,最终得到36 956个视频序列及对应的自然语言描述. 对视频描述进行词云可视化展示,如图3(a)所示;此外,本文还统计了视频序列对应描述的字数长度分布,如图3(b)所示,视频描

述字数长度大致在105~165字.

4 方法

对于NLOT3D任务,本研究提出了一种端到端的解决方法(NLOT3D-TR),在数据集NLOT3D-SPD数据集上有着良好的基线表现.NLOT3D-TR网络结构图如图4所示,模型包括多模态特征提取、自然语言描述驱动的编码、解码及跟踪输出头模块.



(a) 视频序列描述关键词词云 (b) 视频序列描述长度分布

图3 视频描述属性统计

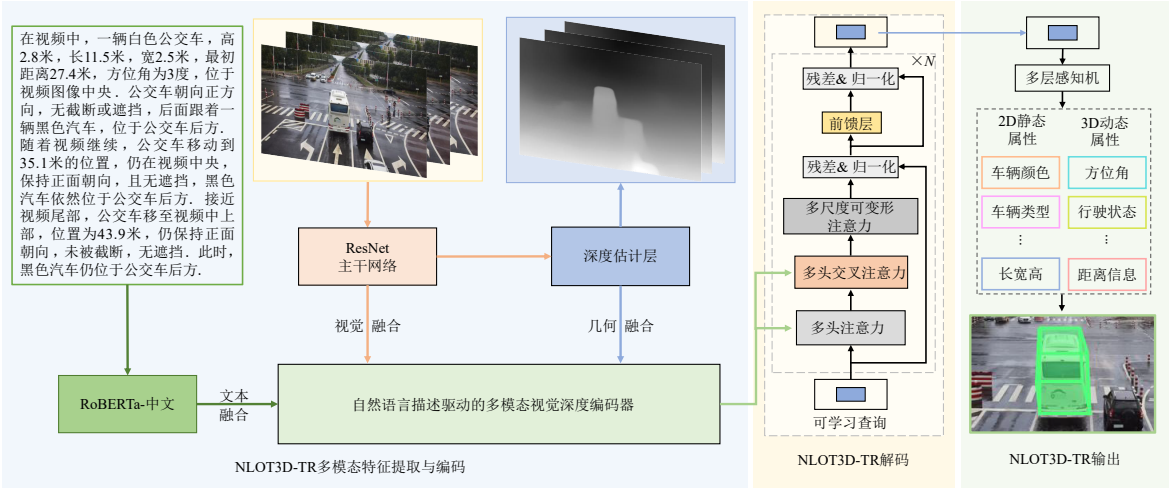


图4 NLOT3D网络结构图

4.1 多模态特征提取器

特征提取器由文本特征提取器和图像特征提取器组成.对于描述文本,采用预训练的RoBERTa for Chinese[32]来进行文本编码,得到文本特征$\boldsymbol{f}_l^t\in\mathbb{R}^{C\times N_l^t}$,其中,$N_l^t$表示输入的文本,$\boldsymbol{f}_l^t$表示经过编码得到的文本特征.对于视频序列每一帧的图像$\boldsymbol{I}^t\in\mathbb{R}^{H\times W\times 3}$,采用ResNet[33]来进行多层次多尺度特征提取,然后通过线性层Linear对维度进行统一,最终得到视频帧特征$\boldsymbol{f}_v^t\in\mathbb{R}^{C\times N_v^t}$,其中$C=256$,表示特征维度,而$N_v^t=HW/8^2+HW/16^2+HW/32^2+HW/64^2$.此外,本文还利用了深度估计层来对ResNet提取到的特征进行深度估计,以弥补缺失的3D深度信息.因此模型在训练阶段需要深度信息来进行监督,使深度预测器学习到如何从视觉特征得到相对深度信息,帮助模型实现更好的3D跟踪定位.在部署推理阶段,只使用视觉和文本信息进行推理,不再使用深度信息.深度估计层采用了轻量级的深度估计网络,得到的深度特征为$\boldsymbol{f}_d^t\in\mathbb{R}^{C\times N_d^t}$,其中$N_d^t=HW/16^2$.多模态特征提取过程如式(1)所示:

$$\begin{aligned}\boldsymbol{f}_l^t&=\mathrm{Linear}\left(\mathrm{RoBERTa\ for\ Chinese}(N_l^t)\right)\\ \boldsymbol{f}_v^t&=\mathrm{Linear}\left(\mathrm{ResNet}(N_v^t)\right)\\ \boldsymbol{f}_d^t&=\mathrm{DepthPredictor}(\boldsymbol{f}_v^t)\end{aligned}\tag{1}$$

4.2 自然语言描述驱动的视觉深度编码器

本文提出了一种融合视觉、深度几何和文本信息的编码器.经过提取的图像特征首先经过自然语言描述驱动的视觉编码模块进行特征融合,随后通过自然语言描述驱动的深度编码模块进行进一步加强.

