文章编号　1004-924X(2025)24-3915-16

跨模态交互学习与迭代融合的3D视觉定位

才　华[1*]，冉　越[1]，张海峰[2]，张高鹏[2]，付　强[3]，孙俊喜[4]

(1. 长春理工大学 电子信息工程学院，吉林 长春 130022；
2. 中国科学院 西安光学精密机械研究所，陕西 西安 710119；
3. 长春理工大学 空间光电技术研究所，吉林 长春 130022；
4. 东北师范大学 信息科学与技术学院，吉林 长春 130117)

摘要：针对现有3D视觉定位方法存在的对单一模态信息依赖过强、视角变化适应性差以及跨模态特征融合效果有限的问题，提出了一种跨模态交互学习与迭代融合的3D视觉定位方法。该方法包括多模态特征提取与跨模态特征融合两个阶段。在特征提取阶段，分别采用点云编码器和文本编码器提取点云与文本特征，并引入点云的类别信息；在特征融合阶段，设计基于Transformer的点云特征增强模块，以提升点云特征的表达能力；通过对称交互学习模块捕捉点云与文本特征之间的深层关联，有效抑制无关特征干扰；跨模态迭代融合模块逐步融合跨模态信息，增强模型在复杂场景下的定位能力。实验结果表明，本文方法在ScanRefer、Nr3D和Sr3D这3个经典的3D视觉定位数据集上均取得了综合的精度提升。在ScanRefer的unique子集上，Acc@0.25和Acc@0.5分别达到了86.19%和69.68%；在Nr3D和Sr3D的easy子集上，定位准确率分别达到了65.20%和74.87%。本文方法在多种3D场景中均展现出稳定的定位能力，验证了其在增强多模态交互与跨模态融合方面的卓越性能。

关　键　词：3D视觉定位；点云；交互学习；迭代融合；Transformer

中图分类号：TN911.73；TP391.41　　**文献标识码**：A

doi：10.37188/OPE.20253324.3915　　**CSTR**：32169.14.OPE.20253324.3915

Cross-modal interactive learning and iterative fusion for 3D visual grounding

CAI Hua[1*], RAN Yue[1], ZHANG Haifeng[2], ZHANG Gaopeng[2], FU Qiang[3], SUN Junxi[4]

(1. *School of Electronical and Information Engineering, Changchun University of Science and Technology, Changchun* 130022, *China*;
2. *Xi'an Institute of Optics and Precision Mechanics of CAS, Xi'an* 710119, *China*;
3. *Institute of Space Opto-Electronic Technology, Changchun University of Science and Technology, Changchun* 130022, *China*;
4. *School of Information Science and Technology, Northeast Normal University, Changchun* 130117, *China*)
* *Corresponding author, E-mail: caihua@cust.edu.cn*

收稿日期：2025-10-11；**修订日期**：2025-11-02.

基金项目：国家自然基金联合基金(No. U2341226)；西安市飞行器光学成像与测量技术重点实验室开放基金课题(No. 2023-013)；吉林省自然科学基金(No. 20260102267JC)

Abstract: To address the problems of existing 3D visual grounding methods—such as excessive dependence on single-modal information, poor adaptability to perspective changes, and limited effectiveness of cross-modal feature fusion-a novel cross-modal interactive learning and iterative fusion approach for 3D visual grounding was proposed. The method comprised two stages: multi-modal feature extraction and cross-modal feature fusion. In the feature extraction stage, point cloud encoder and text encoder were employed to extract point cloud and text features respectively, and the category information of the point cloud was introduced. In the feature fusion stage, a point cloud feature enhancement module based on Transformer was designed to improve the expressive ability of point cloud features. The deep correlation between point cloud and text features was captured by the symmetric interactive learning module, which effectively suppressed the interference of irrelevant features. The cross-modal iterative fusion module gradually integrated cross-modal information to enhance the positioning ability of the model in complex scenes. The experimental results show that the proposed method has achieved comprehensive accuracy improvement on the three classic 3D visual grounding datasets of ScanRefer, Nr3D and Sr3D. On the unique subset of ScanRefer, Acc@0.25 and Acc@0.5 reached 86.19% and 69.68%, respectively. On the easy subset of Nr3D and Sr3D, the localization accuracy reached 65.20% and 74.87%, respectively. The proposed method demonstrates stable localisation capabilities across diverse 3D scenes, validating its exceptional performance in enhancing multimodal interaction and cross-modal fusion.

Key words: 3D Visual Grounding; point cloud; interactive learning; iterative fusion; Transformer

1 引　言

随着传感技术和计算机视觉的持续进步,三维(3D)视觉数据已成为日常生活中不可或缺的组成部分[1-2]。与传统的二维(2D)图像相比,3D场景通常以更加复杂且无序的点云形式呈现,这种表示方式能够更全面地捕捉空间和深度信息。得益于自然语言处理(Natural Language Processing)技术的快速发展,结合自然语言与3D点云的三维视觉定位(3D Visual Grounding)任务应运而生。该任务旨在精确定位与语言描述相关的物体,从而推动计算机辅助室内设计[3]、人机交互[4]和机器人导航[5]等领域的应用。

当前,3D视觉定位任务的研究主要分为两阶段方法[6-10]和一阶段方法[11-18]。两阶段方法采用分步策略,首先从场景中预测出一组候选3D边界框,接着将这些预测边界框与相应的文本描述进行匹配,从而实现目标物体的精确定位。作为该领域的开创性工作,Chen等人[6]最早提出了使用自然语言描述进行3D目标定位的任务,并构建了首个3D视觉定位数据集ScanRefer。在此基础上,为了提高模型识别真实3D场景的能力,Panos等人[7]的ReferIt3D提出了两个具有挑战性的新数据集Nr3D和Sr3D,奠定了后续研究的数据基础,并推动了两阶段方法的研究进展。随后,Zhao等人[8]通过坐标引导的上下文聚合模块提取点云特征,并利用多重注意力模块捕捉文本关键信息,凭借丰富的上下文线索显著消除了定位歧义。Yuan等人[9]将3D视觉定位任务转化为目标匹配问题,通过匹配策略实现显著的定位精度提升。Zhang等人[10]引入动态视觉模块和动态相机定位机制,通过语言引导的空间注意力实现高效的候选框推理。尽管两阶段方法在定位效果上表现优异,但其性能很大程度上依赖于候选边界框的质量,同时也伴随着较高的计算开销与时间成本。

一阶段方法摒弃了两阶段方法中对候选边界框的依赖,直接从视觉与文本输入中提取特征,并将其映射至统一的联合嵌入空间。在该嵌入空间中,模型能够依据文本描述逐步选取点云中的关键点,实现跨模态特征对齐,从而完成端到端的3D视觉定位。作为一阶段方法的典型代表,3D-SPS[11]最早以端到端的方式实现了对3D目标物体的定位,为3D视觉定位研究

提供了一种更加简化的方式,最大程度地减少了中间步骤的需求。Jain等人[12]结合语言引导和基于预训练检测器的对象引导,利用检测头直接解码被指代的物体,在多个3D数据集上均取得了显著的性能提升。Wu等人[13]为了解决3D视觉定位中的细粒度语义丢失问题,提出了一种显式解耦对齐模型,通过解耦文本属性并实现细粒度文本与点云特征的密集对齐,提升了模型的跨模态匹配精度。为了提升模型在不同任务之间的迁移能力,Jin等人[14]将预训练模型引入3D视觉定位领域,提出了一种针对3D场景的通用预训练方法。Qian等人[15]构建了一种多分支协作网络,利用3D视觉定位和3D视觉分割任务的协同互补作用,提高了模型定位的准确率。随后,Shi等人[16]结合视角权重优化模型性能,Guo等人[17]利用点云增强聚合模块弥补点云采样的信息丢失,Huang等人[18]使用语言引导的聚合和分配机制,进一步优化了3D视觉定位的性能。

尽管两阶段与一阶段3D视觉定位方法均已取得显著进展,但仍面临诸多挑战。现有模型普遍受限于点云数据的稀疏性,导致目标识别能力不足;同时,在理解复杂空间关系时表现仍不理想。因此,提升点云数据处理能力并增强模型对复杂空间关系的理解,依然是未来研究的关键方向。

针对上述问题,本文提出了一种跨模态交互学习与迭代融合的3D视觉定位方法(3D Visual Grounding method based on cross-modal interactive learning and iterative fusion,CIF3D)。该方法主要包括多模态特征提取和跨模态特征融合两部分。多模态特征提取部分使用PointNet++[19]提取点云特征,并引入点云的类别信息,同时采用对比语言-图像预训练模型(Contrastive Language-Image Pre-Training,CLIP)的文本编码器[20]提取文本特征。跨模态特征融合由三个模块构成:基于Transformer[21]的点云特征增强模块利用类别信息增强点云特征,缓解点云数据稀疏性对点云目标识别任务的影响;对称交互学习模块通过权重共享策略,以共享交互函数处理跨模态输入,避免模型对单一模态的过度依赖,并且在减少参数量的同时提高模型泛化能力;跨模态迭代融合模块利用迭代更新机制,有效增强语义关联与空间特征融合,提高模型在复杂场景中的定位精度。最终,将融合后的特征输入多层感知器(Multilayer Perceptron,MLP)计算定位框坐标,从而实现精确的3D视觉定位。

