软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software [doi: 10.13328/j.cnki.jos.007187]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于双向拟合掩码重建的多模态自监督点云表示学习*

程浩喆, 祝继华, 史鹏程, 胡乃文, 谢奕凡, 李仕奇

(西安交通大学 软件学院, 陕西 西安 710049)

通信作者: 祝继华, E-mail: zhujh@xjtu.edu.cn

摘　要: 点云自监督表示学习以无标签预训练的方式, 探索三维拓扑几何空间结构关系并捕获特征表示, 可应用至点云分类、分割以及物体探测等下游任务. 为提升预训练模型的泛化性和鲁棒性, 提出基于双向拟合掩码重建的多模态自监督点云表示学习方法, 主要由 3 部分构成: (1) 逆密度尺度指导下的“坏教师”模型通过基于逆密度噪声表示和全局特征表示的双向拟合策略, 加速掩码区域逼近真值. (2) 基于 StyleGAN 的辅助点云生成模型以局部几何信息为基础, 生成风格化点云并与掩码重建结果在阈值约束下融合, 旨在抵抗重建过程噪声对表示学习的不良影响. (3) 多模态教师模型以增强三维特征空间多样性及防止模态信息崩溃为目标, 依靠三重特征对比损失函数, 充分汲取点云-图像-文本样本空间中所蕴含的潜层信息. 所提出的方法在 ModelNet、ScanObjectNN 和 ShapeNet 这 3 种点云数据集上进行微调任务测试. 实验结果表明, 预训练模型在点云分类、线性支持向量机分类、小样本分类、零样本分类以及部件分割等点云识别任务上的效果达到领先水平.

关键词: 三维点云; 自监督表示学习; 多模态特征; 密度尺度; 生成对抗网络

中图法分类号: TP18

中文引用格式: 程浩喆, 祝继华, 史鹏程, 胡乃文, 谢奕凡, 李仕奇. 基于双向拟合掩码重建的多模态自监督点云表示学习. 软件学报. http://www.jos.org.cn/1000-9825/7187.htm

英文引用格式: Cheng HZ, Zhu JH, Shi PC, Hu NW, Xie YF, Li SQ. Multi-modal Self-supervised Point Cloud Representation Learning Based on Bidirectional Fit Mask Reconstruction. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7187.htm

Multi-modal Self-supervised Point Cloud Representation Learning Based on Bidirectional Fit Mask Reconstruction

CHENG Hao-Zhe, ZHU Ji-Hua, SHI Peng-Cheng, HU Nai-Wen, XIE Yi-Fan, LI Shi-Qi

(School of Software Engineering, Xi’an Jiaotong University, Xi’an 710049, China)

Abstract: Point cloud self-supervised representation learning is conducted in an unlabeled pre-training manner, exploring the structural relationships of 3D topological geometric spaces and capturing feature representations. This approach can be applied to downstream tasks, such as point cloud classification, segmentation, and object detection. To enhance the generalization and robustness of the pretrained models, this study proposes a multi-modal self-supervised method for learning point cloud representations. The method is based on bidirectional fit mask reconstruction and comprises three main components: (1) The “bad teacher” model, guided by the inverse density scale, employs a bidirectional fit strategy that utilizes inverse density noise representation and global feature representation to expedite the convergence of the mask region towards the true value. (2) The StyleGAN-based auxiliary point cloud generation model, grounded in local geometric information, generates stylized point clouds and fuses them with mask reconstruction results while adhering to threshold constraints. The objective is to mitigate the adverse effects of noise on representation learning during the reconstruction process. (3) The multi-modal teacher model aims to enhance the diversity of the 3D feature space and prevent the collapse of modal information. It relies on the triple feature contrast loss function to fully extract the latent information contained in the point cloud-image-text sample space. The proposed method is evaluated on ModelNet, ScanObjectNN, and ShapeNet datasets for fine-tuning tasks. Experimental results demonstrate

* 基金项目: 陕西省重点研发项目 (2021GY-025, 2021GXLHZ-097)
收稿时间: 2023-11-02; 修改时间: 2024-03-15; 采用时间: 2024-03-26; jos 在线出版时间: 2024-09-11

that the pretrained model achieves state-of-the-art performance in various point cloud recognition tasks, including point cloud classification, linear support vector machine classification, few-shot classification, zero-shot classification, and part segmentation.

Key words: 3D point cloud; self-supervised representation learning; multi-modal feature; density scale; generative adversarial network (GAN)

随着激光雷达等三维数据采集设备的精度飞速提高, 推动利用点云表征真实场景的发展进程. 现如今, 点云已广泛应用于自动驾驶[1]、虚拟现实以及场景建模[2]等多种前沿智能领域. 基于深度学习的点云处理方法在分类和分割等多个任务上取得令人惊叹的效果. 其中, PointNet[3]直接将点云坐标输入进卷积网络, 通过学习全局特征和对称最大值池化函数完成点云分类和分割工作. 此后, 研究者们从采样方式、卷积层构建等方向, 针对该模型在局部信息探索力度不足等缺陷提出改进方法[4–8], 从而将点云识别精度提升到前所未有的高度.

然而, 点云数据数量庞大, 造成监督学习下网络训练和数据标注的时间成本过高, 阻碍了点云处理领域的进步. 因此, 众多方法开始利用无监督预训练手段获取点云潜层特征表示并将其迁移应用至点云分类[9]、分割[10]、物体探测[11]、以及补全[12]等下游处理工作, 其效果已达到甚至超越同条件下监督学习的精度水平. 作为无监督学习的分支, 点云自监督表示学习预训练以无标签点云为输入的模型, 以挖掘三维空间几何拓扑特征并探索点级语义结构信息. 现有方法以基于点云重建和对比学习的方法为主要发展趋势. 前者将部分点掩码后, 通过教师-学生模型预测并重建完整点云, 旨在学习几何空间特征表示[10]. 然而, 该类方法中点生成模型存在缩放能力差、拟合时间长以及生成新噪声等缺陷. 后者建立点云正负样本对并在嵌入空间中通过相似性对比约束正负样本对之间的度量距离以提高模型特征判别性. 但该方法受到样本对建立、数据过拟合等因素影响, 导致其泛化效果略逊于基于点云重建的方法[11]. 此外, 现有场景描述性信息的获取方式丰富多样, 促进基于点云、图像和文本等多模态表示学习方法的长足发展. 此类方法从多源信息中筛选有助于提高点云表示的特征, 并通过全局特征对齐和补充综合改善表示能力和算法鲁棒性. 然而, 不同模态的数据基础存在较大差异, 直接在度量空间中衡量和判别多模态特征相关性和互益性, 将导致在某模态嵌入空间信息崩溃时大幅降低点云特征表示力[13].

综合上述对现有点云自监督表示学习方法的描述和分析, 本文主要研究问题如下: (1) 现有基于点云重建的方法中模型泛化性差且掩码区域重建时间长. 此外, 完整点云坐标和掩码点云所构成的教师-学生模型中, 重建约束单一, 导致拟合结果非最优化. (2) 掩码重建可能伴随新噪声点生成, 对点云表示空间造成污染. (3) 不加以约束和区别的情况下, 点云、图像和文本模态特征直接对齐融合将导致表示学习效果下降.

