面向中文语义驱动的 3D 视觉定位基准方法

黄永乐[1)], 王天添[1)], 孙士杰[2)*], 胡红利[1)], 刘泽东[1)], 宋翔宇[2)]

[1)] (长安大学信息工程学院　西安　710064)
[2)] (长安大学数据科学与人工智能研究院　西安　710064)
(shijieSun@chd.edu.cn)

摘　要: 3D 视觉定位是多模态学习中的重要研究方向, 旨在理解自然语言描述中的语义需求以提取其在场景中的目标 3D 信息, 在人机交互、自动驾驶等依赖语义指导的领域有巨大的应用前景. 针对现有方法依赖昂贵的高精度设备, 且仅支持英文作为输入, 限制了在中国市场的推广与应用的问题, 提出面向中文语义驱动的 3D 视觉定位基准方法. 首先构建一个中文基准数据集-Traffic3DRefer, 包含 5 148 幅图像和 10 296 条中文自然语言描述, 场景中最远目标的距离可达 175 m, 并涵盖了单目标和多目标 2 种场景类型. 然后提出一个有效的网络框架 Traffic3DVG: 通过多模态特征编码器提取文本特征、目标图像信息和 3D 几何信息; 利用空间感知融合模块将 2D 图像特征与 3D 几何信息融合以学习具有判别性的表征; 通过多模态特征融合模块进行多模态融合以获得鲁棒的多模态表征, 用于视觉-文本匹配. 在 Traffic3DRefer 数据集上的大量实验结果表明, 所提框架在低成本硬件上显著提升了 F1 分数、精确率和召回率, 有效地推动了中文 3D 视觉定位研究的发展与实际应用.

关键词: 自动驾驶; 多模态学习; 多模态融合; 3D 视觉定位; 人机交互
中图法分类号: TP391.4　　**DOI:** 10.3724/SP.J.1089.2025-00047

A Benchmark Method for 3D Visual Grounding Driven by Chinese Semantics

Huang Yongle[1)], Wang Tiantian[1)], Sun Shijie[2)*], Hu Hongli[1)], Liu Zedong[1)], and Song Xiangyu[2)]

[1)] (*School of Information Engineering, Chang'an University, Xi'an*　710064)
[2)] (*School of Data Science and Artificial Intelligence, Chang'an University, Xi'an*　710064)

Abstract: 3D visual grounding is an important research direction in multimodal learning. It aims to understand the semantic requirements in natural language descriptions to extract objects' 3D information in the scene. It has tremendous application potential in fields that rely on semantic guidance, such as human-computer interaction and autonomous driving. To address the fact that existing methods depend on costly high-precision equipment and accept only English input, which limits their promotion and application in the Chinese market, we introduced a Chinese-semantic-driven benchmark for 3D visual grounding. We first constructed a Chinese benchmark dataset, Traffic3DRefer, comprising 5148 images and 10296 Chinese natural language descriptions. The farthest object in a scene is up to 175 meters away, and the dataset covers both single-object and multi-object scene types. Next, we presented an efficient network architecture, Traffic3DVG. It first employs a multimodal encoder to extract textual features, object image information, and 3D geometric information. A spatial-aware fusion module then integrates the 2D image features with the 3D geometric information to learn discriminative representations. Finally, a multimodal feature fusion module is

收稿日期: 2025-02-27; 修回日期: 2025-05-26. **基金项目**: 国家重点研发计划(2023YFB4301800). **黄永乐**(2001—), 男, 硕士研究生, CCF 学生会员, 主要研究方向为计算机视觉、视觉定位; **王天添**(2002—), 男, 硕士研究生, 主要研究方向为计算机视觉、模式识别; **孙士杰**(1989—), 男, 博士, 副教授, 硕士生导师, CCF 会员, 论文通信作者, 主要研究方向为计算机视觉、图像处理、模式识别; **胡红利**(2002—), 女, 硕士研究生, 主要研究方向为计算机视觉、目标检测; **刘泽东**(1999—), 男, 博士研究生, 主要研究方向为知识增强、图学习; **宋翔宇**(1991—), 男, 博士, 副教授, 博士生导师, 主要研究方向为图数据挖掘、深度学习、多模态学习.

employed to perform multimodal integration, obtaining robust multimodal representations that can be utilized for visual-text matching. Extensive experiments on the Traffic3DRefer dataset demonstrate that the proposed framework significantly improves F1 scores, precision, and recall on low-cost hardware, which effectively advances the development and practical application of Chinese 3D visual grounding research.