2　方法架构

本文提出的跨模态交互学习与迭代融合的3D视觉定位(CIF3D)模型的整体架构如图1所示,该模型由多模态特征提取网络和跨模态特征融合网络两部分组成。

多模态特征提取网络用于提取点云、文本和类别特征;跨模态特征融合网络通过点云特征增强模块、对称交互学习模块以及跨模态迭代融合模块,对多模态特征进行处理,最终实现与文本相关的3D目标学习和视觉定位任务。具体流程见算法1。

算法1	CIF3D
输入:	原始点云$\boldsymbol{P}$,原始文本$\boldsymbol{T}$
输出:	3D视觉定位位置坐标(x,y,z,l,w,h)
1:	Function CIF3D($\boldsymbol{P},\boldsymbol{T}$)
2:	//多模态特征提取
3:	$\boldsymbol{P}_{\mathrm{C}},\boldsymbol{P}_{\mathrm{box}}\leftarrow\boldsymbol{P}$//获取输入点云和类别信息
4:	$\boldsymbol{F}_{\mathrm{T}}\leftarrow$CLIPTextEncode($\boldsymbol{T}$);//文本特征提取
5:	$\boldsymbol{F}_{\mathrm{C}}\leftarrow$PointNet++($\boldsymbol{P}_{\mathrm{C}}$);//点云特征提取
6:	$\boldsymbol{F}_{\mathrm{B}}\leftarrow\boldsymbol{P}_{\mathrm{box}}$;//类别特征提取
7:	//点云特征增强模块
8:	$\boldsymbol{F}_{\mathrm{CB}},\boldsymbol{F}_{\mathrm{B}}\leftarrow$PFEM($\boldsymbol{F}_{\mathrm{C}},\boldsymbol{F}_{\mathrm{B}}$);//特征增强
9:	//对称交互学习模块
10:	$\boldsymbol{F}'_{\mathrm{C}},\boldsymbol{F}'_{\mathrm{T}},\boldsymbol{F}_{\mathrm{B}}\leftarrow$SILM($\boldsymbol{F}_{\mathrm{CB}},\boldsymbol{F}_{\mathrm{T}},\boldsymbol{F}_{\mathrm{B}}$);//特征筛选
11:	//跨模态迭代融合模块
12:	$\boldsymbol{F}_{\mathrm{u}}^{0}\leftarrow\boldsymbol{F}'_{\mathrm{C}}$;//迭代融合初始化
13:	For n=1 to 3 do
14:	$\boldsymbol{F}_{\mathrm{u}}^{\mathrm{n}}\leftarrow$CIFM($\boldsymbol{F}_{\mathrm{u}}^{\mathrm{n}-1},\boldsymbol{F}'_{\mathrm{T}},\boldsymbol{F}_{\mathrm{B}}$);//迭代融合
15:	$\boldsymbol{F}_{\mathrm{u}}\leftarrow\boldsymbol{F}_{\mathrm{u}}^{\mathrm{n}}$;//特征更新
16:	预测定位框$(x,y,z,l,w,h)\leftarrow$MLP($\boldsymbol{F}_{\mathrm{u}}$);
17:	计算总损失$L_{\mathrm{total}}=L_{\mathrm{box}}+\alpha L_{\mathrm{text_cls}}+\beta L_{\mathrm{obj_cls}}$;
18:	Return

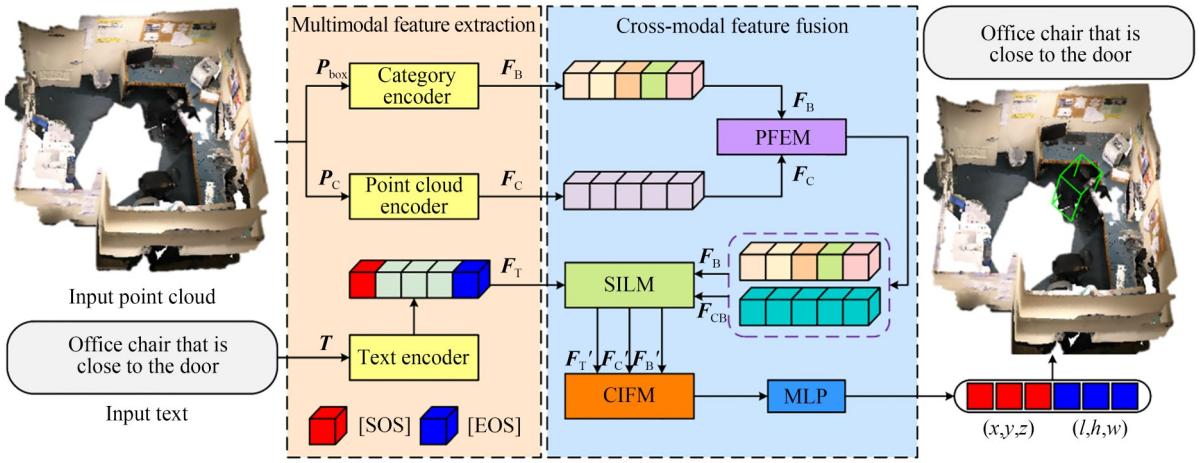

图 1　CIF3D 网络架构图

Fig. 1　CIF3D Network Architecture Diagram

2.1　多模态特征提取网络

本文特征提取部分使用一种基于双编码器和类别提取器的多模态特征提取网络架构，通过两个独立的编码器分别提取输入的点云和文本特征，同时利用类别提取器从点云中提取类别信息构建类别特征。网络架构如图 2 所示。

文本特征提取使用 6 层的 CLIP 文本编码器，该编码器由多层 Transformer 块堆叠组成，每层包含一个多头自注意力（Multi-head Self-Attention，MSA）和一个前馈神经网络（Feedforward Neural Network，FNN），并通过层归一化（Layer Normalization，LN）和残差连接避免梯度爆炸。输入文本首先经过分词器转化为离散的词序列，然后在每个词序列的起始和终止位置分别添加起始标识（Start Of Sequence，[SOS]）和结束标识（End Of Sequence，[EOS]），以指示序列的边界。随后，将词序列逐项映射为词嵌入向量 $\boldsymbol{t}_i$，并与位置编码相 $\boldsymbol{T}_{\text{pos}}$ 加构建文本特征输入，送入到 CLIP 文本编码器中。经过多层 Transformer 处理后，最终得到文本特征矩阵 $\boldsymbol{F}_{\text{T}} \in \boldsymbol{R}^{C_{\text{t}} \times N_{\text{t}}}$，其中 C_{t} 和 N_{t} 表示文本特征的维度和数量。

文本编码器计算过程见式（1）～式（3）：

$$\boldsymbol{W}_0 = (\boldsymbol{t}_{\text{SOS}}, \boldsymbol{t}_1, \cdots, \boldsymbol{t}_{\text{n}}, \boldsymbol{t}_{\text{EOS}}) + \boldsymbol{T}_{\text{pos}}, \boldsymbol{F}_{\text{T}} = \boldsymbol{W}_6, \tag{1}$$

$$\boldsymbol{W}_n' = \text{LayerNorm}(\text{MHA}(\boldsymbol{W}_{n-1})) + \boldsymbol{W}_{n-1}, \tag{2}$$

$$\boldsymbol{W}_n = \text{LayerNorm}(\text{FFN}(\boldsymbol{W}_n')) + \boldsymbol{W}_n', \tag{3}$$

其中，$\boldsymbol{t}_{\text{SOS}}$ 和 $\boldsymbol{t}_{\text{EOS}}$ 是[SOS]和[EOS]的词嵌入向量，$\boldsymbol{W}_0$ 是初始输入的文本特征矩阵，$\boldsymbol{W}_n$ 是第 n 层编码器编码输出的文本特征矩阵。

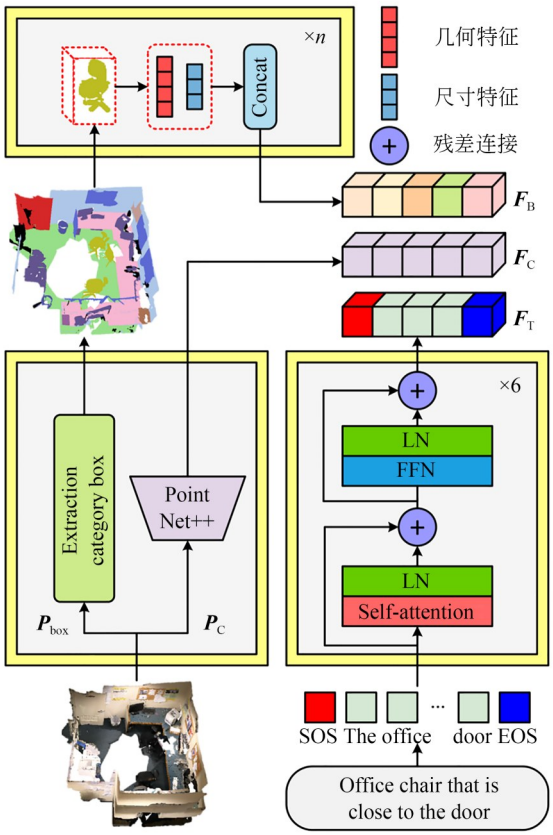

图 2　多模态特征提取网络

Fig. 2　Multimodal feature extraction network

点云特征提取使用 PointNet++网络[19]，通过其层次化的点云特征抽象和局部邻域特征学习能力，显著提升点云特征的提取效率。该网络架构由多个 Set Abstraction（SA）模块构成，多层