针对以上问题, 本文提出基于双向拟合掩码重建的多模态自监督点云表示学习方法. 该方法以点云、图像和文本数据为输入, 联合基于点云重建和对比学习方法为骨干, 通过 3 种简单且有效的新模型提高点云表示学习效果. 首先, 受到信息论中关于信息增益最大化的启发, 本文提出逆密度尺度指导下的“坏教师”模型. 该模型通过对局部邻域真值完成逆密度尺度指导下的高斯噪声偏移, 并提取该模型的特征分布以形成“坏教师”模型. 之后, 在点云坐标和特征分布双重约束下, 被重建点的嵌入空间特征逐步背离“坏教师”模型并向由真值坐标和特征构成的“好教师”模型逼近, 旨在加速重建拟合并提升模型所提取特征的表示力. 针对重建过程生成新噪声点问题, 基于 StyleGAN 的辅助点云生成模型从局部区域特征中获取特定风格, 并生成风格化局部点云. 直接利用风格点云整体替换掩码生成结果将降低模型预训练难度, 不利于微调泛化至下游任务. 因此, 本文将掩码重建和风格生成的两种点云在余弦距离度量下替换或保留. 为了充分探索多模态信息的潜在价值并防止模态信息崩溃, 多模态教师模型将点云特征映射下的多模态标记 (token) 定义为基准样本, 并与图像和文本的预训练特征对齐. 之后, 模态内变量正则化损失函数和协方差正则化损失函数阻止嵌入空间特征归零崩溃, 并切断特征张量通道级关联性以防止信息崩溃效应扩散.

在实证实验中, 本文选择 ModelNet10[14]、ModelNet40[14]、ScanObjectNN[15]及 ShapeNet[16]数据集进行点云分类、线性支持向量机分类、小样本 (few-shot) 分类、零样本 (zero-shot) 分类以及部件分割等下游微调测试, 以验证所提出方法的有效性. 此外, 消融学习、超参数验证及鲁棒性测试的实验结果证明所提出方法包含的子模型均对提升表示学习效果存在贡献, 并且模型所选择的参数和策略均合理. 综上, 本文主要贡献总结如下.

(1) 提出逆密度尺度指导下的“坏教师”模型. 通过设立“逆向”密度噪声偏移的“坏教师”模型和“正向”高斯特征

分布加速重建结果逼近真值.

(2) 提出基于 StyleGAN 的辅助点云生成模型. 通过生成的风格化辅助点云对掩码重建结果二次分析, 减缓重建结果受生成新噪声的不良影响.

(3) 提出多模态教师模型. 该方法考虑多模态信息间的差异性, 通过设计 3 种不同约束加强跨模态特征对齐的有效性并防止模态内嵌入信息崩溃对特征空间造成污染.

(4) 综合分析现有点云自监督表示学习方法存在的挑战后, 提出基于双向拟合掩码重建的多模态自监督点云表示学习方法. 该方法从特征拟合、噪声点处理以及多模态特征捕获和应用等方面入手, 显著提高预训练模型的下游点云识别精度. 实验结果证明了本文方法的有效性.

1 相关工作

1.1 基于点云重建的点云自监督表示学习方法

基于点云重建的点云自监督表示学习方法旨在从缺失点云重建中捕获特征表示. 其中, 掩码预测为该类方法的核心. OcCo[10]从多个相机视角遮挡原始点云, 并使用编解码器补全缺失点云, 以获取初始编码权重. PointBERT[17]设计了基于 Transformer[18]的离散变分自动编码器. 该方法将点云分为若干局部区域并随机掩盖某局部区域后, 通过 Transformer 模型学习特征并恢复被掩盖区域. Point-MAE[19]成功将二维掩码自动编码器的思想应用至点云表示学习领域. 对于 Point-MAE 中存在局部和全局特征关联性未被充分探索的缺陷, Point-M2AE[20]以金字塔模型为基底, 通过多尺度掩码策略生成一致性可见区域, 并利用自注意力机制关注局部区域. MaskPoint[21]将点云表示为离散的占用值, 在掩蔽对象点和采样噪声点之间执行简单的二值分类以完成表示学习任务. 与上述方法不同, 本文所提出的逆密度尺度指导下的“坏教师”模型和基于 StyleGAN 的辅助点云生成模型以提高在掩码重建工作中重建效率、模型泛化表现及鲁棒性为研究目的, 克服点云表示学习领域中现存亟待解决的挑战.

1.2 基于对比学习的点云自监督表示学习方法

基于对比学习的点云自监督表示学习方法通常将经过旋转、缩放等数据增强的点云在样本空间中构建正负样本对. 之后, 在特征空间中训练“拉近”正样本距离并“推远”负样本, 以达到学习潜层表示的目的. PointContrast[11]将二维对比学习方法的思想引入点云场景理解. 该方法通过对比不同视角下的点云图像以捕获点级稠密特征. 由于 PointContrast 中视角映射将损失空间上下文信息, DepthContrast[22]通过对所提取的体素和点级特征进行对比约束以完成自监督任务. ContrastMPCT[23]提出基于掩码 Transformer 的自重建对比学习方法, 利用两种对比损失函数探究局部区域的关联性. ConClu[24]受到二维图像表示学习中孪生网络[25]的启发, 提出由对比和聚类组成的预训练框架, 以最大化被增强数据的全局特征间的相似性. 4DContrast[26]成功将四维点云空间的顺序信息应用至三维表示学习, 建立起空间、时空及顺序的关联性.

1.3 基于多模态的点云自监督表示学习方法

为利用图像和文本等多源信息促进学习点云表示, 研究者们提出众多跨模态预训练方法探索信息有用性. Jing 等人[27]提出将图像-点云特征学习方式扩展至跨模态和跨视角. 该方法将图像和点云输入进二维和三维卷积网络学习特征, 并完成跨模态同类判定. I2P-MAE[28]通过二维图像特征指导三维掩码自动编码器以重构掩码区域. 与随机掩码相比, 二维指导型掩码重建可更好地专注于三维几何结构. Wang 等人[29]提出利用优良的二维图像特征提取器预训练具有几何和色彩保留的三维投影图像. CrossPoint[13]利用对比学习度量点云模态内和点云-图像跨模态特征间的相关性. ACT[30]通过知识蒸馏证明二维图像或自然语言对三维表示学习存在有益性.

除上述图像辅助工作外, 一些文本辅助工作也取得较大进展. PointCLIP[31]受到视觉-文本模型 CLIP[32]的启发, 并对齐基于 CLIP 编码的三维特征和文本特征. 其中, 视间自适应器合并多个全局特征, 并将三维知识迁移至二维模型. 为更深层地挖掘三维数据和语言知识间的潜在关联性, PointCLIP V2[33]不但在视觉端通过形状映射模块改进深度图生成的真实性, 而且在文本端利用 GPT[34]模型产出有利于三维理解的文本信息作为 CLIP 的输入.

ReCon[35]是以图像和文本为辅助信息的点云自监督表示学习方法. 该方法在结构上吸取对比学习和掩码自动编码器的优势, 并利用集成表示蒸馏完成点云表示学习任务. 陈浩楠等人[36]提出基于多模态关系的三维形状识别网络. 所设计的多模态关系模块和门控模块提取并加权局部特征和全局特征, 从而提高三维形状识别的准确率和性能. 与上述方法不同, 本文关注于多模态特征对齐困难及模态信息崩溃的现有挑战, 在图像和文本特征上设置 3 种不同的特征对齐约束, 以保证多模态信息能为点云表示学习能力的提升做出贡献.

