基于深度融合多模态特征的三维点云分类小样本类增量学习

朱晨浠[1]，卢奕南[1]，伍铁如[2]，龚文勇[3]，马锐[2]

(1. 吉林大学计算机科学与技术学院，吉林 长春 130012；
2. 吉林大学人工智能学院，吉林 长春 130012；
3. 暨南大学信息科学技术学院，广东 广州 510632)

摘　　要：传统 3D 点云分类方法在小样本类增量学习(FSCIL)场景下容易出现泛化能力不足和灾难性遗忘等问题。预训练语言-图像模型(CLIP)因具备丰富的 2D 形状先验知识，被证明能够有效提升 3D FSCIL 性能，但现有基于 CLIP 的框架在多模态特征提取与融合过程中仍缺乏灵活性与自适应性，导致增量阶段的分类准确率受限。为解决这些不足，提出了一种深度融合多模态特征的 3D FSCIL 方法，通过引入基于门控单元与残差块的自适应适配器实现多尺度特征对齐与冗余信息抑制，并设计基于自注意力机制的多模态全局特征动态融合模块，根据不同样本特性自适应调整多路特征的权重分配，从而获得更加一致且互补的融合表示。具体地，将点云渲染为多视角深度图，分别利用原始 CLIP 视觉编码器与在深度图上预训练的 CLIP 编码器提取特征，并结合点云几何特征，经自适应适配器处理后送入注意力融合模块，与 CLIP 文本编码器提取的语义特征对齐进行分类。此外，结合对比学习损失、多视角与几何扰动数据增强策略以及记忆回放机制，有效缓解小样本条件下的过拟合与遗忘问题。在 ShapeNet、ModelNet 及 CO3D 数据集上的实验结果表明，与现有主流 3D FSCIL 方法相比，该方法在各增量阶段均取得更高的准确率，且相对准确度下降率与最大阶段跳变率显著降低。

关　键　词：3D 点云；增量学习；小样本学习；3D 分类；预训练模型

中图分类号：TP 391.41；TP 18　　**DOI**：10.11996/JG.j.2095-302X.2026010078

文献标识码：A　　**文 章 编 号**：2095-302X(2026)01-0078-12

Deep fusion of multimodal features for few-shot class-incremental 3D point cloud classification

ZHU Chenxi[1], LU Yinan[1], WU Tieru[2], GONG Wenyong[3], MA Rui[2]

(1. College of Computer Science and Technology, Jilin University, Changchun Jilin 130012, China;
2. School of Artificial Intelligence, Jilin University, Changchun Jilin 130012, China;
3. College of Information Science and Technology, Jinan University, Guangzhou Guangdong 510632, China)

Abstract: Traditional 3D point-cloud classification methods tend to suffer from insufficient generalization and catastrophic forgetting in Few-Shot Class-incremental Learning (FSCIL) scenarios. The pretrained vision-language model CLIP (Contrastive Language-Image Pre-training), which contains rich 2D shape priors, has been shown to effectively enhance 3D FSCIL performance. However, existing CLIP-based frameworks still lack flexibility and adaptability in multimodal feature extraction and fusion, which limits classification accuracy during incremental stages. To address these shortcomings, a 3D FSCIL approach with deeply fused multimodal features was proposed. An adaptive adapter based on gated units and residual blocks was introduced to achieve multi-scale feature alignment and redundancy suppression, and a multimodal global feature dynamic fusion module with self-attention was designed to

收稿日期：2025-06-30；定稿日期：2025-08-23；通信作者：马锐，E-mail：ruim@jlu.edu.cn

Received：30 June，2025；**Finalized**：23 August，2025；**Corresponding author**：MA Rui, E-mail：ruim@jlu.edu.cn

基金项目：国家自然科学基金(62202199)

Foundation items：National Natural Science Foundation of China (62202199)

adaptively adjust the weight allocation of different feature streams according to sample characteristics, thereby obtaining more consistent and complementary fused representations. Specifically, point clouds were rendered into multi-view depth maps, and features were extracted using both the original CLIP visual encoder and a CLIP encoder pretrained on depth maps, combined with point-cloud geometric features. After processing through the adaptive adapter, these features were fed into the attention-based fusion module and aligned with semantic features extracted by the CLIP text encoder for classification. In addition, contrastive learning loss, multi-view and geometric perturbation-based data augmentation strategies, and a memory-replay mechanism were incorporated to effectively mitigate overfitting and forgetting under few-shot conditions. Experiments on ShapeNet, ModelNet, and CO3D demonstrated that the proposed method consistently achieved higher accuracy across incremental stages compared with existing 3D FSCIL approaches, while significantly reducing both relative accuracy drop rates and maximum stage fluctuations.

Keywords: 3D point cloud; incremental learning; few-shot learning; 3D classification; pre-trained model

近年来，随着 3D 传感器、激光扫描及计算机图形学技术的不断进步，数字 3D 资产在各行各业的应用正迅速发展[1-5]。从工业制造到文物保护，再到虚拟现实和智能机器人，海量的 3D 数据正以前所未有的速度增加。与此同时，3D 数据具有类别繁多、分布分散的特点，使得依赖全监督训练的传统 3D 点云识别模型[6-7]在小样本条件下面临巨大挑战，其泛化能力和鲁棒性亟待提升。为了更高效地适应不断扩展的新类别与场景，研究者们开始关注模型在有限标注数据下持续学习的能力，3D 小样本类增量学习(Few-Shot Class-Incremental Learning, FSCIL)[8]在实际应用中变得越来越重要。

一些对 2D 少样本和零样本任务的研究工作[9-10]表明，预训练模型(Pre-Trained Models, PTMs)在增量学习场景下表现优异，且优于不使用 PTMs 的方法。其主要原因为 PTMs 中所蕴含的先验知识增强了下游任务的泛化能力[11-13]。为了在 3D 任务中使用这些 2D 的 PTMs，近期的一些研究工作[14-16]对如何将 3D 表示与语言-图像(Vision-Language, V-L)的 PTMs 知识对齐进行了研究，取得了一些成果，如先将点云渲染成深度图，然后再提取特征。然而，想要将 2D V-L PTMs 应用到 3D FSCIL[17]任务中仍然面临着一些挑战。首先，RGB 图像与 3D 点云渲染出的深度图之间存在域差异，而 2D V-L PTMs 一般在 RGB 图像上进行预训练，所捕获的特征也聚焦于图像的颜色和纹理。其次，真实扫描的点云数据存在着噪声以及一些相似的局部结构，导致投影得到的深度图存在干扰或冗余信息，以致 2D V-L PTMs 提取的特征不够准确，对分类性能造成很大影响。现有的基于 2D V-L PTMs 的 3D FSCIL 方法[18]对这些问题做了一些改进，如另通过在深度图上进行微调 PTMs 来提取特征等。然而，3D FSCIL 方法在特征的提取和融合上仍有所欠缺：提取特征的方法较为粗略，缺乏灵活性，不能很好地剔除样本中的冗余信息与噪声。此外，对于不同预训练策略得到的特征，现有方法[18]仍以相同的权重进行融合，且无法根据不同样本的情况动态调整各部分特征的重要性，使融合过程缺乏灵活性和自适应性，也使得这些 3D FSCIL 方法在性能上仍有提升的空间。

针对上述问题，本文对现有的 3D FSCIL 框架进行了改进。首先，提出了一种基于门控单元和残差块的自适应适配器(Adapter)，在其中引入了门控线性单元(Gated Linear Unit，GLU)，以动态调整不同特征的贡献，抑制冗余信息保留有效信息并结合残差块实现多尺度特征对齐。其次，利用了自注意力融合机制对多模态全局特征进行动态融合，使得模型能够学习到多角度的信息交互，从而可以更加有效地整合特征。此外，本文还针对小样本类增量问题进行了专门的设计：一方面，充分利用 CLIP 作为 V-L PTMs 所蕴含的语义先验，及较强的表示能力缓解下游任务对大规模标注样本的依赖；另一方面，通过融合点云和深度图 2 种模态特征，提升模型在小样本条件下的特征表达能力。同时，本文引入对比学习损失和基于多视角、几何扰动的数据增强策略，在增强特征判别性的同时扩大了特征空间的覆盖范围。最后，结合记忆回放机制以缓解灾难性遗忘，增强模型在各增量阶段对旧类知识的保持能力。上述多层次设计共同构建了一个兼具小样本适应性与类增量扩展性的 3D FSCIL 学习框架。