Key words: autonomous driving; multimodal learning; multimodal fusion; 3D visual grounding; human-computer interaction

当前，随着多模态技术的不断发展，视觉信息与语言信息的结合逐渐成为研究热点. 人类在理解场景时不仅依赖于视觉系统的感知，还需要语言知识的补充和增强，3D 视觉定位任务应运而生，旨在构建自然语言描述与 3D 目标检测技术之间的跨模态关联，并推动了人机交互[1]、自动驾驶[2]以及智能机器人[3]等领域的发展与应用. 目前，针对 3D 视觉定位任务的解决方案分为基于 RGB-D 扫描的方法[4-5]和基于激光雷达点云的方法[6-8]. 尽管这些方法在特定场景中取得了显著成果，但它们对昂贵的深度传感设备的依赖限制了实际应用的普及，因此，开发低成本且易于部署的 3D 视觉定位方案尤为重要. 与依赖高端设备的方法不同，基于单目视觉进行 3D 视觉定位不仅能显著地降低设备成本，还能拓展其应用范围，然而相关研究目前仍然较为匮乏. 针对这一问题，Mono3DVG[9]作为一种创新方法，通过结合自然语言描述和纯视觉图像实现 3D 目标定位. 该方法面临两大挑战：一是仅支持单一目标的 3D 视觉定位，无法满足多目标场景和更复杂应用的实际需求；二是目前所有的 3D 视觉定位方法仅支持英文输入，这种语言壁垒制约着该技术在中国市场的应用.

为了解决这些问题，本文提出面向中文语义驱动的 3D 视觉定位基准方法，首先构建一个中文基准数据集 Traffic3DRefer，然后提出一个简单高效的网络框架 Traffic3DVG.

Traffic3DRefer 数据集为未来基于中文语义表征的 3D 视觉定位的相关研究提供重要的数据支撑，包含 5 148 幅图像和 10 296 条中文自然语言描述，场景中最远目标的距离达到 175 m，符合语义描述的目标个数达到 25 489 个；该数据集涵盖交通场景中的 4 种车辆类别，进一步丰富了其应用范围.

Traffic3DVG 网络框架首先采用多模态特征编码器对中文语义文本进行编码，并提取目标的图像特征与 3D 几何信息；然后利用空间感知融合模块将目标的图像特征与 3D 空间信息进行有效的整合，使得外观相似但几何信息不同的目标能够学习到具有区分性的特征；最后通过多模态特征融合模块进一步结合文本与目标信息以生成鲁棒的多模态表征，并通过视觉-文本匹配任务进行监督训练.

1 相关工作

视觉定位任务通过自然语言查询描述的语义需求得到对应查询目标的位置、结构、深度等信息. 目前，该任务分为基于单目视觉的 2D 视觉定位[10-11]、基于雷达点云的 3D 视觉定位[12-13]与基于单目视觉的 3D 视觉定位[14]3 类方法.

1.1 基于单目视觉的 2D 视觉定位

基于单目视觉的 2D 视觉定位方法通过查询文本，得到 RGB 图像中对应查询目标的 2D 位置信息. 陆庆阳等[15]提出一种基于对比学习的视觉定位方法 CLIPVG，采用 CLIP[16]分别编码图像和文本特征并进行深度融合，最后使用多模态融合特征预测目标的 2D 位置信息；Yang 等[17]提出一个迭代鲁棒视觉定位框架，其基于掩码参考的中心点监督机制和迭代多级视觉语言融合，提升了模型的细粒度定位能力. 尽管这些方法在 2D 视觉方面取得了显著的进展，但实际应用中通常需要对 3D 场景进行理解，而 2D 信息难以捕捉物体的深度、空间关系和立体结构. 引入 3D 信息能够更好地理解物体的几何形态和空间位置，在视觉任务中提供更精确的检测能力. 因此，将 3D 信息融入多模态融合任务对于提升模型的真实感知能力至关重要.

1.2 基于雷达点云的 3D 视觉定位

基于雷达点云的 3D 视觉定位方法结合激光雷达和工业相机，可以实现 3D 场景中的目标细粒度精确定位. Lin 等[18]利用激光雷达和工业相机的多帧同步数据，通过动态视觉编码器和模态交互模块对自然语言、图像和点云数据进行融合，实现在复杂动态场景下的 3D 目标定位；罗寒等[19]提出一

种基于语义一致性约束与局部-全局感知的多模态 3D 视觉定位方法 MM-VG，通过蒸馏 2D 知识促进 3D 模型得到点云-文本语义一致性表征，并设计了一个局部-全局感知模块增强目标表征，以更好地匹配目标. 这些方法虽然在特定场景下取得了一定的成果，但对高成本深度传感设备的依赖性严重地限制了其在实际场景中的广泛应用，难以广泛推广. 因此，探索低成本、易于部署的 3D 视觉定位解决方案至关重要.

1.3　基于单目视觉的 3D 视觉定位

目前，针对单目视觉下的 3D 视觉定位方法较少. Mono3DVG 作为一种具有代表性的方法，展示了通过文本描述和单目视觉在场景中进行目标 3D 定位的潜力，其创新之处在于无需依赖额外信息，仅通过融合自然语言嵌入和视觉特征，即可实现从 2D 图像到 3D 空间的目标定位. 然而，该方法存在明显的局限性，仅支持单一目标的 3D 视觉定位；此外，提出的 Mono3DRefer 数据集仅兼容英文输入，严重地阻碍了其在中国市场的推广与应用.