1.4 基于点云补全的生成模型

现有点云补全生成方法主要包括基于生成对抗网络、基于扩散模型以及基于变分自动编码器. SpareNet[37]提出基于风格的对抗性渲染点生成器补全缺失点云. 该方法首先通过通道注意力指导的边缘卷积探索局部和全局特征. 之后利用 StyleGAN[38]将形状特征编码并利用深度图渲染感知不同视角下的特征. 基于生成对抗网络的方法具有良好的抗噪性, 但生成结果大多非均匀. PVD[39]提出基于扩散模型的三维形状概率生成方法, 用于无条件形状生成和有条件多模式形状补全. 该方法将去噪扩散模型与 3D 形状的混合点-体素表示相结合. 通过一系列去噪将观测点云数据的扩散过程反转为高斯噪声, 并通过优化似然函数的变分下界训练模型. 基于扩散模型的方法在生成结果上相对较佳, 但其训练时间长且鲁棒性较弱. VRCNet[40]提出一种变分自编码器点云补全框架, 通过完整点云和缺失点云间的概率建模及点关系增强策略提高点云补全效果. 基于变分自编码器的方法适用于探索具有层级结构的数据, 因此在点云处理领域得以广泛应用. 本文综合利用自动编码器和生成对抗网络的优势, 即通过基于掩码策略的自动编码器捕获全局和局部点云特征, 进而生成掩码部分. 为了减少掩码生成噪声点的影响, 生成对抗网络的重建结果被有选择地与掩码结果融合, 共同提高表示学习效果.

2 基于双向拟合掩码重建的多模态自监督点云表示学习方法

基于双向拟合掩码重建的多模态自监督点云表示学习网络框架将在第 2.1 节中介绍. 其中, 本文所采用的骨干网络将被详细介绍. 在第 2.2 节中, 逆密度尺度指导下的“坏教师”模型将从逆密度尺度计算和双向拟合重建损失函数两个方面进行描述. 之后, 在第 2.3 节中, 基于 StyleGAN 的辅助点云生成模型将阐述基于局部邻域特征的辅助点云重建过程和相似性融合算法. 在第 2.4 节中, 多模态教师模型将阐明点云、图像及文本信息的特征对齐, 以及三重特征对比损失函数约束的详细过程. 最后, 在第 2.5 节中, 本文的总体损失函数将被给定.

2.1 框架总览

本文所提出的多模态自监督点云表示学习方法的核心在于以图像和文本为辅助, 依赖重建掩码区域和特征相似性对比的方式挖掘点云空间几何拓扑结构所蕴含的特征, 并将其泛化至以捕获点云环境特征为基石的下游预测性工作. 现假设存在由点云、图像和文本组成的数据集, 并定义为 $\mathcal{D}=\{\boldsymbol{P}\in\mathbb{R}^{N\times 3},\boldsymbol{I}\in\mathbb{R}^{H\times W\times 3},\boldsymbol{T}\in\mathbb{R}^{L}\}$, 其中 $\boldsymbol{P}$、$\boldsymbol{I}$ 和 $\boldsymbol{T}$ 分别表征单个形状组成的三维点云集、随机视角下映射的图像及描述点云的文本信息. 其中, N, $H\times W$ 和 L 分别代表点云数、像素数以及文本长度. 网络框架如图 1 所示, 基本包含以下 4 个主要部分.

(1) 基于 Transformer 的掩码重建模型

本文采用基于 Transformer 的掩码重建模型[35]为骨干, 其学习过程如图 2 所示. 内部参数固定的教师模型以三维点云 $\boldsymbol{P}$ 为输入, 分别通过最远点采样 (farthest point sampling, FPS)[4]和 K 最近邻 (K-nearest neighbor, KNN) 查询获取中心点 $\boldsymbol{P}_c\in\mathbb{R}^{N'\times 3}$ 及其局部邻域 $\boldsymbol{P}_n\in\mathbb{R}^{N'\times K\times 3}$, 其中 N' 和 K 分别为采样中心点和邻域点数. 之后, 多层卷积组成的编码器对局部邻域进行特征抽象, 并得到教师模型特征分布 $F_t\in\mathbb{R}^{N'\times C}$. 其中, C 为特征维度数. 教师模型将中心点 $\boldsymbol{P}_c$ 和分布 F_t 送入学生模型中, 以掩码比率 γ 进行点丢弃, 并得到保留部分 $M\in\mathbb{R}^{N_\gamma\times 3}$、丢弃部分 $\tilde{M}\in\mathbb{R}^{N_{1-\gamma}\times 3}$ 以及经过掩码指针 m_{idx} 查询的保留部分特征 $F_M\in\mathbb{R}^{N_\gamma\times C}$. 其中, N_γ 和 $N_{1-\gamma}$ 为保留点和丢弃点数. 为从保留部分和丢弃部分得到更全面的全局特征信息指导, M 和 $\tilde{M}$ 也被线性编码为全局保留位姿特征 $F_M^p\in\mathbb{R}^{N_\gamma\times C}$ 和全局丢弃特征位姿 $F_{\tilde{M}}^p\in\mathbb{R}^{N_{1-\gamma}\times C}$. 由于 Transformer 编解码器以标记 (token) 和位姿 (position) 作为输入, 并且图像和文本模态特征需要与点云域对齐. 因此, 多模态全局查询 (global query) 初始化构建 3 个模态的标记及其位姿: $F_P^t, F_P^p, F_I^t, F_I^p, F_T^t, F_T^p\in\mathbb{R}^C$. 之后, 3 种模态标记和位姿与 F_M 和 F_M^p 进行级联拼接并输入 Transformer 编码器, 以参数共享的方式

将点云模态特征映射至多模态标记和位姿. 在解码重建过程中, 首先初始化掩码标记 $\boldsymbol{F}_{\tilde{M}}^{t} \in \mathbb{R}^{N_{1-\gamma} \times C}$ 并将其与 $\boldsymbol{F}_{M} \in \mathbb{R}^{N_{\gamma} \times C}$ 拼接合并形成与原始张量形状一致的表示. 而全局保留位姿 $\boldsymbol{F}_{M}^{p}$ 和全局丢弃位姿 $\boldsymbol{F}_{\tilde{M}}^{p}$ 被同样拼接合并形成解码部分全局位姿指导. 经过 Transformer 解码器预测, 得到重建结果 $\boldsymbol{P}_{\mathrm{re}} \in \mathbb{R}^{N' \times K \times 3}$.