自然语言描述驱动的视觉深度编码器的结构如图5所示,本文采用多尺度可变形注意力机制(Multi Scale Deformable Attention,MSDA)代替了原有的多头自注意力机制(Multi Head Self Attention,MHSA),减小了多尺度视觉特征注意力机制的计算成本.此外,在MSDA层和前馈神经网络(Feed Forward Network,FFN)之间加入了一个多头交叉注意力层(Multi Head Cross Attention,MHCA),通过引入MHCA来实现模态特征融合,增强了图像目标物体和自然语言描述的关联度[34],得到自然语言描述驱动,更好地指引视觉三维跟踪.然后视觉编码模块对特征进行拆分和组合,利用大小为$HW/16^2$的视觉特征$\boldsymbol{f}_v^{t\prime}$进行查询,随后采用MSDA代替MHA,得到细化后的视觉特征$\boldsymbol{f}_v^{t\prime\prime}$.

自然语言描述驱动深度编码模块是为了充分利用自然语言描述的目标外观特征和三维空间几何关系特征,如图5所示,深度信息编码器由Transformer编码器层组成.深度特征$\boldsymbol{f}_d^t$首先通过MHSA进行自注意力计算,然后通过MHCA进行文本增强,其中Query为经过自注意力计算处理得到的深度信息特征$\boldsymbol{f}_d^{t\prime}$,Key和Value为文本特征$\boldsymbol{f}_l^t$.在此之后加入一个多头注意力层(Multi Head Attention,MHA),将自然语言描述驱动的自注意力应用于图像的深度几何信息中得到增强关联后的深度特征$\boldsymbol{f}_d^{t\prime\prime}$.

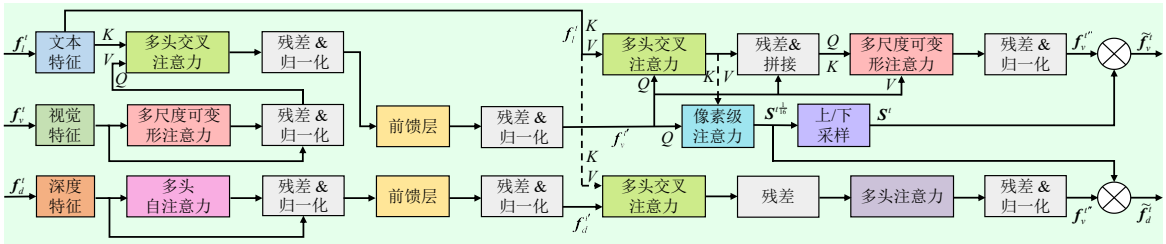


图5　自然语言描述驱动的视觉深度编码器

此后,对MHCA的输出和大小为$HW/16^2$的视觉特征$\boldsymbol{f}_v^{t\prime}$进行投影以得到文本相关特征$\boldsymbol{F}_{\text{text}}^t \in \mathbb{R}^{CHW/16^2}$和原始特征$\boldsymbol{F}_{\text{origin}}^t \in \mathbb{R}^{CHW/16^2}$,如式(2)所示;对文本相关特征和原始视觉特征进行注意力计算得到每个区域的注意力分数$\boldsymbol{s}_{ij} \in \mathbb{R}^{HW/16^2}$,以此来判定文本特征和视觉特征的对齐程度,用于判断特征对齐的注意力分数计算如式(3)所示:

$$\boldsymbol{F}_{\text{text}}^t = \left\| \boldsymbol{F}_{\text{text}}^t \right\|_2, \boldsymbol{F}_{\text{origin}}^t = \left\| \boldsymbol{F}_{\text{origin}}^t \right\|_2 \tag{2}$$

$$\begin{aligned} \boldsymbol{a}_{ij}^c &= \boldsymbol{F}_{\text{text}}^{tc}(i,j) \odot \boldsymbol{F}_{\text{origin}}^{tc}(i,j), \\ \boldsymbol{s}_{ij}^t &= \sum_{c=1}^{C} \boldsymbol{a}_{ij}^{tc} \end{aligned} \tag{3}$$

其中,$\|\cdot\|_2$是L2范数,$\odot$是元素逐项乘积.

