第 14 卷 第 3 期
2025 年 5 月

集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY

Vol. 14 No. 3
May 2025

引文格式：
张胜, 程俊. 基于上下文信息和大语言模型的开放词汇室内三维目标检测 [J]. 集成技术, 2025, 14(3): 51-63.
Zhang S, Cheng J. Contextual information and large language model for open-vocabulary indoor 3D object detection [J]. Journal of Integration Technology, 2025, 14(3): 51-63.

基于上下文信息和大语言模型的开放词汇室内三维目标检测

张 胜[1] 程 俊[1,2*]

[1](中国科学院深圳先进技术研究院 中国科学院人机智能协同系统重点实验室 深圳 518055)
[2](香港中文大学 香港 999077)

摘 要 现有室内三维目标检测算法能检测的目标类别通常有限，这限制其在智能机器人领域的应用。开放词汇目标检测能在不定义目标类别的前提下检测给定场景的所有感兴趣目标，从而解决室内三维目标检测的不足。与此同时，大语言模型的先验知识能显著提升视觉任务的性能。然而现有的开放词汇室内三维目标检测研究存在仅关注目标信息，而忽视了上下文信息的问题。室内三维目标检测的输入数据主要是点云，存在稀疏和噪声问题。仅依赖目标点云信息会对三维目标检测结果产生负面影响。上下文信息包含场景描述，能对目标信息进行补充，从而提升目标检测中类别判定的准确率。为此，本文提出了基于上下文信息和大语言模型的开放词汇室内三维目标检测算法，该算法通过结合上下文信息和大语言模型的思维链推理获取检测结果。此外，该算法在 SUN RGB-D 和 ScanNetV2 数据集上进行了验证，实验结果表明了其有效性。

关键词 大语言模型；室内三维目标检测；开放词汇；上下文信息；思维链
中图分类号 TP183 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20241201003
CSTR: 32239.14.j.issn.2095-3135.20241201003

Contextual Information and Large Language Model for Open-Vocabulary Indoor 3D Object Detection

ZHANG Sheng[1] CHENG Jun[1,2*]

[1](*CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems*, *Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*The Chinese University of Hong Kong*, *Hong Kong* 999077, *China*)
[*]*Corresponding Author*: *jun.cheng@siat.ac.cn*

Abstract Existing indoor three-dimensional (3D) object detection is able to detect a limited number of object categories, thus limiting the application on intelligent robotics. Open-vocabulary object detection is able

收稿日期：2024-12-01 **修回日期：**2025-01-11
基金项目：国家自然科学基金项目 (U21A20487)；深圳市科技计划项目 (JCYJ20220818101206014)
作者简介：张胜，博士，助理研究员，研究方向为大语言模型、计算机视觉等； 程俊 (通讯作者)，博士，研究员，博士生导师，研究方向为大语言模型、计算机视觉、机器人等，E-mail：jun.cheng@siat.ac.cn。

to detect all objects of interest in a given scene without defining object categories, thus solving the shortcomings of indoor 3D object detection. At the same time, the large language model with prior knowledge can significantly improve the performance of visual tasks. However, existing researches on open-vocabulary indoor 3D object detection only focuses on object information and ignores contextual information. The input data for indoor 3D object detection is mainly point cloud, which suffers from sparsity and noise problems. Relying only on the object point cloud can negatively affect the 3D detection results. Contextual information contains scene information, which can complement the object information to promote the recognition on object category. For this reason, this paper proposes an open-vocabulary 3D object detection algorithm based on contextual information assistance. The algorithm integrates contextual information and object information through a large language model, and then performs chain-of-thought reasoning. The proposed algorithm is validated on SUN RGB-D and ScanNetV2 datasets, and the experimental results show the effectiveness of the proposed algorithm.

Keywords large language model; indoor 3D object detection; open-vocabulary; contextual information; chain of thoughts

Funding This work is supported by National Natural Science Foundation of China (U21A20487), Shenzhen Technology Project (JCYJ20220818101206014)

1 引 言

目标检测是计算机视觉领域的一个重要任务，旨在确定目标的位置和识别目标的类别[1]。随着深度学习和三维视觉技术的发展，三维目标检测取得了巨大进步，在工业制造和家居服务中得到广泛应用。三维目标检测的应用场景可分为室内场景和室外场景，其中室内三维目标检测被广泛应用于智能机器人领域。现有的关于室内三维目标检测的算法主要基于点云实现三维目标的定位和识别，其方法包括基于投票的方法、基于三维卷积的方法和基于 Transformer 的方法[2]。但是现有的室内三维目标检测算法通常仅能检测有限的目标类别，这限制了其应用范围。开放词汇目标检测能在不定义目标类别的前提下检测给定场景的所有感兴趣目标，从而解决室内三维目标检测的不足[3]。开放词汇室内三维目标检测能拓展检测类别，具有提升智能机器人感知能力的潜在价值。

点云[4]是室内三维目标检测中广泛应用的数据格式，通过处理点云数据可获取三维目标的位置和类别。与开放词汇二维目标检测相比，基于点云的开放词汇室内三维目标检测研究存在更多困难。一方面，与图片相比，点云存在稀疏问题，导致目标细节损失。另一方面，点云的质量与环境条件息息相关，导致点云数据更易受噪声影响。因此，开放词汇室内三维目标检测仅依赖目标点云信息获取目标类别存在一定困难。在室内场景中，目标与所处场景存在关联，这种关联被称为上下文信息。例如，bed 类别大概率出现在卧室，而不是卫生间。因此，上下文信息辅助能提升目标类别的划分。然而，现有的开放词汇室内三维目标检测忽视了上下文信息的重要性。目标信息是局部信息，与目标特征相关，上下文信息是全局信息，与场景特征相关，仅依赖目标信息会削弱开放词汇室内三维目标检测的性能。

现有算法通过比较目标特征与文本特征的方式确定目标类别，仅考虑目标信息，因此很容易受点云稀疏问题和噪声问题的影响。

为解决上述问题，本文提出了上下文信息和大语言模型辅助的开放词汇室内三维目标检测 (contextual information and large language model for open-vocabulary indoor three-dimensional object detection，CIL3D) 算法。在 CIL3D 算法中，目标信息和上下文信息被同时考虑，且大语言模型 (large language model，LLM)[5]被引入，用于整合目标信息和上下文信息，并进行思维链推理。一方面，来自 LLM 的先验知识能改善模型对模糊的目标特征的分辨，另一方面，LLM 的思维链[6]推理能对检测结果进行筛选。例如，由于点云的稀疏问题和噪声问题，类别为 dresser 的目标被错误地确定为类别 sink。现有算法仅使用目标信息，不能发现错误，从而将目标归类为错误的类别。在本文提出的算法 CIL3D 中，上述错误能通过上下文信息的辅助得到改正。LLM 通过对上下文信息进行处理，认为卧室里出现 sink 的可能性很小，因此认为目标被划分为错误的类别。通过整合目标信息、上下文信息和相关对话内容，被划分为错误类别的目标 sink 被改正为目标 dresser。