总体而言，本文的主要贡献总结如下：

1) 提出了一种基于门控单元和残差块的自适应适配器，可以实现更加有效的全局特征提取以及多模态特征的深度融合。

2) 通过使用基于自注意力机制的方法对多模态全局特征进行动态融合，提高了最终融合特征的表达效果，为模型提供了更强的特征聚合能力。

3) 针对 FSCIL 场景进行了系统性设计，提升了模型在小样本条件下的判别性与泛化能力，缓解了类增量条件下的灾难性遗忘。

4) 与现有的主流 FSCIL 方法相比，本方法在 ShapeNet，ModelNet 和 CO3D 数据集上的实验均有显著的性能提升。

1 相关工作

1.1 3D 点云理解

近年来，出现了许多基于深度学习的方法用于识别点云对象。PointNet[6]是首个直接处理原始点云的研究，其结合多层感知机(Multi-Layer Perceptron，MLP)和对称函数的方式来学习和聚合点云特征，但忽略了点之间的局部空间关系。因此，后续研究方法更关注提取局部和全局特征。如 PointNet++[7]通过分层结构提取局部特征；文献[18-24]基于 3D 点云提出了新的卷积操作。此外，还有一些网络将点云中的点视为图的顶点，并通过邻接点编码局部信息[25-26]。如动态图卷积神经网络(Dynamic Graph CNN，DGCNN[26])提出了 EdgeConv 模块，该模块在特征空间上计算，并连接中心点和其邻域点的特征。

还有一些研究以掩码点建模(Masked point modeling)作为一种 3D 自监督学习策略，并取得了显著成功。例如，Point BERT[27]使用预训练的分词器来预测离散的点标签，而 Point MAE[28]和 Point-M2AE[29]则应用掩码自编码器直接重建被掩盖的 3D 坐标。此外，还有一些研究采用基于 Transformer 的方法[30]进行 3D 点云识别。最近，受 V-L PTMs 的影响，一些研究[14-15]将 CLIP 应用于点云识别，实现了 2D 预训练知识向 3D 领域的迁移，并表现出了优异的性能。

尽管上述方法在 3D 点云理解中取得了显著进展，但大多依赖大规模标注数据进行训练，缺乏在小样本类增量设置下的泛化能力。本文旨在将 V-L PTMs 应用到 3D 点云 FSCIL 任务中，以提升模型的泛化性与鲁棒性。

1.2 小样本类增量学习

文献[8]首次提出了用 Neural Gas 网络来解决 FSCIL 问题。随后，一些研究提出了在嵌入空间中进行向量量化[31]、词向量蒸馏[32]以及参数选择[33]等方法。大多数 FSCIL 方法[33-37]都选择冻结特征提取器，仅训练线性分类器，或采用原型(Prototype)进行分类。持续更新分类器(Continually Evolved Classifier，CEC[36])聚焦于分类器的自适应，设计了一个额外的图结构模型。前向兼容训练(ForwArd Compatible Training，FACT[37])和基于增强角损失函数的增量分类(Augmented Angular Loss Incremental Classification，ALICE[34])则致力于在基础训练阶段学习一个可扩展且紧凑的特征空间。FACT 将同一类别的样本拉近，同时引入虚拟原型进行基础训练。ALICE 则利用余弦相似度和间隔(Margin)来学习一个更适用于横向学习的特征空间。Constrained FSCIL[38]引入了一个可训练的全连接层和一个可重写的记忆模块，并提供了 3 种模型更新方式以实现资源受限下的高效 FSCIL。LIU 等[39]则提出了一种无数据重放(Data-free replay)机制，通过生成器合成数据以代替对历史样本的访问。Microshape[17]首次针对点云分类任务进行小样本类增量方法的研究，将 3D 对象表示为一组预定义的微形状(正交基向量)，从而实现对新类别的增量学习。

近年来，受 2DPTMs 突破性进展的影响，许多研究尝试利用 PTMs 丰富的先验知识来帮助在 FSCIL 任务中学习新概念以及缓解灾难性遗忘[13,40]，如 V-L PTMs 的 CLIP 由于蕴含丰富 2D 形状相关的先验知识，其天然适配 FSCIL 场景。FILP-3D[18]是目前最新的利用 PTMs 的 3D FSCIL 方法，其充分利用了 V-L 模型(CLIP)所蕴含的形状先验知识，并通过引入空间噪声补偿器模块和冗余特征消除器模块，增强了模型对噪声的鲁棒性，使得特征对齐更加准确。然而 FILP-3D 在进行多尺度特征提取以及特征融合时，仍欠缺一定的灵活性和自适应性，模型的泛化能力仍有待提升。本文将沿用 FILP-3D 中利用 CLIP 模型的方法，并引入基于门控单元和残差块的自适应适配器以及基于自注意力机制的多模态全局特征融合模块，实现更加灵活的多模态特征深度融合，使得模型的泛化能力进一步提高。

2 本文方法

本文方法整体架构图如图 1 所示，分别用 3D

编码器和 CLIP 模型提取点云和深度图特征，并通过自适应适配器和自注意力融合模块进行特征的深度融合，最终与 CLIP 提取的文本特征计算相似度并得到最终概率。

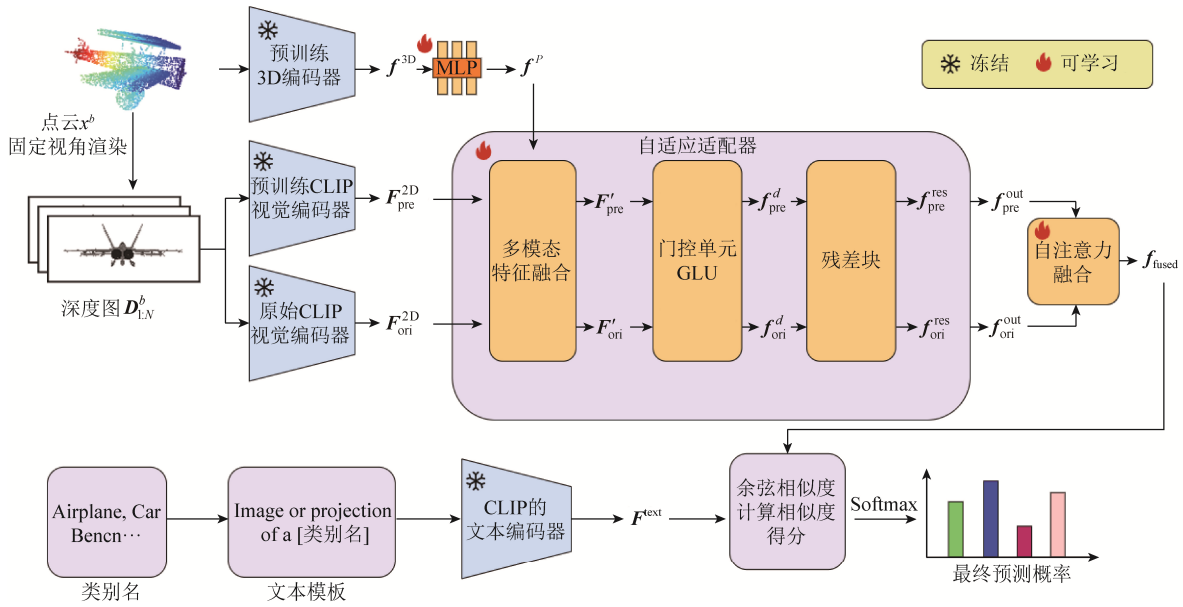

图 1　本文方法架构图

Fig. 1　Architecture of the proposed method

2.1　小样本类增量学习问题定义

给定一系列任务[1,2,⋯,B]，每个任务 b 包含一个带标签的训练集 $T^b=\left\{\left(x_i^b,y_i^b\right)\right\}_{i=1}^{|T^b|}$，其中 $x_i^b\in X$ 表示一个输入点云(或图像等其他数据类型)，$y_i^b\in Y$ 表示其标签，$|\cdot|$ 表示集合的大小。每个任务 b 与其他任务 b'是不相交的，即 $\forall b,b',T^b\cap T^{b'}=\varnothing$。