2　Traffic3DRefer 数据集

目前，3D 视觉定位的研究仍处于初级阶段，相关研究成果和可用数据集有限，制约了该领域的发展，一个主要瓶颈是缺乏高质量的 3D 视觉定位数据集. 如表 1 所示，现有的数据集大多专注于单目标视觉定位，且它们依赖于昂贵的激光雷达和工业相机进行数据采集，极大地限制了其广泛应用的可能性. 此外，目前主流的 3D 视觉定位数据集仅支持英文输入，阻碍了 3D 视觉定位技术在中国市场的推广与落地. 由于缺乏本地化语言支持，相关技术的应用场景和用户体验受到严重制约，亟需开发基于中文语义的 3D 视觉定位数据集，不仅能够推动中文 3D 视觉定位的研究，还能为更智能的 3D 视觉定位技术的发展奠定基础，并最终推动其在人机交互、自动驾驶和机器人导航等领域的广泛应用.

表 1　3D 视觉定位任务数据集的信息统计与对比分析

数据集	出版源	描述平均长度/字数	目标个数	距离/m	视觉组成	标签	目标	语言
ScanRefer[4]	ECCV2020	20.27	11046	10	点云	3D	单目标	英文
SUNRefer[5]	CVPR2021	16.30	7699		RGB-D	3D	单目标	英文
STRefer[18]	ECCV2024		3581	30	点云&RGB	3D	单目标	英文
LifeRefer[18]	ECCV2024		11864	30	点云&RGB	3D	单目标	英文
Mono3DRefer[9]	AAAI2024	**53.24**	8228	102	RGB	2D&3D	单目标	英文
Traffic3DRefer		24.20	**25489**	**175**	RGB	2D&3D	多目标	中文

注. 粗体表示最优值.

2.1　数据集构建

为了推动中文 3D 视觉定位领域的发展并弥补现有数据集的不足，构建了一个名为 Traffic3DRefer 的大规模数据集. 该数据集基于 Rope3D 数据集[20]，通过为场景添加语言描述，构成同时包含单目标与多目标场景的中文基准数据集，为研究更复杂、更贴近真实需求的中文 3D 视觉定位问题提供了有力的数据支撑.

Traffic3DRefer 数据集的流程如图 1 所示. 采用一种基于目标属性的描述生成方法：首先从 Rope3D 数据集中为每个目标对象提取 10 种属性信息，包括颜色、遮挡、截断、位置、距离、朝向、类别、高度、长度和宽度. 其中，颜色属性通过深度学习网络初步识别，并由人工辅助验证和修正；遮挡、截断、类别、高度、长度和宽度直接来自原始标注；位置、朝向和距离则由对原始标注的关键参数进行处理后获得.

图 1 Traffic3DRefer 数据集的流程

为了贴近真实场景的人类描述特征并体现语言的随机性和多样性，对提取的 10 种属性进行随机选择，确保每个场景至少有一个或多个目标符合这些属性；然后将这些属性填入预先设计的提示词模板，如“场景中__米以内的所有__车”；最后利用 ChatGPT 自动生成指代场景中单个或多个目标的自然语言描述. 这种基于随机选择和多属性描述的方式，不仅能模拟真实场景中语言描述的多样性与不确定性，还能有效地避免数据描述的单一和重复；同时，这种策略可帮助模型在训练阶段不断地应对新颖的语义组合和表达方式，提升其理解语言线索和视觉信息的综合能力. 得到的数据集更能反映实际应用需求，并为后续中文 3D 视觉定位模型在复杂、多样化环境中的泛化奠定了坚实的基础.

2.2 数据集概述

Traffic3DRefer 数据集中，平均每条描述长度为 24.2 个字；该数据集同时涵盖单目标与多目标场景描述. 在实际应用中，人们更倾向于以简洁的语言描述单个或多个目标对象，不仅使 Traffic3-DRefer 数据集更加贴近真实场景，也有助于模型学习更具实用性的特征嵌入. Traffic3DRefer 数据集的构建对于中文 3D 视觉定位领域具有重要推动作用，其庞大的数据规模、丰富多样的场景，以及对单目标与多目标情境的全面覆盖，为训练与评估更为强大的中文 3D 视觉定位模型提供了关键的数据支撑.

3 本文方法

Traffic3DVG 网络框架如图 2 所示，由 3 部分组成: (1) 多模态特征编码器; (2) 空间感知融合模块; (3) 多模态特征融合模块.