为更好地提升 CIL3D 的性能，CIL3D 的训练网络和推理网络被独立设计，且 CIL3D 的训练网络包括目标分支和上下文分支。为更好地提升检测性能，本文算法在目标分支和上下文分支中设计了 3 个独立模块。在目标分支中，本文设计了两个模块，分别是目标候选框筛选模块 (object proposal filtering module，OPFM) 和伪标签生成模块 (pseudo labels filtering module，PLFM)。其中，OPFM 用于生成精确的目标候选框，PLFM 用于生成高质量的伪标签。在上下文分支中，本文设计了场景描述生成模块 (scene description generation module，SDGM)，采用多模态 LLM 实现场景描述生成。

本文的主要贡献如下：提出了 CIL3D，通过上下文信息辅助补充目标信息，并采用 LLM 思维链进行推理，提高了开放词汇室内三维目标检测中目标类别判定的准确率。CIL3D 训练网络包括目标分支和上下文分支，在 CIL3D 的目标分支中设计了 OPFM 和 PLFM，其中，OPFM 提高了候选框的质量，PLFM 提高了生成的伪标签质量。

2　国内外研究现状

2.1　室内三维目标检测

作为三维世界感知的一项重要技术，三维目标检测取得了巨大进步，并被广泛用于自动驾驶和机器人领域。三维目标检测的场景包括室内场景和室外场景，本文聚焦于室内三维目标检测。与室外三维目标检测不同，室内场景的目标与所处场景存在一定关联。例如，bed 类别只会出现在卧室，而不会出现在厨房。这种关联信息被称为上下文信息，因此上下文信息在室内三维目标检测中格外重要。VoteNet[7]是一种基于投票的室内三维目标检测算法，主要包括两个步骤：候选框生成和目标分类。其中，候选框生成主要基于 PointNet＋＋[8]网络实现；目标分类主要通过点集投票过程实现。FCAF3D[9]是一种基于三维卷积的室内三维目标检测算法，通过三维卷积处理体素获得检测结果，且整个检测过程不受传统的锚框限制。Transformer[10]技术的发展使其开始被应用于室内三维目标检测。Group-Free[11]通过 Transformer 的注意力机制实现自动学习每个点对每个目标的贡献，避免了手工设计聚合方法的弊端。3DETR[12]构建了一个基于 Transformer 的端到端室内三维检测器。

然而，现有的室内三维目标检测相关研究主要针对闭集的检测。与现有研究不同，本文提出

的 CIL3D 针对开放词汇检测。

2.2 开放词汇目标检测

现有关于开放词汇目标检测的研究主要包括二维目标检测和三维目标检测。二维目标检测的一部分研究采用图片-文本对方法开展开放词汇目标检测，另一部分研究使用 CLIP[13]的文本-图片嵌入空间提升检测性能[14]。例如，RegionCLIP[15]使用 CLIP 匹配描述和图片区域，实现文本特征和图片特征的对齐。GLIP 提出了一种统一物体检测与短语定位的语言-图像预训练方案[16]。Grounding DINO[17]在 DINO[18]的基础上改进得到，融合了文本和图像两个模态的数据，实现了开放集目标检测，即给定一个文本提示，自动框出目标所在，该目标可以是训练集中没有的类别。随着生成式预训练变换器[10]的兴起，生成式 LLM 开始被用于自然语言处理和计算机视觉领域。LLM 从单模态模型发展到多模态模型，例如：MiniGPT-v2[19]构建了一个多模态 LLM，用于完成各种视觉语言任务，包括图像定位和图像标注。

与开放词汇二维目标检测相比，开放词汇三维目标检测的研究大多将点云作为输入数据。OpenSight 提出了一个更先进的基于激光雷达点云的开放词汇检测的 2D-3D 建模框架，用于实现室外开放词汇三维目标检测[20]。在 OV-3DET 算法中，基于点云的三维检测器在二维检测器预测结果的监督下进行学习，并且提出了一种新的去偏三重交叉模态对比学习方法，将图像、点云和文本的模态整合在一起[21]。FM-OV3D[22]提出了一种基于基础模型的跨模态知识混合的开放词汇三维检测方法，通过混合来自多个预训练基础模型的知识，提高三维模型的开放词汇定位和识别能力。CoDA 提出了一种有效的三维新目标定位策略，利用三维框几何先验和二维语义开放词汇先验生成新目标的伪标签。为对新的目标框进行分类，Cao 等[23]设计了跨模态对齐模块，以对齐三维点云和图像/文本模态之间的特征空间。

然而现有的开放词汇目标检测算法聚焦于使用目标特征获得检测结果，忽视了上下文信息，而上下文信息包含场景信息，这限制了了开放词汇目标检测性能的提升。

2.3 大语言模型

随着 LLM 技术的发展，预训练网络参数和预训练数据规模不断增加，同时获得的预训练模型大小也在不断增加。当预训练模型参数突破一定规模时，LLM 便获得了涌现能力，即具备生成类人的自然语言的能力。对于 LLM，GPT[24]系列算法应用较广泛，能通过用户指令实现对话或撰写文章等。为将 LLM 应用到特定领域，参数高效微调方法被引入微调过程，如 LoRA[25]、QLoRA[26]和 P-tuning v2[27]等。为提高 LLM 的推理能力，思维链技术被引入 LLM 的推理过程，以实现 LLM 的逐步推理。随着 LLM 拓展到不同模态，多模态 LLM 被提出。例如，基于 LLM 和视觉编码器的 LLaVA[28]实现了视觉和语言理解。不同模态的引入会在单模态 LLM 的基础上大幅度增加计算成本，为此，MiniCPM-V[29]模型被提出，用于实现模型性能提升与计算成本优化之间的平衡。

本文提出的 CIL3D 通过 LLM 整合全局信息和局部信息，并通过思维链推理获得检测结果。在训练网络的全局分支中使用 MiniCPM-V 实现场景描述的生成。在大模型的选择上，本文充分考虑了性能与计算成本之间的平衡。