第一个任务(b=1)包含大量训练样本，而其余的任务(b>1)仅包含远少于第一个任务的小样本数据(通常每个类仅选取 5 个样本)，即 $\left|T^1\right|\gg\left|T^{b>1}\right|$。为简便起见，本文将任务 1 称为基础任务，其余任务称为增量任务。

对于一个深度神经网络 $g_\phi\left(f_\theta(\cdot)\right)$，其中 $f(\cdot)$ 和 $g(\cdot)$ 分别表示特征提取器和分类器，参数为 θ 和 ϕ，FSCIL 的目标是找到最优参数 θ^* 和 ϕ^*，使得 $g_\phi\left(f_\theta(\cdot)\right)$ 能识别从任务 1 到当前任务 b 的所有学习类别，即 $\{T^1,\cdots,T^b\}$。

2.2　方法概览

最近的研究工作表明，PTMs 在增量学习任务中表现出色，如 V-L PTMs CLIP 由于蕴含丰富 2D 形状相关的先验知识，其天然适配 FSCIL 场景。受此启发，最新的 3D FSCIL 框架[18]通过将点云投影成不同视角的深度图，间接地利用 CLIP 注入形状相关的先验知识。本文将用这种框架。

具体地，给定存在 B 个任务的序列 $D=\{D^1,D^2,\cdots,D^B\}$。其中 D^1 为基础任务，包含大量的点云训练样本。D^2~D^B 则为增量任务，只有少量的点云样本供训练。给定第 b 个任务中的第 i 个训练样本 $\left(\boldsymbol{x}_i^b,y_i^b\right)$，其中 $\boldsymbol{x}_i^b$ 为点云，y_i^b 为对应的标签。本文将点云 $\boldsymbol{x}_i^b$ 渲染成多视角的深度图 $\boldsymbol{D}_{1:N}^b$ 后，沿用 FILP-3D[18]的做法，分别使用原始 CLIP 的视觉编码器[41]以及经过预训练[14]的 CLIP 的视觉编码器来提取深度图特征 $\boldsymbol{F}_{\text{ori}}^{2D}\in\mathbb{R}^{N\times C}$ 和 $\boldsymbol{F}_{\text{pre}}^{2D}\in\mathbb{R}^{N\times C}$。其中 N 为渲染时采用的视角数，C 为视觉编码器的嵌入维度。

此外，为了弥补点云在投影时丢失的几何结构和空间分布信息，以及缓解小样本场景下单一模态可能存在的噪音和缺陷问题，本文还使用了基于图的 3D 模型来提取点云特征 $\boldsymbol{f}^{3D}$，对 2D 多视图特征进行补充。具体地，将沿用遮挡补全(Occlusion Completion，OcCo)[42]中的方法，选择了在 ShapeNet 上经过预训练的 DGCNN 模型。相较于 PointNet[6]和 PointNet++[7]等方法，基于图的 DGCNN 在局部结构建模和特征表达能力上表现更为优越，尤其适合处理真实扫描点云中存在的复杂几何结构与噪声。点云特征 $\boldsymbol{f}^{3D}$ 通过可学习的 MLP 与多视图特征进行维度对齐，得到

$$\boldsymbol{f}^{p}=f_p^2\left(\text{ReLU}\left(f_p^1\left(\boldsymbol{f}^{3D}\right)\right)\right)\tag{1}$$

式中：f_p^1 和 f_p^2 表示可学习的 MLP。

为了充分利用多视图特征与 3D 特征的优势，本文提出一种基于门控单元和残差块的自适应适配器 Adapter 来提取全局融合特征，即

$$\boldsymbol{f}_{\text{ori}}^{\text{out}} = Adapter\left(\boldsymbol{f}^{p}, \boldsymbol{F}_{\text{ori}}^{\text{2D}}\right) \tag{2}$$

$$\boldsymbol{f}_{\text{pre}}^{\text{out}} = Adapter\left(\boldsymbol{f}^{p}, \boldsymbol{F}_{\text{pre}}^{\text{2D}}\right) \tag{3}$$

对于 2 种不同分支的全局特征 $\boldsymbol{f}_{\text{ori}}^{\text{out}}$ 和 $\boldsymbol{f}_{\text{pre}}^{\text{out}}$，本文提出一种基于自注意力机制的方式进行加权融合，得到最终的全局特征，即

$$\boldsymbol{f}^{\text{fused}} = Self\text{-}Attn\left(\boldsymbol{f}_{\text{ori}}^{\text{out}}, \boldsymbol{f}_{\text{pre}}^{\text{out}}\right) \tag{4}$$

对于每个标签为 k 的类别以及其对应的名称 t_k，本文采用如下模板生成文本提示：“image or projection or sketch of a t_k”。为了利用 CLIP 的文本-图像对齐先验知识，本文使用 CLIP 的文本编码器提取文本提示的特征，即

$$\boldsymbol{F}_k^{\text{text}} = Encoder_{\text{text}}\left(prompt(t_k)\right) \tag{5}$$

由于多视图特征已经与 CLIP 的文本嵌入预对齐，可以直接利用全局特征 $\boldsymbol{F}^{G}$ 和文本特征 $\boldsymbol{F}_k^t$ 之间的余弦相似度计算出第 k 类的 logit，即

$$l_k = \cos\left(\boldsymbol{f}^{g}, \boldsymbol{F}_k^{t}\right) \tag{6}$$

最终，预测概率为

$$\boldsymbol{p} = \text{Softmax}\left([l_1, \cdots, l_K]\right) \tag{7}$$

2.3 基于门控单元和残差块的自适应适配器

为了更好地融合多视角特征并实现多尺度特征对齐，本文提出了一种基于门控单元和残差块的自适应适配器模块 Adapter。该模块主要设计思想在于利用 GLU 动态调整多模态特征贡献，同时结合残差块实现多尺度特征对齐，抑制冗余信息从而获得更为鲁棒和具有判别性的全局特征表示。自适应适配器作为轻量化的结构模块，在类增量学习场景中具备显著优势：其允许模型在主干参数冻结的前提下，仅通过微调少量适配器参数吸收新类知识，避免对旧类表示的干扰。Adapter 以经过维度对齐的点云特征 $\boldsymbol{f}^{p}$ 和多视图特征 $\boldsymbol{F}_{1:N}^{\text{2D}}$ 之和 $\boldsymbol{F}_{1:N}^{g}$ 作为输入，输出为最终的全局特征，即

$$\boldsymbol{F}_{1:N}^{g} = \boldsymbol{f}^{p} + \boldsymbol{F}_{1:N}^{\text{2D}} \tag{8}$$

$$\boldsymbol{f}^{\text{out}} = Adapter\left(\boldsymbol{F}_{1:N}^{g}\right) \tag{9}$$

具体实现如下：

1) 对于输入的多视角融合特征进行加权融合，即

$$\boldsymbol{F}' = \boldsymbol{F}_{1:N}^{g} \odot \boldsymbol{\alpha} \tag{10}$$

式中：$\boldsymbol{\alpha}$表示可学习的视角加权系数向量，在训练中倾向于对信息量少的视角(如遮挡严重的视角)分配低权重，从而减少无效特征对融合结果的干扰。

