面向单目图像下三维目标检测与空间语义描述的统一建模

王天添[1],郭柯宇[1],罗函轲[1],孙士杰[2*],程惠泽[1],孙张龙[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: 单目三维目标检测为三维感知提供了低成本解决方案,但现有方法难以生成可供人类直观理解的场景描述,从而限制了它们在人机交互、自动驾驶和其他需要丰富语义理解的场景中的适用性。视觉描述作为人类语言智能的直接体现,提供了理想的沟通媒介,赋予机器直观"讲述"场景的能力。但现有的视觉描述方法主要聚焦于单目图像内容,仅能表述物体间的二维拓扑关系,缺乏对三维几何信息(如精确距离、空间位置与运动状态)的精确建模与表达能力。若采用"先进行三维检测,再借助大模型生成描述"的两阶段方法,则存在系统效率低、信息一致性差的问题。而大模型的描述内容也只能局限于物体间的拓扑关系,无法准确反映三维几何信息,且因其固有的"幻觉"现象,也会导致空间信息的不准确并伴随冗余描述。为此,本文首次提出了单目三维视觉检测与空间语义描述(Monocular 3D Detection and Captioning,Mono3DDC)这一新颖任务。该任务旨在统一单目三维目标检测与描述生成,要求模型同时学习深度感知的视觉特征与语言语义,通过端到端的网络架构使生成的描述能够准确地学习到一致的三维空间信息,确保描述的几何准确性与三维目标检测的高精度。为支撑该交叉研究领域的深入探索,本文构建了首个支持中文语义的单目三维视觉描述基准数据集KITTI3DDC。该数据集基于KITTI数据集设计了一套高效的自动化数据生成流程,通过大语言模型与结构化验证模板的协同机制,在保证描述多样性与语言流畅性的同时,严格控制了空间信息的几何一致性,为后续研究提供了高质量的多模态监督信号。此外,本文设计了一个统一的新型框架,即Mono3DDC-TR(Monocular 3D Detection and Captioning based on Transformer)。该框架通过深度融合优化后的几何与视觉特征,在生成描述准确率及多类别三维检测效果上均展现出显著优势,在构建的KITTI3DDC基准上取得了最优性能,验证了该框架在几何信息与语言语义联合建模方面的有效性。本文为Mono3DDC任务提供了全面的基准测试,有效地推动了该任务的发展与实际应用。

关键词: 多模态学习;场景理解;三维目标检测;三维视觉描述;单目视觉;端到端学习

基金项目: 国家重点研发计划(No.2023YFB4301800);国家自然科学基金(No.62576050);江西省青年基金(No.S2024QNJJL0062)

中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2026)03-1105-13

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20251129

Towards Unified Modeling of 3D Detection and Spatial Semantic Captioning with Monocular Images

WANG Tiantian[1], GUO Keyu[1], LUO Hanke[1], SUN Shijie[2*], CHENG Huize[1], SUN Zhanglong[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: Monocular 3D object detection provides a low-cost solution for 3D perception. However, existing methods struggle to generate scene descriptions that are intuitively understandable to humans, which limits their applicability in scenarios requiring rich semantic understanding, such as human-computer interaction and autonomous driving. As a direct manifestation of human linguistic intelligence, visual captioning offers an ideal communication medium, endowing machines with the ability to intuitively "narrate" a scene. However, existing visual captioning methods primarily focus on monocular image content and can only describe two-dimensional topological relationships between objects, lacking the capability to accurately model and express 3D geometric information (e.g., precise distance, spatial location, and motion state). If a two-stage approach is adopted—first performing 3D detection and then leveraging a large model to generate descriptions—it suffers from low system efficiency and poor information consistency. Moreover, the descriptions generated by large models are limited to topological relationships and fail to accurately reflect 3D geometric information. In addition, the inherent "halluci-

收稿日期:2026-02-02;录用日期:2026-03-07;责任编委:王栋
*通信作者:孙士杰

nation" phenomenon of large models often leads to inaccurate spatial information accompanied by redundant descriptions. To address these issues, this paper proposes for the first time a novel task: monocular 3D detection and captioning (Mono3DDC). This task aims to unify monocular 3D object detection with caption generation, requiring the model to simultaneously learn depth-aware visual features and linguistic semantics. Through an end-to-end network architecture, the generated descriptions are enabled to accurately capture consistent 3D spatial information, ensuring both geometric accuracy in the descriptions and high precision in 3D object detection. To support in-depth exploration in this interdisciplinary research area, we construct the first benchmark dataset for monocular 3D visual captioning with Chinese semantic support, named KITTI3DDC. Based on the KITTI dataset, this dataset employs an efficient automated data generation pipeline that leverages the synergy between a large language model and structured verification templates. While ensuring diversity and linguistic fluency, the pipeline strictly maintains geometric consistency in spatial information, providing high-quality multimodal supervision signals for subsequent research. Furthermore, we design a novel unified framework, Mono3DDC-TR (Monocular 3D Detection and Captioning based on Transformer). By deeply integrating optimized geometric and visual features, this framework achieves significant advantages in both caption generation accuracy and multi-category 3D detection performance. It attains state-of-the-art results on the constructed KITTI3DDC benchmark, validating the effectiveness of the proposed end-to-end unified framework in jointly modeling geometric information and linguistic semantics. This paper provides a comprehensive benchmark for the Mono3DDC task, effectively promoting its development and practical application.

Keywords: multimodal learning; scene understanding; 3D object detection; 3D visual captioning; monocular vision; end-to-end learning

Foundation Item(s): National Key Research and Development Program of China (No.2023YFB4301800); National Natural Science Foundation of China (No.62576050); The Jiangxi Province Science Foundation Fund (No.S2024QNJJL0062)

0 引言

三维目标检测是实现自动驾驶[1-2]、机器人精细化操作[3]等领域的关键技术。传统的基于激光雷达的检测方法[4-6]虽精度较高,但成本昂贵,难以大规模普及。在此背景下,单目三维目标检测[7-10]应运而生,如图1①所示,它仅需单个摄像头即可从二维图像中推断三维空间结构,以其低成本和易部署的优势成为研究前沿,并涌现出诸多提升性能的创新方法[11-13]。

然而,当前研究存在一个明显的局限:仅依赖三维目标检测的单模态数据精度,无法保证自动驾驶等领域的安全性能。现有方法专注于几何属性的回归(如边界框中心、尺寸、朝向等),其输出是一系列抽象的数值。这种"机器友好"但"人类不友好"的输出形式,极大地限制了技术在需要人机协同决策的场景[14]中的应用。例如,在L2级智能驾驶接管场景中,当系统面临失效风险时,向驾驶员报告"左后方20 m有一辆快速接近的卡车"远比在屏幕上闪烁一个抽象红框更能迅速唤起驾驶员的空间感知与警觉。同样,在盲人辅助或机器人交互等场景中,自然语言是人机交互最高效的带宽,具备空间感知的描述能帮助视障人士建立环境心理地图。

视觉描述技术[15]作为连接视觉感知与人类语言的桥梁,能够将视觉内容转化为流畅的自然语言句子,为机器赋予"讲述"场景的能力,是实现自然人机交互的关键。但该技术主要应用在二维图像描述领域,如图1②所示,侧重于对图片进行整体性描述,缺乏对场景中每个独立目标的三维几何属性(如精确距离、空间位置和运动状态)的深度理解与描述能力。若能将该技术拓展至三维空间理解领域,则可以实现上文所提及的直观预警与交互功能。

为此,本文提出了一个全新的任务:单目图像下三维目标检测与描述生成任务的统一(Monocular 3D Detection and Captioning,Mono3DDC)。如图1③所示,该任务要求模型同步学习深度感知的视觉特征与语言语义,以实现高精度的三维目标检测,并确保生成描述中三维空间信息的几何准确性。为了支撑这一交叉研究领域,本文基于KITTI数据集,构建了一个同时支持高质量三维检测与中文视觉描述的数据集KITTI3DDC。该数据集由DeepSeek生成,结合自动化验证流程,有效克服了传统标注方法在规模扩展与质量把控上的局限性,显著降低了大型语言模型在生成过程中常见的描述失真现象,从而在根本上保证了三维空间信息与文本描述之间严格的几何一致性。这一基准研究可以为实时驾驶辅助、交互式系统引导和安全关键决策等具体应用提供支持。