使用高斯函数来对每个像素特征和文本特征之间的语义相似度$\boldsymbol{S}^{1/16}$进行建模,如式(4)所示,其中α为缩放因子,σ为标准差,两者均是可学习的参数.

$$\boldsymbol{S}^{t/16} = \boldsymbol{\alpha} \cdot \exp\left(-\frac{(1-\boldsymbol{s}_{ij}^t)^2}{2\sigma^2} \right) \tag{4}$$

得到语义相似度$\boldsymbol{S}^{t/16}$后通过双线性插值法来对其进行采样,再通过最大池化层进行下采样,再将其进行拼接,得到综合的多尺度注意力得分$\boldsymbol{S}^t \in \mathbb{R}^{N_v}$,计算如式(5)所示:

$$\boldsymbol{S}^t = \text{Concat}\left[\text{Up}(\boldsymbol{S}^{t/16}), \boldsymbol{S}^{t/16}, \text{Down}(\boldsymbol{S}^{t/16}) \right] \tag{5}$$

通过上述逐像素的多尺度注意力得分,将视觉外观特征和深度几何特征关联到文本描述相关的区域,然后将深度特征$\boldsymbol{f}_d^{t\prime\prime}$和视觉特征$\boldsymbol{f}_v^{t\prime\prime}$与注意力得分进行逐元素相乘,获得文本引导的目标特征,如式(6)所示.这些特征能更精确地聚焦于与文本描述高度相关的区域,最终提升模型的目标跟踪精度.

$$\widetilde{\boldsymbol{f}}_v^t = \boldsymbol{f}_v^{t\prime\prime} \cdot \boldsymbol{S}^t, \widetilde{\boldsymbol{f}}_d^t = \boldsymbol{f}_d^{t\prime\prime} \cdot \boldsymbol{S}^{t/16} \tag{6}$$

4.3　自然语言描述驱动的解码器

自然语言描述驱动的解码器负责对编码器编码的特征进行进一步处理,通过结合自然语言描述驱动的视觉特征和几何特征,准确预测目标的几何和外观属性.利用深度、文本、视觉信息的堆叠注意力机制,自适应地融合目标的几何深度特征和视觉特征,最终实现自然语言描述驱动的三维目标跟踪.

自然语言描述驱动的解码器包括N层解码器层,每层解码器层由MHA、MHCA和MSDA组成.如图4中的解码器所示,可学习查询$\tilde{\boldsymbol{Q}}^t \in \mathbb{R}^{C\times 1}$聚合初始几何信息和文本增强的视觉特征,然后将精确的外观视觉特征和几何特征融合到可学习的查询$\tilde{\boldsymbol{Q}}^t$中.

模型的跟踪输出头通过利用多个多层感知机来实现预测视频每一帧的属性.解码器的输出$\tilde{\boldsymbol{Q}}^t \in \mathbb{R}^{C\times 1}$,表示不同帧的可学习的查询.通过三层感知机来预测物体类别,估算2D边界框(l,r,b,t)和3D边界框的2D投影中心坐标$(x_{3\text{D}}, y_{3\text{D}})$,其中$(l,r,b,t)$表示了2D投影中心点距离目标物体边界框左侧、右侧、底部和顶部之间的距离.同时使用双层感知机来预测3D框的尺寸、方向和

深度($h_{3D}, w_{3D}, l_{3D}, \theta, d_{reg}$),其中$h_{3D}$、$w_{3D}$、$l_{3D}$表示3D框的高宽长,$\theta$表示物体的旋转角,$d_{reg}$表示深度信息.

在损失函数的设计中,误差分为类别误差、2D平面属性误差和3D空间属性误差. 类别误差包括车辆颜色误差和车辆类型误差,如式(7)所示,采用多分类交叉熵损失进行计算. 2D属性误差包括2D框的大小和3D投影的2D中心点,2D框的损失通过GIOU计算,3D中心点的2D投影和中心点到4个边的距离通过L1损失来优化. 总体损失公式如式(8)所示. 3D属性包括3D框的尺寸、方向和深度,3D损失如式(9)所示,L_{size3D}采用了3D IoU定向损失,L_{orient}采用MultiBin Loss作为损失计算函数,L_{depth}使用了拉普拉斯任意不确定性损失(laplacian aleatoric uncertainty loss).

$$L_{class}=\alpha_1 L_{ColorCls}+\alpha_2 L_{CarCls} \tag{7}$$

$$L_{2D}=\lambda_1 L_{class}+\lambda_2 L_{GIoU}+\lambda_3 L_{3Dxy}+\lambda_4 L_{lrtb} \tag{8}$$

$$L_{3D}=L_{size3D}+L_{depth}+L_{orient} \tag{9}$$

此外,本文使用了额外的Focal Loss来作为深度图的预测损失$L_{depthMap}$,总体的损失L_{all}如式(10)所示:

$$L_{all}=L_{2D}+L_{3D}+\beta L_{depthMap} \tag{10}$$

5 实验

5.1 实验配置

NLOT3D任务对应的数据集共包括了36 956条文本描述和对应的视频序列,按照7:1:2的数据集划分数据集,得到训练集25 869条视频序列,验证集和测试集分别为3 696和7 391条视频序列. 实验中的学习率和权重衰减均设置为1×10^{-4},优化器选择AdamW. 实验在2张Tesla V100 32 GB GPU上进行,训练周期为30个epoch.