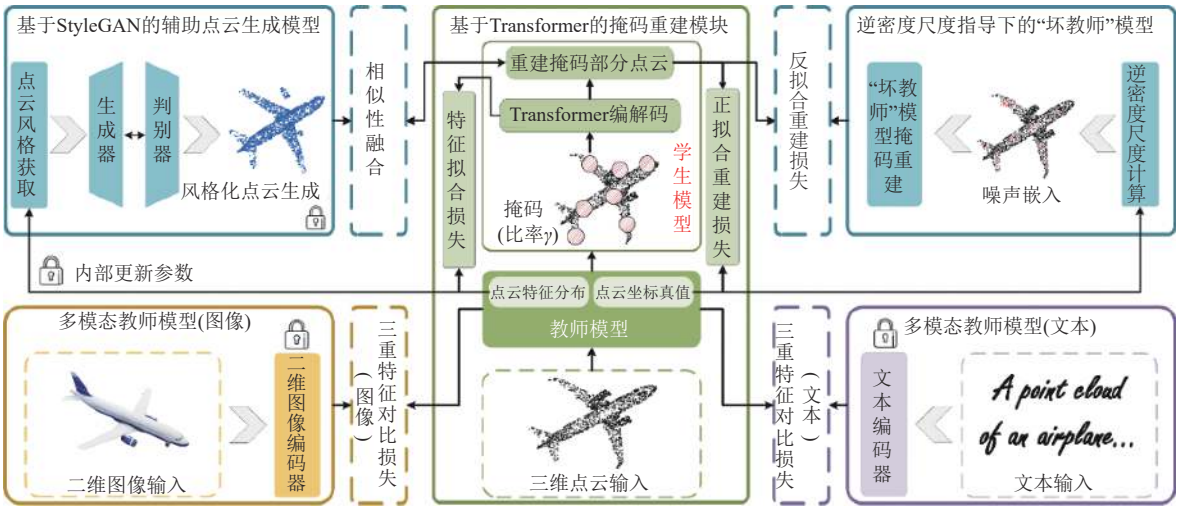


图 1　基于双向拟合掩码重建的多模态自监督点云表示学习网络框架

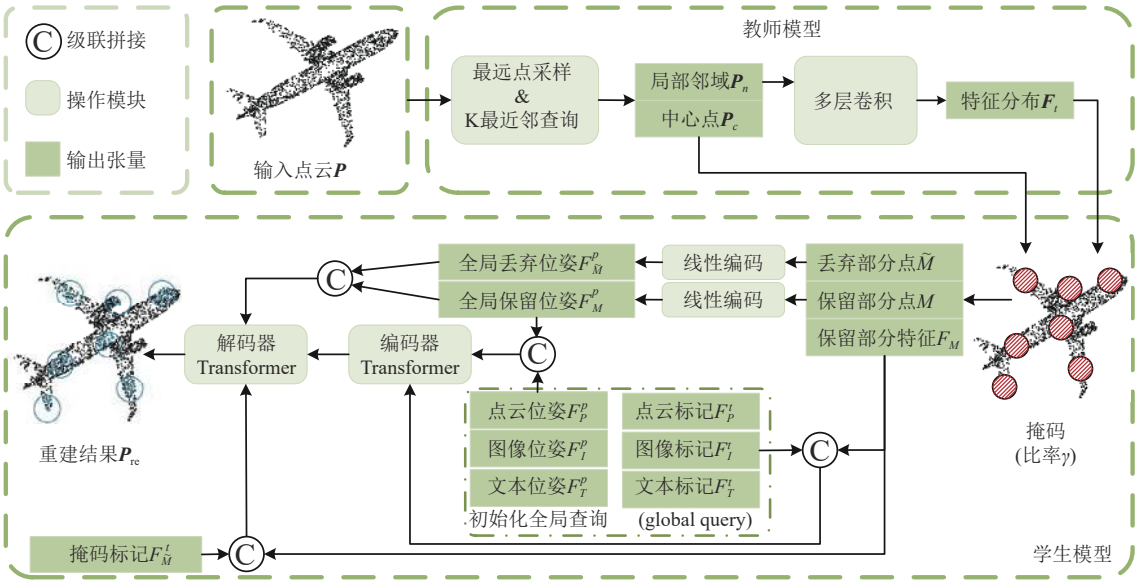


图 2　基于 Transformer 的掩码重建模型

(2) 逆密度尺度指导下的“坏教师”模型

提高丢弃部分的生成效率和质量是掩码重建工作的重点. 基于该原则, 为加速保留部分特征表示向真值逼近, 并提升模型所提取特征的表示力, “坏教师”模型在考虑点云局部几何密度结构下, 通过对添加高斯噪声的邻域特征真值进行特征抽象, 以构建“坏教师”特征分布. 其中, “坏教师”特征分布将被局部逆密度尺度重新加权, 以达到控制噪声偏移范围的目的. 之后, 通过本文所设计的双向拟合重建损失函数从坐标和特征分布两种层面上使重建结果逼近原始真值.

(3) 基于 StyleGAN 的辅助点云生成模型

在掩码重建过程中, 模型不可避免地生成不利于特征拟合的噪声点. 为了防止此类点影响后续特征表示学习效果, 网络模型需要在生成过程中对其进行迭代纠正和剔除. 为此, 本文对 StyleGAN 模型在面向点云生成方向改进, 以提高重建效果并隔绝噪声影响. 该模型将教师模型中局部点云作为基础风格的真实值输入进 StyleGAN 网络. 经过生成器和判别器不断地对抗生成, 产生出与局部点云风格相似的输出. 值得注意的是, 该生成对抗过程仅在内部完成, 即不参与全局参数更新, 以保证生成结果的纯净性. 最后, 具有阈值控制的相似性融合模块计算辅助点云与掩码重建结果的相似性得分. 若满足阈值条件则保留原始掩码重建结果, 反之将其定义为噪声并将辅助点云弥补替换进掩码重建结果.

(4) 多模态教师模型

为从其他包括图像和文本等模态获取更多有利于点云表示学习的信息, 现有方法[13,35]多数以对比学习的方式完成相似性聚类估计. 然而, 不同模态的数据基础存在本质差异, 经过参数非共享的不同网络学习表示后所产生出的特征存在跨模态信息冲突或单模态崩溃[41]. 因此, 通过挖掘并对齐点云-图像-文本三模态的相关特征以提高点云表示能力是多模态教师模型的核心任务. 一方面, 多模态教师模型通过 ViT-B[42]和 CLIP[32]模型获取图像和文本特征. 另一方面, 基于 Transformer 的掩码重建模型中初始化的图像和文本标记经过与点云模态参数共享的学习网络处理后, 两类标记被迭代地映射上点云模态的全局特征, 旨在建立跨模态参数传递桥梁. 之后, 图像和文本模态的特征和标记经过三重特征对比损失函数的不断迭代学习后, 能更为精确地对齐多模态特征并提高点云模态表示和判别能力.

2.2 逆密度尺度指导下的“坏教师”模型

现有教师-学生模型通过监督学习模型将重建结果向真值逼近以获取表示. 但该类模型在面向掩码重建任务上存在拟合速度慢且模型鲁棒性差的缺陷. 因此, 本文期望通过合理设计噪声模型反向逼近拟合并正确导向嵌入空间参数更新. 然而, 标准高斯分布对局部区域真值偏移的实用效果欠佳. 其原因在于点云分布存在非均匀性. 若统一添加某固定范围变化内的噪声将打破原有三维几何结构, 致使反向拟合效果降低. 基于此, 本文提出逆密度尺度指导下的“坏教师”模型, 如图 3 所示.

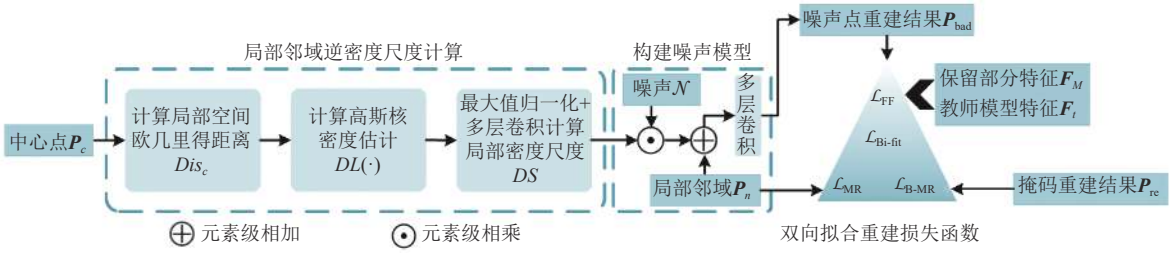


图 3 逆密度尺度指导下的“坏教师”模型

“坏教师”模型首先考虑点云的密度属性, 因此本文利用基于高斯核密度估计的逆密度重加权方法[8]衡量局部密度尺度并约束局部噪声变化范围. 模型对经过最远点采样的中心点 $\boldsymbol{P}_c \in \mathbb{R}^{N'\times 3}$ 计算欧几里得距离, 并得到局部空间距离张量 $Dis_c \in \mathbb{R}^{N'\times N'}$. 其次, 高斯核密度估计 $DL(\cdot)$ 局部空间距离张量 $Dis_c \in \mathbb{R}^{N'\times N'}$ 计算逆密度:

$$DL(Dis_c) = 1\Big/\frac{1}{K}\sum_{i=1}^{K}\frac{1}{h^3}g\left(\frac{(Dis_c)}{h}\right) \tag{1}$$

其中, K 为局部邻域点数, h 为带宽, $g(u)=\left(\frac{1}{\sqrt{2\pi}}\right)^3 \mathrm{e}^{-\frac{1}{2}u^2}$ 为多元高斯核函数. 之后, 逆密度尺度被最大值归一化, 并送入多层卷积组成的密度网络学习局部密度特征, 其输出结果被定义为局部逆密度尺度 $DS \in \mathbb{R}^{N'\times K\times 1}$.