此外,本文提出了一个端到端统一网络框架Mono3DDC-TR。相比于由独立检测模型与通用语言模型组成的两阶段系统架构,Mono3DDC-TR参数更少、推理路径更短,在保持高效的同时,通过内部闭环优化获得了更优的几何一致性。其核心由异构感

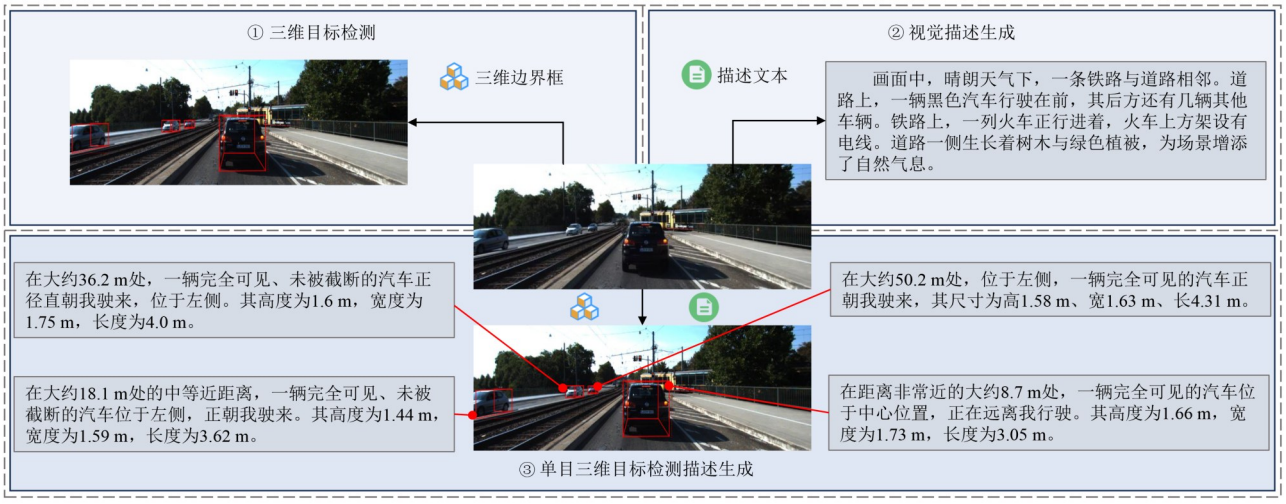

图1　Mono3DDC任务与其他任务的对比图

Figure 1　Comparison chart of Mono3DDC task with other tasks

知几何编码器与空间对齐融合模块构成。其中,异构感知几何编码器通过深度分布注意力机制,协同处理由物体中心网络提取的实例级深度先验与场景理解网络生成的布局级深度线索,从而构建出鲁棒的三维场景几何表示。在此基础上,空间对齐融合模块通过交叉注意力机制动态整合强化后的三维几何特征与细粒度的二维表观特征,为生成兼具视觉准确性与空间精确性的描述奠定基础。为确保生成文本与三维场景的严格一致,本文专门设计了几何约束描述头。该模块通过引入场景专用的符号化词典与基于三维空间关系的语义推理,有效规避了通用生成模型常见的描述失真问题,在提升推理效率的同时显著降低了空间指向的模糊性。

本文的主要贡献如下。

(1)提出Mono3DDC新任务,首次将单目三维目标检测与描述生成相统一,要求模型同时输出精确的三维边界框和包含几何准确性的场景描述。

(2)提出一个高效的自动化数据生成流程,并构建出首个支持中文语义的单目三维视觉描述基准数据集KITTI3DDC,填补了该交叉研究领域的数据空白。

(3)提出Mono3DDC-TR端到端统一网络框架,实现三维几何特征与二维视觉线索的深度耦合,在保持高推理效率的同时显著提升了描述准确性与三维检测性能。

(4)为Mono3DDC任务建立完整的基准体系,并通过系统的实验验证,证明了所提出的Mono3DDC-TR模型在检测精度与描述质量上均优于现有方法。

1　相关工作

本文的研究属于计算机视觉与自然语言处理的交叉领域,核心在于实现单目视觉中三维几何空间与语言语义空间的精确映射与对齐。相关工作主要从以下三个紧密关联的方向展开。

1.1　单目三维目标检测

传统三维目标检测方法通常依赖激光雷达或多传感器点云数据[16-18],虽然几何精度较高,但硬件成本昂贵、部署复杂,限制了其在实际场景中的大规模应用。单目三维目标检测旨在从单张二维图像中推断出场景中物体的三维边界框,因其硬件成本低廉而成为研究热点。早期工作如MonoCon[9]通过预测一系列三维关键点属性来估计物体姿态。MonoDETR[7]将检测视为集合预测问题,利用Transformer架构提升了性能。尽管这些方法在几何估计精度上不断提升,但其输出始终是缺乏语义的抽象数值(如中心点、尺寸),无法生成可供人类直观理解的场景描述,形成了“机器友好”但“人类不友好”的语义鸿沟,极大地限制了其在人机交互等场景中的应用。

1.2　视觉描述生成

视觉描述生成任务旨在为图像或视频生成连贯的文本描述。早期的图像描述方法[19-20]主要基于CNN-RNN架构。随着Transformer的普及,诸如Captioning Transformer等模型[21-22]在多项基准测试中取得了成功。然而,这些工作主要聚焦于描述图像的二维视觉内容(如物体属性、表面动作),缺乏对三维空间几何信息(如精确距离、深度关系)的精确建模与描述能力。因此,现有描述模型难以直接应用于对空间精度要求极高的自动驾驶等场景,无法生成如“一辆轿车在右前方20米处”等具备几何准确性的描述。

1.3 图像空间关系理解

近年来，视觉与语言的交叉研究备受关注。以Visual Genome[23]、SpatialSense[24]等为代表的工作，致力于研究图像中物体间的二维空间关系理解（如“在…之上”“在…左边”）。其本质是基于外观的拓扑关系分类或短语定位任务，无需恢复物体在三维世界中的度量信息。而ReferIt3D[25]、Mono3DVG[26]、NLOT3D[27]等任务将自然语言指代与三维场景中的特定物体关联起来，属于定位式任务，即根据语言输入确定一个目标的位置。本文提出的Mono3DDC任务与上述工作均存在本质区别：它不仅是三维感知的，需要确保描述与三维空间几何严格一致，而且是任务统一的，要求模型同步完成高精度的三维目标检测与高质量的描述生成，对模型的几何理解与语言生成能力提出了前所未有的统一性要求。

2 KITTI3DDC数据集构建

当前，高质量多模态三维数据集的构建面临显著挑战。传统依赖人工标注的方案效率低下且成本高昂，而基于大语言模型的自动化生成方法又难以保证描述与三维空间信息的一致性，普遍存在描述失真问题，在确保准确性方面仍需大量人工干预。这些局限性严重制约了三维场景理解研究的深入发展。

本文受到基于模板的描述补全生成方法的启发，提出了一种自动化数据构建流程，如图2所示。该流程通过大模型进行描述生成，再构建对比模板进行质量验证。通过它们的紧密协作，将原始的三维物体标注自动化转换为准确、自然且具备空间感知的描述文本。系统采用模块化架构，通过标准化的API（Application Programming Interface）接口进行数据交换，在确保描述准确性的同时显著提升了构建效率。