5.2 评估指标

实验采用成功率(Success Rate,SR)和精确率(Precision Ratio,PR)作为评估指标,其中SR@IOU表示在不同交并比(Intersection over Union,IoU)阈值下,达到该阈值的帧的数量占总帧数的百分比. 平均重叠率(Average Overlap Rate,AOR)用于衡量所有帧的3D边界框预测值与真实值之间的IoU重叠程度,反映了模型在3D目标定位上的准确性. PR@d表示所有帧中预测框与真实框中心点距离小于等于d的帧数占总帧数的百分比,测量模型在不同空间误差范围内的跟踪表现. 平均中心误差(Average Center Error,ACE)用于衡量预测的3D边界框中心与实际边界框中心之间的平均距离,反映了模型3D目标中心定位上的精度. 通过这些评估指标,能够全面地评估模型在3D视觉语言跟踪任务中的表现.

5.3 实验基准

为了全面评估本研究提出的NLOT3D任务,本文设计了一系列的基准实验,并以标准化的评估方法对各类参考方法的有效性进行评估. 基准实验中,使用多种基于文本指导的2D视觉感知方法,比如ZSGNet[35],FAOA[36],ReSC[37]以及TransVG[38]方法,将其中的2D视觉感知模型通过逆投影的方法与目标跟踪算法进行结合以实现文本指导的3D单目标跟踪. 此外,模型Mono3DVG[25]作为3D文本感知识别算法,其结合跟踪算法同样可以实现基于文本指导的单目标跟踪. 将这些基线同本文提出的NLOT3D-TR进行比较参考,分析其他模型和本文模型的性能差异,证明本文提出模型的有效性.

5.4 结果分析

在NLOT3D-SPD数据集上进行的实验测试结果如表2所示. 从实验结果可以看出,NLOT3D-TR的AOR为66.39%,且在所有IoU阈值下的SR均高于基线模型,证明了该模型在性能上优于其他对比模型. 这些指标表明,NLOT3D-TR能够有效预测目标的3D边界框. 此外,模型的ACE指标为0.413,保持较低的误差距离,表明其对目标3D中心点的预测准确. PR值在各个阈值下的表现也优于其他方法,进一步验证了模型的优势.

表2 基线实验对比结果

方法	SR@0.5/% ↑	SR@0.9/% ↑	AOR/% ↑	PR@1.0/% ↑	PR@0.5/% ↑	ACE ↓
ZSGNet + backproj[35]	36.64	21.69	39.93	42.69	30.58	1.137
FAOA + backproj[36]	34.40	20.22	40.53	41.28	30.82	1.072
ReSC + backproj[37]	45.09	39.89	50.24	47.56	35.75	0.762
TransVG + backproj[38]	43.23	37.05	47.87	47.43	34.84	0.796
Mono3DVG[25]	<u>55.81</u>	<u>42.45</u>	<u>60.88</u>	<u>59.23</u>	<u>47.67</u>	<u>0.491</u>
NLOT3D-TR	**61.56**(+5.75)	**47.53**(+5.08)	**66.39**(+5.51)	**62.48**(+3.25)	**52.26**(+4.59)	**0.413**(-0.078)

注:↑箭头表示值越高模型性能越佳,↓箭头则表示值越低模型性能越佳.加粗表示性能最佳,下划线表示次之,括号内表示本文模型相比次优方法指标提升幅度.

与本文方法相比,基于2D视觉感知与逆投影技术的方法在SR和PR指标上表现均比较差. 这些方法将2D跟踪结果逆投影到3D边界框,使得最终预测结果极大依赖于2D预测的准确性. 同时,逆投影过程也会导致3D边界框的不准确定位,进一步影响整体性能,导致效果不佳. 而Mono3DVG直接预测目标3D边界

框,不再受到2D逆投影的固有限制,其性能得到了提升.

相比之下,NLOT3D-TR直接预测3D跟踪边界框的位置,并且结合了深度信息编码器,利用深度推理模型得到图像的深度信息进一步提升了3D预测的准确性.这些额外的深度信息使得NLOT3D-TR在3D定位任务上展现出显著的优势.同时,NLOT3D-TR通过多模态融合机制整合了文本描述、视频图像及图像深度几何信息,这些丰富的特征输入使得模型在跟踪精度和鲁棒性方面相比其他方法具有明显的提升.

5.5　可视化

在NLOT3D任务中,NLOT3D-TR模型展现了显著的性能优势.为了更直观地展示模型的跟踪效果,本文使用OpenCV可视化结果并进一步分析模型在不同场景下的表现.通过可视化,不仅能够直观地观察到模型在目标跟踪过程中的精准度和鲁棒性,还可以深入理解模型在3D空间中的决策过程及其对不同输入信息的响应,可视化对比如图6所示.