3 上下文信息辅助的开放词汇室内三维目标检测算法

3.1 符号定义

基于点云的开放词汇室内三维目标检测算法训练样本包括点云 P、图片 I 和投影矩阵 $\boldsymbol{M}$ 等 3 部分。点云 P 是三维点的集合 $\{(x_i, y_i, z_i)\}_{i=1}^{N_p}$，其

中 N_p 为点的数量。二维的 RGB 图片 $I \in \mathbb{R}^{h\times w\times 3}$ 和 P 是匹配的。按照基于点云的开放词汇三维目标检测的设置，图片只出现在训练过程中，在推理过程中只有点云数据作为输入。投影矩阵 $\boldsymbol{M}$ 用于二维边界框与三维边界框之间的转换。此外，开放词汇目标检测采用无监督学习的方式进行训练，训练数据均为无标注数据。本文将边界框定义为 (x,y,z,w,h,l,θ)。其中，(x,y,z) 为边界框中心；(w,h,l) 为边界框尺寸；θ 为航向角。本文通过三维骨干网络提取点云的目标特征 $f_{\text{obj}} \in \mathbb{R}^{N_q\times D_p}$ 和上下文特征 $f_{\text{con}} \in \mathbb{R}^{1\times D_p}$，$f_{\text{obj}}$ 经过 OPFM 处理得到筛选后的目标特征 $f_{\text{obj_re}}$。其中，N_q 为三维骨干网络 Transformer 的查询数量；D_p 为三维特征维度。

3.2　CIL3D 网络结构

在基于点云的开放词汇室内三维目标检测中，目标信息和上下文信息的整合能提高检测性能。本文提出的 CIL3D 可分为训练网络和推理网络。CIL3D 的训练网络主要包括目标分支和上下文分支，如图 1 所示。

对于 CIL3D 训练网络，点云输入到三维骨干网络用于提取目标特征 f_{obj} 和上下文特征 f_{con}。由于开放词汇室内三维目标检测采用无监督学习，因此缺乏监督信号，导致目标分支和上下文分支中均包含了伪标签获取。对于目标分支，边界框通过 OPFM 获取，目标类别通过 LLM 进行预测。PLFM 用于对生成的伪标签进行筛选，从而提高伪标签的质量。在上下文分支中，场景描述通过 SDGM 生成作为上下文分支的监督信号。如图 1 所示，在训练过程中，红色标注的模块的参数固定不变，包括两个 LLM 编码器、LLM、PLFM 和 SDGM。为平衡性能与计算成本，CIL3D 中的大模型选用 MiniCPM[30]、SIGLIP[31] 和 MiniCPM-V[29]。其中，MiniCPM 用于构建 LLM，SIGLIP 用于构建 PLFM，MiniCPM-V 用于构建 SDGM。

CIL3D 推理网络结构如图 2 所示。首先，LLM 生成场景描述并预测场景类型。然后，初步的检测结果通过目标分支被获得。最后，思维链提示被用于指导 LLM 推理，且检测结果被按

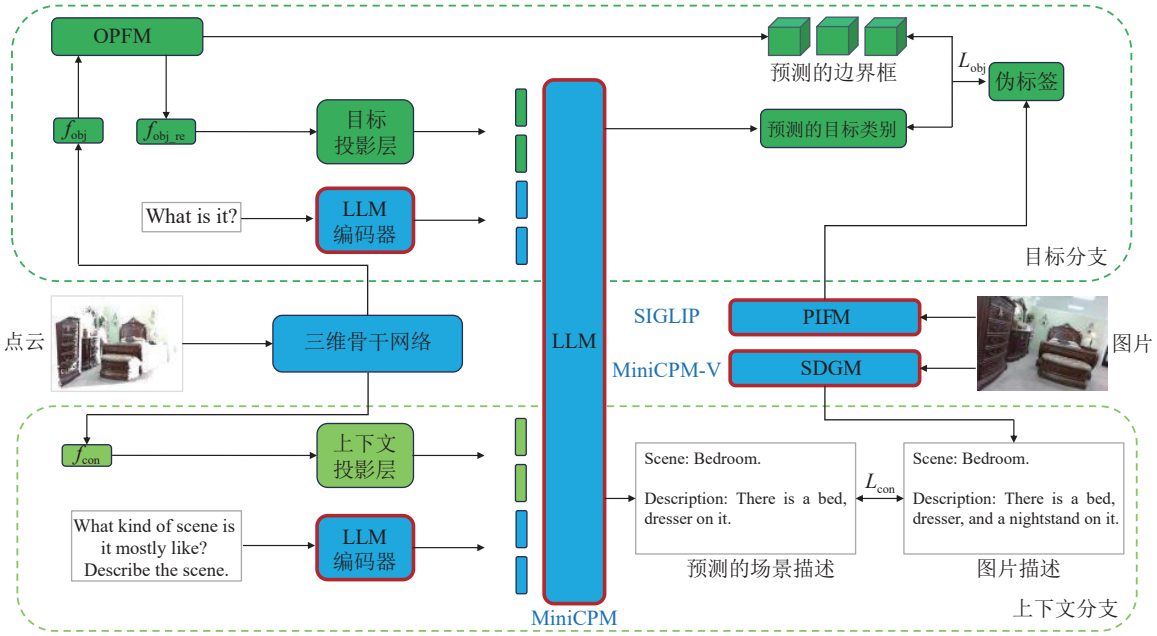
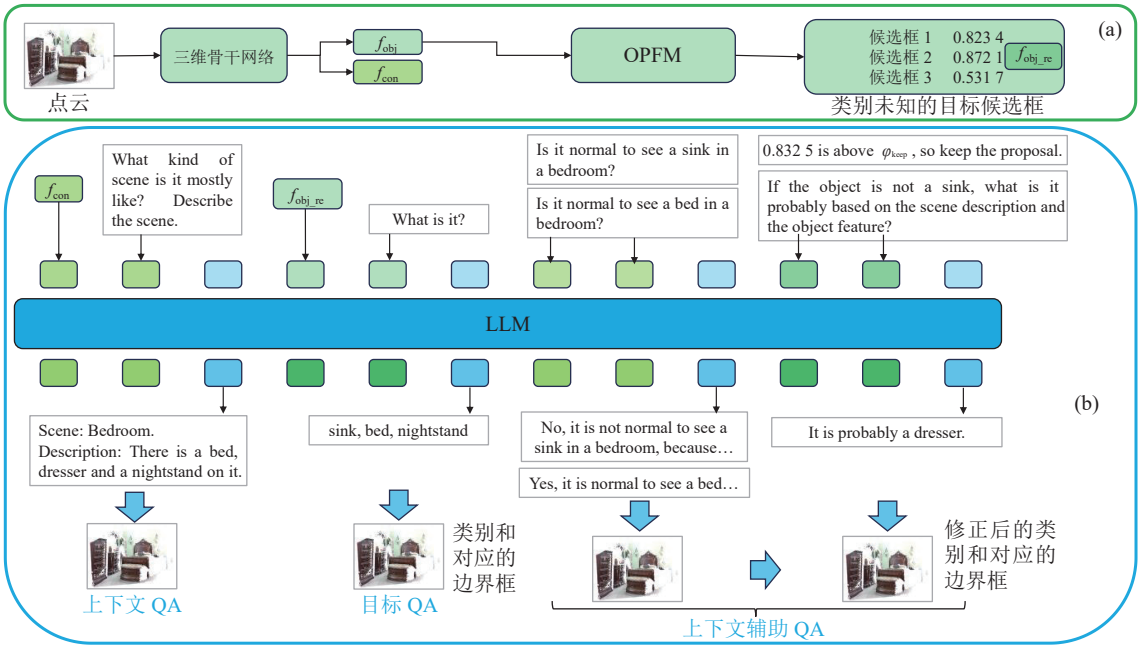