次的SA模块逐步对点云进行采样，并提取局部和全局特征，最终生成高维度的点云特征矩阵 $F_C \in R^{C_P \times N_P}$，其中，$C_P$ 和 N_P 分别表示点云特征的维度和类别总数。

考虑到点云数据具有稀疏性这一特点，仅依靠PointNet++提取的点云特征可能无法充分捕捉数据中的细节信息，导致对点云中各定位实体的特征提取不完整，从而影响后续结合文本特征进行跨模态学习时的模型定位精度和性能。针对这一局限，本文在多模态特征提取网络中引入了点云的类别信息，作为第三类输入特征参与后续处理。

具体来说，首先从输入点云中获取空间几何信息 P_{box}，并依据类别对其进行分类。对于每一类点云，计算其对应的三维边界框，从中抽取出几何特征与尺寸特征。其中，几何特征包括三维边界框的中心点坐标和体积，尺寸特征为三维边界框的长宽高。随后，将几何特征与尺寸特征依次拼接，构建一个7维的特征张量 z^n，其中n代表第 n 个类别对象。在此基础上，假设所有场景中点云类别的最大数为 N_P，按照点云对象的类别顺序，将点云中所有类别对应的特征张量逐行排列，构成一个 N_P 行的类别特征矩阵 $F_B \in R^{7 \times N_P}$。若输入点云中的实际类别数少于 N_P，则使用0向量填充至 N_P，以确保类别特征的类别总数与点云特征一致。

类别特征计算过程如式(4)，式(5)：

$$z^n = \mathrm{Concat}((x_n, y_n, z_n, v_n), (l_n, w_n, h_n)), \quad (4)$$

$$F_B = (z^1, z^2, \cdots, z^{n-1}, z^n)^T. \quad (5)$$

2.2　基于Transformer的点云特征增强

在多模态特征提取网络中，PointNet++依赖局部感受野的点云特征提取特性限制了其对点云全局上下文信息的理解，而点云固有稀疏性进一步损害了点云特征的完整性，导致后续融合与目标定位的性能下降。受Transformer在2D图像特征提取中的成功启发[22]，本文提出一种基于Transformer的点云特征增强模块(Point Cloud Feature Enhancement Module，PFEM)。该模块通过多头自注意力建模点云间的依赖关系，融合全局与局部点云的几何信息，有效增强点云特征的表达能力。PFEM的主体架构是一个3层的Transformer编码器，其输入由点云特征与类别特征共同构成。编码器通过自注意力层捕捉点云中不同区域之间的上下文关联，实现对点云特征的深层增强。该模块整体架构如图3所示。

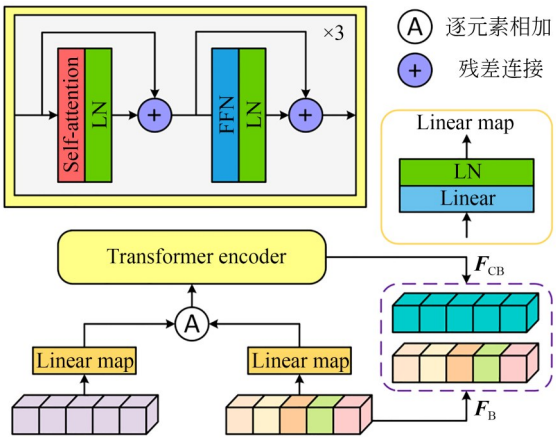

图3　点云特征增强模块

Fig. 3　Point cloud feature enhancement module

首先，将点云特征 F_C 和点云的类别特征 F_B 分别送入两个独立但结构相同的线性映射层(Linear map)。线性映射层由线性层和层归一化构成，旨在将两个特征映射至同一维度空间 C_P。对于点云特征，由于其原始维度已为目标维度，线性映射仅用于引入额外参数增强特征的表达能力。而对于点云的类别特征，则通过线性层映射到高维空间，并利用层归一化防止过拟合，该过程如式(6)，式(7)所示：

$$Z_C = \mathrm{LN}(\mathrm{Linear}(F_C)), \quad (6)$$

$$Z_B = \mathrm{LN}(\mathrm{Linear}(F_B)). \quad (7)$$

随后，采用简单的逐元素相加方法，将映射后的点云特征与类别特征进行融合，构建Transformer编码器的特征输入。该方法能够在保留点云原始语义完整性的基础上，有效引入了类别特征中的空间几何信息，得到兼具形状与空间位置信息的增强特征表示 $M_0 \in R^{C_P \times N_P}$，计算过程如式(8)所示：

$$M_0 = Z_C + Z_B. \quad (8)$$

接着，将 M_0 馈送至3层的Transformer编码器。该编码器通过多头自注意力机制(MSA)建模点云中各个对象之间的语义与空间依赖关系，得到增强的点云特征 $F_{CB} \in R^{C_P \times N_P}$。

增强的点云特征计算过程如式(9)～式(11)所示。

$$\boldsymbol{M}_n'=\mathrm{LayerNorm}(\mathrm{MHA}(\boldsymbol{M}_{n-1}))+\boldsymbol{M}_{n-1},\quad(9)$$

$$\boldsymbol{M}_n=\mathrm{LayerNorm}(\mathrm{FFN}(\boldsymbol{M}_n'))+\boldsymbol{M}_n',\quad(10)$$

$$\boldsymbol{F}_{\mathrm{CB}}=\boldsymbol{M}_3.\quad(11)$$

最终，上述增强的点云特征与类别特征共同构成了 PFEM 模块的输出。

2.3　基于权重共享的对称交互学习

在多模态任务中，单一模态所提供的信息往往存在局限性[23]，例如在点云和文本特征对中，文本特征通常缺乏三维点云空间的精确几何信息，而点云特征则缺失定位目标对象的语义信息。此外，两种模态独立提取的特征中均包含与任务无关的冗余信息，这些信息的存在会对模型判断产生干扰，影响目标定位的精度。针对以上问题，本文在点云特征增强模块之后，引入了一个基于权重共享的对称交互学习模块(Symmetric Interactive Learning Module，SILM)，其核心架构如图 4 所示。权重共享机制的引入，旨在迫使文本与点云两种模态在同一个特征子空间内进行对称的交互与学习，从而确保模态间能够实现深层信息互补，并从机制上抑制过拟合，增强了模型的泛化能力。

考虑到跨模态交互学习的本质是一个对称学习过程，点云需要理解文本特征，文本也需要理解点云特征，二者应遵循统一的学习机制。因此，本模块以权重共享为核心，设计一个对称交互函数 F。该函数被点云与文本两种模态复用以完成各自的特征更新，确保显著降低参数规模的同时，避免模型对单一模态的过度依赖。

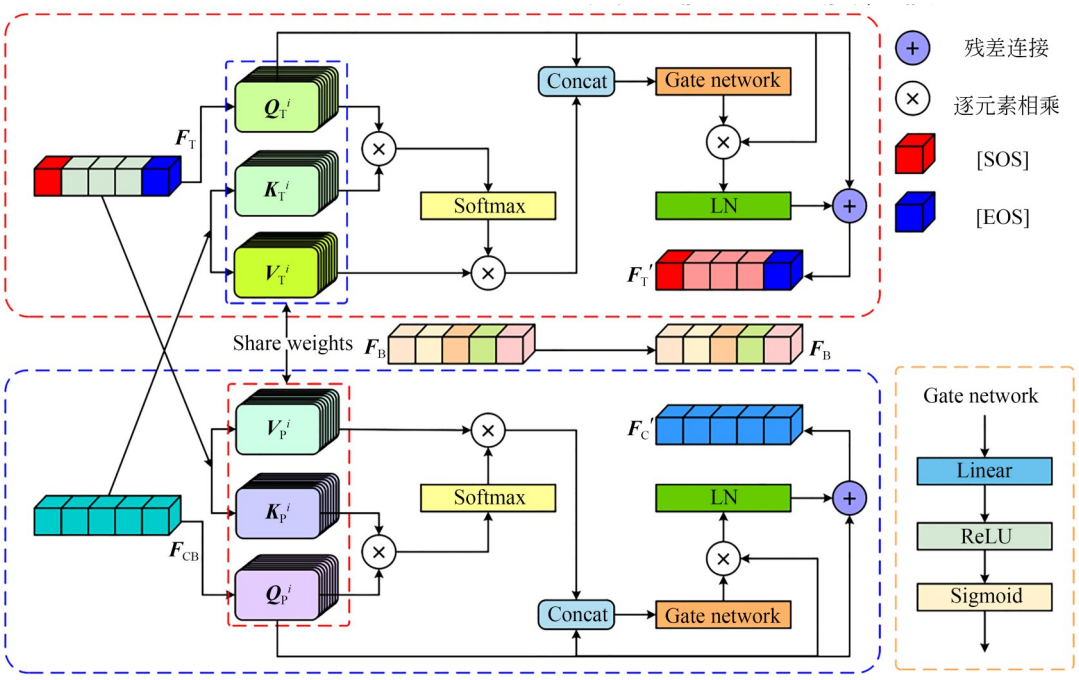

图 4　对称交互学习模块

Fig. 4　Symmetric Interactive Learning Module

在该模块中，点云的类别特征 $\boldsymbol{F}_{\mathrm{B}}$ 直接输出，而文本特征和点云特征则分别作为查询向量 $\boldsymbol{Q}$ 和键值对向量 $\boldsymbol{K}$ 与 $\boldsymbol{V}$ 输入。随后，根据多头注意力机制将其拆分为 h 个头，并采用权重共享策略，使文本和点云特征多头注意力的对应头共享同一组查询与键值向量，如式(12)～式(14)所示：

$$\boldsymbol{Q}_{\mathrm{T}}=\boldsymbol{Q}_P=(\boldsymbol{Q}^1,\boldsymbol{Q}^2,\cdots,\boldsymbol{Q}^i)^{\mathrm{T}},\quad(12)$$

$$\boldsymbol{K}_{\mathrm{T}}=\boldsymbol{K}_P=(\boldsymbol{K}^1,\boldsymbol{K}^2,\cdots,\boldsymbol{K}^i)^{\mathrm{T}},\quad(13)$$

$$\boldsymbol{V}_{\mathrm{T}}=\boldsymbol{V}_P=(\boldsymbol{V}^1,\boldsymbol{V}^2,\cdots,\boldsymbol{V}^i)^{\mathrm{T}}.\quad(14)$$