在“坏教师”模型中, 本文首先初始化噪声张量 $\mathcal{N}$. 其中, 各子噪声的变化范围限制于各局部邻域点距离的最大和最小值. 其次, 局部逆密度尺度 DS 与噪声 $\mathcal{N}$ 逐元素相乘以控制变化范围. 之后, 基于 Transformer 的掩码重建模型中的局部邻域 $\boldsymbol{P}_n \in \mathbb{R}^{N'\times K\times 3}$ 与逆密度尺度加权下的噪声 $\mathcal{N}$ 加和, 完成邻域点噪声偏移. 最终, 偏移结果经过多层卷积完成特征抽象并重建估计噪声点, 得到与掩码重建结果形状相同的“坏教师”重建结果 $\boldsymbol{P}_{bad} \in \mathbb{R}^{N'\times K\times 3}$.

为将嵌入空间参数向正确方向更新, 掩码重建结果 $\boldsymbol{P}_{\text{re}}$ 需在背离“坏教师”重建结果 $\boldsymbol{P}_{\text{bad}}$ 的同时, 加速向真实值 $\boldsymbol{P}_n$ 靠近. 因此, 在损失函数的设定上, 本文采用重建任务中的倒角距离 l_{CD} (chamfer-distance loss), 可被称之为正向拟合重建损失函数 $\mathcal{L}_{\text{MR}}$:

$$\mathcal{L}_{\text{MR}} = l_{\text{CD}}\left(\boldsymbol{P}_{\text{re}}[m_{\text{idx}}], \boldsymbol{P}_n[m_{\text{idx}}]\right) = \sum\left[\frac{1}{|\boldsymbol{P}_{\text{re}}[m_{\text{idx}}]|}\sum_{re\in P_{\text{re}}[m_{\text{idx}}]}\min_{gt\in P_n[m_{\text{idx}}]}\|re-gt\|_2^2 + \sum_{gt\in P_n[m_{\text{idx}}]}\min_{re\in P_{\text{re}}[m_{\text{idx}}]}\|re-gt\|_2^2\right] \tag{2}$$

其中, re 和 gt 分别为掩码丢弃部分的重建结果和真值点. “坏教师”模型同样以倒角距离 l_{CD} 反向拟合. 因此, 存在“坏教师”掩码重建损失函数 $\mathcal{L}_{\text{B-MR}}$:

$$\mathcal{L}_{\text{B-MR}} = \max\left(0, l_{\text{CD}}\left(\boldsymbol{P}_{\text{bad}}[m_{\text{idx}}], \boldsymbol{P}_{\text{re}}[m_{\text{idx}}]\right) - \mathcal{L}_{\text{MR}}\right) \tag{3}$$

该损失函数的设定旨在拉进掩码重建结果与真值距离, 并强制推远与“坏教师”噪声重建结果的距离, 达到加速重建的研究目的. 除上述坐标层面上的重建逼近外, 特征空间的关联性变化也需要被约束. 因此, 为构建保留部分特征 $F_M \in \mathbb{R}^{N_v \times C}$ 和教师模型特征 $F_t \in \mathbb{R}^{N' \times C}$ 之间的信息传递, 本文预定义一种高斯分布的条件先验 $p(H)$, 即存在基于 H 的两种分布 $q_\phi(H|F_t)$ 和 $p_\varphi(H|F_M)$. 由于研究目标为通过基于保留部分特征的条件分布 $p_\varphi(H|F_M)$ 拟合基于教师模型特征的潜在分布 $q_\phi(H|F_t)$、 $q_\phi(H|F_t)$ 和 $p_\varphi(H|F_M)$ 被分别定义为先验分布和条件概率分布. 因此, 特征拟合损失函数 $\mathcal{L}_{\text{FF}}$ 可表示为:

$$\mathcal{L}_{\text{FF}} = -KL[q_\phi(H|F_t)\|p(H)] - KL[p_\varphi(H|F_M)\|q_\phi(H|F_t)] \tag{4}$$

其中, KL 为 KL 散度 (Kullback-Leibler divergence, 或称相对熵). 综上, 基于 Transformer 的掩码重建模型和“坏教师”模型联合构成的双向拟合重建损失函数可表示为:

$$\mathcal{L}_{\text{Bi-fit}} = \mathcal{L}_{\text{MR}} + \mathcal{L}_{\text{FF}} + \mathcal{L}_{\text{B-MR}} \tag{5}$$

2.3 基于 StyleGAN 的辅助点云生成模型

掩码生成模型的重建结果在初始阶段包含噪声点. 尽管优化过程中能检测并剔除部分噪声, 但在学习过程中, 噪声对点云全局查询标记和特征嵌入空间的完整性及表示力造成负面影响, 进而阻碍后续多模态特征对齐等多项操作的正常运行. 基于上述分析, 本文认为点云自监督表示学习方法需要在学习过程中, 对噪声加以额外限制, 并提出基于 StyleGAN 的辅助点云生成模型, 如图 4 所示. 为使辅助生成结果更加契合局部邻域真值, 该模型需要获取局部点云的空间结构风格信息. 局部邻域 $\boldsymbol{P}_n \in \mathbb{R}^{N' \times K \times 3}$ 先通过多层卷积提取特征并将结果与原始局部点级联拼接, 输出初始局部风格化点云特征 $\boldsymbol{P}_n^{\text{style}} \in \mathbb{R}^{N' \times K \times (3+C)}$. 之后, 该特征再次被卷积抽象至坐标风格 $\boldsymbol{P}_n^{p\text{-style}} \in \mathbb{R}^{N' \times K \times 3}$ 并作为 StyleGAN 网络的输入.

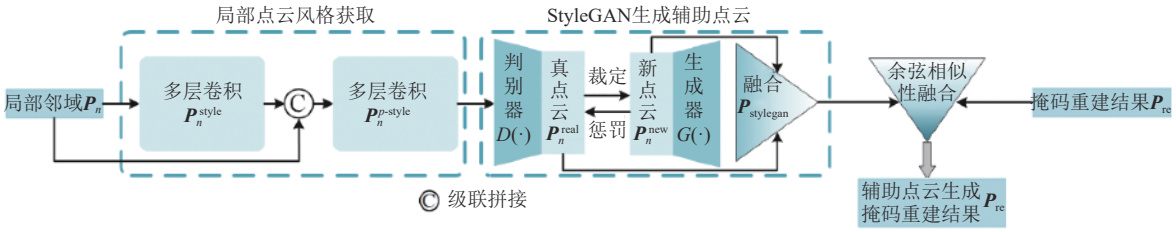


图 4 基于 StyleGAN 的辅助点云生成模型

然而, StyleGAN 模型主要基于二维图像的生成任务, 直接引入或将造成模型崩溃. 因此, 本文的改进方向如下: (1) 加深特征探索的网络层级并将卷积核大小固定为 1×1. (2) 保留噪声并使用低通滤波器对生成结果进行优化. 在改进的 StyleGAN 网络中, 先定义局部风格点云 $\boldsymbol{P}_n^{p\text{-style}} \in \mathbb{R}^{N' \times K \times 3}$ 为真值 $\boldsymbol{P}_n^{\text{real}}$. 生成器 $G(\cdot)$ 不断生成新点云 $\boldsymbol{P}_n^{\text{new}}$. 之后, 网络将 $\boldsymbol{P}_n^{\text{new}}$ 在判别器 $D(\cdot)$ 中裁定和惩罚性优化. 最终, 真实点云 $\boldsymbol{P}_n^{\text{real}}$ 与新点云 $\boldsymbol{P}_n^{\text{new}}$ 有权地融合, 得到风格化辅助点云输出 $\boldsymbol{P}_{\text{stylegan}}$:

$$\boldsymbol{P}_{\text{stylegan}} = \omega \boldsymbol{P}_n^{\text{real}} + (1-\omega)\boldsymbol{P}_n^{\text{new}} \tag{6}$$