图 1　CIL3D 训练网络结构

Fig. 1　Pipeline of training network for CIL3D

步进行筛选。筛选过程如下：(1) 同时获得来自上下文分支的场景描述和来自目标分支的类别与边界框，经过上下文分支得到场景描述的过程被称为上下文问题-答案 (question-answer，QA)，经过目标分支得到的类别和目标边界框过程被称为目标 QA；(2) 将这两部分信息整合，获得最终的检测结果，该过程也被称为上下文辅助 QA。



(a) 推理过程；(b)QA 过程

图 2 CIL3D 推理网络结构

Fig. 2 Pipeline of inference network for CIL3D

3.3 CIL3D 训练网络

3.3.1 目标分支

目标分支的作用是生成初步的检测结果。首先，OPFM 通过目标特征 f_{obj} 提取目标候选框 $\{b_i, f^i_{obj_re}\}_{i=1}^{N_{obj}}$，其过程如下：

$$\{b_i, f^i_{obj_re}\}_{i=1}^{N_{obj}} = \mathrm{OPFM}(f_{obj}) \tag{1}$$

其中，b_i 为第 i 个目标的边界框；$f^i_{obj_re}$ 为第 i 个目标的特征；N_{obj} 为目标数。

由于点云中存在噪声，检测器可能将前景和背景混淆，从而输出错误的目标候选框。因此，为筛选目标候选框，本文提出了 OPFM。在 OPFM 中，类别不确定的目标候选框 $\{\hat{b}^i_{3D}, \hat{o}_i\}_{i=1}^{N_q}$ 通过一系列检测头对目标特征 f_{obj} 进行处理获得。其中，$\hat{b}^i_{3D}$ 为第 i 个边界框；$\hat{O}_i$ 为第 i 个候选框的置信度。置信度低于阈值 φ_{obj} 时被移除，最终得到了与类别无关的检测结果 $\{b^i_{3D}, c_i\}_{i=1}^{N_{obj}}$。其中，$b^i_{3D}$ 是类别无关检测结果中的检测框，包括检测框的长、宽和高；c_i 是类别。

在闭集检测中，预测置信度能通过真实标签学习得到。然而，在开放词汇目标检测中，监督信号通过手动设置获得。首先，基于交并比构建候选框 $\{\hat{b}^i_{3D}\}_{i=1}^{N_q}$ 和伪标签 $\{\tilde{b}^i_{3D}\}_{i=1}^{\tilde{N}}$ 之间的二分匹配，$\tilde{N}$ 为伪标签的个数[32]。然后所有的候选框通过式 (2) 进行标签匹配：

$$y_i = \begin{cases} 1 \exists j,\ b_i \text{ is matched with } b_j, \\ 0,\ \text{otherwise} \end{cases} \tag{2}$$

其中，$y_i=1$ 表示第 i 个候选框为正，即为前景目标；$y_i=0$ 表示第 i 个候选框为负，即为背景

目标。

目标类别 c_i 通过 LLM 进行预测，其预测过程如下：

$$c_i=\mathrm{LLM}\left(t_{\mathrm{obj}},\mathrm{projector}_{\mathrm{obj}}\left(f^i_{\mathrm{obj_re}}\right)\right) \tag{3}$$

其中，c_i 为第 i 个目标的预测类别；t_{obj} 为 LLM 的提示，本文使用“What is it ?”作为 t_{obj}。目标投影层 $\mathrm{projector}_{\mathrm{obj}}(\cdot)$ 是一个线性层，用于对齐目标特征 $\{f^i_{\mathrm{obj_re}}\}_{i=1}^{N_{\mathrm{obj}}}$ 和 LLM 的嵌入空间。OPFM 和目标投影层的训练受 PLFM 生成的伪标签的监督。

作为一个无监督学习任务，开放词汇目标检测未将标注数据用于模型训练。为解决该问题，现有的研究工作首先获取二维伪标签，然后通过二维-三维投影矩阵转换得到三维伪标签。然而，二维检测器的检测能力有限，在开放词汇目标检测过程中可能生成错误的标签，会对开放词汇三维检测器的训练造成一定负面影响。为解决上述问题，本文设计了 PLFM，以移除伪标签中类别错误的伪标签。首先，YOLOv10[33]被用于处理 RGB 图片，并生成二维伪标签；然后，SIGLIP 被用于筛选上述伪标签。

通过 YOLOv10 生成的初始二维标签为 $\{\bar{b}^i_{\mathrm{2D}},\bar{c}_i\}_{i=1}^{\bar{N}}$。其中，$\bar{b}^i_{\mathrm{2D}}$ 为第 i 个二维边界框；$\bar{c}_i$ 为第 i 个目标类别；$\bar{N}$ 为二维标签的数目。然后按照 $\{\bar{b}^i_{\mathrm{2D}}\}_{i=1}^{\bar{N}}$ 对图片 I 进行裁剪，得到相应的图片块 $\{p_i\}_{i=1}^{\bar{N}}$。应用 SIGLIP 对二维标签进行过滤，并设置了两个提示模版，表示如下：

$$\begin{cases} t^+(\mathrm{class}):\ \text{“This is a\{class\}.”} \\ t^-(\mathrm{class}):\ \text{“This is not a\{class\}.”} \end{cases} \tag{4}$$