针对每一个头 i，计算查询向量与键向量的点积注意力，并对其结果进行归一化，得到注意力权重。然后，通过将这些权重与值向量加权求和，得到最终的注意力输出，如式(15)所示：

$$\boldsymbol{A}_i=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}^i(\boldsymbol{K}^i)^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\cdot\boldsymbol{V}^i,\quad(15)$$

其中：d_k为特征的维度，A_i为第i个头的注意力输出值，所有头的输出值按顺序拼接即可得到多头注意力的输出权重矩阵A。

为了对不同的输入特征进行差异化选择，本模块在多头注意力之后引入门控网络，该网络由线性层、RELU非线性激活函数和Sigmoid激活函数构成。通过门控机制，模型能够动态调节信息流动，自动决定保留多少原始特征以及融入多少注意力更新内容，有效提升特征间交互学习的灵活性。

具体而言，根据已知的查询向量$\boldsymbol{Q}$，将其与求得的权重矩阵$\boldsymbol{A}$拼接，构造门控网络的特征输入$\boldsymbol{F}_{\mathrm{QA}}$，如式(16)所示：

$$\boldsymbol{F}_{\mathrm{QA}}=\mathrm{Concat}(\boldsymbol{Q},\boldsymbol{A})=[\boldsymbol{Q};\boldsymbol{A}]. \tag{16}$$

随后，将$\boldsymbol{F}_{\mathrm{QA}}$送入门控网络(Gate Network)，基于门控机制学习能够选择性控制特征更新强度与范围的门控系数，如式(17)所示：

$$\boldsymbol{G}=\sigma(\mathrm{ReLU}(\boldsymbol{F}_{\mathrm{QA}}\cdot\boldsymbol{W}_{\mathrm{g}}+b_{\mathrm{g}})), \tag{17}$$

其中：$\boldsymbol{G}$中每个位置的元素表示原始特征中对应位置的特征保留比例，σ代表Sigmoid激活函数，$\boldsymbol{W}_{\mathrm{g}}$和$b_{\mathrm{g}}$是门控网络的可训练参数。

最后，将门控网络的输出$\boldsymbol{G}$与查询$\boldsymbol{Q}$逐元素相乘，经过层归一化处理后，再与输入对应的查询$\boldsymbol{Q}$进行残差连接，这样就得到了权重共享的对称交互函数F，如式(18)所示：

$$F(\boldsymbol{Q})=\mathrm{LN}(\boldsymbol{G}\cdot\boldsymbol{Q})+\boldsymbol{Q}. \tag{18}$$

至此，分别以文本特征$\boldsymbol{F}_{\mathrm{T}}$或增强的点云特征$\boldsymbol{F}_{\mathrm{CB}}$作为查询，以另一特征作为键值对代入函数F中，即可得到特征更新后的点云特征$\boldsymbol{F}'_{\mathrm{C}}$和文本特征$\boldsymbol{F}'_{\mathrm{T}}$，如式(19)，式(20)所示：

$$\boldsymbol{F}'_{\mathrm{C}}=F(\boldsymbol{F}_{\mathrm{CB}}), \tag{19}$$

$$\boldsymbol{F}'_{\mathrm{T}}=F(\boldsymbol{F}_{\mathrm{T}}). \tag{20}$$

2.4　跨模态迭代融合模块

为解决现有模型因理解复杂3D点云空间关系能力不足而导致的融合效果有限问题，本文对上述提取的文本、点云与类别三类特征进行深度融合，提出了跨模态迭代融合模块(Cross-modal Iterative Fusion Module，CIFM)。该模块采用训练-推理解耦设计：在训练阶段引入多视图空间增强技术，以提升模型对复杂点云结构的语义理解能力；在推理阶段则仅基于输入视角实现高效且准确的跨模态融合。模块的迭代核心由两个多头交叉注意力层(Cross Attention)和一个多头空间聚合层(Sa Layer)构成，交叉注意力层负责挖掘不同模态间的语义关联，空间聚合层用于空间结构特征融合，通过构建可迭代的特征更新机制，使模块在多轮迭代过程中交互式地整合多源信息，从而增强模型在复杂空间环境中的定位精度。具体流程如图5所示。

在首轮迭代开始之前，根据任务训练(Train)或推理(Inference)对输入类别特征$\boldsymbol{F}_{\mathrm{B}}$进

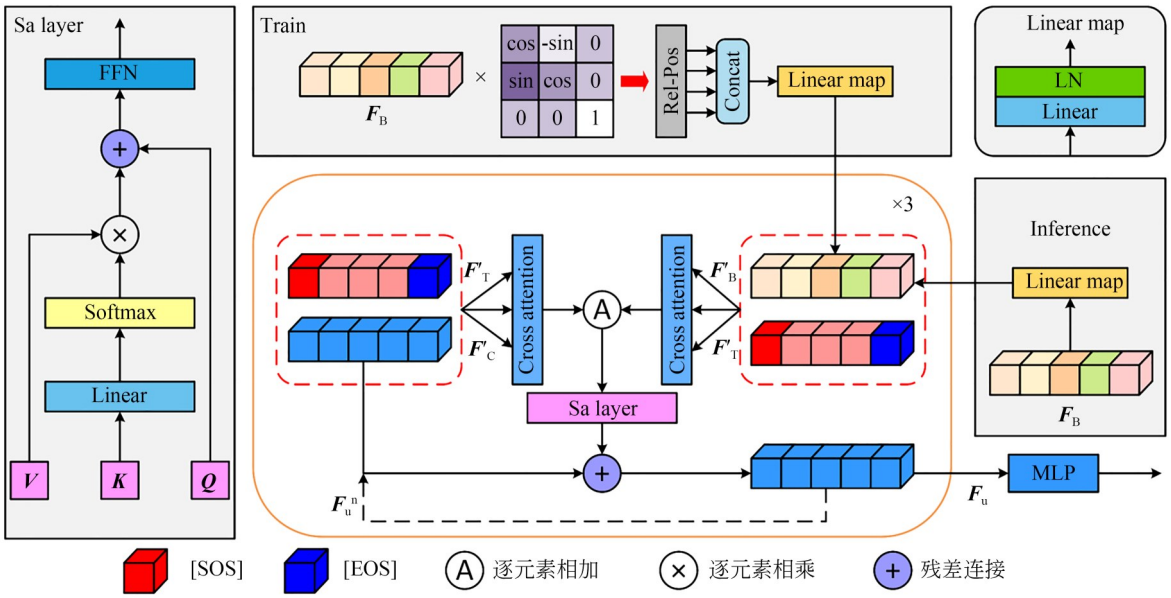

图5　跨模态迭代融合模块

Fig. 5　Cross-modal Iterative Fusion Module Modul

行差异化处理。在推理任务中，类别特征直接通过线性层和层归一化组成的线性映射层映射到与文本和点云特征相同的维度空间 C_{P}，得到高维类别特征 $\boldsymbol{F}'_{\mathrm{B}}$，如式(21)所示：

$$\boldsymbol{F}'_{\mathrm{B}}=\mathrm{LN}(\mathrm{Linear}(\boldsymbol{F}_{\mathrm{B}})). \tag{21}$$

在训练任务中，为了增强模型对复杂点云的语义理解能力及其对视角变化的感知，本文引入多视图空间增强方法。受点云旋转不变性的启发[24]，本文在保持点云 z 轴方向不变的情况下，采用均匀采样的旋转矩阵对输入点云的类别特征进行随机旋转，生成 0°，90°，180°和 270°共 4 组旋转点云类别特征。这种方法能够在确保构建完整对称的点云视角覆盖和避免过高计算开销的基础上，促使模型学习更具泛化性的 3D 点云特征，使得模型对不同空间朝向具有全面而均衡的感知能力。均匀采样的旋转矩阵 $\boldsymbol{R}_{z}(\theta)$ 如式(22)所示：

$$\boldsymbol{R}_{z}(\theta)=\begin{bmatrix}\cos\theta & -\sin\theta & 0\\ \sin\theta & \cos\theta & 0\\ 0 & 0 & 1\end{bmatrix}. \tag{22}$$