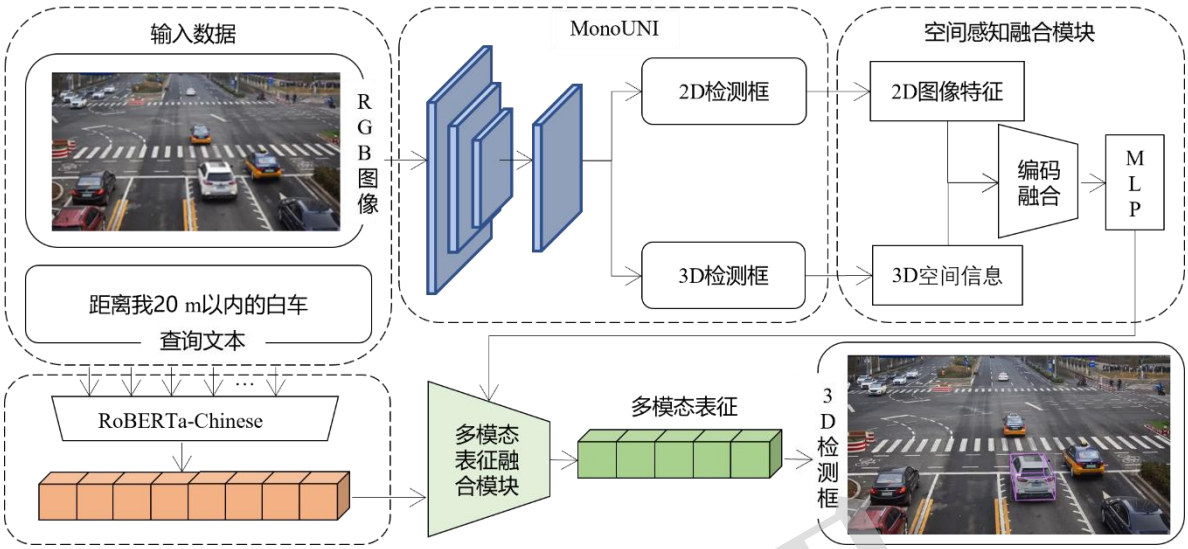

图 2　Traffic3DVG 网络框架

3.1　多模态特征编码器

为了更准确地捕捉输入查询文本的深层语义，采用预训练的 RoBERTa-Chinese 模型[21]作为文本编码器. 该模型基于 Transformer 架构，在大规模中文语料上训练而成，能有效地捕捉中文文本的语义和上下文信息. 将文本输入到该模型中，中文文本序列编码为高维嵌入向量表征 $\boldsymbol{f}_{\mathrm{t}} \in \mathbb{R}^{C}$，其中，$C=512$. 这些向量中的分类标记(classification token, CLS)向量 $\boldsymbol{f}_{\mathrm{CLS}} \in \mathbb{R}^{C}$ 不仅包含对单词语义特征的捕捉，也凝练了上下文关联与语义结构，为后续任务提供更丰富、精确的中文语义支持.

对于输入的RGB图像，采用最新的3D目标检测器 MonoUNI[22]作为视觉检测器，能同时为每个检测到的目标生成 2D 边界框和相应的 3D 几何信息. 如图 3 所示，MonoUNI 采用 CenterNet[23]作为主干网络生成多尺度的特征图，能够在不同尺度下有效地捕捉物体的空间信息；然后设计多个检测头预测目标的不同属性，包括中心点热力图、2D 边界框、3D 偏移量、目标尺度、目标方向、3D 归一化立方体深度、偏差深度以及深度不确定性. 在推理过程中, MonoUNI 利用预测的 3D 归一化立方体深度与偏差深度，并结合相机参数(如焦距)等信息，有效地修正深度误差，精确地估计目标的 3D 位置.

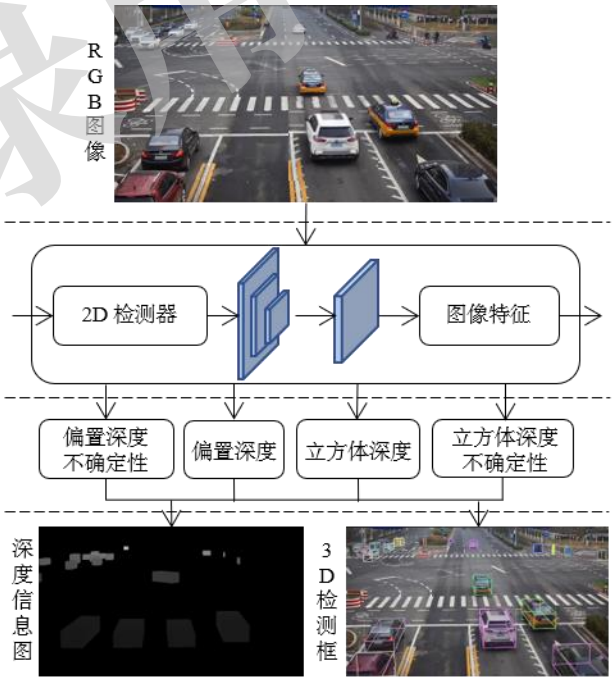

图 3　MonoUNI 的结构

3.2　空间感知融合模块

为了使外观相似但几何信息不同的目标能学到更具区分性的表征，设计了一个空间感知融合模块，将目标的 3D 几何信息与图像特征进行有效的融合. 首先将 MonoUNI 输出的目标 3D 信息整合为一个空间提示向量 $\boldsymbol{v}=[x, y, z, h, w, l, r_{\sin}, r_{\cos}, d]$，其中，$x$, y 和 z 对应目标的 3D 坐标，h, w 和 l 对应 3D 目标的高度、宽度和长度，$r_{\sin}$ 和 $r_{\cos}$ 分别对应物体在相机坐标系中全局方位角的正弦值和余弦值，d 表示目标与相机之间的距离. 然后根据 MonoUNI 输出的目标 2D 坐标，从图像中裁剪出目标区域，并通过预训练的 ResNet50[24]得到全局平