其中，图片块 $\{p_i\}_{i=1}^{\bar{N}}$ 和类别 $\{c_i\}_{i=1}^{\bar{N}}$ 被送入 SIGLIP，用于计算置信度得分，计算过程如下：

$$[\varphi_i^+,\varphi_i^-]=\mathrm{Softmax}(\mathrm{SIGLIP}(t^+(\bar{c}_i),p_i),\ \mathrm{SIGLIP}((t^-(\bar{c}_i),p_i)) \tag{5}$$

其中，φ_i^+ 为 p_i 属于 c_i 的置信度得分；φ_i^- 为 p_i 不属于 c_i 的置信度得分。本文将 φ_i^+ 对应的标签中高于预定义阈值 $\varphi_{\mathrm{SIGLIP}}$ 的标签进行保留，然后将这些保留下来的二维标签 $\{\tilde{b}^i_{\mathrm{2D}},\tilde{c}_i\}_{i=1}^{\tilde{N}}$ 通过投影矩阵转换为三维伪标签 $\{\tilde{b}^i_{\mathrm{3D}},\tilde{c}_i\}_{i=1}^{\tilde{N}}$，其中，$\tilde{c}_i$ 为伪标签类别。

3.3.2 上下文分支

在上下文分支中，基于上下文特征 f_{con} 得到预测的场景类型 s 和生成的场景描述 d。对 LLM 输入的提示文本如下：

$$t_{\mathrm{con}}:\ \text{“What kind of scene is it mostly like? Describe the scene.”} \tag{6}$$

其中，t_{con} 为上下文提示词模板。

本文采用上下文投影层 $\mathrm{projector}_{\mathrm{con}}(\cdot)$ 对齐全局特征 f_{con} 和 LLM 嵌入空间。对齐过程表示如下：

$$s,d=\mathrm{LLM}(t_{\mathrm{con}},\mathrm{projector}_{\mathrm{con}}(f_{\mathrm{con}})) \tag{7}$$

为了对上述 s、d 进行监督，本文设计了 SDGM，用于对输入的图像 I 进行处理，从而获得场景类型和场景描述。与单模态 LLM 不同，多模态 LLM 能同时处理来自多个模态的特征输入。因此多模态 LLM 的计算成本远高于相应的单模态 LLM 模型。为此，需要对多模态 LLM 计算成本予以考虑。SDGM 采用高效的多模态 LLM (MiniCPM-V) 构建，能很好地平衡性能和计算成本。MiniCPM-V 能从对应的图片 I 中生成场景类型标签 $\tilde{s}$ 和场景描述标签 $\tilde{d}$，并用于后续的损失函数计算，$\tilde{s}$ 和 $\tilde{d}$ 的转换过程如下：

$$\tilde{s},\tilde{d}=\mathrm{SDGM}(I) \tag{8}$$

3.4　CIL3D 推理网络

CIL3D 推理网络如图 2 所示。如图 2 (a) 所示，首先，通过三维骨干网络提取上下文特征 f_{con} 和目标特征 f_{obj}；然后，通过 OPFM 获得类别无关的目标候选框、相应的置信度得分和三维特征 $f_{\mathrm{obj_re}}$。

LLM 可以将上文提取的上下文和目标特征进行整合，然后通过思维链推理按步对检测结果进行筛选。如图 2 (b) 所示，思维链[34-36]提示被用于 LLM 推理过程中。整个推理过程被分为上下

文 QA、目标 QA 和上下文辅助 QA 等 3 个阶段。在上下文 QA 中，LLM 被要求基于上下文特征 f_{con} 预测场景类型和描述场景内容。在目标 QA 中，目标特征 f_{obj} 被用于预测目标候选框的类别。

在完成上下文 QA 和目标 QA 后，初步的检测结果被得到。该检测结果在上下文辅助 QA 中经过思维链推理，并按步进行筛选。对于每一个预测类别，提示模版“Is it normal to see a {class} in a {scene type}?”被用于检测预测类别的合理性。如果类别 c 被认为在预测的场景类型中是合理的，则属于类别 c 的目标都会被保留下来。如果类别 c 被认为不可能属于预测场景类型中，则属于类别 c 的目标会被进一步筛选。对于每一个不合理的目标，如果置信度低于阈值 $\varphi_{\text{keep}}=0.78$，则该目标会被自动移除，否则，该目标会被保留。例如，在图 2 (b) 中，sink 类别的置信度大于阈值 φ_{keep}，被保留了下来。为得到正确的候选框类别，进一步对 LLM 输入下述提示“If the object is not a sink, what is it probably based on the scene description and the object feature?”。然后，LLM 根据目标信息和上下文信息得到结果，最终根据 LLM 的回答，目标类别被改正。

3.5 训练目标

训练损失包括边界框回归损失 L_{bbox}、置信度预测损失 L_{conf}、目标分类损失 L_{cls} 和场景理解损失 L_{scene} 等 4 部分。本文使用 3DETR 中的回归损失函数 loss_{reg} 计算边界框回归损失 L_{bbox}，表示如下：

$$L_{\text{bbox}}=\text{loss}_{\text{reg}}(\{b_i\}_{i=1}^{N_{\text{obj}}},\{\tilde{b}_i\}_{i=1}^{\tilde{N}}) \tag{9}$$

置信度损失 L_{conf} 的计算过程如下：

$$L_{\text{conf}}=-\frac{1}{N_{\text{q}}}\sum_{i=1}^{N_{\text{q}}}[y_i\log\hat{o}_i+\xi_{\text{conf}}(1-y_i)\log(1-\hat{o}_i)] \tag{10}$$

其中，ξ_{conf} 为一个权值系数，用于损失函数计算中各部分保持平衡。目标类别通过 LLM 预测得到，L_{cls} 通过最大化标签文本标记 (token) 的概率得到。假设标签文本是一个标记序列 $t_{\text{ser}}=(\text{ser}_1,\text{ser}_2,\cdots,\text{ser}_l)$，则每个 token 预测的概率为 $p(t_{\text{ser}})=[p(\text{ser}_1),p(\text{ser}_2),\cdots,p(\text{ser}_l)]$，文本损失 L_{text} 定义如下：

$$L_{\text{text}}(p(t_{\text{ser}}))=-\sum_{i=1}^{l}\log p(\text{ser}_i) \tag{11}$$

目标分类损失 L_{cls} 表示如下：

$$L_{\text{cls}}=\frac{1}{\tilde{N}}\sum_{i=1}^{\tilde{N}}L_{\text{text}}(p_{\text{obj}}(\tilde{c}_i)) \tag{12}$$