依据类别特征 $\boldsymbol{F}_{\mathrm{B}}$ 的几何特征、尺寸特征以及旋转角度，计算旋转后各视角下对象间的相对位置(Rel-Pos)，随后将 4 组视角特征拼接，并通过线性映射层统一至与文本和点云特征相同的维度空间 C_{P}，得到高维类别特征 $\boldsymbol{F}'_{\mathrm{B}}$，计算过程如式(23)，式(24)所示：

$$\boldsymbol{Z}_{\mathrm{B}}=\mathrm{Concat}\left(\boldsymbol{F}_{\mathrm{B}}\times\sum_{k=0}^{3}\boldsymbol{R}_{z}\left(\frac{2k\pi}{4}\right)\right), \tag{23}$$

$$\boldsymbol{F}'_{\mathrm{B}}=\mathrm{LayerNorm}(\mathrm{Linear}(\boldsymbol{Z}_{\mathrm{B}})). \tag{24}$$

随后，迭代融合特征 $\boldsymbol{F}_{\mathrm{u}}^{0}$ 初始化为更新后的点云特征 $\boldsymbol{F}'_{\mathrm{C}}$，然后以高维类别特征 $\boldsymbol{F}'_{\mathrm{B}}$ 和迭代融合特征 $\boldsymbol{F}_{\mathrm{u}}^{0}$ 作为查询向量，更新后的文本特征 $\boldsymbol{F}'_{\mathrm{T}}$ 作为键向量和值向量，分别利用两个独立的多头交叉注意力层进行特征融合，得到各自对应的交叉融合结果 $\boldsymbol{M}_{\mathrm{P}}$ 和 $\boldsymbol{M}_{\mathrm{B}}$，如式(25)，式(26)所示：

$$\boldsymbol{M}_{\mathrm{P}}=\mathrm{MCA}(\boldsymbol{F}_{\mathrm{u}}^{n},\boldsymbol{F}'_{\mathrm{T}}), \tag{25}$$

$$\boldsymbol{M}_{\mathrm{B}}=\mathrm{MCA}(\boldsymbol{F}'_{\mathrm{B}},\boldsymbol{F}'_{\mathrm{T}}), \tag{26}$$

其中：$\boldsymbol{F}_{\mathrm{u}}^{n}$ 是每一轮的迭代融合特征，$n=0$ 时代表初始化迭代融合特征 $\boldsymbol{F}_{\mathrm{u}}^{0}$。

为了确保输入至多头空间聚合层的特征同时蕴含丰富的语义信息与空间关系，首先将两个交叉注意力模块的输出特征沿通道维度进行逐元素相加，并以此作为聚合层的输入。在空间聚合层中，键向量经线性层映射为标量分数，再通过 Softmax 函数归一化为注意力权重，然后与值向量相乘后与查询向量进行残差连接，最终送入前馈层。上述计算过程如式(27)～式(29)所示：

$$\boldsymbol{W}_{\mathrm{Q}},\boldsymbol{W}_{\mathrm{K}},\boldsymbol{W}_{\mathrm{V}}=\boldsymbol{M}_{\mathrm{P}}+\boldsymbol{M}_{\mathrm{B}}, \tag{27}$$

$$\boldsymbol{F}'_{\mathrm{u}}=\mathrm{Softmax}(\mathrm{Linear}(\boldsymbol{W}_{\mathrm{K}}))\cdot\boldsymbol{W}_{\mathrm{V}}, \tag{28}$$

$$\boldsymbol{F}_{\mathrm{u}}^{n}=\mathrm{FFN}(\boldsymbol{F}'_{\mathrm{u}}+\boldsymbol{W}_{\mathrm{Q}}). \tag{29}$$

当一轮迭代执行完毕后，上一轮迭代融合特征 $\boldsymbol{F}_{\mathrm{u}}^{n-1}$ 更新为当前轮次的最终输出 $\boldsymbol{F}_{\mathrm{u}}^{n}$，随后执行下一轮迭代，直至完成全部轮次，获得最终融合输出 $\boldsymbol{F}_{\mathrm{u}}$。

最后，将融合输出 $\boldsymbol{F}_{\mathrm{u}}$ 送入一个层数为 3 输出维度为 6 的多层感知机中，计算得到 6 维的预测坐标 (x,y,z,l,w,h)。其中，(x,y,z) 表示预测矩形框的几何中心点坐标，(l,w,h) 表示预测矩形框的长宽高。

2.5 损失函数

本文网络的损失函数 L_{total} 由 3D 边界框回归损失、文本分类损失和点云分类损失 3 部分组成。其中，3D 边界框回归损失为主要损失，文本分类损失和点云分类损失用来辅助训练，如式(30)所示：

$$L_{\mathrm{total}}=L_{\mathrm{box}}+\alpha L_{\mathrm{text\_cls}}+\beta L_{\mathrm{obj\_cls}}, \tag{30}$$

其中：L_{box} 表示 3D 边界框回归损失，该损失为评估模型在 3D 点云图中预测框与真实框的距离与交并比损失；$L_{\mathrm{text\_cls}}$ 表示文本分类损失，该损失为评估模型从给定输入文本中正确分类出文本实际描述目标物体的交叉熵损失；$L_{\mathrm{obj\_cls}}$ 表示点云分类损失，该损失为从 3D 点云图中正确分类出所有物体的交叉熵损失，α 和 β 为平衡各损失值的加权系数。遵循 ScanRefer[6] 的实验设置，本文采用对称权重，即将加权系数均设为 0.5，以确保各损失项在训练阶段提供均衡的梯度贡献，防止单个损失主导梯度更新。若采用非对称权重，模型在跨模态融合过程中会过度偏向高权重模态，从而降低其联合建模能力。

3　实验分析

本节详细介绍了CIF3D在3个经典3D视觉定位数据集上的实验与测试结果,并与一些经典的3D视觉定位模型进行了全面比较,展示了本文模型在各个数据集上的优异表现。同时,对模型进行了可视化分析,进一步验证了本算法的有效性。

3.1　3D视觉定位数据集

本文在ScanRefer[6]、Nr3D[7]和Sr3D[7]这3个经典3D视觉定位数据集上进行实验,所有数据集均基于大规模室内场景数据集ScanNet[25]构建。

ScanRefer数据集包含800个ScanNet场景和51 583个文本描述,涵盖约11 046个3D对象,数据集划分为3部分。其中训练集(train)有36 665组点云文本对,验证集(val)有9 508组点云文本对,测试集(test)有5 410组点云文本对。根据是否存在干扰对象,测试集进一步划分为“Unique”,“Multiple”和“Overall”。其中,“Unique”不含干扰对象,“Multiple”子集存在干扰对象,两个子集统一为“Overall”子集。遵循ScanRefer[6],评价指标使用Acc@IoU=0.25和Acc@IoU=0.5。

Nr3D数据集包含707个ScanNet场景和41 503个文本描述,数据集划分为训练集(train)和测试集(test)。其中,训练集有28 716组点云文本对,测试集有7 485组点云文本对,数据集划分为5部分,分别为“Easy”,“Hard”,“View-Dep”,“View-Indep”和“Overall”。其中,“Easy”和“Hard”子集的定义标准与ScanRefer中“Unique”和“Multiple”相同,“View-Dep”子集的点云文本对依赖相机的视角,“View-Indep”子集的点云文本对不依赖相机的视角,四个子集统一为“Overall”子集。遵循Referit3d[7],评价指标使用二分类准确率。

Sr3D数据集包含1 513个ScanNet场景和83 572个文本描述,数据集划分为训练集(train)和测试集(test),其中训练集有65 844组点云文本对,测试集有17 726组点云文本对,评价指标与Nr3D一致。

3.2　实验设置

本文所有实验均在Ubuntu 22.04系统环境下进行,软件配置包括python 3.12,pytorch 2.5.1,CUDA 12.4和torchvision 0.20.1,硬件配置为一张32GB vGPU显卡。模型训练采用AdamW优化器[26],使用预训练CLIP(base)[20]作为文本编码器,训练时权重不冻结,输入文本最大词长度为24,点云最大类别数N_P为51,点云采样数为50 000,PointNet++中SA模块数量为3,训练批次为8,训练周期为200轮。除文本编码器、点云特征增强模块和跨模态迭代融合模块的初始学习率为0.000 05外,其余部分统一设置为0.000 5,所有模块的dropout率均为0.2。训练过程中,前30轮保持学习率不变,30轮之后每10轮以权重衰减系数0.05对学习率进行衰减。

3.3　对比实验

为了验证本文方法的有效性,本文选取了3D视觉定位领域共11种模型,在3个经典的3D视觉定位数据集上进行了广泛的对比实验,实验结果如表1和表2所示。

表1展示了本文方法在ScanRefer数据集上的对比实验结果,其中第2列“输入”表示模型推理时除文本外所需的数据类型,“3D”代表需传入3D点云,“3D+2D”表示需传入3D点云与2D多视图。由表中数据可见,本文方法在ScanRefer的各测试子集上均取得了最优准确率。在“Overall”子集中,当以ACC@0.5为评价指标时,本文CIF3D模型相比排名第二的VPP-Net提升了1.42%,在所有子集中提升最大。综合所有子集,本文CIF3D方法实现了平均0.94%的准确率提升。

表2展示了本文方法在Nr3D和Sr3D数据集上的对比实验结果。由表可知,本文方法在除Nr3D的“Overall”子集和Sr3D的“View-Indep”子集外的所有子集上均取得最优准确率。在上述两个子集上,本文方法排名第2,与最优模型MCLN[15]和GNL3D[18]仅相差0.34%和0.04%。总体而言,本文方法在Nr3D与Sr3D数据集上实现了定位准确率的有效提升。