均池化层输出的 2D 图像特征 $\boldsymbol{f}_{\mathrm{o}} \in \mathbb{R}^{D}$，其中，$D=2\,048$；同时，一个多层感知机(multi-layer perception, MLP)将目标的空间提示向量 $\boldsymbol{v}$ 编码为 3D 空间信息特征 $\boldsymbol{f}_{\mathrm{p}} \in \mathbb{R}^{C}$. 最后将图像特征与 3D 空间信息特征进行有效的融合，获得最终的目标视觉表征 $\boldsymbol{f}_{\mathrm{v}} \in \mathbb{R}^{C}$，公式为

$$\boldsymbol{f}_{\mathrm{v}}=\mathrm{MLP}\left(\boldsymbol{f}_{\mathrm{o}} \odot W \boldsymbol{f}_{\mathrm{p}}\right).$$

其中，W 表示线性层；$\odot$ 表示哈达玛积操作，其将 2 个向量对应位置元素相乘，结果是一个向量且该向量的维度与原始向量相同.

3.3 多模态特征融合模块

为了获得更鲁棒的多模态表征用于视觉-文本匹配，基于先对齐后融合[25]的思路设计了一个多模态特征融合模块，其结构如图 4 所示.

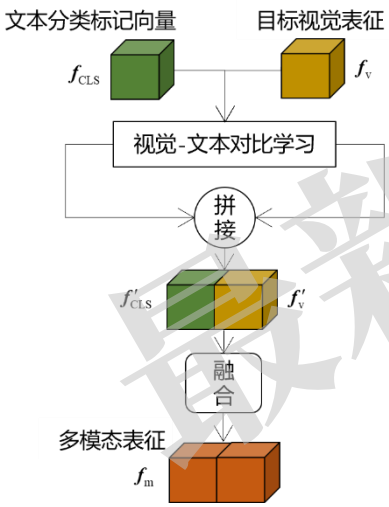

图 4 多模态特征融合模块的结构

首先将文本表征的分类标记向量 $\boldsymbol{f}_{\mathrm{CLS}}$ 与目标视觉表征 $\boldsymbol{f}_{\mathrm{v}}$ 进行跨模态对齐，通过视觉-文本对比学习更好地捕捉跨模态的信息关联；然后将对齐过后的文本表征 $\boldsymbol{f}'_{\mathrm{CLS}} \in \mathbb{R}^{C}$ 与视觉表征 $\boldsymbol{f}'_{\mathrm{v}} \in \mathbb{R}^{C}$ 的拼接进行多模态融合，得到最终的多模态表征 $\boldsymbol{f}_{\mathrm{m}} \in \mathbb{R}^{C}$，公式为

$$\boldsymbol{f}_{\mathrm{m}}=\mathrm{MLP}\left(\boldsymbol{f}'_{\mathrm{CLS}} \oplus \boldsymbol{f}'_{\mathrm{v}}\right).$$

其中，$\oplus$ 表示沿维度方向拼接操作.

3.4 损失函数

为了在多模态融合之前获得更好的跨模态对齐特征，采用双向三元组损失函数[26]作为监督信号. 其中，通过鼓励匹配的图文对具有更高的相似性得分，不匹配的图文对则具有更低的相似性得分，实现有效的跨模态对齐. 该函数定义为

$$L_1=\max\left(\left[m-s\left(\boldsymbol{f}_{\mathrm{CLS}}, \boldsymbol{f}_{\mathrm{v}}\right)+s\left(\boldsymbol{f}_{\mathrm{CLS}}, \boldsymbol{f}_{\mathrm{v}-}\right)\right], 0\right)+$$

$$\max\left(\left[m-s\left(\boldsymbol{f}_{\mathrm{CLS}}, \boldsymbol{f}_{\mathrm{v}}\right)+s\left(\boldsymbol{f}_{\mathrm{CLS}-}, \boldsymbol{f}_{\mathrm{v}}\right)\right], 0\right) \quad (1)$$

其中，m 表示用于设置相似性得分最小间隔的边界阈值超参数；$\boldsymbol{f}_{\mathrm{v}-}$ 和 $\boldsymbol{f}_{\mathrm{CLS}-}$ 表示同一个批次中的负样本；$s(\cdot)$ 表示衡量图文对之间相似度的度量方式，本文采用余弦相似度作为度量方式.