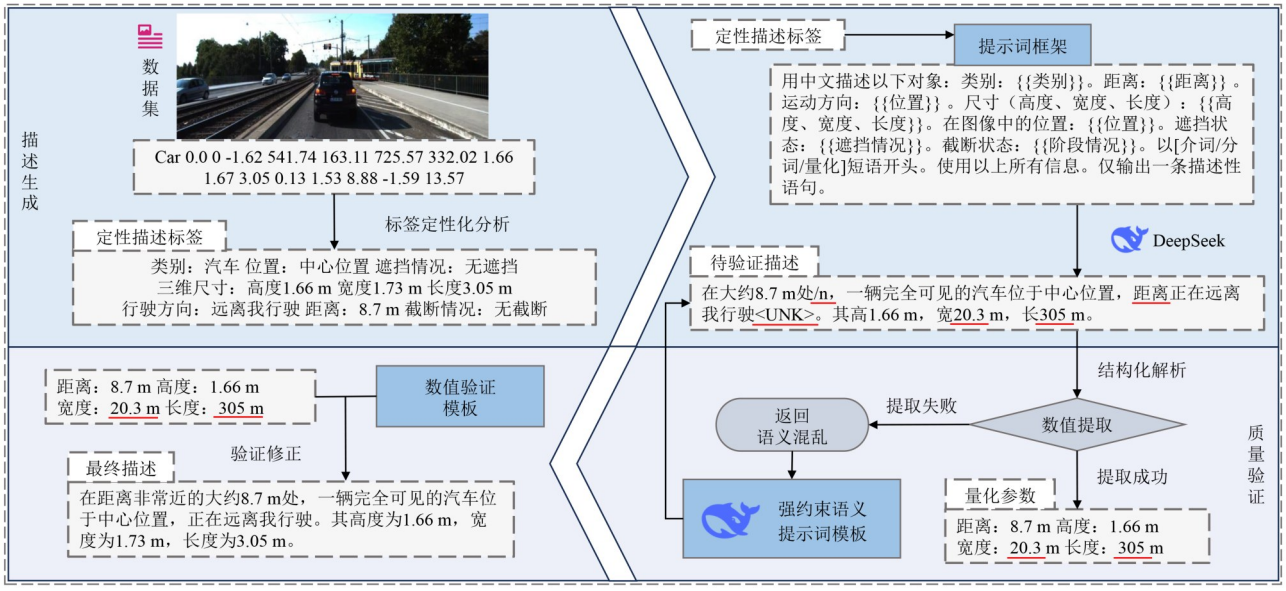

图2 KITTI3DDC数据集构建流程

Figure 2 KITTI3DDC dataset construction pipeline

2.1 描述生成

描述生成的核心目标是将结构化的物体属性转化为流畅的自然语言描述。该流程起始于数据解析阶段，系统从原始标注中精确提取物体的三维位置、物理尺寸、全局方向角等关键属性。有些属性并不直观，需要将其转化为类别、遮挡状态、位置、方向等人类易于理解的语义单元。接着，在精心设计的提示词框架指导下，系统调用DeepSeek模型[28]将这些离散的语义单元融合为连贯的场景描述。为确保后续验证的可靠性，系统会同步生成包含关键数值的验证模板，为质量评估提供明确依据。

为确保生成流程的鲁棒性与模型无关性，本文在开发阶段对包括DeepSeek-V3、GPT-3.5-Turbo和豆包在内的多种大语言模型进行了小规模对比实验。实验表明，在相同的结构化输入和约束性提示词下，不同模型均能有效完成从属性到描述的转换任务，其主要差异在于语言风格的丰富度与核心信息的忠实度。具体而言，GPT-3.5-Turbo生成的描述在连贯性上表现良好，但其原生训练语料以英文为主，导致生成的中文描述在词汇选择与表达逻辑上有所不足；豆包模型在语言表达的丰富性与流畅度上具有优势，但在严格的数值与空间关系约束下，其生成结果中偶尔会出现“空间幻觉”，例如对方向或相对位置的描述出现模糊或偏差。因此，在平衡信息准确性、语言质量、生

成效率与成本等方面的考量后,本文最终选用DeepSeek-V3作为核心生成引擎,以确保数据构建流程在可靠性、质量与可扩展性上达到最优平衡。

2.2　质量验证

质量验证流程负责确保生成描述的空间准确性和语义可靠性,其核心是一个严格的自动化验证与过滤机制,专门用于控制和消除大模型可能产生的"幻觉"。该流程首先对描述文本进行结构化解析,从中提取距离、尺寸、相对位置等关键量化参数。此解析过程通过检测文本是否符合预设逻辑规则,自动完成对语义清晰度与一致性的初步检查。随后,系统将这些提取出的参数与验证模板中的基准真值进行自动化比对。对于未通过流程验证的描述(约占总生成量的5%),系统会将其判定为"失败案例",并自动触发优化指令。分析表明,失败案例主要分为四类:数值篡改、信息遗漏或矛盾、语法结构异常、逻辑一致性冲突。这些案例不会直接进入数据集,而是进行优化,直至输出结果完全符合精度要求。

2.3　数据集统计与分析

表1展示了所构建的KITTI3DDC数据集的统计信息。该数据集包含来自KITTI[29]训练集的3 712张真实道路场景图像,涵盖19 083个标注物体,每个物体都配有对应的指代描述。数据集支持多类别与多目标场景,数据集的空间覆盖范围达到102米,充分体现了实际道路场景的深度多样性。在语言描述方面,数据集构建了包含1 854个核心词汇的词典,平均每条描述由41.71个词汇组成,确保了描述内容的丰富性和语言的自然流畅。所有描述均通过提出的自动化流程生成,在保证高多样性和低成本的同时,严格保证了描述与三维空间信息的一致性。KITTI3DDC提供了全面的二维、三维边界框和语义丰富的描述文本,为单目三维视觉描述研究提供了高质量的多模态监督信号。

表1　三维单目视觉相关任务的数据集统计比较

Table 1　Statistical comparison of datasets for 3D monocular vision related tasks

数据集	目标个数	距离/m	视觉组成	标签
Sr3d[25]	8 863	10	点云	3D
Multi3DRefer[30]	11 609	10	点云	3D
SUNRefer[31]	7 699	—	RGB-D	3D
LifeRefer[32]	11 864	30	点云&RGB	3D
Mono3DRefer[26]	8 228	102	RGB	2D&3D
KITTI3DDC (ours)	19 083	102	RGB	2D&3D&描述文本

3　本文方法

如图3所示,本文提出的Mono3DDC-TR框架采用端到端的统一架构,主要由三个核心组件构成:多尺度视觉特征编码器、多模态感知解码器、多任务协同学习。该框架允许模型在统一且连续的特征空间内同时优化检测与描述任务,通过跨模态梯度传递,实现了底层几何感知与高层语义生成的深度耦合。