表2展示了本文方法在Nr3D和Sr3D数据集上的对比实验结果。由表可知,本文方法在除Nr3D的“Overall”子集和Sr3D的“View-Indep”

表 1　ScanRefer 数据集对比实验

Tab. 1　Comparative experiments on the Scan Refer dataset　(%)

Model	Input	Unique		Multiple		Overall	
		ACC@0.25	ACC@0.5	ACC@0.25	ACC@0.5	ACC@0.25	ACC@0.5
ScanRefer[6]	3D	67.64	46.19	32.06	21.26	38.97	26.10
Referit3D[7]	3D	53.75	37.47	21.03	12.83	26.44	16.90
3DVG[8]	3D+2D	81.93	60.64	39.30	28.42	47.57	34.67
InstanceRefer[9]	3D	77.45	66.83	31.27	24.77	40.23	32.93
3D-SPS[11]	3D+2D	84.12	66.72	40.32	29.82	48.82	36.98
BUTD-DETR[12]	3D	81.47	61.24	44.20	32.81	49.76	37.05
EDA[13]	3D	85.76	68.57	49.13	37.64	54.59	42.26
3D-VLP[14]	3D+2D	84.23	64.61	43.51	33.41	51.41	39.46
MCLN[15]	3D	84.43	68.36	49.72	38.41	54.30	42.64
VPP-Net[16]	3D	86.05	67.09	50.32	39.03	55.65	43.29
CFAM[17]	3D	84.50	66.60	41.80	31.22	49.63	37.72
CIF3D(our)	3D	**86.19**	**69.68**	**51.33**	**40.17**	**56.52**	**44.71**

表 2　Nr3D 和 Sr3D 数据集对比实验

Tab. 2　Comparison experiments of Nr3D and Sr3D datasets　(%)

Model	Input	Nr3D					Sr3D				
		Easy	Hard	VI	VD	Overall	Easy	Hard	VI	VD	Overall
Referit3D[7]	3D	43.60	27.90	37.10	32.50	35.60	44.70	31.50	40.80	39.20	40.80
3DVG[8]	3D+2D	48.50	34.80	43.70	34.80	40.80	60.50	50.20	57.70	49.90	57.40
InstanceRefer[9]	3D	46.00	31.80	41.90	34.50	38.80	51.10	40.50	48.10	45.40	48.00
Multi3Drefer[10]	3D	55.60	43.40	52.90	42.30	49.40	—	—	—	—	—
3D-SPS[11]	3D+2D	58.10	45.10	53.20	48.00	51.50	56.20	65.40	63.20	49.20	62.60
BUTD-DETR[12]	3D	60.70	48.40	58.00	46.00	54.60	68.60	63.20	67.60	53.00	67.00
EDA[13]	3D	—	—	—	—	52.10	—	—	—	—	68.10
MCLN[15]	3D	—	—	—	—	**59.80**	—	—	—	—	68.40
VPP-Net[16]	3D	—	—	—	—	56.90	—	—	—	—	56.90
CFAM[17]	3D	62.30	48.20	56.10	54.90	55.70	67.60	59.50	65.10	59.00	65.30
GNL3D[18]	3D	—	—	—	—	—	72.80	64.00	**70.60**	58.00	70.10
CIF3D(our)	3D	**65.20**	**52.65**	**59.23**	**58.01**	59.46	**74.87**	**66.33**	70.56	**63.51**	**70.21**

"—" indicates not detected; "VI": View-Indep; "VD": View-Dep.

子集外的所有子集上均取得最优准确率。在上述两个子集上，本文方法排名第 2，与最优模型 MCLN[15] 和 GNL3D[18] 仅相差 0.34% 和 0.04%。总体而言，本文方法在 Nr3D 与 Sr3D 数据集上实现了定位准确率的有效提升。

3.4　消融实验

为分析各模块对定位准确率的影响，本文在 Nr3D 数据集上进行了消融实验，实验包括 CIF3D 整体架构的组合消融、文本编码器层数、点云特征增强模块编码器层数、对称交互学习模块权重以及跨模态迭代融合模块迭代次数。

3.4.1　组合消融实验

为了验证 CIF3D 各模块的定位有效性，本实验设计了 4 组模型，分别命名为 CIF3D(1) 至 CIF3D(4)。实验中，点云特征增强模块的编码器层数为 3，对称交互学习模块启动权重共享，跨模态迭代融合模块的迭代次数为 3，实验结果如表 3 所示。

表3　CIF3D组合消融实验结果

Tab. 3　CIF3D ablation experiment results　　(%)

Model name	Module combination			Nr3D					Parm
	PFEM	SILM	CIFM	Easy	Hard	VI	VD	Overall	/M
CIF3D(1)			√	58.41	44.21	52.19	51.27	51.74	72.76
CIF3D(2)	√		√	63.67	51.06	58.22	56.21	57.79	82.22
CIF3D(3)		√	√	61.03	48.17	55.01	53.75	54.92	74.34
CIF3D(4)	√	√	√	**65.20**	**52.65**	**59.23**	**58.01**	**59.46**	83.78

"VI": View-Indep; "VD": View-Dep.

对比CIF3D(1)与CIF3D(2)的实验结果可知,引入点云特征增强模块(PFEM)后,模型在Nr3D测试集上的定位准确率显著提升,其中"Hard"子集提升达6.85%,说明该模块有效增强了复杂场景理解能力。进一步对比CIF3D(1)至CIF3D(3)可发现,对称交互学习模块(SILM)虽整体增幅不及PFEM,但仅以1.58 M参数量即在"Hard"子集上带来3.96%的性能提升,表现出优异的参数效率。最后,CIF3D(4)在同时引入PFEM与SILM后达到最优性能,验证了各模块设计的合理性与互补性,在性能与复杂度之间取得了良好平衡。

3.4.2　文本编码器层数实验

文本编码器是文本特征提取的核心组件,其层数增加通常能带来更强的特征表示能力,但也会导致参数量上升影响训练与推理效率,为寻求性能与效率的平衡,本文通过设置不同的层数进行消融实验,除改变文本编码器层数外,其余配置均与组合消融实验一致,训练和推理时间采用10组共80对点云文本对的平均时间,实验结果如表4所示。

对比表中数据可知,随着文本编码器层数增加,模型在Nr3D数据集上的定位准确率逐步提升,并在12层时达到最高。然而,6层至12层间的性能提升幅度明显低于1至6层。与此同时,层数的增加显著提升了模型参数量,每新增一层都会引入约3.15 M的参数,极大增加训练和推理的时间成本。综合考虑性能增幅和模型复杂度,本文采用6层的文本编码器更为合理。

表4　文本编码器层数消融实验结果

Tab. 4　Text encoder layers ablation experiment results　　(%)

Layers	Nr3D					Parm	Train time	Inference time
	Easy	Hard	VI	VD	Overall	/M	/ms	/ms
1	53.88	41.33	47.91	46.69	48.14	68.01	37.71	19.67
2	59.71	47.16	53.74	52.52	53.97	71.17	39.03	20.80
4	63.83	51.28	57.86	56.64	58.09	77.47	41.79	22.24
6	65.20	52.65	59.23	58.01	59.46	83.78	44.08	24.33
8	65.26	52.71	59.29	58.07	59.52	90.08	47.97	26.01
12	65.38	52.83	59.41	58.19	59.64	102.69	53.59	29.46

"VI": View-Indep; "VD": View-Dep.

3.4.3　点云特征增强模块编码器层数实验

点云特征增强模块(PFEM)本质上是一个多层Transformer编码器,为研究编码器层数对点云特征增强效果及定位准确率的影响,本文在保持其他模块配置不变的前提下,对PFEM中Transformer编码器的层数进行了消融实验,层数设置范围为0~4层,实验结果如表5所示。

根据实验结果可知,仅引入单层编码器的PFEM模块即可显著提升模型定位准确率,在"Easy"子集上提升了2.07%。随着编码器层数的增加,模型参数量逐渐增大,PFEM模块的性能增幅趋于平缓,在3层时达到最大。然而,当层

表5　PFEM编码器层数消融实验结果

Tab. 5　PFEM encoder layers ablation experiment results　(%)

Layers	Nr3D					Parm
	Easy	Hard	VI	VD	Overall	/M
0	61.03	48.17	55.01	53.75	54.92	74.34
1	63.10	49.75	56.98	55.47	56.59	77.49
2	64.68	51.54	58.61	57.17	58.33	80.64
3	**65.20**	**52.65**	**59.23**	**58.01**	**59.46**	83.78
4	64.96	51.63	58.79	57.64	58.74	86.94

"VI": View-Indep; "VD": View-Dep.