由于本文的最终任务是利用多模态表征进行有效的视觉-文本匹配，即判断中文查询文本是否与所检测的 3D 目标匹配，因此，多模态特征融合模块输出的多模态表征 $\boldsymbol{f}_{\mathrm{m}}$ 将被输入到一个 MLP 中预测一个二分类概率值 $p \in (0,1)$. 则本文的视觉-文本匹配损失函数定义为

$$L_2=\frac{1}{N}\left[y \cdot \log p+(1-y) \cdot \log (1-p)\right].$$

其中，y 表示独热编码的真实标签，N 表示训练阶段批次的大小.

最终，本文的损失函数为 $L=L_1+L_2$.

4 实验结果及分析

通过与现有的基于不同模态输入的 3D 视觉定位方法进行实验，证明本文方法的有效性.

4.1 实验环境及相关设置

本文实验中，超参数设置如下：训练总轮数为 50，热身轮数为 10；初始学习率为 3E−4，权重衰减为 2E−3；批次大小为 128；梯度裁剪阈值为 2.0；式(1)中的边界阈值超参数 m 为 0.3. 对数据集进行随机打乱处理，并按照 7∶1∶1 的比例划分为训练集、验证集和测试集. 此外，Traffic3DVG 网络框架基于 Pytorch 框架实现；参数优化器 AdamW 的学习率调整策略采用热身与正余弦退火相结合的方法，首先通过热身阶段逐步增加学习率，然后利用正余弦函数调整学习率，实现最终的优化效果. 采用真实边界框与预测边界框 2 种输入方式进行评估.

4.2 评估指标

为了全面地评估检测和匹配的准确性，结合 F1 分数 F_1 与交并比(intersection over union, IoU)分数 $\mathrm{IoU}_{3\mathrm{D}}$ 作为评估指标. 其中，$\mathrm{IoU}_{3\mathrm{D}}$ 通过计算 2 个 3D 边界框的交集体积与并集体积之比，量化预测边界框 B_{pred} 与真实边界框 B_{gt} 之间的重叠程度. $\mathrm{IoU}_{3\mathrm{D}}$ 的计算公式为

$$\mathrm{IoU}_{3D}=\frac{V\left(B_{\mathrm{pred}}\cap B_{\mathrm{gt}}\right)}{V\left(B_{\mathrm{pred}}\cup B_{\mathrm{gt}}\right)}.$$

其中，$V\left(B_{\mathrm{pred}}\cap B_{\mathrm{gt}}\right)$表示预测边界框与真实边界框的交集体积，$V\left(B_{\mathrm{pred}}\cup B_{\mathrm{gt}}\right)$表示预测边界框与真实边界框的并集体积. 当预测边界框与真实边界框的 IoU_{3D} 超过预定义的阈值时，该预测将被归类为真阳性(true positive, TP)，确保只有高度匹配的预测才被视为准确匹配. 基于 IoU_{3D} 进行匹配后，采用 F1 分数评估检测的准确性，其综合考虑了精确率 P 召回率 R，旨在全面衡量模型的整体性能.

P 用于衡量模型在所有被预测为正类的样本中实际为正类的比例，计算公式为

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}.$$

其中，TP 指模型正确预测为正类的样本数量，即实际为正类的样本被准备识别为正类；假阳性(false positive, FP)指模型错误地将负类样本预测为正类的数量，即实际为负类的样本被误判为正类. P 越大，表示模型在预测正类时错误较少，即误报率低.

R 用于衡量模型在所有实际为正类的样本中正确预测为正类的比例，计算公式为

$$R=\frac{\mathrm{TP}}{\mathrm{TP+FN}}.$$

其中，假阴性(false negative, FN)指模型错误地将正类样本预测为负类的数量，即实际为正类的样本被误判为负类. R 越大，表示模型能够识别出大多数的正类样本，即漏报率低.

F1 分数作为 P 与 R 的调和平均值，旨在全面衡量模型的整体性能，计算公式为

$$F_1=2\times\frac{P\times R}{P+R}.$$

F_1 的取值为 0~1，其值越高，表示模型的综合性能越优. 在 P 和 R 存在权衡关系的情况下，F_1 提供了一个统一的衡量标准，有助于评估模型在不同方面的表现.

4.3　对比实验

在 Traffic3DRefer 测试集上，分别采用真实边界框和预测边界框 2 种输入方式，将本文方法与其他方法进行实验，包括 ScanRefer[4]，ReferIt3D[27]，InstanceRefer[28]，3DVG-Trans[6]，Multi-View Transformer[29]和 M3DRef-CLIP[30]，结果如表 2 和表 3 所示. 使用真实边界框可以消除目标检测阶段的误差影响，更精确地评估模型在理想条件下将自然语言描述与 3D 目标相匹配的能力；使用预测边界框则能反映模型在真实应用中对检测误差的处理能力，评估其在更复杂条件下的鲁棒性. 为了确保实验的公平性，对其他方法中的点云或深度分支进行裁剪，避免其利用额外的深度信息.