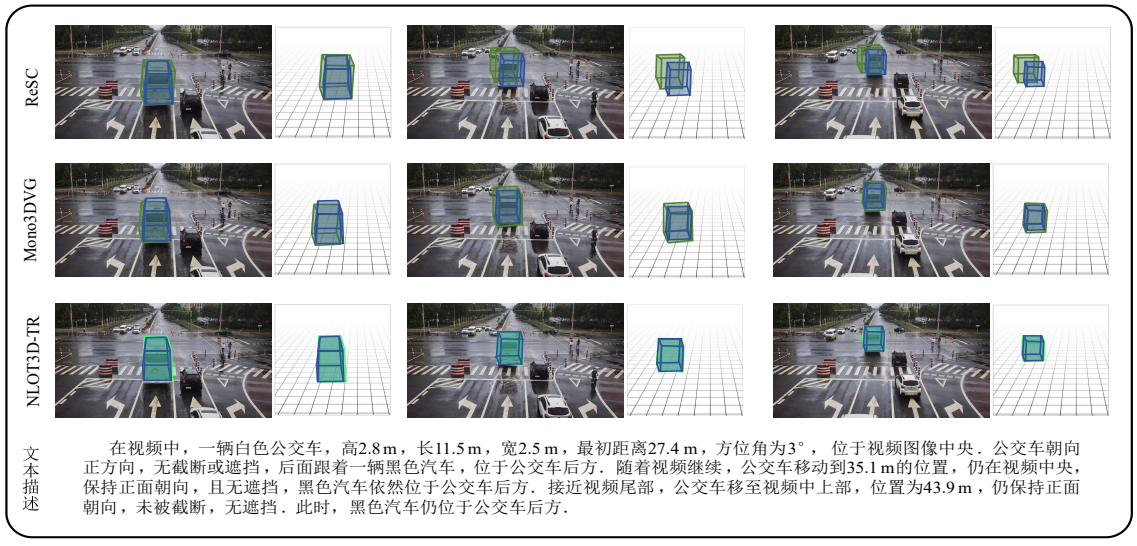


图6　基线方法可视化展示(蓝色为实际边界框,绿色为模型预测框)

图6中展示了不同基线模型与NLOT3D-TR的可视化效果.可视化结果表明NLOT3D-TR在3D跟踪任务中的表现优于其他模型.无论是边界框的大小、3D框的中心位置,还是目标框的旋转角度,本文的模型预测的结果都与实际边界框信息接近.与此相比,基线方法ReSC结合逆投影的方法在视频帧的初始阶段能够精确地定位目标物体,但在跟踪的中后期,预测的边界框位置出现较大偏差,逐渐偏离了实际的目标位置,这表明该模型在长期跟踪任务中的稳定性较差,难以应对长时间序列的跟踪挑战.方法Mono3DVG的表现优于方法ReSC,能够持续跟踪目标物体,且在整个跟踪过程中,预测的中心点与实际中心点的差距较小,但该模型在边界框的大小上存在预测不准确的情况,未能完全反映目标物体的实际尺度.

相比之下,NLOT3D-TR在整个跟踪过程中表现出更高的精度.本文的模型不仅能够准确地跟踪目标的中心点,而且在目标物体的大小预测上也展现了较高的精度,保证了不同场景下跟踪任务的稳定性和准确性.

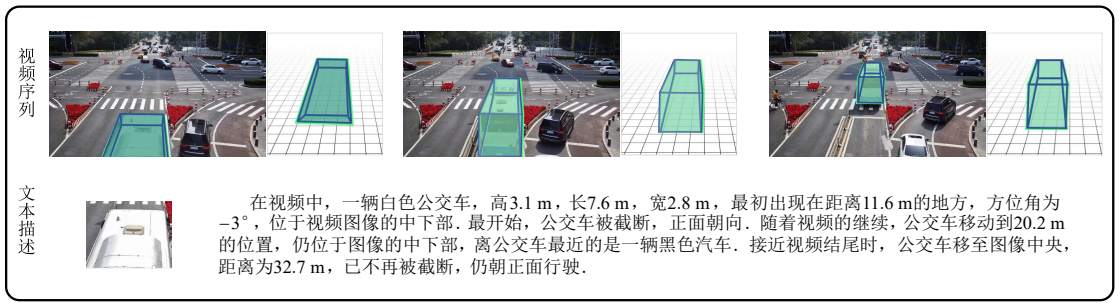
在更多的数据上进行测试,测试可视化结果如图7所示,其中图7(a)展示了复杂场景下大型公交车在初始状态被截断场景下跟踪可视化效果,其跟踪效果良好.但本文在其他数据可视化时观察到初始被截断的小目标车辆在跟踪后期无法得到准确的车辆三维边界框.图7(b)中展示了复杂场景下大型货运车辆跟踪可视化效果,图7(c)中展示了复杂场景下小型车辆的跟踪可视化效果.