其中，p_{obj} 为 LLM 在目标分支上预测标记的概率。

与目标分类损失计算方法类似，基于上下文信息的场景理解损失 L_{scene} 表示如下：

$$L_{\text{scene}}=L_{\text{text}}(p_{\text{con}}(\tilde{s}))+L_{\text{text}}(p_{\text{con}}(\tilde{d})) \tag{13}$$

其中，p_{con} 为 LLM 在上下文分支上预测标记的概率。

综上所述，总的损失函数 L_{sum} 表示如下：

$$L_{\text{sum}}=\xi_1L_{\text{bbox}}+\xi_2L_{\text{conf}}+\xi_3L_{\text{cls}}+\xi_4L_{\text{scene}} \tag{14}$$

其中，ξ_1、ξ_2、ξ_3 和 ξ_4 为总的损失函数中的权值系数。

4 实验结果

4.1 数据集和度量指标

本文在 SUN RGB-D[37]和 ScanNetV2[38]两个室内数据集上进行了实验评估。SUN RGB-D 包括 10 335 个场景，其中 5 285 个场景用于训练，5 050 个场景用于测试，数据集约有 800 个目标；ScanNetV2 包括 1 513 个场景，其中 1 201 个场景用于训练，312 个场景用于测试，数据集拥有超过 200 个目标。

本文使用 IoU 阈值为 0.25 的全类平均正确率 (mAP) 作为检测性能的评估度量。本文参照 OV-3DET[21]进行实验评估，分别选择 SUN RGB-D 和 ScanNetV2 数据集上 top-20 类别进行评估。为与其他算法[22, 39]进行对比，本文还

进行了 top-10 类别的评估实验。为便于后文实验结果描述，top-20 和 top-10 度量指标分别表述为 mAP_{25}^{20cls} 和 mAP_{25}^{10cls}。

4.2　实施细节

训练过程分为两个阶段：第一阶段训练三维骨干网络和边界框预测头，持续 400 个 epochs，共使用 6 张 GPU 显卡进行训练，每张显卡上的批大小为 4；第二阶段训练目标置信度预测头、目标分支投影层和上下文分支投影层，持续 50 个 epochs，共使用 6 张显卡，每张显卡上的批大小为 2。训练过程中，基础学习率为 1×10^{-4}，损失函数的权重系数 ξ_{conf}、ξ_1、ξ_2、ξ_3 和 ξ_4 分别设为 0.3、3.0、9.0、1.0 和 1.0，阈值 φ_{obj}、φ_{SIGLIP}、φ^-、φ^+ 和 φ_{keep} 分别设为 0.10、0.50、0.25、0.60 和 0.78。本文利用 3DETR 构建三维骨干网络和边界框预测检测头，单模态 MiniCPM2 构建 LLM，多模态 SIGLIP 构建 PLFM，多模态 MiniCPM-V 2.6 构建 SDGM。本文实验均在 6 张 NVIDIA A800 GPU 上进行。

4.3　与现有方法的比较

CIL3D 与现有算法在 SUN RGB-D 数据集上的比较结果如表 1 和表 2 所示。在 top-20 类别上，与 OV-3DET 相比，mAP_{25}^{20cls} 从 20.46 提至 21.56，特别是 sink 类别，提升了 10.17，证明了 CIL3D 的有效性。由于表格篇幅有限，因此本文仅列举了 20 个类别中的 10 个类别。在 top-10 类别上，与 OV-3DET 相比，mAP_{25}^{10cls} 从 31.06 提至 32.65。

CIL3D 与现有算法在 ScanNetV2 数据集上的比较结果如表 3 和表 4 所示。与现有方法 CoDA 相比，CIL3D 的 mAP_{25}^{20cls} 从 19.32 提至 20.95，CIL3D 的 mAP_{25}^{10cls} 从 28.76 提至 31.34。CIL3D 在单个类别上的检测性能提升很明显，具体如下：chair 提升了 10.90，toilet 提升了 1.92，table 提升了 3.26。由于表格篇幅有限，因此本文仅列举了 20 个类别中的 10 个类别。

4.4　消融实验

为分析每个模块对 CIL3D 的影响，本文在 SUN RGB-D 数据集上对 CIL3D 进行了消融实验。消融实验分为基础设置、增加 PLFM、增加 OPFM 和增加上下文信息 4 组，为表述方便，表 5 中分别用 Base setting、＋OPFM、＋PLFM

表 1　CIL3D 和现有算法在 SUN RGB-D 数据集上的 20 类比较结果

Table 1　Top 20 comparison results of CIL3D and existing algorithms on SUN RGB-D dataset

算法	mAP_{25}^{20cls}	table	stand	cabinet	counter	bin	booksh	pillow	micro	sink	stool
OV-3DET[21]	20.46	23.31	2.75	3.40	0.75	23.52	9.83	10.27	1.98	18.57	4.10
CIL3D	21.56	20.29	4.96	3.87	1.71	23.34	13.77	16.90	6.42	28.74	1.96
算法	mAP_{25}^{10cls}	toilet	bed	chair	bathtub	sofa	dresser	scanner	fridge	lamp	desk
OV-3DET[21]	—	72.64	66.13	34.80	44.74	42.10	11.52	0.29	12.57	14.64	11.21
CIL3D	—	73.70	67.65	36.60	47.44	42.60	12.55	3.31	10.19	18.22	14.26

表 2　CIL3D 和现有算法在 SUN RGB-D 数据集上的 10 类比较结果

Table 2　Top 10 comparison results of CIL3D and existing algorithms on SUN RGB-D dataset

算法	mAP_{25}^{10cls}	toilet	bed	chair	bathtub	sofa	dresser	scanner	fridge	lamp	desk
OV-3DETIC[36]	13.03	43.97	6.17	0.89	45.75	2.26	8.22	0.02	8.32	0.07	14.60
FM-OV3D[22]	21.47	55.00	38.80	19.20	41.91	23.82	3.52	0.36	5.95	17.40	8.77
CIL3D	32.65	73.70	67.65	36.60	47.44	42.60	12.55	3.31	10.19	18.22	14.26

表 3 CIL3D 和现有算法在 ScanNetV2 数据集上的 20 类比较结果

Table 3 Top 20 comparison results of CIL3D and existing algorithms on ScanNetV2 dataset