2) 将加权融合后的特征进行展平拼接，即

$$\boldsymbol{F}_{\text{flat}} = Flatten(\boldsymbol{F}') \tag{11}$$

展平后的特征输入到全连接层进行非线性映射。与传统的线性映射不同，本方法先使用全连接层将展平后的特征映射到 2D 维空间，以适应后续的门控操作，即

$$\boldsymbol{H} = \boldsymbol{W}_1 \boldsymbol{F}_{\text{flat}} + \boldsymbol{b}_1 \tag{12}$$

3) 采用 GLU 将二维的特征自动分为 2 部分，一部分作为输入信号，另一部分作为门控权重，通过逐元素相乘实现信息筛选，从而得到最终 d 维的特征，即

$$GLU(\boldsymbol{H}) = \boldsymbol{H}_1 \otimes \text{Sigmoid}(\boldsymbol{H}_2) \tag{13}$$

$$\boldsymbol{f}^{d} = \boldsymbol{W}_2 GLU(\boldsymbol{H}) + \boldsymbol{b}_2 \tag{14}$$

式中：$\boldsymbol{H}_1$ 和 $\boldsymbol{H}_2$ 分别表示一半的特征。若 $\boldsymbol{H}_2$ 的某些维度权重趋近于 0，对应 $\boldsymbol{H}_1$ 的特征会被抑制。如点云投影中重复的局部结构或噪声对应的特征通道可能被低权重过滤。

为了进一步提升特征的多尺度对齐能力，模块还引入了残差块。具体而言，在得到全局融合特征后，通过一个残差块进行特征变换，提取跨尺度的信息，浅层特征保留细节(如边缘)，深层特征聚合全局上下文(如物体整体形状)。最后，将变换后的特征与原始全局特征进行残差连接，模型自动选择跨尺度的判别性特征，避免冗余局部特征的过度累积，从而实现更稳健的特征融合，即

$$\boldsymbol{f}_{\text{res}} = R(\boldsymbol{f}^{d}) \tag{15}$$

$$\boldsymbol{f}^{\text{out}} = \boldsymbol{f}_{\text{res}} + \boldsymbol{f}^{d} \tag{16}$$

式中：$R(\cdot)$表示残差模块；$\boldsymbol{f}_{\text{res}}$ 表示残差输出；$\boldsymbol{f}^{\text{out}}$ 表示最终融合特征。

整体上，该适配器通过引入 GLU 和残差机制，有效地克服了传统加权融合方法难以捕捉复杂非线性关系的问题，抑制了冗余信息并在多视角加权融合过程中实现了对齐不同尺度特征的目标，从而使得模型在下游任务(如图像与点云融合、多模态分类等)中展现出更高的鲁棒性和准确性。

2.4 基于自注意力机制的多模态全局特征的动态融合

现有的 3D FSCIL 框架[18]采用了 2 种 CLIP 模

型对多视角特征进行提取：①原始的 CLIP 模型，能保留更一般化的视觉表示，捕获低级别和全局信息。②经过预训练[14]后的 CLIP 模型，更侧重于与下游任务相关的细粒度信息。然而，对于 2 种经过不同预训练策略得到的图像点云联合全局特征 $\boldsymbol{f}_{\text{ori}}^{\text{out}}$ 和 $\boldsymbol{f}_{\text{pre}}^{\text{out}}$，现有的 3D FSCIL 框架[18]直接以相同的权重融合，即

$$\boldsymbol{f}_{\text{fused}}=\left(\boldsymbol{f}_{\text{ori}}^{\text{out}}+\boldsymbol{f}_{\text{pre}}^{\text{out}}\right)/2 \tag{17}$$

这种固定的融合规则，难以根据不同样本的情况动态调整各部分特征的重要性。由此导致在 2 种特征之间的互补信息无法被充分挖掘和利用，其融合过程缺乏灵活性和自适应性。

针对此问题，本文提出了基于自注意力机制的融合模块，对 2 路特征进行动态加权，自动捕捉其间的相互依赖关系，从而更有效地整合各自的优势。

1) 对输入的 2 路特征进行逐元素相加，形成初步的融合输入，即

$$\boldsymbol{f}_{\text{fusion}}=\boldsymbol{f}_{\text{ori}}^{\text{out}}+\boldsymbol{f}_{\text{pre}}^{\text{out}} \tag{18}$$

2) 利用多头自注意力(Multihead self-attention)层对融合输入进行处理，使得模型能够从不同注意力头中学习到多角度的信息交互。具体来说，融合输入首先被分别投影为查询矩阵 $\boldsymbol{Q}=\boldsymbol{f}_{\text{fusion}}\boldsymbol{W}_Q$，键矩阵 $\boldsymbol{K}=\boldsymbol{f}_{\text{fusion}}\boldsymbol{W}_K$，值矩阵 $\boldsymbol{V}=\boldsymbol{f}_{\text{fusion}}\boldsymbol{W}_V$。其中 $\boldsymbol{W}_Q$，$\boldsymbol{W}_K$ 和 $\boldsymbol{W}_V$ 分别为可学习的线性变换矩阵。对于每个注意力头(本文设置头数量为 2)，注意力输出为

$$head_h=\text{Softmax}\left(\boldsymbol{Q}_h\boldsymbol{K}_h^{\text{T}}/\sqrt{d_h}\right)\boldsymbol{V}_h \tag{19}$$

式中：d_h 表示每个头的特征维度。

3) 将所有头的注意力输出拼接后再进行线性变换，得到最终注意力输出，即最终动态融合结果为

$$\boldsymbol{f}_{\text{fused}}=Concat\left(head_1,head_2\right)\boldsymbol{W}_O \tag{20}$$

式中：$\boldsymbol{W}_O$ 表示输出映射矩阵。

这样，每个特征的贡献不再是预先固定的，而是由数据自适应所决定，且有助于消除传统平均或固定权重融合方法在处理模态差异时的不足，从而提高最终融合特征的表达效果。

2.5　损失函数

对于第 b 个训练任务，使用交叉熵计算分类损失，即

$$L_{\text{cls}}^b=\frac{1}{\left|D^b\right|}\sum_{i=1}^{\left|D^b\right|}L_{\text{ce}}\left(\boldsymbol{p}_i^b,y_i^b\right) \tag{21}$$

式中：L_{ce} 表示交叉熵损失函数；$\boldsymbol{p}_i^b$ 表示第 b 个任务中第 i 个样本的预测概率；$\left|D^b\right|$ 表示第 b 个任务中所有样本的数量。

在增量学习阶段，由于新类别可供训练的样本数量少，类别特征混淆的情况严重。为此，本文采用了对比学习的方法对提取的特征进行优化，使其更加贴近正确类别的原型。此外，为了防止小样本场景下模型发生过拟合以及提升对比学习效果，本文还使用了数据增强的方法来扩充样本。具体地，在增量阶段中对所有原始样本进行了数据增强，具体增强方式为将点云沿坐标轴随机旋转以及渲染时对相机视距进行随机变化，每个原始样本产生 7 个增强样本。

对于第 b 个任务中的第 i 个样本，本文将其对应类别的原型特征 $\boldsymbol{F}_{y_i^b}^t$ 作为正样本，将所有其他已见类别的原型特征 $\boldsymbol{F}_{\text{res}_i^b}^t$ 作为负样本。

本文使用 InfoNCE 损失函数[43]作为对比学习的损失函数。第 b 个任务的对比学习损失为

$$L_{\text{cont}}^b=\frac{1}{\left|D^b\right|}\sum_{i=1}^{\left|D^b\right|}\sum_{j=1}^{N_{\text{Aug}}}L_{\text{InfoNCE}}\left(\boldsymbol{f}_{i,j}^g,\boldsymbol{F}_{y_i^b}^t,\boldsymbol{F}_{\text{res}_i^b}^t\right) \tag{22}$$