数增加至4层时,参数量过大导致模型过拟合,准确率开始下降。结果表明,PFEM模块在性能与模型复杂度之间存在平衡,需要合理控制编码器层数,以获得最佳的综合性能。

3.4.4　对称交互学习模块实验

在对称交互学习模块(SILM)中,文本与点云模态通过权重共享机制构建了统一的学习规则,实现了高效的特征更新,并显著降低了模型参数量。为探究权重共享策略对模块性能的影响,本文设置了三组实验,分别为无SILM模块、权重不共享的SILM模块和权重共享的SILM模块。其中,权重不共享指在文本和点云特征的多头注意力中,每个头对应的查询向量与键值向量相互独立,实验结果如表6所示。

表6　SILM权重共享/不共享消融实验结果

Tab. 6　SILM weight sharing/Non-Sharing ablation experiment results　(%)

SILM	Nr3D					Parm
	Easy	Hard	VI	VD	Overall	/M
No	63.67	51.06	58.22	56.21	57.79	82.22
Sharing	**65.20**	52.65	**59.23**	**58.01**	**59.46**	83.78
Non-sharing	64.91	**52.86**	59.07	57.79	59.17	85.37

"VI": View-Indep; "VD": View-Dep.

综合分析表内数据可知,无论是否采用权重共享,引入SILM模块均能提升模型的定位准确率。进一步对比第2与第3行结果可知,权重不共享方法凭借更大的参数量增强了对复杂点云的理解能力,在“Hard”子集上表现更优。然而综合考虑整体性能与模型复杂度,权重共享方法更具优势,因为其参数量仅为83.79 M,相较于权重不共享方法减少了1.58 M,且在其余4个子集上均取得更高准确率,实现了性能与复杂度的有效平衡。

3.4.5　跨模态迭代融合模块实验

本文跨模态迭代融合模块(CIFM)采用训练-推理解耦的设计,在训练阶段引入多视图空间增强,提高模型对复杂点云的理解能力,在推理阶段取消多视图空间增强,实现高效且准确的跨模态融合。通过交叉注意力与空间聚合层深度融合文本、点云与类别特征,利用多轮迭代为目标定位提供更加精准的融合特征。为了研究迭代次数对模型定位准确率的影响,本文设计了4组实验,实验结果如表7所示。

由实验结果可知,随着迭代次数增加,模型定位准确率逐步提高,并在第3次迭代时达到峰值。然而,每次迭代会引入约8.98 M的参数量,到第4次迭代时,总参数量达到92.72 M,显著增加了模型复杂度,进而导致模型过拟合,定位准确率下降。为了更加直观展示CIFM模块的跨模态融合效果,本文基于Nr3D数据集随机选取4组点云文本对,利用Open3D库对每轮迭代过程中的融合特征进行了三维可视化,如图6所示。

表7　CIFM迭代次数消融实验结果

Tab. 7　CIFM iterations count ablation experiment results　　(%)

Iterations	Nr3D					Parm
	Easy	Hard	VI	VD	Overall	/M
1	59.46	45.79	53.26	52.08	52.97	65.90
2	63.02	49.61	56.75	55.86	56.17	74.84
3	**65.20**	**52.65**	**59.23**	**58.01**	**59.46**	83.78
4	64.97	52.14	58.97	57.62	58.76	92.72

"VI": View-Indep; "VD": View-Dep.

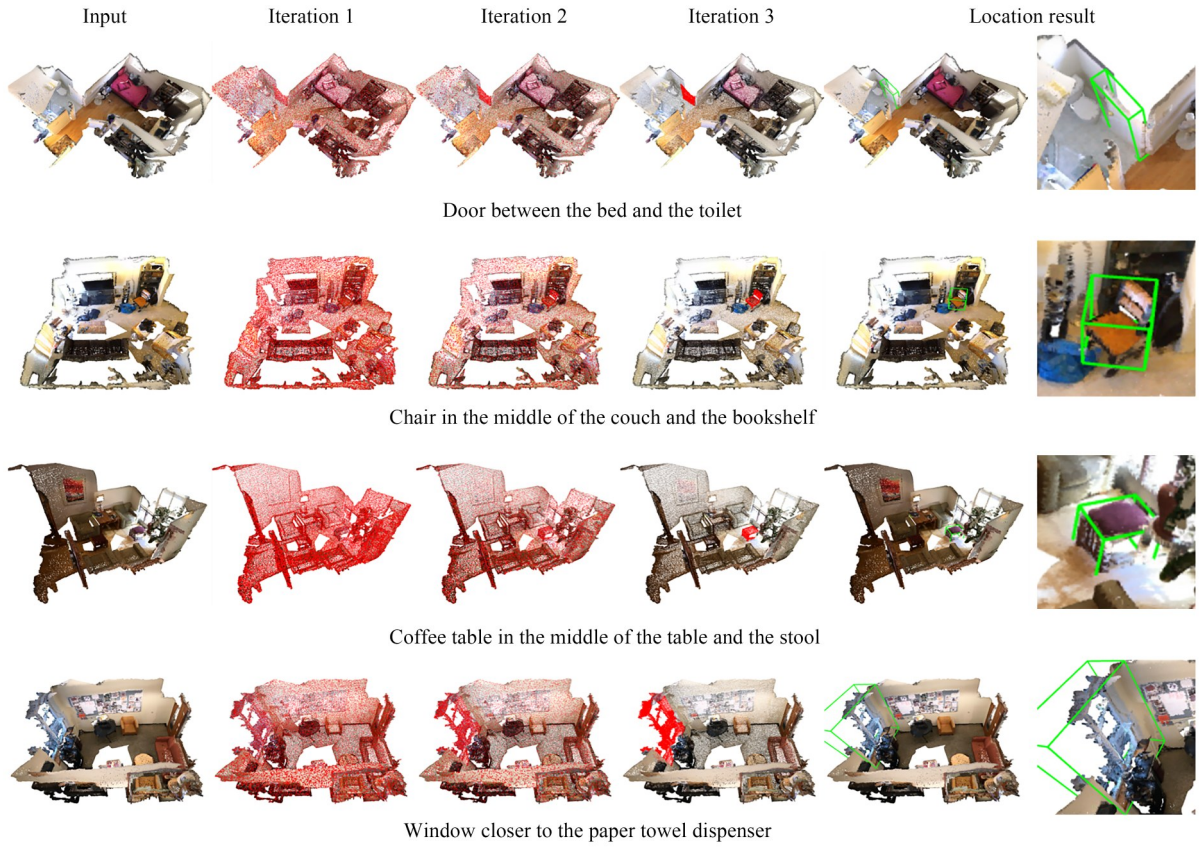

图6　CIFM迭代特征可视化结果

Fig. 6　CIFM iterative feature visualization results

图示结果表明,随着迭代次数增加,点云图像中的冗余背景信息被逐渐滤除,模型对文本描述相关目标内容的关注度得到显著增强,这一现象验证了本文所提出CIFM模块的设计有效性。该模块能够有效利用不同的多模态特征输入,通过多层交叉注意力与空间聚合层捕捉跨模态的语义关联,并借助迭代优化机制不断抑制无关的点云信息,最终为定位任务提供质量更高且判别性更强的特征表示。

3.5　可视化

图7展示了本文算法在12组点云文本对上的定位可视化结果,涵盖了"Easy","Hard","View-Indep"和"View-Dep"共4种划分方式,以及"一对多"和"多对一"两种实验类型,按顺序分别对应图中1~6行。其中,"一对多"指同一文本输入定位多个不同的点云,"多对一"指不同文本输入至同一点云中进行定位。

结果表明,得益于多模态特征融合模块

(Cross-modal Information Fusion Module,CIFM)的迭代融合机制,本文方法在处理不同类型点云文本对时均具有稳定的定位能力。针对干扰物较多的"Hard"类型或依赖视角的"View-Dep"类型,本文方法均能够通过迭代机制逐步降低定位歧义,成功识别并定位文本所描述的目标对象。