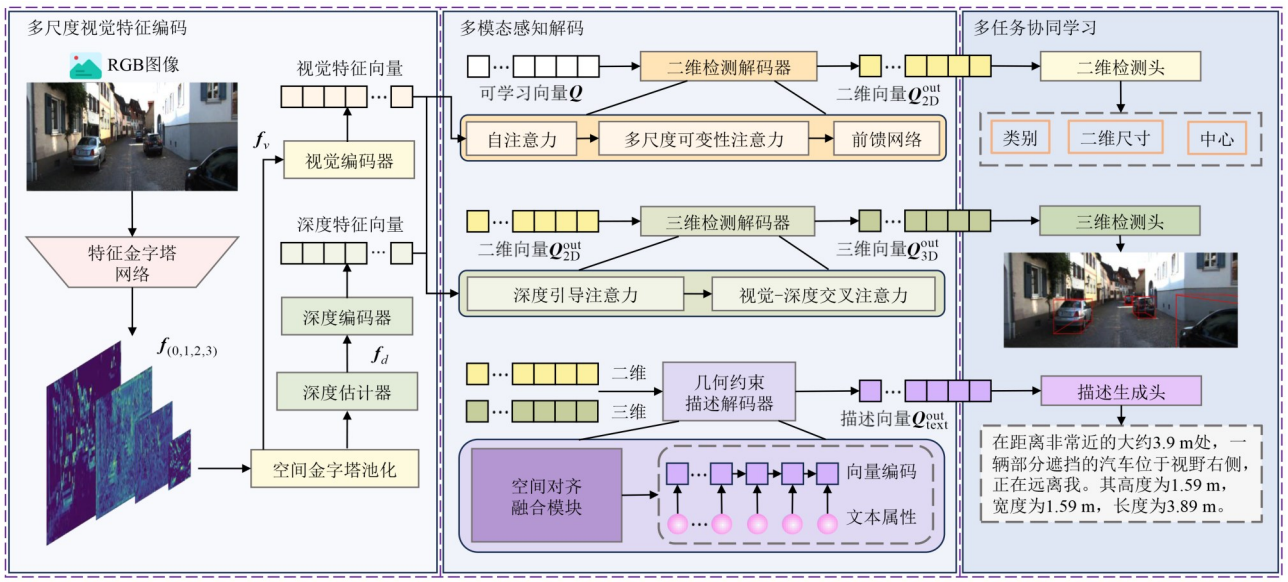

图3　Mono3DDC-TR网络框架图

Figure 3　Mono3DDC-TR network architecture diagram

3.1 多尺度视觉特征编码器

本文多尺度视觉特征编码器采用层次化特征提取策略，通过逐层融合不同尺度的特征，实现对视觉信息与深度特征的多尺度统一表示。给定输入图像 $\boldsymbol{I}\in\mathbb{R}^{H\times W\times 3}$，编码器首先通过特征金字塔网络提取四个尺度的特征映射 $\boldsymbol{f}_i\in\mathbb{R}^{\frac{H}{2^{i+3}}\times\frac{W}{2^{i+3}}\times C}$，其中 $i=0,1,2,3$。

在每个尺度上，本文设计的异构几何感知深度特征增强方法并行执行局部几何感知和全局上下文建模两个关键操作。对于局部几何感知，模块通过物体中心网络定位潜在物体区域，并利用深度回归器估计实例级深度分布，形成精确的物体空间约束：

$$\boldsymbol{D}^{(i)}_{\text{instance}}=\text{MiDas}\left(\text{Conv}_{3\times 3}\left(\boldsymbol{f}^{(i)}\right)\right)\odot\boldsymbol{M}^{(i)}_{\text{object}} \tag{1}$$

其中，$\boldsymbol{M}_{\text{object}}$ 表示通过物体中心网络生成的对象掩码，MiDas 为深度估计器[33]。同时，全局上下文建模分支通过空间金字塔池化捕获场景的布局信息，生成与尺度相适应的全局深度先验：

$$\boldsymbol{D}^{(i)}_{\text{layout}}=\text{SPP}\left(\boldsymbol{f}^{(i)}\right)=\text{Concat}\begin{pmatrix}\text{Pool}_{1\times 1}\left(\boldsymbol{f}^{(i)}\right)\\ \text{Pool}_{2\times 2}\left(\boldsymbol{f}^{(i)}\right)\\ \text{Pool}_{4\times 4}\left(\boldsymbol{f}^{(i)}\right)\end{pmatrix} \tag{2}$$

两种深度线索通过提出的深度分布注意力机制进行融合。该机制首先在通道维度对特征进行重组：

$$\boldsymbol{f}_r^{(i)}=\text{Reshape}\left(\boldsymbol{f}^{(i)}\right)\cdot\boldsymbol{W}_r \tag{3}$$

其中，$\boldsymbol{W}_r\in\mathbb{R}^{C\times C}$ 为可学习的通道重组矩阵。通过交叉参考加权单元计算通道间和空间位置的相互依赖关系：

$$\boldsymbol{A}_c=\text{Softmax}\left(\frac{\boldsymbol{Q}_c\boldsymbol{K}_c^{\text{T}}}{\sqrt{d_c}}\right),\boldsymbol{A}_s=\text{Softmax}\left(\frac{\boldsymbol{Q}_s\boldsymbol{K}_s^{\text{T}}}{\sqrt{d_s}}\right) \tag{4}$$

其中，$\boldsymbol{Q}_c$、$\boldsymbol{K}_c$、$\boldsymbol{Q}_s$、$\boldsymbol{K}_s$ 分别表示通道和空间维度的查询和键向量，d_c 和 d_s 为对应的维度缩放因子。深度分布注意力实现深度感知的特征增强的完整计算过程为

$$\begin{aligned}\boldsymbol{f}_d^{(i)}&=\text{DAttn}\left(\boldsymbol{D}^{(i)}_{\text{instance}},\boldsymbol{D}^{(i)}_{\text{layout}}\right)\\&=\gamma\cdot\left(\boldsymbol{A}_c\cdot\boldsymbol{f}_r^{(i)}\right)+(1-\gamma)\cdot\left(\boldsymbol{A}_s\odot\boldsymbol{f}_r^{(i)}\right)\end{aligned} \tag{5}$$

其中，γ 为自适应平衡参数。

随后对二维图像特征进行增强，在每个尺度上应用通道加权的注意力机制，通过像素级相乘深度处理得到的物体关键区域掩码来增强目标关键特征并抑制环境干扰：

$$\boldsymbol{f}_v^{(i)}=\text{Sigmoid}\left(\text{Conv}_{1\times 1}\left(\boldsymbol{f}^{(i)}\right)\right)\odot\boldsymbol{M}^{(i)}_{\text{instance}} \tag{6}$$

高层视觉特征经过上采样后与底层细节特征逐元素相加，同时引入门控机制控制信息流动，最终输出增强的多尺度视觉特征：

$$\boldsymbol{f}_v=\boldsymbol{f}_v^{(i)}+\boldsymbol{\gamma}^{(i)}_{\text{gate}}\cdot\text{Upsample}\left(\boldsymbol{f}_v^{(i+1)}\right) \tag{7}$$

为后续的检测与描述任务提供丰富的几何感知表示。

3.2 多模态感知解码器

多模态感知解码器负责将视觉特征和深度特征转化为具体的检测结果和描述内容，其核心在于建立视觉外观与空间几何的深度关联。本文设计了结构化特征解耦策略，获得 $\boldsymbol{f}_v$ 和 $\boldsymbol{f}_d$ 后，检测解码器分为两个分支以减少几何深度线索对二维检测的干扰。

二维感知分支专注于物体的视觉外观和类别信息。该分支以可学习查询向量 $\boldsymbol{Q}\in\mathbb{R}^{C\times N}$ 作为输入，通过多尺度可变形注意力机制从 $\boldsymbol{f}_v$ 中提取外观特征，其中 N 为检测过程中最大的查询数量。具体而言，查询向量与不同尺度的特征图进行交互，通过预测参考点和注意力权重，实现对不同大小物体的自适应感知：

$$\begin{aligned}&\boldsymbol{A}=\text{Softmax}\left(\boldsymbol{W}_a\boldsymbol{Q}_{\text{2D}}\right),\Delta\boldsymbol{p}=\text{MLP}\left(\boldsymbol{Q}_{\text{2D}}\right)\\&\boldsymbol{Q}^{\text{out}}_{\text{2D}}=\text{MSDeformAttn}\left(\boldsymbol{Q}_{\text{2D}},\boldsymbol{f}_v\right)=\sum_{i=1}^{N}\boldsymbol{A}_i\cdot\boldsymbol{f}_v\left(\boldsymbol{p}+\Delta\boldsymbol{p}_i\right)\end{aligned} \tag{8}$$

该分支输出包含物体类别、二维边界框的完整表征 $\boldsymbol{Q}^{\text{out}}_{\text{2D}}$。

三维几何分支致力于恢复物体的精确空间位置。将 $\boldsymbol{Q}^{\text{out}}_{\text{2D}}$ 作为此三维检测分支的先验指导专注于从深度感知特征中挖掘几何信息，同时利用增强的视觉特征 $\boldsymbol{f}_v$ 和深度特征 $\boldsymbol{f}_d$ 进行几何推理。分支首先通过深度引导的注意力机制定位潜在的物体中心：

$$\boldsymbol{Q}^{\text{init}}_{\text{3D}}=\text{DepthAttn}\left(\boldsymbol{Q}^{\text{out}}_{\text{2D}},\boldsymbol{f}_d\right) \tag{9}$$