算法	mAP_{25}^{20cls}	bathtub	refridge	desk	nightst	counter	door	curtain	box	lamp	bag
OV-3DET[21]	18.02	56.28	10.99	19.72	0.77	0.31	9.59	10.53	3.78	2.11	2.71
CoDA[23]	19.32	50.51	6.55	12.42	15.15	0.68	7.95	0.01	2.94	0.51	2.02
CIL3D	20.95	51.33	6.88	13.91	14.74	0.21	1.07	8.91	2.44	3.85	2.05
算法	mAP_{25}^{10cls}	toilet	bed	chair	sofa	dresser	table	cabinet	booksh	pillow	sink
OV-3DET[21]	—	57.29	42.26	27.06	31.50	8.21	14.17	2.98	5.56	23.00	31.60
CoDA[23]	—	68.09	44.04	28.72	44.57	3.41	20.23	5.32	0.03	27.95	45.26
CIL3D	—	70.01	43.80	39.62	42.52	8.20	23.49	8.03	8.42	25.74	43.62

表 4 CIL3D 和现有算法在 ScanNetV2 数据集上的 10 类比较结果

Table 4 Top 10 comparison results of CIL3D and existing algorithms on ScanNetV2 dataset

算法	mAP_{25}^{10cls}	toilet	bed	chair	sofa	dresser	table	cabinet	booksh	pillow	sink
OV-3DETIC [36]	12.65	48.99	2.63	7.27	18.64	2.77	14.34	2.35	4.54	3.93	21.08
FM-OV3D [22]	21.53	62.32	41.97	22.24	31.80	1.89	10.73	1.38	0.11	12.26	30.62
CIL3D	31.34	70.01	43.80	39.62	42.52	8.20	23.49	8.03	8.42	25.74	43.62

表 5 CIL3D 在 SUN RGB-D 数据集上的消融实验

Table 5 Ablation experiment of CIL3D on SUN RGB-D dataset

方法	mAP_{25}^{20cls}	ΔmAP_{25}^{20cls}	mAP_{25}^{10cls}	ΔmAP_{25}^{10cls}
Base setting	19.85	—	30.08	—
＋OPFM	20.30	＋0.45	30.95	＋0.87
＋PLFM	20.64	＋0.34	31.41	＋0.46
＋Context	21.56	＋0.92	32.65	＋1.24

和＋Context 表示。Base setting 组与 CIL3D 的区别是 PLFM 从 YOLOv10 得到的所有伪标签均被保留；OPFM 未使用阈值对候选框进行筛选，所有目标特征均被保留；上下文信息未被应用到思维链推理。＋OPFM 组的 mAP_{25}^{20cls} 从 19.85 提至 20.30，mAP_{25}^{10cls} 从 30.08 提至 30.95，表明 OPFM 能消除点云噪声对检测结果的影响，提高了检测性能。随着 PLFM 模块的加入，＋PLFM 组的 mAP_{25}^{20cls} 从 20.30 提至 20.64，mAP_{25}^{10cls} 从 30.95 提至 31.41，表明 PLFM 能过滤低质量的伪标签，从而获得高质量的监督信号，进而提高检测性能。与＋PLFM 组相比，＋Context 组的 mAP_{25}^{20cls} 提升了 0.92，mAP_{25}^{10cls} 提升了 1.24。对于 CIL3D，上下文信息的辅助能有效避免一些错误目标类别划分，从而提升检测性能。

5 可视化分析

CIL3D 的可视化分析结果如图 3 所示。场景 1 是 bedroom，CIL3D 通过结合上下文信息和目标信息获取的检测目标是 dresser、bed 和

Scene 1: bedroom
Objects: dresser, bed, nightstand

Scene 2: dining room
Objects: table, chair

图 3 CIL3D 的可视化分析

Fig. 3 Visualization analysis of CIL3D

nightstand。场景 2 是 dining room，CIL3D 通过结合上下文信息和目标信息获取的检测目标是 table 和 chair。室内场景的点云通过 RGB-D 相机获取，存在噪声和稀疏问题。仅依赖点云的目标信息进行开放词汇室内三维目标检测的推理可能会导致目标类别错误。CIL3D 通过大语言模型处理上下文信息和目标信息，并进行思维链推理，消除了与场景不匹配的错误类别，进而提高了开放词汇室内三维目标检测的性能。

6　讨论和分析

开放词汇室内三维检测涉及多模态融合、大语言模型和三维视觉，算法的训练和评估需要高昂的算力成本。CIL3D 采用更高效的大语言模型组合实现计算成本与性能之间的平衡，算法采用的大语言模型包括：MiniCPM，SIGLIP 和 MiniCPM-V。其中，MiniCPM 用于构建 LLM，SIGLIP 用于构建 PLFM，MiniCPM-V 用于构建 SDGM。在算法的运行条件方面，训练和评估实验需要 6 张 Nvidia A800 GPU。开放词汇室内三维目标检测算法多以点云为主要输入数据。一方面，与图片相比，点云存在稀疏问题，导致目标细节损失；另一方面，点云的质量与环境条件息息相关，导致点云数据更容易受噪声影响。仅依靠从点云获取的目标信息作为检测器的输入可能会出现类别错误的问题，为此需要在检测过程中考虑上下文信息。在算法的判断条件方面，CIL3D 通过结合上下文 QA 和目标 QA，实现上下文辅助 QA，进而在检测过程中融合上下文信息和目标信息，提高开放词汇室内三维目标检测的性能。

7　结　论

对于开放词汇室内三维目标检测算法，检测过程存在仅依赖目标信息而忽视上下文信息的问题，为此本文提出了 CIL3D，通过结合上下文信息和大语言模型进行检测。CIL3D 包括目标分支和上下文分支。在目标分支中，上下文信息被加入到基于 LLM 的思维链推理过程，用于提高检测性能。为进一步改善目标信息的质量，在目标分支中设计了 OPFM 和 PLFM 两个模块。其中，OPFM 用于降低噪声对候选框的影响，PLFM 用于提高伪标签的质量。为降低计算成本，本文用 MiniCPM 构建 LLM，用 SIGLIP 构建 PLFM，用 MiniCPM-V 构建 SDGM。为验证 CIL3D 的性能，本文分别在 SUN RGB-D 和 ScanNetV2 数据集上进行了对比实验和消融实验。实验结果表明，CIL3D 能提高开放词汇室内三维目标检测的性能。

参 考 文 献

[1] 孙毅, 吴斯曼, 方伟, 等. 基于 ResNet 的安全监控目标检测 [J]. 集成技术, 2024, 13(6): 44-52.
Sun Y, Wu SM, Fang W, et al. Object detection of security monitoring based on ResNet [J]. Journal of Integration Technology, 2024, 13(6): 44-52.