式中：N_{Aug} 表示数据增强样本的数量；$\boldsymbol{f}_{i,j}^g$ 表示全局特征。

对于第 b 个训练任务，最终的损失函数为

$$L^b=L_{\text{cls}}^b+\alpha L_{\text{cont}}^b \tag{23}$$

式中：b=1 时，α=0；$b\geqslant 2$ 时，α=1。

3　实验及结果分析

3.1　数据集与评价指标

本文在3种不同的3D物体数据集上进行了充分的实验评估，包括 2 个合成数据集(ModelNet40[44]和ShapeNet55[1])和一个真实扫描数据集(CO3D[3])。ModelNet40 数据集是点云分析中最广泛使用的基准之一，其包含来自 40 个类别(如飞机、汽车、植物、灯具等)的12 311个CAD生成的三维网格模型，其中 9 843 个(约占 80%)用于训练，2 468 个(约占 20%)用于测试。ShapeNet55 是完整 ShapeNet 数据集的一个子集，包含单一、干净的 3D 模型，并配有经过人工验证的类别与姿态(对齐)标注。该子集涵盖了 55 个常见的物体类别，共有 52 470 个独特的 3D 模型，其中 41 592 个(约占 80%)用于训练，10 518 个(约占 20%)用于测试。CO3D 是由 Meta AI 于 2021 年发布的一个大规模真实世界 3D 数据集，其中包含来自近 19 000 个视频序列的 150 万帧图像，涵盖了 50 个常见物体类别(源自

MS-COCO 数据集[45])，并提供相机位姿和三维点云的真实标注。

为了充分评估本方法的有效性，对于 3 种数据集，本文参照 FILP-3D[18]中的任务设定，构建了 2 个跨领域的 FSCIL 任务，即模型在基础阶段和增量阶段使用不同的数据集进行训练。具体而言，2 个跨领域的 FSCIL 任务分别为：在基础阶段使用 ShapeNet55 数据集进行训练，增量阶段则使用 ModelNet40 数据集训练以及在基础阶段使用 ShapeNet55 数据集进行训练，增量阶段则使用 CO3D 数据集训练。对于 ShapeNet55 到 ModelNet40 任务，在基础训练阶段中，保留了 ShapeNet55 中的全部 55 个类别，而在增量任务中，首先排除掉 ModelNet40 中与基础任务中重复的 16 个类别，剩下的 24 个独有类别被平均划分到 6 个增量阶段中，每个类随机选择 5 个样本进行训练。对于 ShapeNe55 到 CO3D，同样保留了 ShapeNet55 中的全部类作为基础任务，而在增量任务中，则排除了 CO3D 中与 ShapeNet55 重复的 9 个类，剩下 41 类分配到 11 个增量阶段中，对于每个新类仍随机选择 5 个样本进行训练(前 10 个增量阶段中，每个阶段包含 4 个新类，最后一个阶段只包含 1 个新类)。

实验中所用的定量指标为准确率 Acc 以及相对准确度下降率Δ和最大阶段跳变率Δ′。在每个阶段的训练完成之后，本文在基类和所有的新类上进行测试，计算准确率 Acc。此外，在所有增量阶段完成后，可根据每个阶段的准确率计算出相对准确度下降率 $\Delta = \left| Acc_T - Acc_0 \right| / Acc_0$，以及最大阶段跳变率 $\Delta' = \mathrm{MAX}\left(\left| Acc_t - Acc_{t-1} \right| / Acc_t \right)$。其中 Acc_T 和 Acc_0 表示最终阶段和基础阶段的准确率，Acc_t 表示第 t 阶段的准确率。相对准确率下降率作为衡量方法性能的综合指标，较低的相对准确率下降率意味着更优的模型表现。最大阶段跳变率则用以衡量模型在增量学习过程中阶段间性能变化的最剧烈程度，较低的最大阶段跳变率意味着模型有着更强的抗遗忘能力。

3.2 实验细节

本实验在配备 NVIDIA RTX 4090D 24 GB 显存的单张 GPU 上运行。本文选择 32 图像块的基础 Vision Transformer (ViT-B/32)[41]模型作为原始 CLIP 的视觉编码器。预训练后的 CLIP 模型则使用 CLIP2Point[14]中提出的在 ShapeNet 上经过预训练的深度图编码器。对于 2 种不同的视觉编码器得到的特征输入，其自适应适配器权重单独训练。本文所使用的 ShapeNet55，ModelNet40 以及 CO3D 数据集均为经过后处理的点云数据集，其对原始数据集中的 3D 模型数据采用了最远点采样的方式在物体模型表面提取出 1 024 个点作为点云数据。所采样的点为 *XYZ* 坐标，不含颜色和法向量信息。对于得到的点云数据，本文还使用了点云中心化以及单位球归一化对其进行预处理。本文采用 CLIP2Point 中提出的渲染方法将点云渲染成深度图：采用固定位姿的视角进行渲染，视角数可选为 6 或 10。当视角数为 6 时，方位角分别为：0, 90, 180, 270, 0, 180；仰角分别为：0, 0, 0, 0, 90, –90；所有视角的距离均设为 1。当视角数为 10 时，方位角分别为：0, 0, 0, 0, –45, 45, –45, 45, –90, 90；仰角分别为：0, 90, 180, 270, 225, 225, 315, 315, 0, 0；所有视角距离均设为 1。在本实验中渲染视角数选为 10。使用 OcCo[43]在 ShapeNet 上预训练的 DGCNN[26]模型作为 3D 编码器。本文中图像特征以及文本特征的维度均设置为 512。InfoNCE 损失的温度系数设置为0.1。在训练过程中，将使用 Adam 优化器[46]。且将学习率设置为 1×10^{-3}，权重衰减设置为 1×10^{-4}。对于基础任务和增量任务，训练轮数分别设置为 10 和 40。在增量任务中，对于每个新类，分别随机选择 5 个样本进行训练，并从先前的阶段中随机选取一个样本作为历史样本以缓解灾难性遗忘问题。所有训练的批次大小设置为 32。

3.3 对比实验

为了验证本方法的有效性，本文选取了以下几种模型进行了比较：语言-图像-点云统一表征模型(Unified Representation of Language, Images, and Point Clouds，ULIP[47])，FACT[37]，Microshape[17]和 FILP-3D[18]。ULIP 是目前达到 SOTA 的 3D 小样本学习方法，其提出了一种统一语言、图像和点云模态的多模态预训练框架，用于提升 3D 理解能力。该方法借助 CLIP 等大型图文 PTMs 的通用语义空间，通过构造图像-文本-点云三元组，将 3D 特征对齐到已有的图文特征空间。ULIP 冻结图像和文本编码器，仅训练 3D 编码器以实现跨模态特征对齐。FACT 是一种针对二维图像分类的 FSCIL 方法，其在基础训练阶段就为未来可能加入的新类预留嵌入空间。具体而言，FACT 在特征空间中引入虚拟原型(Virtual prototypes)，将已知类的特征压缩到更紧凑的区域，从而为未来的新类预留空间，并通过特征混合(Manifold mixup)模拟新类样本，提升模型面向未来类别的适应能力。此外，虚拟原型还

在推理阶段作为嵌入空间的基底，加强分类器的判别能力。为了将 FACT 应用到点云分类任务中，本文沿用 FILP-3D 的方法，将 FACT 中使用的 CNN 替换为 CLIP2Point 的深度图编码器。Microshape 是面向点云分类的 FSCIL 方法，该方法将 3D 对象表示为一组预定义的微形状表达(正交基向量)，从而实现对新类别的增量学习。也可利用少量的样本在保持原有知识的同时学习新类，有效缓解了训练基类和增量新类之间的域差异带来的性能下降问题。FILP-3D 是目前 SOTA 的 3D FSCIL 方法，其充分利用 V-L 模型 CLIP 所蕴含的形状先验知识，并引入空间噪声补偿器模块用于增强 3D 特征的稳定性，冗余特征消除器模块用于在投影后的高维空间中对齐和压缩特征表示。表 1 为 4 种方法及本方法在 ShapeNet 到 ModelNet 任务上的实验结果。