其中, ω 为 StyleGAN 融合因子. 在风格化辅助点云与掩码重建的生成点云融合中, 本文采用余弦距离 $Dis(\cdot)$ 衡量二者:

$$Dis(\boldsymbol{P}_{\text{stylegan}},\boldsymbol{P}_{\text{re}})=1-\cos(\boldsymbol{P}_{\text{stylegan}},\boldsymbol{P}_{\text{re}})=1-\frac{\boldsymbol{P}_{\text{stylegan}}\cdot\boldsymbol{P}_{\text{re}}}{\|\boldsymbol{P}_{\text{stylegan}}\|\cdot\|\boldsymbol{P}_{\text{re}}\|} \tag{7}$$

最终, 当满足距离阈值 τ 时, 保留掩码重建结果. 反之, 辅助点云替换重建结果. 最终得到经过辅助点云去噪的掩码重建输出结果:

$$\boldsymbol{P}_{\text{re}}\Leftarrow\begin{cases}\boldsymbol{P}_{\text{re}}, & 0<Dis(\boldsymbol{P}_{\text{stylegan}},\boldsymbol{P}_{\text{re}})<\tau\\ \boldsymbol{P}_{\text{stylegan}}, & Dis(\boldsymbol{P}_{\text{stylegan}},\boldsymbol{P}_{\text{re}})>\tau\end{cases} \tag{8}$$

经过辅助点云替换的重建结果 P_{re} 进一步在“坏教师”模型中发挥作用.

2.4 多模态教师模型

为充分挖掘多模态信息中有助于提升点云表示力的潜在特征, 并防止多模态特征对齐过程中存在某模态内信息归零所导致的模型崩溃, 本文受到二维图像正则化方法 VICReg[41]的启发, 通过多模态教师模型在三重特征对比损失函数的约束下改善上述挑战. 首先, 与点云 $\boldsymbol{P}$ 相对应的图像 $\boldsymbol{I}$ 和文本 $\boldsymbol{T}$ 数据被分别送入预训练模型 CLIP 或 ViT-B 进行编码 $f(\cdot)$ 并得到图像特征 $\{F_I=f_I(\boldsymbol{I})|F_I\in\mathbb{R}^C\}$ 和文本特征 $\{F_T=f_T(\boldsymbol{T})|F_T\in\mathbb{R}^C\}$. 在基于 Transformer 的掩码重建模型中, 参数共享映射所得到的图像标记 $F_{I\text{-token}}$ 和文本标记 $F_{T\text{-token}}$ 经过线性编码得到 $F_{I\text{-}E},F_{T\text{-}E}$. 之后, 与模态特征 F_I,F_T 进行跨模态不变性特征对齐 (由于计算方式相同, 并且为了避免混淆, 模态标记 I,T 被暂时省略), 其损失函数 $\mathcal{L}_s$ 可被表示为:

$$\mathcal{L}_s(F,F_E)=Smooth_{L1}(F,F_E) \tag{9}$$

在模态内变量正则化损失函数 $\mathcal{L}_v$ 中, 本文通过具有阈值约束的铰链损失函数来预防嵌入向量归零崩溃:

$$\mathcal{L}_v(F,F_E)=\frac{1}{B}\sum_i^B\left\{\max\left[0,1-\sqrt{Var(F_i)+\varepsilon}\right]\right\}+\frac{1}{B}\sum_i^B\left\{\max\left[0,1-\sqrt{Var(F_{E,i})+\varepsilon}\right]\right\} \tag{10}$$

其中, B 为整体批次大小, ε 为常量, 一般定义为 1E–4. 假设某嵌入向量信息归零崩溃, 则在对比学习的梯度下降过程中会连带性地影响多个嵌入向量. 因此, 需要协方差正则化损失函数 $\mathcal{L}_c$ 解除向量之间的关联性:

$$\begin{aligned}\mathcal{L}_c(F,F_E)=&\frac{1}{B}\sum_{i\neq j}\left\{\frac{1}{C-1}\sum_{j=1}^C[F_j-mean(F_j)][F_j-mean(F_j)]^{\mathrm{T}}\right\}_{i,j}^2\\&+\frac{1}{B}\sum_{i\neq j}\left\{\frac{1}{C-1}\sum_{j=1}^C[F_{E,j}-mean(F_{E,j})][F_{E,j}-mean(F_{E,j})]^{\mathrm{T}}\right\}_{i,j}^2\end{aligned} \tag{11}$$

其中, C 为嵌入特征通道数, $mean(\cdot)$ 表示通道级均值化操作. 综上所述, 带有模态标记的多模态三重特征对比损失函数为如下:

$$\begin{aligned}\mathcal{L}_{\text{MM}}(F_I,F_T,F_{I\text{-}E},F_{T\text{-}E})=&\alpha[\mathcal{L}_s(F_I,F_{I\text{-}E})+\mathcal{L}_s(F_T,F_{T\text{-}E})]+\beta[\mathcal{L}_v(F_I,F_{I\text{-}E})+\mathcal{L}_v(F_T,F_{T\text{-}E})]\\&+\gamma[\mathcal{L}_c(F_I,F_{I\text{-}E})+\mathcal{L}_c(F_T,F_{T\text{-}E})]\end{aligned} \tag{12}$$

其中, α、β 和 γ 为可学习平衡参数.

2.5 总体损失函数

综合上述对本文所提出方法的描述, 总体损失函数共包含两个部分: (1) 逆密度尺度指导下的“坏教师”模型中双向拟合重建损失函数 $\mathcal{L}_{\text{Bi-fit}}$. (2) 多模态教师模型中的三重特征对比损失函数 $\mathcal{L}_{\text{MM}}$. 因此, 本文总体损失函数 $\mathcal{L}_{\text{overall}}$ 可表示为:

$$\mathcal{L}_{\text{overall}}=\mathcal{L}_{\text{Bi-fit}}+\mathcal{L}_{\text{MM}} \tag{13}$$

3 实 验

3.1 预训练和应用细节

本文以基于三维坐标的点云、基于 RGB 的图像及基于类别描述的文本作为输入. ShapeNet[16]是点云模型的预训练数据集, 由 CAD 模型组成且包含 55 个人工合成形状. 其扩展版 ShapeNetPart 含有归属于 16 个类别的

16 881 个物体. 经过映射转化后的 RGB 图像共有 43 783 张. 在预训练阶段, 基于 Transformer 的掩码重建模型的超参数设定与 ReCon[35]相同, 即中心点和邻域点数分别为 64 和 32, 掩码率为 60%, 特征通道数为 384. 此外, Transformer 的编码深度和头数分别为 12 和 6, 解码深度和头数为 4 和 6. 在多模态教师模型中, 图像模态和文本模态的特征不参与反向传播, 仅通过 ViT-B[42]和 CLIP[32]两种优良模型完成模态信息提取工作, 并与点云模态中多模态标记进行对比. 最终, 基于 Transformer 的点云识别网络利用预训练特征完成包括分类和分割等下游微调识别工作. 在下游基于 ModelNet 数据集的分类任务中, 类别数、输入点数、编码深度、头数、中心点数及邻域点数分别为 40、1 024、12、6、64 和 32. 而对于 ScanObjectNN 数据集, 上述超参数分别设定为 15、2 048、12、6、128 和 32. 小样本和零样本物体分类任务的超参数设定与 ModelNet 数据集分类任务相同, 而部件分割的超参数设定则与 ScanObjectNN 相同. 本文网络由深度学习框架 PyTorch 为基础搭建, 并且涉及的所有实验均基于搭配两张 Nvidia GeForce RTX3090 的 Ubuntu 22.04 系统上完成. 网络预训练中参数优化器、学习率、权重衰减、批次数训练轮次分别为 AdamW、5E–4、5E–5、128 以及 300. 在实验结果的展示中, 加粗代表最高值.