同时，分支引入视觉-几何交叉注意力机制，将增强的视觉特征 $\boldsymbol{f}_v$ 与深度特征 $\boldsymbol{f}_d$ 进行融合，利用深度特征提供的几何信息和视觉特征提供的语义信息，为后续的三维属性回归提供丰富的特征表示。最终的三维几何表征通过前馈网络进一步精炼得到：

$$\boldsymbol{Q}^{\text{out}}_{\text{3D}}=\text{FFN}\left(\text{CrossAttn}\left(\boldsymbol{Q}^{\text{init}}_{\text{3D}},\boldsymbol{f}_v\right)\right) \tag{10}$$

为了将稀疏的二维和三维特征与稠密的语言语义进行端到端对齐，本文设计了空间对齐融合模块来连接二维外观与三维几何特征，如图 4 所示。

模块采用双路径编码架构，分别处理 $\boldsymbol{Q}^{\text{out}}_{\text{2D}}$ 和 $\boldsymbol{Q}^{\text{out}}_{\text{3D}}$。在每条路径内部，首先通过自注意力机制增强特征的内聚性：

$$\begin{aligned}\boldsymbol{Q}^{\text{self}}_{\text{2D}}&=\text{SelfAttn}\left(\boldsymbol{Q}^{\text{out}}_{\text{2D}}\right)\\\boldsymbol{Q}^{\text{self}}_{\text{3D}}&=\text{SelfAttn}\left(\boldsymbol{Q}^{\text{out}}_{\text{3D}}\right)\end{aligned} \tag{11}$$

随后，通过双向交叉注意力建立外观与几何的对应关系。视觉到几何的注意力路径使三维空间感知物体的外观特征，而几何到视觉的注意力路径则让外观表征感知空间约束。两个路径的输出通过自适应

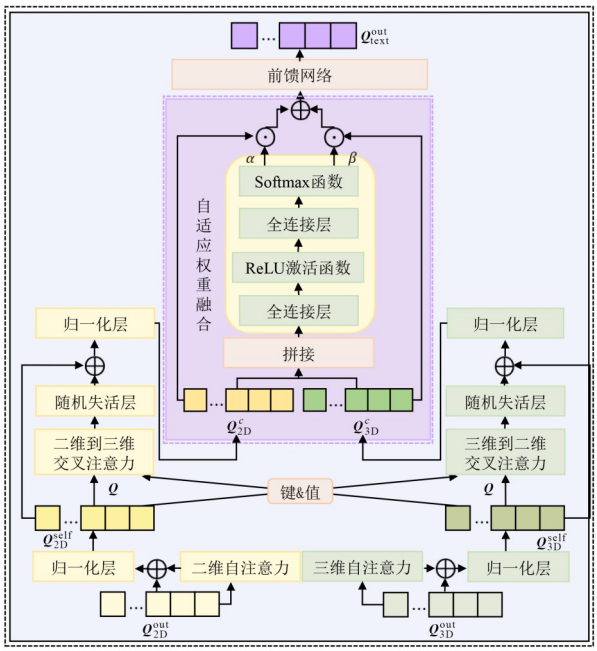

图4　空间对齐融合模块网络架构

Figure 4　Spatial alignment fusion module network architecture

权重进行融合:

$$[\alpha,\beta]=\mathrm{Softmax}\left(\mathrm{MLP}\left(\left[\boldsymbol{Q}_{2D}^{\mathrm{cross}},\boldsymbol{Q}_{3D}^{\mathrm{cross}}\right]\right)\right)$$
$$\boldsymbol{Q}_{\mathrm{text}}=\alpha\cdot\boldsymbol{Q}_{2D}^{\mathrm{self}}+\beta\cdot\boldsymbol{Q}_{3D}^{\mathrm{self}} \tag{12}$$

最终,融合后的特征经过前馈网络进一步精炼,输出统一的多模态表征 $\boldsymbol{Q}_{\mathrm{text}}^{\mathrm{out}}$,为描述生成提供兼具视觉区分度和几何准确性的特征基础。

3.3　多任务协同学习

本文采用多任务协同学习框架,通过共享特征表示和任务特异性处理的结合,实现检测与描述的统一优化。框架包含多个预测头,每个头针对特定任务进行优化,同时通过联合训练促进任务间的正向迁移。

深度感知预测头统一处理实例级深度分布和布局级深度先验表示,从视觉特征中提取关键的深度线索,构建场景的连贯深度表示。具体而言,头部分析特征图的空间响应,识别包含丰富深度信息的区域,并通过轻量级卷积网络回归深度值及其置信度。为避免深度估计的尺度模糊性,头部分支预测尺度感知的深度参数,确保预测结果与真实物理尺度一致。

检测头采用统一架构处理二维和三维检测任务。对于二维检测,头部分支预测物体类别、二维边界框参数 (l,r,t,b)。对于三维检测,在二维检测的基础上,额外回归物体的三维属性,包括相对于图像平面的三维中心 (x_{3D},y_{3D})、物理尺寸 (h_{3D},w_{3D},l_{3D})、朝向角 θ(通过多区间分类-回归联合优化)以及绝对深度 d_{pred}。为提升深度估计的鲁棒性,三维检测分支引入拉普拉斯不确定度建模,为每个预测提供可靠性评估。

为确保语言描述与预测几何严格一致,本文设计了几何约束描述头将多模态特征转化为自然语言描述。该头基于Transformer解码器架构,将空间对齐融合模块输出的已对齐几何信息特征 $\boldsymbol{Q}_{\mathrm{text}}^{\mathrm{out}}$ 作为条件输入,通过掩码自注意力机制自回归地生成单词序列。为确保描述的空间准确性,头内部集成了几何约束模块。该模块通过查询三维检测结果,将空间关系转化为具体的方位描述。此外,头内部设计了一个场景专用的符号词典,将连续的三维空间关系离散化为语义概念,有效避免描述中的空间模糊性。

联合优化目标将各任务的损失函数统一为多任务学习框架。总体损失函数定义为

$$L_{\mathrm{total}}=\lambda_1 L_{\mathrm{depth}}+\lambda_2 L_{2D}+\lambda_3 L_{3D}+\lambda_4 L_{\mathrm{caption}} \tag{13}$$

其中,二维检测损失 L_{2D} 整合了焦点损失(类别)、L1损失(边界框)和GIoU损失(空间重叠),三维检测损失 L_{3D} 包含三维IoU导向的尺寸损失、多区间朝向损失和不确定性深度损失,深度预测损失 L_{depth} 采用尺度不变对数误差,描述生成损失 L_{caption} 使用标签平滑的交叉熵。通过端到端的联合训练,各任务间形成正向促进,最终实现几何感知与语义生成的统一优化。

4　实验

4.1　实验配置

本研究在KITTI3DDC基准数据集上进行全面验证,该数据集包含7 481张真实道路场景图像。按照其他三维目标检测任务的标准划分,保证实验结果对比时的公平性。训练集包含3 712张图像,验证集包含3 769张图像,对应的指代描述共有19 083条。数据集覆盖汽车、行人、骑行者三个主要交通参与类别,并按视觉遮挡程度和边界框尺寸划分为三个难度等级。所有实验均在PyTorch框架下使用单张NVIDIA Tesla V100 GPU进行,使用Transformer架构处理三维目标检测和文本生成任务。训练过程持续250个周期,批次规模为16。优化器选用AdamW,初始学习率设为 2×10^{-4},权重衰减系数为 1×10^{-4}。学习率调度采用分段衰减策略,每完成50个训练周期后将学习率减半。为平衡训练效率与生成质量,设定置信度阈值为0.2,仅对超过该阈值的检测目标进行描述生成。