表 1　在 ShapeNet 到 ModelNet 数据集上的实验结果分析

Table 1　Analysis of experimental results from ShapeNet to ModelNet dataset

方法	Acc↑							Δ↓	Δ′↓
	1	2	3	4	5	6	7		
ULIP	86.3	83.3	80.3	75.8	72.8	62.9	65.1	24.6	9.9
FACT	82.6	77.0	72.4	69.8	68.4	67.7	67.3	18.5	5.6
Microshape	86.9	84.6	82.8	78.3	78.5	71.5	68.6	21.1	7.0
FILP-3D	90.5	87.1	84.5	81.8	80.9	80.2	77.6	14.3	3.4
本文方法	**90.7**	**88.8**	**87.0**	**84.8**	**84.7**	**83.6**	**82.6**	**8.9**	**2.2**

注：↑表示数值越大越好；↓表示数值越小越好；ULIP，FACT，Microshape 实验结果来自文献[18]；FILP-3D 表示使用官方代码的本地复现结果；第 1 阶段表示基础阶段，其余表示增量阶段；Acc 表示模型在当前阶段的准确率；Δ为相对准确度下降率，表示最终阶段相对于基础阶段的准确率下降比例；Δ'为最大阶段跳变率，表示模型在所有增量阶段中的最大准确率跳变幅度；加粗数据表示最优值。

从表 1 可以看出，本方法在基础阶段以及各个增量阶段的准确率最高，并且相对准确度下降率也最低。具体地，相较于面向零样本或小样本设计的 ULIP，本方法在增量阶段的准确率平均上升了 17.0%，相对准确度下降率降低了 15.7%，最大阶段跳变率下降了 7.7%。说明本方法对于增量学习中产生的灾难性遗忘问题有了较好地解决。相较于 FACT，本方法在每个阶段的准确率均有高于 10.0% 的提升，相对准确度下降率降低了 9.6%，最大阶段跳变率下降了 3.4%。FACT 准确度较低的原因可能在于其使用的虚拟原型在 3D 点云中无法有效模拟类间结构，导致模型难以预留合理的特征空间用于增量学习。此外，FACT 原本的推理机制是基于图像特征的分布假设，与点云特征的分布差异较大。而本方法由于是面向点云分类任务设计的，并且未使用虚拟原型的方法，所以未出现上述问题。相较于 Microshape，本文在各阶段的准确率平均提升了 4.5%，相对准确度下降率降低了 12.2%，最大阶段跳变率下降了 4.8%。由于 Microshape 是专门面向点云分类任务的 FSCIL 方法，其性能相对 FACT 有所提升。然而，在后期的增量阶段(如阶段 5~6，阶段 6~7)，Microshape 的准确率明显出现了较大幅度的下滑。其主要原因在于该方法早期所学习到的正交基数量有限，随着新类别增多，可用的表示能力趋于饱和，多个类别共享近似的形状组合，区分度降低。在后期增量阶段，模型经过多轮参数微调，之前学习到的表征空间会逐步受到干扰，微形状表达被扭曲，使得整体性能下降。随着新类别逐步引入，问题进一步叠加，模型泛化能力减弱，从而导致后期准确率显著下滑。而本方法所使用的 CLIP 模型在大量的图像-文本对上进行了预训练，其蕴含丰富的先验知识，可直接用来提取深度图特征，且不需要另外学习类似微形状的特征表达，所以受到新增类别的干扰较小。本方法与目前 SOTA 的 3D FSCIL 方法 FILP-3D 相比，增量阶段的准确率平均仍提升了 3.6%，且相对准确度下降率降低了 5.4%，最大阶段跳变率下降了 1.2%。此外，在增量阶段中，FILP-3D 出现了几次较大幅度的性能下降(如从阶段 1~2 准确度下降了 3.4%；从阶段 3~4 下降了 2.7%等)，而本模型性能相对稳定。这主要是由于 FILP-3D 方法在进行多尺度特征提取以及在进行特征融合时，所采用的方法仍欠缺一定的灵活性和自适应性，导致模型的鲁棒性有所欠缺。而本方法通过引入基于门控单元和残差块的自适应适配器以及基于自注意力机制的多模态全局特征的动态融合机制，可以更好地提取多模态特征并进行特征的深度融合，从而提高模型最终的特征表达效果，增强了模型的稳定性。

表 2 为 3 种方法及本方法在 ShapeNet 到 CO3D 任务上的实验结果。由于 CO3D 是真实扫描的点云数据集，其与合成数据集 ShapeNet 的域差异更大，且存在噪声影响。因此，合成数据集到真实数据集

的任务更具有挑战性，且更符合现实应用的场景，以此要求模型从有限的样本中泛化的能力更强。从表 2 可以看到，本方法在各个阶段仍有着最优的准确率以及最低的相对准确度下降率和最大阶段跳变率。而 Microshape 方法受到域差异的影响尤为显著，其准确度相对下降率达到了 30.3%，最大阶段跳变率达到了 12.1%。这可能是由于 Microshape 在干净、规则的合成数据集上学习到的微形状表达无法很好地在真实点云中进行泛化。同时，真实数据中物体外观变化大、视角多样，进一步加剧了微形状表达空间的模糊性和类别间混淆，从而导致整体性能下降更严重。相比之下，FACT 受到的影响较小，甚至在某些阶段有一定的性能提升。分析原因可能是由于 FACT 使用的虚拟原型和嵌入空间预留机制对特征分布有着更强的结构约束与抗过拟合能力，这种设计在迁移过程中提供了类间判别性的缓冲区域，使模型在面对真实数据的结构复杂性和偏移时，仍能保持甚至提升性能。受益于 CLIP 的 PTMs，FILP-3D 方法仍保持着相对较高的准确度。但在各阶段的准确度本方法仍比 FILP-3D 有所提升，且有着更低的Δ和Δ′值。这说明本方法在面对真实点云数据时，仍可依靠自适应适配器以及自注意力融合模块，更好地提取多模态特征并进行深度融合。

3.4 消融实验

为了验证本方法引入模块的有效性，本文设计了如下消融实验：在完整模型的基础上，分别或全部移除基于门控单元和残差块的自适应适配器以及基于自注意力机制的多模态特征融合模块，以评估各自作用。①消融 1：移除自适应适配器，使用 FILP-3D 中采用的原始适配器，即不在适配器中使用本文提出的门控机制和残差块；②消融 2：移除自注意力多模态特征融合模块，直接以相同权重对 2 种多模态全局特征进行融合；③消融 3：将自适应适配器和注意力融合模块全部移除。表 3 展示了在 ShapeNet 到 ModelNet 任务中的消融实验结果。

表 2 在 ShapeNet 到 CO3D 数据集上的实验结果分析

Table 2 Analysis of experimental results from ShapeNet to CO3D dataset

方法	Acc↑												Δ↓	Δ′↓
	1	2	3	4	5	6	7	8	9	10	11	12		
ULIP	86.3	85.6	81.7	74.0	71.7	68.1	67.6	64.5	59.5	58.4	55.2	57.5	28.8	7.7
FACT	82.4	77.2	74.5	73.1	71.3	70.4	67.2	65.2	63.8	61.8	59.9	59.8	27.4	5.2
Microshape	85.2	78.6	71.0	72.0	75.2	68.8	56.1	58.5	62.9	59.1	52.2	59.4	30.3	12.1
FILP-3D	89.9	84.9	84.9	83.2	81.8	80.6	78.6	77.1	76.1	74.8	73.5	72.2	19.7	5.0
本文方法	**90.5**	**88.0**	**86.7**	**84.4**	**83.9**	**82.1**	**79.4**	**78.5**	**77.7**	**76.6**	**74.8**	**74.0**	**18.2**	**2.7**