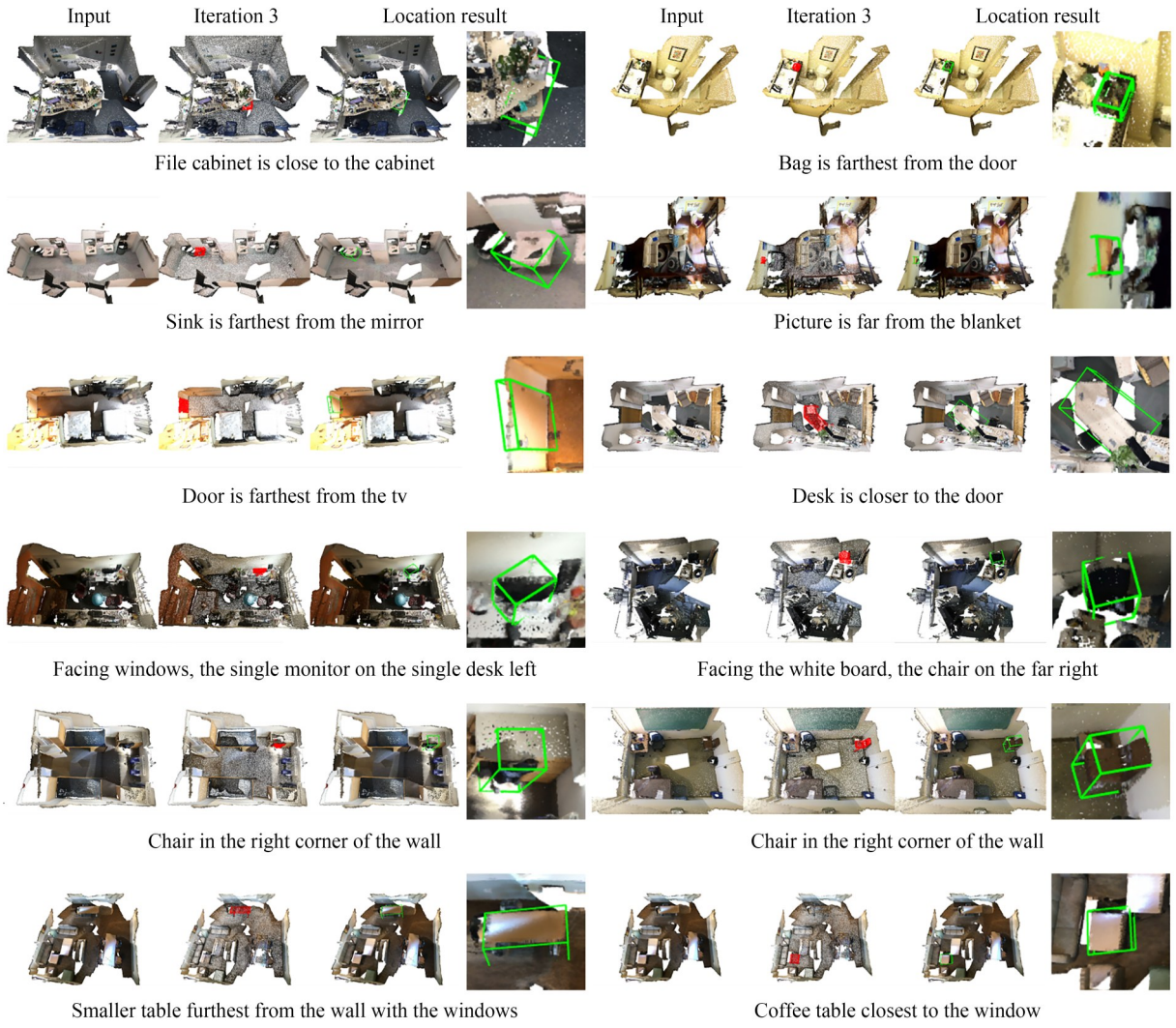

图 7　3D 视觉定位可视化结果

Fig. 7　3D Visual Grounding Visualization Results

4　结　论

本文提出跨模态交互学习与迭代融合的CIF3D方法,有效缓解了现有方法视角变化适应性差以及跨模态融合效果有限的问题。通过引入点云类别信息并设计点云特征增强模块,显著提升了点云特征的质量,对称交互学习模块和跨模态迭代融合模块有效更新并融合多模态特征,增强了模型在复杂3D场景中的目标定位能力。实验结果表明,本文方法在多个数据集上均实现了精度提升。其中,在 ScanRefer 的"Unique","Multiple"和"Overall"这3个子集中,本文方法相比于VPP-Net在ACC@0.25指标上分别提升了0.14%,1.01%和0.87%;在Nr3D和Sr3D的"Easy"子集中,本文方法相比于CFAM和GNL3D分别提升了2.9%和2.07%。尽管如此,本方法在处理极端视角时仍存在局限性,且跨模态特征融合的效率与精度仍有待提升,未来研究可通过结合3D大模型技术改善

模型的定位精度与泛化能力,从而推动智能导航、机器人辅助室内设计等领域的应用。

作者贡献声明:

才华:方法设计,论文构思和撰写;

冉越:实验设计,数据整理与结果可视化;

张海峰:论文审核与编辑,实验数据分析;

张高鹏:论文指导与审核,实验数据分析;

付强:论文指导,资源支持;

孙俊喜:方法可行性评估与分析。

参考文献:

[1] 周肖剑,王晓红,李炜,等. 结合双注意力和加权动态图卷积的三维点云分类与分割[J]. 光学精密工程,2024,32(18):2823-2835.
ZHOU X J, WANG X H, LI W, *et al*. 3D point cloud classification and segmentation based on dual attention and weighted dynamic graph convolution [J]. *Opt. Precision Eng.*, 2024, 32(18): 2823-2835. (in Chinese)

[2] LIU Y, ZHANG C S, DONG X J, *et al*. Point cloud-based deep learning in industrial production: a survey [J]. *ACM Computing Surveys*, 2025, 57(7): 1-36.

[3] GANIN Y, BARTUNOV S, LI Y, *et al*. Computer -aided design as language[J]. *Neural Information Processing Systems*, 2021, 34: 5885-5897.

[4] LI Y L, LIU X P, LU H, *et al*. Detailed 2D-3D Joint Representation for Human-Object Interaction [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020. *Seattle*, *WA*, *USA*. IEEE, 2020.

[5] 郝雯,张雯静,梁玮,等. 面向三维点云的场景识别方法综述[J]. 光学 精密工程, 2022, 30(16): 1988-2005.
HAO W, ZHANG W J, LIANG W, *et al*. Scene recognition for 3D point clouds: a review[J]. *Opt. Precision Eng.*, 2022, 30(16): 1988-2005. (in Chinese)

[6] CHEN D Z, CHANG A X, NIEßNER M. ScanRefer: 3D object localization in RGB-D scans using natural language [C]. *Computer Vision-ECCV* 2020. *Cham*: *Springer International Publishing*, 2020: 202-221.

[7] ACHLIOPTAS P, ABDELREHEEM A, XIA F, *et al*. ReferIt3D: neural listeners for fine-grained 3D object identification in real-world scenes[C]. *Computer Vision-ECCV* 2020. *Cham*: *Springer International Publishing*, 2020: 422-440.

[8] ZHAO L C, CAI D G, SHENG L, *et al*. 3DVG-Transformer: relation modeling for visual grounding on point clouds[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2021: 2908-2917.

[9] 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]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2021: 1771-1780.

[10] ZHANG Y M, GONG Z M, CHANG A X. Multi3DRefer: grounding text description to multiple 3D objects[C]. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 1-6, 2023, *Paris*, *France*. IEEE, 2023: 15179.

[11] LUO J Y, FU J H, KONG X H, *et al*. 3D-SPS: single-stage 3D Visual grounding *Via* referred point progressive selection[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 16433-16442.

[12] JAIN A, GKANATSIOS N, MEDIRATTA I, *et al*. Bottom UpTop DownDetection transformers forlanguage grounding in images and point clouds [C]. *Computer Vision-ECCV* 2022. *Cham*: *Springer Nature Switzerland*, 2022: 417-433.

[13] WU Y M, CHENG X H, ZHANG R R, *et al*. EDA: explicit text-decoupling and dense alignment for 3D visual grounding [C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 19231-19242.

[14] JIN Z, HAYAT M, YANG Y W, *et al*. Context-aware alignment and mutual masking for 3D-language Pre-Training[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*,

Canada. IEEE, 2023: 10984-10994.

[15] QIAN Z P, MA Y W, LIN Z K, *et al*. Multi-branch collaborative learning network for3D visual grounding [C]. *Computer Vision-ECCV* 2024. *Cham*: *Springer Nature Switzerland*, 2025: 381-398.

[16] SHI X X, WU Z H, LEE S. Viewpoint-aware visual grounding in 3D scenes [C]. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 16-22, 2024, *Seattle*, *WA*, *USA*. IEEE, 2024: 14056-14065.

[17] GUO P, ZHU H Y, YE H C, *et al*. Revisiting 3D visual grounding with context-aware feature aggregation [J]. *Neurocomputing*, 2024, 601: 128195.

[18] HUANG W C, LIU D Z, HU W. Advancing 3D object grounding beyond a single 3D scene [C]. *Proceedings of the* 32*nd ACM International Conference on Multimedia*. *Melbourne VIC Australia*. ACM, 2024: 7995-8004.

[19] QI C R, YI L, SU H, *et al*. Pointnet++: Deep hierarchical feature learning on point sets in a metric space [J]. *arXiv preprint* arXiv: 1706.02413, 2017.

[20] RADFORD A, KIM J W, HALLACY C, *et al*. Learning transferable visual models from natural language supervision [J]. *arXiv preprint* arXiv: 2103.00020, 2021.

[21] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need [J]. *arXiv preprint* arXiv: 1706.03762, 2017.

[22] 才华, 易亚希, 付强, 等. 基于跨模态引导和对齐的多模态预训练方法[J]. 电子学报, 2024, 52(10): 3368-3381.
CAI H, YI Y X, FU Q, *et al*. Multimodal pre-training with cross-modal guidance and alignment [J]. *Acta Electronica Sinica*, 2024, 52(10): 3368-3381. (in Chinese)

[23] GAO Z X, JIANG X, XU X, *et al*. Embracing unimodal aleatoric uncertainty for robust multimodal fusion [C]. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 16-22, 2024, *Seattle*, *WA*, *USA*. IEEE, 2024: 26866-26875.

[24] WANG Z X, YU Y L, LI X Z. Rethinking local-to-global representation learning for rotation-invariant point cloud analysis [J]. *Pattern Recognition*, 2024, 154: 110624.

[25] DAI A, CHANG A X, SAVVA M, *et al*. ScanNet: richly-annotated 3D reconstructions of indoor scenes [C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *July* 21-26, 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 2432-2443.

[26] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C]. *International Conference on Learning Representations*, 2019

通讯作者：

才　华(1977—),男,吉林长春人,现为长春理工大学电子信息工程学院教授,博士生导师,主要从事机器视觉、自然语言处理与视觉语言多模态方面的研究。E-mail: caihua@cust.edu.cn

作者简介：

冉　越(2000—),男,河南郑州人,现为长春理工大学电子信息工程学院硕士研究生,主要从事自然语言处理与视觉语言多模态方面的研究。E-mail:2023100683@mails.cust.edu.cn