5.6　消融实验

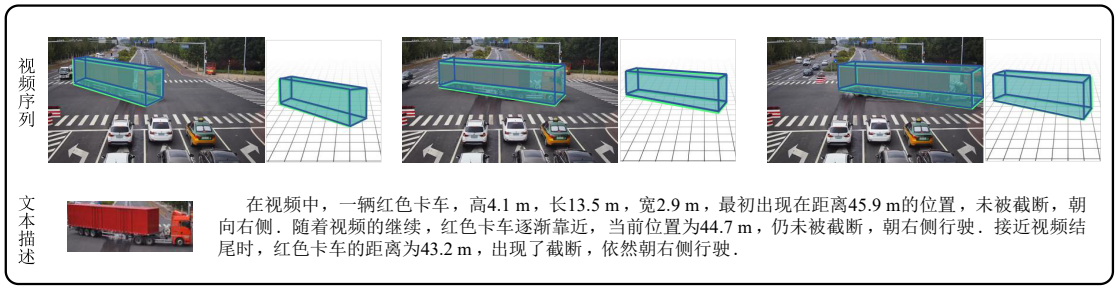
为了进一步验证NLOT3D-TR模型各个组件对整体性能的贡献,本文设计了消融实验,系统评估了各部分对单目标三维跟踪效果的影响.通过调整不同的编码器配置(例如自然语言描述驱动的视觉深度编码器层数设置),以及自然语言描述驱动的解码器模态堆叠顺序进行相关实验,进一步证明了这些组件在提升模型性能中的有效性.

(1)自然语言描述驱动的视觉深度编码器配置分析

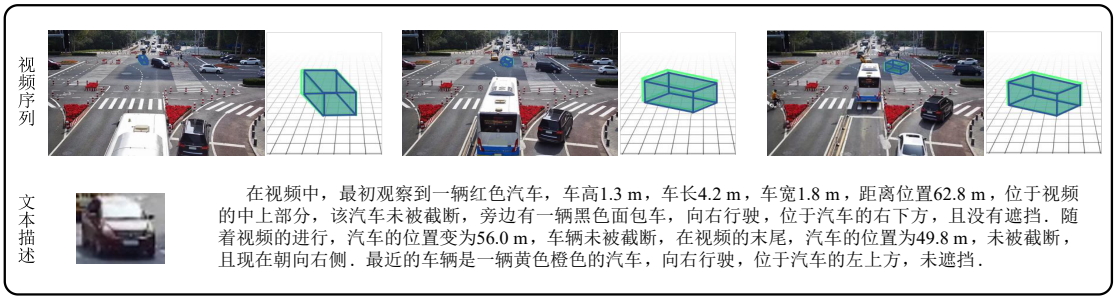
在自然语言描述驱动视觉深度编码器中,视觉特征和深度特征通过自然语言描述进行引导和增强,这有效捕捉了多模态视觉和深度几何信息.为了评估不同



(*a*) 大型截断车辆场景跟踪



(*b*) 大型未被截断车辆场景跟踪



(*c*) 小型车辆复杂场景跟踪

图7 模型NLOT3D-TR的多场景可视化展示

编码模块层数对模型跟踪性能的影响，本文设计了多种编码模块配置并进行了消融实验，实验结果如表3所示．结果表明，当视觉编码模块为2层、深度编码模块为1层时，模型在各项指标上达到了最优性能．这表明，适当的视觉深度编码器层数能够更有效地提取与自然语言描述相关的视觉和深度几何特征，更好地驱动多模态信息的融合，为下游跟踪任务提供更精准的指导．

(2)解码器模态堆叠分析

解码器通过对编码器输出的特征进行进一步处理，且不同模态特征的堆叠顺序对跟踪任务的性能具有显著影响．本文尝试了将文本特征(T)、视觉特征(V)和深度特征(D)按不同顺序进行堆叠，并进行了相关实验，结果如表4所示．实验表明，D-T-V的堆叠顺序实现了最佳性能．这表明，将深度和文本特征的注意力模块置于视觉特征处理之前，有助于更有效地融合多模态信息，提升模型的整体性能．

表3 不同视觉深度编码器层数配置性能对比结果

模块	层数	SR@0.5/% ↑	AOR/% ↑	PR@1.0/% ↑	ACE ↓
自然语言描述驱动的视觉编码模块	*L*=1	58.45	63.22	60.03	0.562
	***L*=2**	**61.56**	**66.39**	**62.48**	**0.413**
	L=3	59.12	63.62	60.29	0.545
自然语言描述驱动的深度编码模块	***D*=1**	**61.56**	**66.39**	**62.48**	**0.413**
	D=2	57.23	61.94	59.32	0.586
	D=3	56.94	59.82	58.06	0.591

注：↑箭头表示值越高模型性能越佳，↓箭头则表示值越低模型性能越佳，加粗字体表示对应参数模型的各项指标最佳．

6 结论

如何让机器像人类一样感知真实的三维复杂场景，是人工智能领域一个具有挑战性的问题．针对这一问题，本文从人类通过视觉系统与自然语言描述相结合的方式理解并感知三维世界的机制出发，提出了单目视角下自然语言描述驱动的三维目标跟踪(NLOT3D)任务，

表4　不同模态堆叠次序对跟踪性能影响

解码器堆叠顺序	SR@0.5/% ↑	AOR/% ↑	PR@1.0/% ↑	ACE ↓
T→D→V	58.23	63.67	59.41	0.563
D→T→V	**61.56**	**66.39**	**62.48**	**0.413**
D→V→T	59.64	63.58	60.64	0.541

注:↑箭头表示值越高模型性能越佳,↓箭头则表示值越低模型性能越佳,加粗字体表示对应解码器堆叠顺序的模型各项指标最佳.