注：↑表示数值越大越好；↓表示数值越小越好；ULIP，FACT，Microshape 实验结果来自文献[18]；FILP-3D 表示使用官方代码的本地复现结果；第 1 阶段表示基础阶段，其余表示增量阶段；Acc 表示模型在当前阶段的准确率；Δ为相对准确度下降率，表示最终阶段相对于基础阶段的准确率下降比例；Δ'为最大阶段跳变率，表示模型在所有增量阶段中的最大准确率跳变幅度；加粗数据表示最优值。

表 3 模块有效性消融实验结果

Table 3 Module effectiveness ablation results

方法	Acc↑							Δ↓	Δ′↓
	1	2	3	4	5	6	7		
无自适应适配器	90.5	87.9	85.5	83.0	82.7	81.9	80.8	10.7	2.6
无注意力融合	**90.8**	87.8	85.7	83.0	82.6	77.8	77.1	15.0	4.8
均无	90.5	87.1	84.5	81.8	80.9	80.2	77.6	14.3	3.4
本文方法	90.7	**88.8**	**87.0**	**84.8**	**84.7**	**83.6**	**82.6**	**8.9**	**2.2**

注：↑表示数值越大越好；↓表示数值越小越好；第 1 阶段表示基础阶段，其余表示增量阶段；Acc 表示模型在当前阶段的准确率；Δ为相对准确度下降率，表示最终阶段相对于基础阶段的准确率下降比例；Δ'为最大阶段跳变率，表示模型在所有增量阶段中的最大准确率跳变幅度；加粗数据表示最优值。

由表 3 可知，移除任何一个模块都会导致模型性能退化。其中，当移除自适应适配器时，模型在每个增量阶段上的准确度均有下降，平均下降了 1.6%，Δ上升了 1.8%，Δ'上升了 0.4%。这表明了没有了门控单元和残差块来动态调整并对齐多尺度特征后，模型更易受到点云中多余的几何信息以及相似的局部结构等冗余信息影响，且难以捕获复杂依赖关系，从而导致模型性能下降。相比之下，移除注意力特征融合模块对模型的影响更大，模型在各个增量阶段准确率均有所下降，特别是在后几个阶段上性能下滑严重，准确率平均下降了 2.9%，Δ上升了 6.1%，Δ'上升了 2.6%。这说明了采用固定的融合规则，难以根据不同样本的情况动态调整各部分特征的重要性，特征之间的互补信息无法被充

分挖掘和利用，从而导致模型的性能下降较大。而将自适应适配器和注意力融合模块均移除后，模型在前几个增量阶段的准确率进一步降低。值得注意的是，此时模型相对无注意力融合模块，在第 6 和 7 阶段的准确率却有略微提升，并因此具有更低的 Δ和Δ'值。这可能是由于自适应适配器在缺乏注意力融合模块支持时，在后期阶段有较多新类的情况下，出现了放大局部噪声的情况，难以维持其提取的全局特征的判别性。由此可见，本文所提出的各模块在不同方面均对模型性能提升有所贡献，体现了其设计的必要性和有效性。

此外，为了验证本文在自适应模块中提出的门控单元和残差块的有效性，还对其单独进行了消融实验：在完整模型的基础上，分别移除自适应适配器中的 GLU 门控机制，保留残差块以及保留自适应适配器中的 GLU 门控机制，移除残差块。表 4 展示了相关消融实验结果。

从表 4 可知，无论是移除 GLU 门控还是移除残差块，都会导致模型在各增量阶段的准确率下降，相对准确度下降率和最大阶段跳变率提高。这说明了自适应适配器中 GLU 门控和残差块的有效性。

为了验证本文在损失函数部分提出的在增量阶段使用对比学习优化提取特征的方法的有效性，对其单独进行了消融实验：在模型和其他训练条件一致的情况下，不考虑对比学习损失。表 5 展示了相关消融实验结果。

从表 5 可以看到，当没有对比学习损失时，模型在各个阶段的准确率均出现了下降，相对准确度下降率和最大阶段跳变率均有所提高，说明了在增量阶段采用对比学习损失的有效性。

为了验证本文采用的记忆回放在缓解灾难性遗忘问题上的有效性，本文单独对其进行了消融实验：在其他条件相同的情况下，不再从先前的阶段中随机选取一个样本作为历史样本进行训练。表 6 展示了消融实验结果。

表 4　GLU 门控机制与残差块有效性消融实验结果

Table 4　Effectiveness ablation results of GLU gating mechanism and residual blocks

方法	Acc↑							Δ↓	Δ′↓
	1	2	3	4	5	6	7		
无 GLU 门控机制	90.2	81.2	84.5	83.3	82.0	79.2	77.9	12.3	3.3
无残差块	90.6	**89.0**	86.0	83.9	83.7	82.9	81.4	9.2	3.0
本文方法	**90.7**	88.8	**87.0**	**84.8**	**84.7**	**83.6**	**82.6**	**8.9**	**2.2**

注：↑表示数值越大越好；↓表示数值越小越好；第 1 阶段表示基础阶段，其余表示增量阶段；Acc 表示模型在当前阶段的准确率；Δ为相对准确度下降率，表示最终阶段相对于基础阶段的准确率下降比例；Δ'为最大阶段跳变率，表示模型在所有增量阶段中的最大准确率跳变幅度；加粗数据表示最优值。

表 5　对比学习有效性消融实验结果

Table 5　Effectiveness ablation results of contrastive learning

方法	Acc↑							Δ↓	Δ′↓
	1	2	3	4	5	6	7		
无对比学习	90.5	86.2	83.1	78.6	76.7	76.2	75.8	14.7	4.5
本文方法	**90.7**	**88.8**	**87.0**	**84.8**	**84.7**	**83.6**	**82.6**	**8.9**	**2.2**

注：↑表示数值越大越好；↓表示数值越小越好；第 1 阶段表示基础阶段，其余表示增量阶段；Acc 表示模型在当前阶段的准确率；Δ为相对准确度下降率，表示最终阶段相对于基础阶段的准确率下降比例；Δ'为最大阶段跳变率，表示模型在所有增量阶段中的最大准确率跳变幅度；加粗数据表示最优值。

表 6　记忆回放有效性消融实验结果

Table 6　Effectiveness ablation results of memory replay

方法	Acc↑							Δ↓	Δ′↓
	1	2	3	4	5	6	7		
无记忆回放	90.7	84.7	79.2	79.2	77.9	76.9	69.5	21.2	7.4
本文方法	**90.7**	**88.8**	**87.0**	**84.8**	**84.7**	**83.6**	**82.6**	**8.9**	**2.2**

注：↑表示数值越大越好；↓表示数值越小越好；第 1 阶段表示基础阶段，其余表示增量阶段；Acc 表示模型在当前阶段的准确率；Δ为相对准确度下降率，表示最终阶段相对于基础阶段的准确率下降比例；Δ'为最大阶段跳变率，表示模型在所有增量阶段中的最大准确率跳变幅度；加粗数据表示最优值。

从表 6 中可以看到，在不采用历史样本记忆回放的情况下，模型在各个阶段的准确率均出现了下降，相对准确度下降率和最大阶段跳变率均提高，也即表现出了更严重的遗忘问题。

4 结束语

本文提出了一种基于深度融合多模态特征的 3D FSCIL 方法，通过引入基于门控单元和残差块的自适应适配器以及基于自注意力机制的多模态全局特征动态融合模块，实现了多模态特征的深度融合与动态权重调整。此外，结合对比学习损失、多视角与几何扰动数据增强策略以及记忆回放机制，有效缓解了小样本条件下的过拟合与遗忘问题。在 ShapeNet，ModelNet 和 CO3D 数据集上的实验表明，该方法在各增量阶段均取得显著的性能提升，相对准确度下降率与最大阶段跳变率均明显低于现有主流方法，验证了其在提升小样本条件下模型泛化能力与鲁棒性方面的有效性。当前方法的局限性在于数据域适应性有限：实验主要集中在常见的合成数据集和部分真实扫描数据上，对于工业检测、医学影像等高噪声且分布复杂的真实场景，模型的泛化性能仍需进一步验证。未来将着重考虑扩展数据集规模与多样性以增强跨域适应性，及研究可学习视角或基于神经渲染的动态投影方法，以更充分保留点云的几何细节与结构信息，增强特征表达能力。

参考文献 (References)

[1] CHANG A X, FUNKHOUSER T, GUIBAS L, et al. ShapeNet: an information-rich 3D model repository[EB/OL]. [2025-04-30]. https://arxiv.org/abs/1512.03012.pdf.