3.2 三维点云分类

(1) 人工物体分类

ModelNet[14]是由 CAD 图像合成的大型三维物体数据集, 共包含 12 311 个人工物体. 根据其类别数不同, 该数据集可被分为包含 10 类物体的 ModelNet10 和 40 类物体的 ModelNet40. 为验证所提出方法对三维人工数据的分类效果, ModelNet40 被作为分类器的输入. 其中, 分类网络对 ModelNet40 进行不同数量的下采样, 采样点的个数分别为 1 024 和 8 192. 本文方法与现有方法在 ModelNet40 数据集上的形状分类对比结如表 1 所示, 分类指标由总体精度 (overall accuracy, OA) 表示. 表 1 中各类方法通过发表期刊或会议、发表时间、方法类别以及输入点的个数进行标识. 从结果可得, 本文方法在 ModelNet40 数据集上的分类精度达到了最高水平. 其中, 在输入点个数分别为 1 024 和 8 192 下, 分类精度达到 94.6% 和 94.8%. 该结果不但完全优于众多监督学习优秀算法, 并且与其他基于掩码重建、基于多模态的最新算法相比, 也展现出优良的分类效果.

表 1 ModelNet40 数据集上各种方法的点云形状分类对比结果 (%)

方法	期刊或会议/年份	方法类别	输入点个数	
			1 024	8 192
PointNet[3]	CVPR/2017	监督学习	89.2	90.8
PointNet++[4]	NIPS/2017		90.7	91.9
DGCNN[5]	TOG/2019		92.9	—
PointMLP[43]	ICLR/2022		94.5	—
PointNeXt[44]	NIPS/2022		94.0	—
P2P-RN101[29]	NIPS/2022		93.1	—
P2P-HorNet[29]	NIPS/2022		94.0	—
文献[36]	软件学报/2023		93.8	—
Point-BERT[17]	CVPR/2022	自监督学习	93.2	93.8
MaskPoint[21]	ECCV/2022		93.8	—
Point-MAE[19]	ECCV/2022		93.8	94.0
Point-M2AE[20]	NIPS/2022		94.0	—
ACT[30]	ICLR/2023		93.7	94.0
ReCon[35]	ICML/2023		94.5	94.7
I2P-MAE[28]	CVPR/2023		94.1	—
本文方法	—		**94.6**	**94.8**

为进一步验证本文所提出方法的类别判别能力, 本文在 ModelNet40 数据集上采用线性支持向量机 (SVM) 分类测试完成方法评价. 本文方法与现有方法的线性支持向量机分类对比结果如表 2 所示. 表中各类方法采用不同

骨干网络进行预训练或直接训练得到分类结果. 本文所提出的方法达到 93.6%, 为最高线性分类结果. 在与当前点云-图像[28]或点云-图像-文本[35]的多模态方法相比, 本文结果展现出优越性.

表 2 ModelNet40 数据集上各种方法的点云线性支持向量机分类对比结果 (%)

方法	期刊或会议/年份	骨干网络	分类结果 (OA)
Jigsaw3D[9]	NIPS/2019	DGCNN	90.6
Info3D[45]	ECCV/2020	PointNet	89.8
ACD[46]	ECCV/2020	PointNet++	89.8
CMCV[27]	CVPR/2021	DGCNN	89.8
OcCo[10]	ICCV/2021	DGCNN	89.2
CrossPoint[13]	CVPR/2022	DGCNN	91.2
Point-BERT[17]	CVPR/2022	Transformer	87.4
Point-MAE[19]	ECCV/2022	Transformer	91.0
Point-M2AE[20]	NIPS/2022	Transformer	92.9
ReCon[35]	ICML/2023	Transformer	93.4
I2P-MAE[28]	CVPR/2023	Transformer	93.4
MCIB[47]	KBS/2024	DGCNN	91.6
本文方法	—	Transformer	**93.6**

(2) 真实世界物体分类

ScanObjectNN[15]是从真实世界收集而来的物体级点云数据集. 其中, 室内场景中的 2092 个物体被划分为 15 个类别. 相比人工物体数据集 ModelNet, ScanObjectNN 数据集含有诸多噪声和异常点, 对点云预训练模型造成挑战. 本文方法与现有方法在真实世界数据集 ScanObjectNN 上的形状分类对比结果 (OA) 如表 3 所示. 其中, 背景级拆分部分 (OBJ_BG)、物体级拆分部分 (OBJ_ONLY) 及最难级拆分部分 (PB_T50_RS) 为官方给定的 3 种不同数据集拆分类型. 从结果可得, 在 3 种不同拆分类别数据上, 本文方法均表现出最优的分类结果, 分别达到 96.22%、94.10% 及 92.44%, 明显高于最新方法 ACT[30]、ReCon[35]以及 I2P-MAE[28]. 在官方给定最难拆分部分 PB_T50_RS 中, 本文方法也具有卓越的分类表现. 与 ModelNet 数据集相同, 本文同样在 ScanObjectNN 数据集 (OBJ_BG 拆分部分) 上完成线性支持向量机分类测试. 其结果与现有方法对比结果如表 4 所示. 与最新的期刊或会议方法相比, 本文方法的结果达到 87.8%, 为当前最高线性支持向量机分类结果.

表 3 ScanObjectNN 数据集上各种方法的形状分类对比结果 (%)

方法	期刊或会议/年份	方法类别	OBJ_BG	OBJ_ONLY	PB_T50_RS
PointNet[3]	CVPR/2017	监督学习	73.3	79.2	68.0
PointNet++[4]	NIPS/2017		82.3	84.3	77.9
DGCNN[5]	TOG/2019		82.8	86.2	78.1
PointMLP[43]	ICLR/2022		—	—	85.4±0.3
PointNeXt[44]	NIPS/2022		—	—	87.7±0.4
P2P-RN101[29]	NIPS/2022		—	—	87.4
P2P-HorNet[29]	NIPS/2022		—	—	89.3
Point-BERT[17]	CVPR/2022	自监督学习	87.43	88.12	83.07
MaskPoint[21]	ECCV/2022		89.30	88.10	84.30
Point-MAE[19]	ECCV/2022		90.02	88.29	85.18
Point-M2AE[20]	NIPS/2022		91.22	88.81	86.43
ACT[30]	ICLR/2023		93.29	91.91	88.21
ReCon[35]	ICML/2023		95.35	93.63	91.26
I2P-MAE[28]	CVPR/2023		94.15	91.57	90.11
本文方法	—		**96.22**	**94.10**	**92.44**

表 4　ScanObjectNN 数据集上各种方法的线性支持向量机分类对比结果 (%)