该任务通过融合自然语言描述文本信息与视觉信息,实现精准的单目标三维跟踪. 为了支撑这一任务,本文还提出了大规模单目视角下自然语言描述驱动的三维目标跟踪数据集(NLOT3D-SPD),为该任务提供了真实场景下多模态视觉与文本数据,确保了任务的可行性. 此外,本文还提出了NLOT3D-TR模型,并通过一系列实验评估了其在实际跟踪中的性能. 实验结果表明,NLOT3D-TR模型在精度和鲁棒性方面优于现有模型,证明了其在多模态三维目标跟踪领域的优越性能.

尽管本文构建了NLOT3D任务的数据集,提出对应模型并进行基准测试,但这些工作仍存在一定局限性. 首先NLOT3D-SPD数据集场景有限,主要包含室外白天场景,缺乏雨雪天气和夜晚等特殊场景数据,这会使得网络模型泛化性不足,无法用于特殊场景. 此外,NLOT3D-TR模型处理视频帧中被截断的小目标车辆时无法保证预测3D框的准确性,且模型在长视频序列的处理上也需进一步优化. 下一步的工作将重点针对这些问题,扩充更加多样化的场景数据集,改进模型特征提取和时序建模能力,以提高模型在复杂场景中的鲁棒性和准确性,推动三维感知技术的发展.

以人类的视觉感知和跟踪能力作为评估机器智能的标准,将推动机器视觉技术向更接近人类真实水平的方向发展. NLOT3D任务的提出能够推动多模态感知技术的进一步发展,为机器视觉、具身智能及自动驾驶等前沿领域提供新的研究方向与技术解决方案. 此外,这一任务的实现也将促进人工智能技术在实际应用中的跨越式进步,为相关领域的创新和发展注入新的动力.

参考文献

[1] 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 J, JIANG B T, PENG W, 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)

[2] 孟琭, 杨旭. 目标跟踪算法综述[J]. 自动化学报, 2019, 45(7): 1244-1260.
MENG L, YANG X. A survey of object tracking algorithms[J]. Acta Automatica Sinica, 2019, 45(7): 1244-1260. (in Chinese)

[3] LUO W H, XING J L, MILAN A, et al. Multiple object tracking: A literature review[J]. Artificial Intelligence, 2021, 293: 103448.

[4] MARVASTI-ZADEH S M, CHENG L, GHANEI-YAKHDAN H, et al. Deep learning for visual tracking: A comprehensive survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(5): 3943-3968.

[5] JIAO L C, WANG D, BAI Y D, et al. Deep learning in visual tracking: A review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(9): 5497-5516.

[6] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 4293-4302.

[7] LI B, YAN J J, WU W, et al. High performance visual tracking with Siamese Region proposal network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8971-8980.

[8] LI B, WU W, WANG Q, et al. SiamRPN++: Evolution of Siamese visual tracking with very deep networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 4282-4291.

[9] 何俊, 张彩庆, 李小珍, 等. 面向深度学习的多模态融合技术研究综述[J]. 计算机工程, 2020, 46(5): 1-11.
HE J, ZHANG C Q, LI X Z, et al. Survey of research on multimodal fusion technology for deep learning[J]. Computer Engineering, 2020, 46(5): 1-11. (in Chinese)

[10] 郭宗洋, 刘立东, 蒋东华, 等. 基于语义引导神经网络的人体动作识别算法[J]. 图学学报, 2024, 45(1): 26-34.
GUO Z Y, LIU L D, JIANG D H, et al. Human action recognition algorithm based on semantics guided neural networks[J]. Journal of Graphics, 2024, 45(1): 26-34. (in Chinese)

[11] KIM G, KWON T, YE J C. DiffusionCLIP: Text-guided diffusion models for robust image manipulation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 2416-2425.

[12] 陆庆阳, 袁广林, 朱虹, 等. 一种基于对比学习大模型的视觉定位方法[J]. 电子学报, 2024, 52(10): 3448-3458.
LU Q Y, YUAN G L, ZHU H, et al. A Visual grounding method with contrastive learning large model[J]. Acta Electronica Sinica, 2024, 52(10): 3448-3458. (in Chinese)

[13] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[C]//International Conference on Machine Learning. San Diego: PMLR, 2021: 8821-8831.