[2] DEITKE M, SCHWENK D, SALVADOR J, et al. Objaverse: a universe of annotated 3D objects[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2023: 13142-13153.

[3] REIZENSTEIN J, SHAPOVALOV R, HENZLER P, et al. Common objects in 3D: large-scale learning and evaluation of real-life 3D category reconstruction[C]//2021 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2021: 10881-10891.

[4] UY M A, PHAM Q H, HUA B S, et al. Revisiting point cloud classification: a new benchmark dataset and classification model on real-world data[C]//2019 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2019: 1588-1597.

[5] WU T, ZHANG J R, FU X, et al. OmniObject3D: large-vocabulary 3D object dataset for realistic perception, reconstruction and generation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2023: 803-814.

[6] QI C R, SU H, MO K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 77-85.

[7] QI C R, YI L, SU H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[C]//The 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 5105-5114.

[8] TAO X Y, HONG X P, CHANG X Y, et al. Few-shot class-incremental learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 12180-12189.

[9] QIN C W, JOTY S. Continual few-shot relation learning via embedding space regularization and data augmentation[C]// The 60th Annual Meeting of the Association for Computational Linguistics. New York: Association for Computational Linguistics, 2022: 2776-2789.

[10] ZHOU D W, CAI Z W, YE H J, et al. Revisiting class-incremental learning with pre-trained models: generalizability and adaptivity are all you need[J]. International Journal of Computer Vision, 2025, 133(3): 1012-1032.

[11] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// 2019 Conference of the North American Chapter of the Association for Computational Linguistics. New York: Association for Computational Linguistics, 2019: 4171-4186.

[12] HE K M, CHEN X L, XIE S N, et al. Masked autoencoders are scalable vision learners[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2022: 15979-15988.

[13] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. [2025-04-30]. http://proceedings.mlr.press/v139/radford21a.html.

[14] HUANG T Y, DONG B W, YANG Y H, et al. CLIP2Point: transfer CLIP to point cloud classification with image-depth pre-training[C]//2023 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2023: 22100-22110.

[15] ZENG Y H, JIANG C H, MAO J G, et al. CLIP^2: contrastive language-image-point pretraining from real-world point cloud data[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2023: 15244-15253.

[16] ZHANG R R, GUO Z Y, ZHANG W, et al. PointCLIP: point cloud understanding by clip[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2022: 8542-8552.

[17] CHOWDHURY T, CHERAGHIAN A, RAMASINGHE S, et al. Few-shot class-incremental learning for 3D point cloud objects[C]//The 17th European Conference on Computer Vision. Cham: Springer, 2022: 204-220.

[18] XU W, HUANG T Y, QU T Y, et al. FILP-3D: enhancing 3D few-shot class-incremental learning with pre-trained vision-language models[J]. Pattern Recognition, 2025, 165: 111558.

[19] LI Y Y, BU R, SUN M C, et al. PointCNN: convolution on *X*-transformed points[C]//The 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2018: 828-838.

[20] LIU Y C, FAN B, XIANG S M, et al. Relation-shape convolutional neural network for point cloud analysis[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2019: 8887-8896.

[21] POULENARD A, RAKOTOSAONA M J, PONTY Y, et al. Effective rotation-invariant point CNN with spherical harmonics kernels[C]//2019 International Conference on 3D Vision. New York: IEEE Press, 2019: 47-56.

[22] RAO Y M, LU J W, ZHOU J. Spherical fractal convolutional neural networks for point cloud recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2019: 452-460.

[23] WU W X, QI Z G, LI F X. PointConv: deep convolutional networks on 3D point clouds[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE

Press, 2019: 9613-9622.
[24] XU Y F, FAN T Q, XU M Y, et al. SpiderCNN: deep learning on point sets with parameterized convolutional filters[C]//The 15th European Conference on Computer Vision. Cham: Springer, 2018: 90-105.
[25] LI G C, MÜLLER M, THABET A, et al. DeepGCNs: can GCNs go as deep as CNNs?[C]//2019 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2019: 9266-9275.
[26] WANG Y, SUN Y B, LIU Z W, et al. Dynamic graph CNN for learning on point clouds[J]. ACM Transactions on Graphics , 2019, 38(5): 146.
[27] YU X M, TANG L L, RAO Y M, et al. Point-BERT: pre-training 3D point cloud transformers with masked point modeling[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2022: 19291-19300.
[28] PANG Y T, WANG W X, TAY F E H, et al. Masked autoencoders for point cloud self-supervised learning[C]//The 17th European Conference on Computer Vision. Cham: Springer, 2022: 604-621.
[29] ZHANG R R, GUO Z Y, FANG R Y, et al. Point-M2AE: multi-scale masked autoencoders for hierarchical point cloud pre-training[C]//The 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2022: 1962.
[30] ZHAO H S, JIANG L, JIA J Y, et al. Point transformer[C]// 2021 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2021: 16239-16248.
[31] CHEN K L, LEE C G. Incremental few-shot learning via vector quantization in deep embedded space[EB/OL]. [2025-04-30]. https://dblp.org/db/conf/iclr/iclr2021.html#ChenL21.
[32] CHERAGHIAN A, RAHMAN S, FANG P F, et al. Semantic-aware knowledge distillation for few-shot class-incremental learning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2021: 2534-2543.
[33] MAZUMDER P, SINGH P, RAI P. Few-shot lifelong learning[C]//The 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 2337-2345.
[34] PENG C, ZHAO K, WANG T R, et al. Few-shot class-incremental learning from an open-set perspective[C]//The 17th European Conference on Computer Vision. Cham: Springer, 2022: 382-397.
[35] XIANG X, TAN Y W, WAN Q, et al. Coarse-to-fine incremental few-shot learning[C]//The 17th European Conference on Computer Vision. Cham: Springer, 2022: 205-222.
[36] ZHANG C, SONG N, LIN G S, et al. Few-shot incremental learning with continually evolved classifiers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2021: 12450-12459.
[37] ZHOU D W, WANG F Y, YE H J, et al. Forward compatible few-shot class-incremental learning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2022: 9036-9046.
[38] HERSCHE M, KARUNARATNE G, CHERUBINI G, et al. Constrained few-shot class-incremental learning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2022: 9047-9057.
[39] LIU H, GU L, CHI Z X, et al. Few-shot class-incremental learning via entropy-regularized data-free replay[C]//The 17th European Conference on Computer Vision. Cham: Springer, 2022: 146-162.
[40] WANG R Q, DUAN X Y, KANG G L, et al. AttriCLIP: a non-incremental learner for incremental knowledge learning[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2023: 3654-3663.
[41] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. [2025-04-22]. https://arxiv.org/abs/2010.11929.pdf.
[42] WANG H C, LIU Q, YUE X Y, et al. Unsupervised point cloud pre-training via occlusion completion[C]//2021 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2021: 9762-9772.
[43] VAN DEN OORD A, LI Y Z, VINYALS O. Representation learning with contrastive predictive coding[EB/OL]. [2025-03-10]. https://arxiv.org/abs/1807.03748.pdf.
[44] WU Z R, SONG S R, KHOSLA A, et al. 3D ShapeNets: a deep representation for volumetric shapes[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2015: 1912-1920.
[45] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//The 13th European Conference on Computer Vision. Cham: Springer, 2014: 740-755.
[46] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[EB/OL]. [2025-02-14]. https://arxiv.org/abs/1711.05101.pdf.
[47] XUE L, GAO M F, XING C, et al. ULIP: learning a unified representation of language, images, and point clouds for 3D understanding[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2023: 1179-1189.