4.2　评估方法

为全面评估Mono3DDC任务的性能,本研究构建了多维度的评价体系,涵盖几何感知准确性和语义生成质量两大方面。

三维检测评价指标采用平均精度均值(mean Average Precision,mAP)。该指标通过计算不同交并比阈值下的精度-召回曲线面积,综合反映检测算法的

定位准确性和分类能力。对于单目三维检测任务，本文重点关注三种难度等级下的$AP_{3D|R40}$（基于40个召回点采样的三维平均精度）。语义生成质量的评估需要先对描述进行标准化处理。具体而言，将所有数字转换为中文汉字（例如将“20米”转换为“二十米”）。这种转换确保了数值信息被纳入文本匹配的考量范围。由于参考描述由真实三维标注通过确定性规则生成，具备几何真值，后面所使用的文本指标能直接反映生成描述在数值准确性上的表现。BLEU指标通过比较生成描述与参考描述之间的n-gram重叠度，评估生成文本的词汇选择和流畅性。经过数字转换预处理后，BLEU（BiLingual Evaluation Understudy）指标不仅能衡量一般词汇的匹配程度，还能直接捕捉数值表述的精确匹配情况。例如，“一点七八米”与“二十米”完全不匹配，而与“一点七一米”部分匹配。其计算基于精确度加权几何平均：

$$\mathrm{BLEU}=\mathrm{BP}\cdot\exp\left(\sum_{n=1}^{N}w_n\log p_n\right) \tag{14}$$

其中，p_n表示n-gram精确度，BP为长度惩罚因子。ROUGE-L指标则侧重评估生成描述的语义覆盖度和内容完整性，通过计算最长公共子序列来度量生成文本与参考文本的语义相似性：

$$\begin{aligned}R_{\mathrm{lcs}}&=\frac{\mathrm{LCS}(X,Y)}{m}\\P_{\mathrm{lcs}}&=\frac{\mathrm{LCS}(X,Y)}{n}\\F_{\mathrm{lcs}}&=\frac{(1+\beta^2)R_{\mathrm{lcs}}P_{\mathrm{lcs}}}{R_{\mathrm{lcs}}+\beta^2P_{\mathrm{lcs}}}\end{aligned} \tag{15}$$

其中，$\mathrm{LCS}(X,Y)$表示序列X和Y的最长公共子序列长度。经过预处理后，包含准确数值的最长公共子序列将被有效识别，使ROUGE-L指标同样能反映数值信息的完整性和顺序正确性。通过这两个指标的互补评估，可以确保生成描述既自然流畅又能在空间信息上准确可靠。

考虑到Mono3DDC是全新提出的任务，现有方法均未直接支持检测与描述的联合输出。为建立公平的对比基准，本文采用统一的适配框架将传统三维检测方法扩展至描述生成任务。具体而言，每个基准方法（包括基于激光雷达、深度图和单目视觉的检测器）提取其倒数第二层的特征表示，接入统一的通用VLM模型[34-35]生成描述。在相同的KITTI3DDC数据集上训练，确保描述生成模块的一致性。通过这种方式，各基准方法能够在相同条件下生成场景描述，进而使用BLEU和ROUGE-L指标进行量化比较。这样既可以保留各检测方法在几何感知方面的特性差异，也可以看出本文设计的描述生成头和一般通用VLM模型的差别，从而确保比较的公平性和可靠性。

4.3 实验结果分析

为验证所提出方法的有效性，本文在KITTI测试集上开展了系统性的定量评估，将Mono3DDC-TR与近年来具有代表性的单目三维检测方法进行了全面比较。如表2所示，Mono3DDC-TR在三类目标（车辆、行人、骑行者）及全部三个难度等级上均取得了最优或极具竞争力的$AP_{3D|R40}$表现。其中，在车辆类别上，Mono3DDC-TR在简单、中等和困难条件下分别达到27.36、19.63和15.54，相较于当前最优基线MonoDGP[12]实现了稳定且持续的性能提升，验证了本文提出的跨模态三维一致性建模策略在强几何约束任务中的有效性。在行人和骑行者两类目标上，Mono3DDC-TR亦展现出一致优势：行人在简单、中等和困难条件下分别达到14.17、10.68和8.68，骑行者则分别达到12.32、5.19和4.46。在困难场景下的显著提升尤其值得注意，说明该方法在应对尺度变化大、姿态复杂等挑战性场景时具备更高的鲁棒性与泛化能力。

除三维检测性能外，本文进一步评估了模型在联合描述任务中的语言生成质量。Mono3DDC-TR在BLEU和ROUGE-L指标上分别取得0.249 1与0.503 0的最佳成绩，明显优于所有使用通用VLM模型生成描述文本的对比方法。这表明所提出的三维语义引导模块能够有效建模场景空间结构，本文设计的描述生成头比通用VLM模型能够生成更加精细、连贯且与几何一致的文本描述。

4.4 实验结果可视化

为深入评估本文所提方法的实际效能，本文在KITTI验证集上分别从相机视图与雷达视图综合对比了不同方法的检测性能与描述生成质量。如图5所示，视觉对比结果清晰地揭示了Mono3DDC-TR模型在多方面均展现出显著优势。

从三维检测可视化结果来看，Mono3DDC-TR在多个关键维度表现突出。在边界框精度方面，Mono3DDC-TR预测的3D边界框在尺寸、位置和方向等多个几何属性上与真实标注高度吻合。特别是在远处小目标、严重遮挡物体等挑战性场景中，模型仍能保持稳定的检测性能。相比之下，MonoDETR[7]在这些场景中无法正确检测出目标物体。而MonoDGP[12]方法虽然没有出现漏检的情况，但在深度估计方面存在明显偏差，其预测的边界框在Z轴方向上出现系统性误差。

在描述生成方面，不同方法之间存在显著差异。基于通用VLM的基线模型生成的描述常常出现距离幻觉、空间关系误判、运动状态描述不准确等问题。例如，在用MonoDGP方法处理第一张图片时，VLM模型将“正在靠近的车辆”错误描述为“远离的车辆”，

表2　不同单目三维检测方法在KITTI数据集上的性能比较

Table 2　Performance comparison of different monocular 3D detection methods on the KITTI dataset

方法	车辆 $AP_{3D\|R40}$			行人 $AP_{3D\|R40}$			骑行者 $AP_{3D\|R40}$			BLEU	ROUGE-L
	简单	中等	困难	简单	中等	困难	简单	中等	困难		
OccupancyM3D[36]	25.55	17.02	14.79	14.68	9.15	7.90	7.37	3.56	2.84	0.095 9	0.382 4
MonoPGC[37]	24.68	17.17	14.14	14.16	9.67	8.26	5.88	3.30	2.85	0.094 7	0.378 4
MonoJSG[10]	24.69	16.14	13.64	11.02	7.49	6.41	5.45	3.21	2.57	0.096 2	0.381 5
MonoGround[8]	21.37	14.36	12.62	12.37	7.89	7.13	4.62	2.68	2.53	0.095 9	0.384 2
MonoCon[9]	22.50	16.46	13.95	13.10	8.41	6.94	2.80	1.92	1.55	0.096 0	0.384 1
MonoDTR[38]	21.99	15.39	12.73	**15.33**	10.18	8.61	5.05	3.27	3.19	0.024 2	0.280 7
MonoDETR[7]	25.00	16.47	13.58	12.65	7.19	6.72	5.12	2.74	2.02	0.099 5	0.386 1
MonoDDE[39]	24.93	17.14	15.10	11.13	7.32	6.67	5.94	3.78	3.33	0.093 5	0.375 4
MonoDGP[12]	26.35	18.72	**15.97**	15.04	9.89	8.38	5.28	3.61	3.22	0.097 2	0.379 2
Mono3DDC-TR (ours)	**27.36** (+0.99)	**19.63** (+0.91)	15.54 (-0.43)	14.17 (-1.16)	**10.68** (+0.50)	**8.68** (+0.07)	**12.32** (+4.95)	**5.19** (+1.41)	**4.46** (+1.13)	**0.249 1** (+0.149 6)	**0.503 0** (+0.116 9)