[14] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. (2021-02-26)[2025-05-20]. https://arxiv.org/abs/2103.00020v1.

[15] 樊琳, 龚勋, 郑岑洋. 基于文本引导下的多模态医学图像分析算法[J]. 电子学报, 2024, 52(7): 2341-2355.
FAN L, GONG X, ZHENG C Y. A multi-modal medical image analysis algorithm based on text guidance[J]. Acta Electronica Sinica, 2024, 52(7): 2341-2355. (in Chinese)

[16] XIA W H, YANG Y J, XUE J H, et al. TediGAN: Text-guided diverse face image generation and manipulation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 2256-2265.

[17] KOCASARı U, DIRIK A, TIFTIKCI M, et al. StyleMC: Multi-channel based fast text-guided image generation and manipulation[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2022: 3441-3450.

[18] ABDAL R, ZHU P H, FEMIANI J, et al. CLIP2StyleGAN: Unsupervised extraction of StyleGAN edit directions[C]//Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings. New York: ACM, 2022: 1-9.

[19] LOU X D, LIU Y G, LI X W. TeCM-CLIP: TeCM-CLIP: Text-Based Controllable Multi-Attribute Face Image Manipulation[M]// Computer Vision-ACCV 2022. Cham: Springer Nature Switzerland, 2023: 71-87.

[20] LIU S L, ZENG Z Y, REN T H, et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection[M]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 38-55.

[21] LI L H, ZHANG P C, ZHANG H T, et al. Grounded language-image pre-training[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 10955-10965.

[22] CHENG T H, SONG L, GE Y X, et al. YOLO-world: Real-time open-vocabulary object detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 16901-16911.

[23] JIANG Q, LI F, ZENG Z Y, et al. T-Rex2: Towards generic object detection via text-visual prompt synergy[M]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 38-57.

[24] KAMATH A, SINGH M, LECUN Y, et al. MDETR - modulated detection for end-to-end multi-modal understanding[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 1760-1770.

[25] ZHAN Y, YUAN Y, XIONG Z T. Mono3DVG: 3D visual grounding in monocular images[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(7): 6988-6996.

[26] YANG L, YUAN C F, ZHANG Z Q, et al. Exploiting contextual objects and relations for 3D visual grounding[J]. Advances in Neural Information Processing Systems, 2024, 1: 36.

[27] CHEN D Z, CHANG A X, NIEßNER M. ScanRefer: 3D Object Localization in RGB-D Scans Using Natural Language[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 202-221.

[28] LIN Z X, PENG X D, CONG P S, et al. WildRefer: 3D Object Localization in Large-Scale Dynamic Scenes with Multi-Modal Visual Data and Natural Language[M]// Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 456-473.

[29] YU H B, YANG W X, RUAN H Z, 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 (CVPR). Piscataway: IEEE, 2023: 5486-5495.

[30] YU H B, LUO Y Z, SHU M, et al. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 21329-21338.

[31] ZENG A H, XU B, WANG B W,et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools[EB/OL]. (2024-07-30)[2025-05-20]. https://arxiv.org/abs/2406.12793v2.

[32] CUI Y M, CHE W X, LIU T, 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.

[33] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[34] LAN M, RONG F, JIAO H Z, et al. Language query-based transformer with multiscale cross-modal alignment for visual

grounding on remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5626513.
[35] SADHU A, CHEN K, NEVATIA R. Zero-shot grounding of objects from natural language queries[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 4694-4703.
[36] YANG Z Y, GONG B Q, WANG L W, et al. A fast and accurate one-stage approach to visual grounding[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 4683-4693.
[37] YANG Z Y, CHEN T L, WANG L W, et al. Improving One-Stage Visual Grounding by Recursive Sub-Query Construction[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 387-404.
[38] DENG J J, YANG Z Y, CHEN T L, et al. TransVG: End-to-end visual grounding with transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 1749-1759.

作者简介

杨　洋　男,2000年12月生,陕西汉中人.长安大学信息工程学院硕士研究生.主要研究方向为视觉定位、多模态融合与计算机视觉.
E-mail: yangyy@chd.edu.cn

胡红利　女,2002年4月生,河南濮阳人.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、位姿估计、三维检测.
E-mail: hhlhu@chd.edu.cn

魏弘凯　男,2001年5月生,福建南平人.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

郭柯宇　男,1999年9月生,贵州黔南人.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

孙士杰　男,1989年10月生,河南商丘人.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

宋焕生　男,1964年10月生,内蒙古赤峰人.长安大学信息工程学院博士生导师、二级教授,国务院政府特殊津贴专家.主要研究方向为基于机器视觉的交通感知及交通预警.
E-mail: hshsong@chd.edu.cn

宋翔宇　男,1991年3月生,陕西西安人.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