[2] Rukhovich D, Vorontsova A, Konushin A. TR3D: towards real-time indoor 3D object detection [C] // Proceedings of the IEEE International Conference on Image Processing, 2023: 281-285.

[3] Zhu CY, Chen L. A survey on open-vocabulary detection and segmentation: past, present, and future [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 8954-8975.

[4] 郑泽凡, 谷飞飞, 王思成, 等. 基于三维视觉的机器人安全预警系统 [J]. 集成技术, 2022, 11(4): 80-91.
Zheng ZF, Gu FF, Wang SC, et al. A robot safety warning system based on 3D vision [J]. Journal of Integration Technology, 2022, 11(4): 80-91.

[5] 王耀祖, 李擎, 戴张杰, 等. 大语言模型研究现状与趋势 [J]. 工程科学学报, 2024, 46(8): 1411-

1425.
Wang YZ, Li Q, Dai ZJ, et al. Current status and trends in large language modeling research [J]. Chinese Journal of Engineering, 2024, 46(8): 1411-1425.

[6] Wei J, Wang XZ, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models [C] // Proceedings of the Advances in Neural Information Processing Systems, 2022: 24824-24837.

[7] Qi CR, Litany O, He KM, et al. Deep hough voting for 3D object detection in point clouds [C] // Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, 2019: 9277-9286.

[8] Qi CR, Yi L, Su H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space [C] // Proceedings of the Advances in Neural Information Processing Systems, 2017: 1-10.

[9] Rukhovich D, Vorontsova A, Konushin A, et al. FCAF3D: fully convolutional anchor-free 3D object detection [C] // Proceedings of the European Conference on Computer Vision, 2022: 477-493.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Proceedings of the Advances in Neural Information Processing Systems, 2017: 1-11.

[11] Liu Z, Zhang Z, Cao Y, et al. Group-Free 3D object detection via Transformers [C] // Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, 2021: 2949-2958.

[12] Misra I, Girdhar R, Joulin A, et al. An end-to-end Transformer model for 3D object detection [C] // Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, 2021: 2906-2917.

[13] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] // Proceedings of the International Conference on Machine Learning, 2021: 8748-8763.

[14] Zareian A, Rosa KD, Hu DH, et al. Open-vocabulary object detection using captions [C] // Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 14393-14402.

[15] Zhong YW, Yang JW, Zhang PC, et al. RegionCLIP: region-based language-image pretraining [C] // Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16793-16803.

[16] Li LH, Zhang PC, Zhang HT, et al. Grounded language-image pre-training [C] // Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 10965-10975.

[17] Liu SL, Zeng ZY, Ren TH, et al. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection [Z/OL]. arXiv Preprint, arXiv: 2303.05499, 2023.

[18] Zhang H, Li F, Liu SL, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection [Z/OL]. arXiv Preprint, arXiv: 2203.03605, 2022.

[19] Chen J, Zhu DY, Shen XQ, et al. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning [Z/OL]. arXiv Preprint, arXiv: 2310.09478, 2023.

[20] Zhang H, Xu J, Tang T, et al. OpenSight: a simple open-vocabulary framework for LiDAR-based object detection [Z/OL]. arXiv Preprint, arXiv: 2312.08876, 2023.

[21] Lu YH, Xu CF, Wei X, et al. Open-vocabulary point-cloud object detection without 3D annotation [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 1190-1199.

[22] Zhang DM, Li C, Zhang R, et al. FM-OV3D: foundation model-based cross-modal knowledge blending for open-vocabulary 3D detection [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 16723-16731.

[23] Cao Y, Zeng YH, Xu H, et al. CoDA: collaborative novel box discovery and cross-modal alignment for open-vocabulary 3D object detection [C] // Proceedings of the Advances in Neural Informa-

tion Processing Systems, 2024: 1-12.

[24] OpenAI. GPT-4 technical report [Z/OL]. arXiv Preprint, arXiv: 2303.08774, 2023.

[25] Hu EJ, Shen YL, Wallis P, et al. LoRA: low-rank adaptation of large language models [Z/OL]. arXiv Preprint, arXiv: 2106.09685, 2021.

[26] Dettmers T, Pagnoni A, Holtzman A, et al. QLoRA: efficient finetuning of quantized LLMs [C] // Proceedings of the Advances in Neural Information Processing Systems, 2024: 1-26.

[27] Liu X, Ji KX, Fu YC, el al. P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks [Z/OL]. arXiv Preprint, arXiv: 2110.07602, 2021.

[28] Liu H, Li C, Wu Q, et al. Visual instruction tuning [C] // Proceedings of the Advances in Neural Information Processing Systems, 2024: 1-25.

[29] Yao Y, Yu TY, Zhang A, et al. MiniCPM-V: a GPT-4V level MLLM on your phone [Z/OL]. arXiv Preprint, arXiv: 2408.01800, 2024.

[30] Hu SD, Tu YG, Han X, et al. MiniCPM: unveiling the potential of small language models with scalable training strategies [Z/OL]. arXiv Preprint, arXiv: 2404.06395, 2024.

[31] Zhai XH, Mustafa B, Kolesnikov A, et al. Sigmoid loss for language image pre-training [Z/OL]. arXiv Preprint, arXiv: 2303.15343, 2023.

[32] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with Transformers [C] // Proceedings of the European Conference on Computer Vision, 2020: 213-229.

[33] Wang A, Chen H, Liu LH, et al. YOLOv10: real-time end-to-end object detection [Z/OL]. arXiv Preprint, arXiv: 2405.14458, 2024.

[34] Chu Z, Chen J, Chen Q, et al. A survey of chain of thought reasoning: advances, frontiers and future [Z/OL]. arXiv Preprint, arXiv: 2309.15402, 2023.

[35] Feng G, Zhang B, Gu Y, et al. Towards revealing the mystery behind chain of thought: a theoretical perspective [C] // Proceedings of the Advances in Neural Information Processing Systems, 2024: 1-38.

[36] Merrill W, Sabharwal A. The expresssive power of transformers with chain of thought [Z/OL]. arXiv Preprint, arXiv: 2310.07923, 2023.

[37] Song SR, Lichtenberg SP, Xiao JX. SUN RGB-D: a RGB-D scene understanding benchmark suite [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015: 567-576.

[38] Dai A, Chang AX, Savva M, et al. ScanNet: richly-annotated 3D reconstructions of indoor scene [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017: 5828-5839.

[39] Lu YH, Xu CF, Wei XB, et al. Open-vocabulary 3D detection via image-level class and debiased cross-modal contrastive learning [Z/OL]. arXiv Preprint, arXiv: 2207.01987, 2022.