注: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。

显示出对动态场景理解的局限性。相较之下,几何约束描述头生成的描述在准确概括场景内容与目标属性方面表现更佳,语义连贯性更强,体现出对场景更深层次的理解。

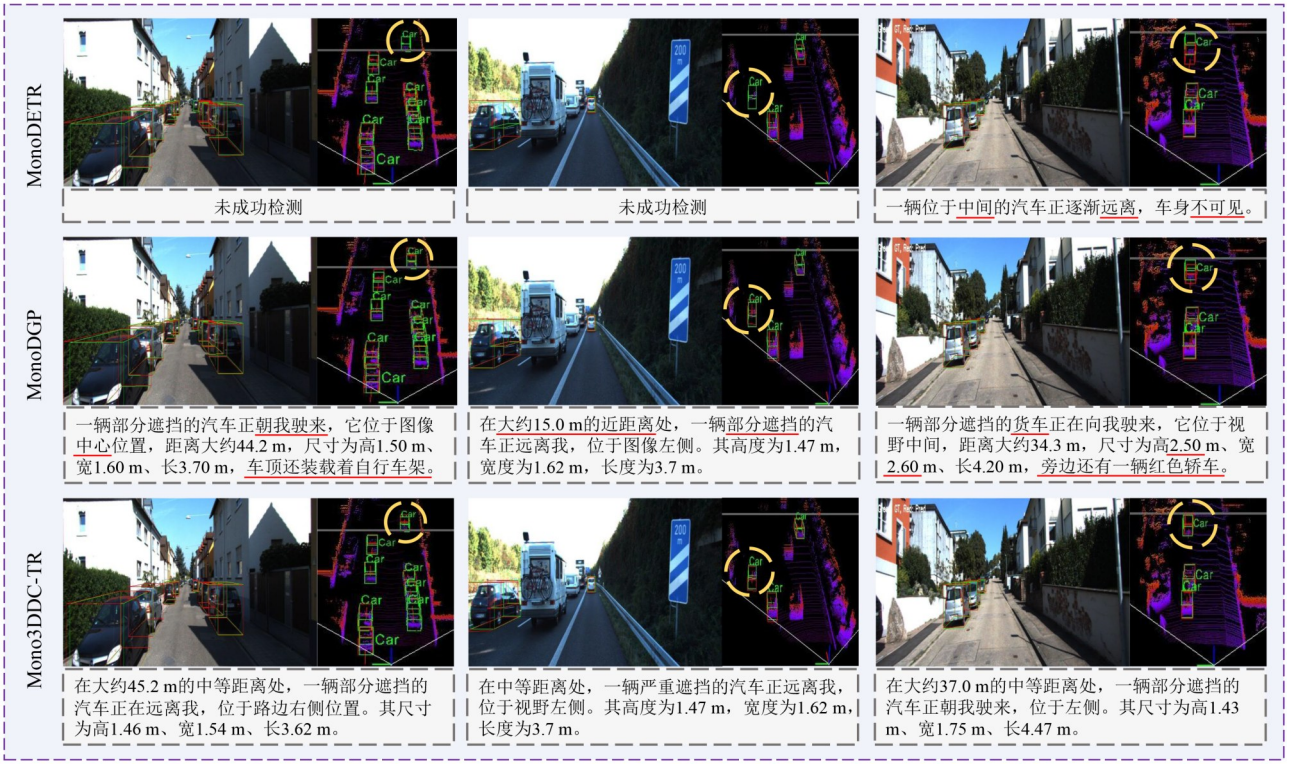

图5　Mono3DDC-TR与其他方法可视化对比结果图

Figure 5　Visualization comparison results of Mono3DDC-TR and other methods

4.5　消融实验

本文进一步对模型中的关键组件进行了系统性的消融实验,结果如表3所示。消融内容包括实例级深度、布局级深度、空间对齐融合模块、几何约束描述头对整体性能的影响。所有实验均在KITTI验证集上进行,并展示车辆类别在不同难度下的 $AP_{3D|R40}$、BLEU与ROUGE-L指标。

实例级与布局级深度。在不使用深度信息的情况下,模型的基线性能为25.57、18.16、15.39。加入实例级深度后,$AP_{3D|R40}$ 明显提升,说明实例层面的几何

表3 关键组件在KITTI验证集上的消融实验结果

Table 3 Ablation study results of key components on the KITTI validation set

实例级深度	布局级深度	空间对齐融合模块	▲几何约束描述头 ■通用VLM模型	车辆 $AP_{3D\|R40}$			BLEU	ROUGE-L
				简单	中等	困难		
×	×	×	■	25.57	18.16	15.39	0.095 1	0.380 8
√	×	×	■	26.69	19.47	15.63	0.097 5	0.384 0
√	√	×	■	27.26	19.98	16.37	0.094 3	0.377 7
√	√	√	■	28.97	21.46	17.48	0.109 6	0.401 6
√	√	×	▲	28.46	21.05	17.26	0.235 3	0.497 0
√	×	√	▲	28.21	20.83	16.92	0.249 1	0.503 0
×	√	√	▲	27.94	21.12	17.05	0.258 5	0.504 4
√	√	√	▲	**30.15**	**22.36**	**18.54**	**0.263 2**	**0.515 9**

注：加粗表示每个指标的最佳性能值。

约束能够为目标尺度与姿态提供更准确的先验。进一步引入布局级深度后，模型在三种难度下均获得稳定增益，这表明全局结构信息对三维空间推理同样至关重要。

空间对齐融合模块。在引入空间对齐融合模块后，模型在BLEU和ROUGE-L分数上均实现稳定提升，进一步证明显式的跨尺度空间对齐对三维几何结构建模具有关键作用。该模块通过整合多尺度区域特征，使网络能够更准确地捕获图像特征与几何特征的关系，从而有效增强描述文本的几何准确性。

几何约束描述头。在不同深度配置下加入通用VLM模型后，能够得到一个基本的描述生成文本，表明通用视觉语言模型能够提供额外的语义先验。然而，该模块并不能保证生成描述的几何准确性，说明仅依赖大模型的语义理解难以精确描述空间结构。相比之下，加入几何约束描述头后，模型在语言指标上获得很大提升，其中BLEU和ROUGE-L分别达到0.263 2和0.515 9。这说明几何一致性的跨模态监督能够提升对目标三维描述的准确性，其对结构化场景的建模作用远强于通用VLM模型。

完整模型性能。当实例级深度、布局级深度、空间对齐模块与几何约束描述头全部启用后，模型在车辆类别上取得最佳结果：$AP_{3D|R40}$分别为30.15、22.36、18.54，BLEU和ROUGE-L分别达到0.263 2和0.515 9。相较于初始基线，在三种难度下分别提升了4.58、4.20和3.15，充分验证了各模块的有效性和互补性。

在描述生成过程中，如果为每一个检测到的目标都生成文本描述，将会显著降低训练与推理的效率。因此，本文采用匈牙利匹配算法，仅为检测置信度高于预设阈值的目标生成描述。为分析该置信度阈值对最终性能的影响，本文在KITTI验证集上进行了如表4所示的消融实验。实验结果表明，当阈值设置为0.2时，在所有难度等级上均取得最佳表现。当阈值过低时，过量的匹配目标会引入噪声，从而在训练过程中传播错误信息。相反，当阈值设置过高时，可生成描述的目标数量减少，导致训练信号不足，限制模型的全面学习能力。因此，该阈值在一定程度上控制了视觉-文本对齐的可靠性，并对描述质量产生直接影响。

表4 几何约束描述头置信度阈值在KITTI验证集上的消融实验结果

Table 4 Ablation study results of confidence threshold for geometric constraint description head on the KITTI validation set

几何约束描述头置信度阈值	车辆 $AP_{3D\|R40}$			BLEU	ROUGE-L
	简单	中等	困难		
0.1	29.36	21.67	17.53	0.243 2	0.501 5
0.2	**30.15**	**22.36**	**18.54**	**0.263 2**	**0.515 9**
0.3	29.13	21.82	17.79	0.256 0	0.507 5
0.4	28.71	21.03	16.97	0.238 5	0.491 4

注：加粗表示每个指标的最佳性能值。

5 结论

本文提出了Mono3DDC这一创新任务，首次将单目三维目标检测与视觉描述生成相统一。为突破该研究领域的数据瓶颈，本文设计了一套高效的自动化数据生成流程，构建了首个支持中文语义的单目三维视觉描述基准数据集，命名为KITTI3DDC。针对任务特性，本文提出了Mono3DDC-TR统一网络框架，通过异构感知几何编码器和空间对齐融合模块的协同设计，实现了三维几何特征与二维视觉线索的深度耦合。实验结果表明，该方法在三维检测精度和语义描述质量方面均取得了显著提升。在KITTI3DDC基准上的综合评估显示，本文的方法在保持高检测精度的同时，生成的描述文本在流畅性和几何准确性方面均优于现有方法。特别是在多类别检测任务中，模型展现出了强大的泛化能力，为复杂场景下的环境感知提供了新的解决方案。

本研究为单目视觉理解与自然语言生成的交叉融合提供了新的思路,具有较强的理论意义和应用价值。未来工作将围绕以下几个方向展开:首先,探索更加高效的几何特征表示方法,进一步提升模型在极端场景下的鲁棒性;其次,扩展数据集的规模和多样性,涵盖更复杂的交通场景和天气条件;最后,推动该方法在实时自动驾驶系统和智能人机交互中的实际应用,为实现更加智能的环境感知与决策提供技术支持。本文的研究为三维视觉与语言理解的深度融合开辟了新的途径,在自动驾驶、智能交通等领域具有广阔的应用前景。

参考文献

[1] 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.