表 2　在 Traffic3DRefer 测试集上采用真实边界框作为输入时的结果对比

方法	F_1↑	P↑	R↑	TP↑	FP↓	FN↓
ScanRefer[4]	48.99	36.10	76.21	46194	81779	14420
ReferIt3D[27]	50.12	37.09	77.26	47487	80542	13977
InstanceRefer[28]	50.26	37.08	77.96	47621	80811	13460
3DVG-Trans[6]	52.01	38.60	79.73	49736	79124	12648
Multi-View Transformer[29]	53.89	40.11	82.11	51872	77457	11304
M3DRef-CLIP[30]	54.00	40.06	82.85	52233	78165	10813
Traffic3DVG	**58.24**	**43.79**	**86.92**	**58095**	**74562**	**8742**

注. 粗体表示最优值.

表 3　在 Traffic3DRefer 测试集上采用预测边界框作为输入时的结果对比

方法	F_1↑	P↑	R↑	TP↑	FP↓	FN↓
ScanRefer[4]	39.96	32.25	52.50	33114	69568	29958
ReferIt3D[27]	41.52	33.61	54.28	34518	68177	29075
InstanceRefer[28]	41.87	33.91	54.70	34809	67828	28824
3DVG-Trans[6]	42.83	34.48	56.49	35373	67204	27241
Multi-View Transformer[29]	45.03	36.37	59.11	37841	66198	26174
M3DRef-CLIP[30]	45.44	36.72	59.59	38142	65724	25864
Traffic3DVG	**50.11**	**40.78**	**64.98**	**43427**	**63055**	**23407**

注. 粗体表示最优值.

从表 2 可以看出，Traffic3DVG 在各项评估指标上均表现优异；在 F_1 上，比 M3DRef-CLIP 方法提高 4.24，表明该方法在综合 P 和 R 上的表现更加均衡；在 P 上，比 M3DRef-CLIP 方法提升 3.73，证明该方法能够更准确地识别目标；在 R 上，比 M3DRef-CLIP 方法提升 4.07，说明该方法在漏检控制方面更加有效，能够识别更多的实际目标. 实验结果证明了 Traffic3DVG 在 3D 视觉定位任务上的鲁棒性和优越性.

从表 3 可以看出，Traffic3DVG 的性能有所下降，这可能是由于第 1 阶段的 MonoUNI 检测中产生的错误对后续视觉-文本匹配模块造成影响；尽管存在检测误差，Traffic3DVG 仍然展现了出色的性能，比 M3DRef-CLIP 方法的 F_1 提高 4.67，P 提升 4.06，R 提升 5.39. 实验结果表明，即便在存在一定检测问题的条件下，Traffic3DVG 依然具有良好的稳定性与鲁棒性.

为了更直观地展示本文方法的有效性，在 Traffic3DRefer 数据集上进行定量对比实验，结果如图 5 所示. 可以看出，当输入方式为真实边界框时，Traffic3DVG 实现了精准的 3D 视觉定位效果，这一优异表现得益于该网络框架在多模态融合过程中充分利用目标的外观信息与几何信息，提升了定位精度；然而当输入方式为预测边界框时，Traffic3DVG 在部分情境下未能实现成功的定位，这是由于该网络框架对第 1 阶段 MonoUNI 检测器的性能有较强依赖，当检测器的表现不佳时，错误的预测边界框将影响后续定位任务，整体定位效果下降.

图 5 Traffic3DVG 的定量结果

4.4 消融实验

通过一系列消融实验，评估 Traffic3DVG 网络框架中各模块的有效性. 通过使用真实边界框，可在排除目标检测误差的基础上更准确地衡量各模块对整体性能的影响. 逐步将空间感知融合模块与多模态特征融合模块集成到模型中，观察它们对性能的影响，结果如表 4 所示，可以看出，每当引入一个新模块，模型的指标均显著提升，充分证明各个模块在性能提升中发挥了关键作用，验证了本文设计的合理性和有效性.

表 4 在 Traffic3DRefer 测试集上的消融实验结果

融合模块		F_1 ↑	P ↑	R ↑
空间感知	多模态特征			
×	×	49.13	35.97	77.47
×	√	51.43	38.00	79.54
√	×	56.66	42.30	85.80
√	√	**58.24**	**43.79**	**86.92**

注. 粗体表示最优值.

5 结 语

本文构建了大规模中文 3D 视觉定位基准数据集 Traffic3DRefer，为中文 3D 视觉定位任务提供高质量的数据支撑，支持研究人员进行更加深入的中文 3D 视觉定位研究；进一步，提出网络框架 Traffic3DVG，在多目标场景下实现了高效的中文 3D 视觉定位．实验结果表明，Traffic3DVG 在中文 3D 视觉定位任务中具有较高的精度和鲁棒性，为未来相关研究提供了重要的参考价值．本文代码和数据集链接为：https://github.com/xl010405/Traffic3DVG.

参考文献(References):

[1] Legaspi R, Xu W Z, Konishi T, *et al*. The sense of agency in human–AI interactions[J]. Knowledge-Based Systems, 2024, 286: Article No.111298

[2] Yuan Z X, Song X, Bai L, *et al*. Temporal-channel transformer for 3D lidar-based video object detection for autonomous driving[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(4): 2068-2078

[3] Chen Z Y, Wu K, Wu J, *et al*. Residual shrinkage transformer relation network for intelligent fault detection of industrial robot with zero-fault samples[J]. Knowledge-Based Systems, 2023, 268: Article No.110452

[4] Chen D Z, Chang A X, Nießner M. ScanRefer: 3D object localization in RGB-D scans using natural language[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 202-221

[5] Liu H L, Lin A R, Han X G, *et al*. Refer-it-in-RGBD: a bottom-up approach for 3D visual grounding in RGBD images[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 6032-6041

[6] Zhao L C, Cai D G, Sheng L, *et al*. 3DVG-transformer: relation modeling for visual grounding on point clouds[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 2928-2937

[7] Miyanishi T, Azuma D, Kurita S, *et al*. Cross3DVG: cross-dataset 3D visual grounding on different RGB-D scans[C] //Proceedings of the International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2024: 717-727

[8] Zhang Yuqi, Luo Han, Yang Yuwei, *et al*. Weakly supervised 3D visual grounding based on pseudo-text query generation and position awareness[J]. Modern Computer, 2024, 30(11): 35-39(in Chinese)
(张宇琦，罗寒，杨昱威，等. 基于伪文本查询生成及位置感知的弱监督 3D 视觉定位方法[J]. 现代计算机, 2024, 30(11): 35-39)

[9] Zhan Y, Yuan Y, Xiong Z T. Mono3DVG: 3D visual grounding in monocular images[C] //Proceedings of the 38th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2024: 6988-6996

[10] Deng J J, Yang Z Y, Chen T L, *et al*. TransVG: end-to-end visual grounding with transformers[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 1769-1779

[11] Chen S J, Li B C. Multi-modal dynamic graph transformer for visual grounding[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 15534-15543

[12] Cui K Y, Shen L M, Zheng Y Q, *et al*. Talk2Radar: talking to mmWave radars via smartphone speaker[C] //Proceedings of the IEEE Conference on Computer Communications. Los Alamitos: IEEE Computer Society Press, 2024: 2358-2367

[13] Yang L, Yuan C F, Zhang Z Q, *et al*. Exploiting contextual objects and relations for 3D visual grounding[C] //Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2023: Article No.2156

[14] Lei Q, Sun S J, Song X Y, *et al*. Bootstrapping vision–language transformer for monocular 3D visual grounding[J]. IET Image Processing, 2025, 19(1): Article No.e13315.

[15] Lu Qingyang, Yuan Guanglin, Zhu Hong, *et al*. A visual grounding method with contrastive learning large model[J]. Acta Electronica Sinica, 2024, 52(10): 3448-3458(in Chinese)
(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)

[16] Radford A, Kim J W, Hallacy C, *et al*. Learning transferable visual models from natural language supervision[C] //Proceedings of the 38th International Conference on Machine Learning. Maastricht: ML Research Press, 2021: 8748-8763

[17] Yang L, Xu Y, Yuan C F, *et al*. Improving visual grounding with visual-linguistic verification and iterative reasoning[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 9499-9508

[18] 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[C] //Proceedings of the 18th European Conference on Computer Vision. Heidelberg: Springer, 2025: 456-473

[19] Luo Han, Ma Haotong, Liu Jie, *et al*. Semantic consistency constrain and local-global aware multi-modal 3D visual grounding[J]. Application Research of Computers, 2024, 41(7): 2203-2208(in Chinese)
(罗寒，马浩统，刘杰，等. 基于语义一致性约束与局部-全局感知的多模态 3D 视觉定位[J]. 计算机应用研究, 2024, 41(7): 2203-2208)

[20] Ye X Q, Shu M, Li H Y, *et al*. Rope3D: the roadside perception dataset for autonomous driving and monocular 3D object detection task[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 21341-21350

[21] Liu Y X, Liu H, Wong L P, *et al*. A hybrid neural network RBERT-C based on pre-trained RoBERTa and CNN for user intent classification[C] //Proceedings of the 1st International

Conference on Neural Computing for Advanced Applications. Heidelberg: Springer, 2020: 306-319

[22] Jia J R, Li Z J, Shi Y F. MonoUNI: a unified vehicle and infrastructure-side monocular 3D object detection network with sufficient depth clues[C] //Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2023: Article No.514

[23] Duan K W, Bai S, Xie L X, *et al*. CenterNet: keypoint triplets for object detection[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 6569-6578

[24] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[25] Li J N, 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. Red Hook: Curran Associates Inc, 2021: Article No.742

[26] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3128-3137

[27] Achlioptas P, Abdelreheem A, Xia F, *et al*. ReferIt3D: neural listeners for fine-grained 3D object identification in real-world scenes[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 422-440

[28] Yuan Z H, Yan X, Liao Y H, *et al*. InstanceRefer: cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 1791-1800

[29] Huang S J, Chen Y L, Jia J Y, *et al*. Multi-view transformer for 3D visual grounding[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 15524-15533

[30] Zhang Y M, Gong Z M, Chang A X. Multi3DRefer: grounding text description to multiple 3D objects[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 15225-15236