[2] 李昌财, 陈刚, 侯作勋, 等. 自动驾驶中的三维目标检测算法研究综述[J]. 中国图象图形学报, 2024, 29(11): 3238-3264.
Li Changcai, Chen Gang, Hou Zuoxun, et al. Survey of 3D object detection algorithms for autonomous driving[J]. Journal of Image and Graphics, 2024, 29(11): 3238-3264. (in Chinese)

[3] Han D, Mulyana B, Stankovic V, et al. A survey on deep reinforcement learning algorithms for robotic manipulation[J]. Sensors, 2023, 23(7): 3762.

[4] Qin Y R, Wang C Q, Kang Z J, et al. SupFusion: Supervised LiDAR-camera fusion for 3D object detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 22014-22024.

[5] Yin J B, Zhou D F, Zhang L J, et al. ProposalContrast: Unsupervised pre-training for LiDAR-based 3D object detection[M]//Computer Vision - ECCV 2022. Cham: Springer Nature Switzerland, 2022: 17-33.

[6] Shi P C, Liu Z Q, Dong X L, et al. CL-fusionBEV: 3D object detection method with camera-LiDAR fusion in Bird's Eye View[J]. Complex & Intelligent Systems, 2024, 10(6): 7681-7696.

[7] Zhang R R, Qiu H, Wang T, et al. MonoDETR: Depth-guided transformer for monocular 3D object detection[C]// Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 9155-9166.

[8] Qin Z Q, Li X. MonoGround: Detecting monocular 3D objects from the ground[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 3793-3802.

[9] Liu X P, Xue N, Wu T F. Learning auxiliary monocular contexts helps monocular 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 1810-1818.

[10] Lian Q, Li P L, Chen X Z. MonoJSG: Joint semantic and geometric cost volume for monocular 3D object detection [C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 1070-1079.

[11] Yan L F, Yan P, Xiong S Z, et al. MonoCD: Monocular 3D object detection with complementary depths[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 10248-10257.

[12] Pu F Q, Wang Y F, Deng J R, et al. MonoDGP: Monocular 3D object detection with decoupled-query and geometry-error priors[C]//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 6520-6530.

[13] Wu Z Z, Gan Y Z, Wu Y Z, et al. FD3D: Exploiting foreground depth map for feature-supervised monocular 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(6): 6189-6197.

[14] Legaspi R, Xu W Z, Konishi T, et al. The sense of agency in human-AI interactions[J]. Knowledge-Based Systems, 2024, 286: 111298.

[15] 魏忠钰, 范智昊, 王瑞泽, 等. 从视觉到文本: 图像描述生成的研究进展综述[J]. 中文信息学报, 2020, 34(7): 19-29.
Wei Zhongyu, Fan Zhihao, Wang Ruize, et al. From vision to text: A brief survey for image captioning[J]. Journal of Chinese Information Processing, 2020, 34(7): 19-29. (in Chinese)

[16] 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)

[17] 葛同澳, 李辉, 郭颖, 等. 基于双融合框架的多模态3D目标检测算法[J]. 电子学报, 2023, 51(11): 3100-3110.
Ge Tongao, Li Hui, Guo Ying, et al. A multimodal 3D object detection method based on double-fusion framework[J]. Acta Electronica Sinica, 2023, 51(11): 3100-3110. (in

Chinese)

[18] 周治国, 马文浩. 一种多层多模态融合 3D 目标检测方法[J]. 电子学报, 2024, 52(3): 696-708.
Zhou Zhiguo, Ma Wenhao. 3D object detection based on multilayer multimodal fusion[J]. Acta Electronica Sinica, 2024, 52(3): 696-708. (in Chinese)

[19] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille: PMLR, 2015: 2048-2057.

[20] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 3156-3164.

[21] Zhang J, Xie Y S, Ding W C, et al. Cross on cross attention: Deep fusion transformer for image captioning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4257-4268.

[22] Wang Y Y, Xu J G, Sun Y F. End-to-end transformer based model for image captioning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 2585-2594.

[23] Krishna R, Zhu Y K, Groth O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73.

[24] Yang K Y, Russakovsky O, Deng J. SpatialSense: An adversarially crowdsourced benchmark for spatial relation recognition[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 2051-2060.

[25] 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.

[26] 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.

[27] 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)

[28] Liu Aixin, Feng Bei, Xue Bing, et al. DeepSeek-V3 technical report[PP/OL]. V2. arXiv (2024-12-27) [2025-11-08]. https://arXiv.org/abs/2412.19437.

[29] 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.

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

[31] 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 6032-6041.

[32] Lin Z X, Peng X D, Cong P S, et al. WildRefer: 3D object localization in large-scale dynamic scenes with multimodal visual data and natural language[C]//Proceedings of the 18th European Conference on Computer Vision. Heidelberg: Springer, 2025: 456-473.

[33] Ranftl R, Lasinger K, Hafner D, et al. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(3): 1623-1637.

[34] Wei H R, Kong L Y, Chen J Y, et al. Vary: Scaling up the vision vocabulary for large vision-language model[C]//Proceedings of the 18th European Conference on Computer Vision. Heidelberg: Springer, 2025: 408-424.

[35] Chu X X, Qiao L M, Zhang X Y, et al. MobileVLM V2: Faster and stronger baseline for vision language model[PP/OL]. V1. arXiv (2024-02-06)[2025-12-07]. https://arXiv.org/abs/2402.03766.

[36] Peng L, Xu J K, Cheng H R, et al. Learning occupancy for monocular 3D object detection[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 10281-10292.

[37] Wu Z Z, Gan Y Z, Wang L, et al. MonoPGC: Monocular 3D object detection with pixel geometry contexts[C]//Proceedings of 2023 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2023: 4842-4849.

[38] Huang K C, Wu T H, Su H T, et al. MonoDTR: Monocular 3D object detection with depth-aware transformer[C]//

Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 4012-4021.

[39] Li Z L, Qu Z, Zhou Y, et al. Diversity matters: Fully exploiting depth clues for reliable monocular 3D object detection[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 2791-2800.

作者简介

王天添　男,2002年8月出生于湖南省娄底市。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多目标检测、多模态学习。

E-mail: wangtiantian@chd.edu.cn

郭柯宇　男,1999年9月出生于贵州省黔南布依族苗族自治州。现为长安大学信息工程学院博士研究生。主要研究方向为视觉定位、多模态学习。

E-mail: nymph@uestc.edu.cn

罗函轲　女,2002年3月出生于陕西省咸阳市。现为长安大学信息工程学院硕士研究生。主要研究方向为3D目标检测、文本/图像到3D生成。

E-mail: snowwhite@uestc.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

程惠泽　男,2002年11月出生于陕西省西安市。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和多模态研究。

E-mail: hzcheng@chd.edu.cn

孙张龙　男,2003年1月出生于安徽省安庆市。现为长安大学信息工程学院硕士研究生。主要研究方向为视觉语言分割。

E-mail: 2024224109@chd.edu.cn