方法	期刊或会议/年份	骨干网络	分类结果 (OA)
Jigsaw3D[9]	NIPS/2019	PointNet DGCNN	55.2 59.5
OcCo[10]	ICCV/2021	PointNet DGCNN DGCNN	69.5 78.3 77.9
CrossPoint[13]	CVPR/2022	PointNet DGCNN	75.6 81.7
Point-MAE[19]	ECCV/2022	Transformer	77.7
Point-M2AE[20]	NIPS/2022	Transformer	84.1
CrossNet[48]	TMM/2023	PointNet DGCNN	76.8 83.9
I2P-MAE[28]	CVPR/2023	Transformer	87.1
本文方法	—	Transformer	**87.8**

3.3　小样本物体分类

小样本 (few-shot) 物体分类旨在通过有限的数据验证自监督预训练模型的分类能力. 其测试方法为在现有数据集中挑选 N 个类别, 每个类别中包含 K 个样本以完成分类评价. 因此, 小样本分类测试分组可划分表示为 N-way/K-shot. 与多数现有自监督表示学习方法设定相同, 本文从 ModelNet40 数据集中挑选两种类别数和两种样本数, 即 5-way/10-shot、5-way/20-shot、10-way/10-shot 及 10-way/20-shot. 本文方法与现有方法在 ModelNet40 数据集上的小样本物体分类对比结果 (OA) 如表 5 所示. 其中, 本文采用 10 轮测试结果的平均数和标准差展示分类结果. 显而易见地, 本文方法在 4 种不同的小样本物体分类任务设定下, 取得了最好的分类效果, 即 97.6% ± 2.0%, 98.9% ± 1.0%, 93.6% ± 3.8%, 96.0% ± 2.2%. 与最新的先进方法 ACT[30]、ReCon[35]和 GPr-Net[49]相比, 本文的小样本物体分类效果展示出强劲的竞争力.

表 5　ModelNet40 数据集上各种方法的小样本物体分类对比结果 (%)

方法	期刊或会议/年份	骨干网络	5-way		10-way	
			10-shot	20-shot	10-shot	20-shot
PointNet++[4]	NIPS/2017	—	38.5±16.0	42.4±4.5	23.1±2.2	18.8±1.7
DGCNN[5]	TOG/2019	—	31.6±9.0	40.8±14.6	19.9±6.5	16.9±1.5
Jigsaw3D[9]	NIPS/2019	PointNet DGCNN	66.5±2.5 34.3±1.3	69.2±2.4 42.2±3.5	56.9±2.5 26.0±2.4	66.5±1.4 29.9±2.6
cTree[50]	NIPS/2020	PointNet DGCNN	63.2±3.4 60.0±2.8	68.9±3.0 65.7±2.6	49.2±1.9 48.5±1.8	50.1±1.6 53.0±1.3
OcCo[10]	ICCV/2021	PointNet DGCNN	89.7±1.9 90.6±2.8	92.4±1.6 92.5±1.9	83.9±1.8 82.9±1.3	89.7±1.5 86.5±2.2
CrossPoint[13]	CVPR/2022	PointNet DGCNN	90.9±4.8 92.5±3.0	93.5±4.4 94.9±2.1	84.6±4.7 83.6±5.3	90.2±2.2 87.9±4.2
Point-BERT[17]	CVPR/2022	Transformer	94.6±3.1	96.3±2.7	91.0±5.4	92.7±5.1
MaskPoint[21]	ECCV/2022	Transformer	95.0±3.7	97.2±1.7	91.4±4.0	93.4±3.5
Point-MAE[19]	ECCV/2022	Transformer	96.3±2.5	97.8±1.8	92.6±4.1	95.0±3.0
Point-M2AE[20]	NIPS/2022	Transformer	96.8±1.8	98.3±1.4	92.3±4.5	95.0±3.0
ACT[30]	ICLR/2023	Transformer	96.8±2.3	98.0±1.4	93.3±4.0	95.6±2.8
ReCon[35]	ICML/2023	Transformer	97.3±1.9	98.9±1.2	93.3±3.9	95.8±3.0
GPr-Net[49]	CVPR/2023	PNet[51]	80.4±0.6	82.0±0.9	70.4±1.8	72.8±1.8
本文方法	—	Transformer	**97.6±2.0**	**98.9±1.0**	**93.6±3.8**	**96.0±2.2**

3.4 零样本物体分类

零样本 (zero-shot) 物体分类是将预训练模型泛化至未曾训练过的测试集中, 验证其分类效果. 基于此, 本文以人工合成的 ModelNet10 和 ModelNet40 数据集作为输入, 验证预训练模型的泛化能力. 值得注意的是, 零样本学习与微调学习的预训练模型存在细微差异, 即图像和文本编码器均采用 ViT-B[42]且传播冻结 (frozen). 本文方法与现有 4 种点云-图像-文本三模态自监督表示学习方法在零样本物体分类对比结果 (OA) 如表 6 所示. PointCLIP[31]通过多视角映射图像来聚合三维特征, 并且依靠文本信息辅助以获取三维点云表示. 该方法优势主要体现于无需三维点云模态训练, 但其零样本分类表现欠佳. 为了更好地吸取图像和文本模态的有用信息, CLIP2Point[52]、ReCon[35]及 PointCLIP V2[33]这 3 种现有先进方法做出不同的改进, 即通过点云-深度图像映射聚合、联合掩码重建和多模态对比学习以及 GPT-3 文本特征增强等方式提升点云表示的丰富性. 与上述 4 种方法相比, 本文方法则将研究重点落于掩码重建中的鲁棒性生成以及多模态特征对齐. 该方法的基本思想和预训练方法相对简单, 并且多模态特征仅需 CLIP 和 ViT-B 预训练模型获取图像和文本特征. 从表 6 可知, 本文方法在 ModelNet10 和 ModelNet40 的零样本分类结果分别达到 82.4% 和 67.4%, 均超过所对比的最新方法, 证明本文方法在零样本物体分类下游任务上的有效性.

表 6 ModelNet10 和 ModelNet40 数据集上各种方法的零样本物体分类对比结果 (%)

方法	期刊或会议/年份	骨干网络	ModelNet10	ModelNet40
PointCLIP[31]	CVPR/2022	ResNet-50	30.2	20.2
CLIP2Point[52]	ICCV/2023	Transformer	66.6	49.4
ReCon[35]	ICML/2023	Transformer	81.6	66.8
PointCLIP V2[33]	ICCV/2023	Transformer	73.1	64.2
本文方法	—	Transformer	**82.4**	**67.4**

3.5 三维点云部件分割

相比点云分类, 点云分割侧重于检测点云预训练模型对三维几何信息的深层探索能力. 因此, 本文采用 ShapeNetPart 数据集完成部件分割实验. 为了公平比较, 实验所选用的分割头与 Point-MAE[19]和 ReCon[35]相同. 此外, 评价指标选择广泛应用的交并比 (intersection-over-union, IoU). 实验中, 平均 IoU (mIoU) 从两种层面被计算且将其定义为两种评价类型: 类别 mIoU 和实例 mIoU. 本文方法和现有方法在 ShapeNetPart 数据集上的部件分割对比结果如表 7 所示. 此外, 本文同样展示每个类别的 IoU% 结果, 如表 8 所示. 从表 7 和表 8 中的对比结果可以得出, 本文方法在类别 mIoU 和实例 mIoU 两种指标上分别达到 85.1% 和 86.6%, 均超越其他监督和无监督策略下基于卷积神经网络和基于 Transformer 的众多先进方法. 此外, 在 16 项单个类别分割结果中, 本文方法在 10 项上具有突出的领先优势. 除了定量结果展示, 本文将部件分割结果进行定性可视化, 如图 5 所示. 由于分割结果与真值 (ground truth) 在视觉上差异过小, 因此本文展示分割结果与真值的对比结果. 在图 5 中, 红色点表示预测错误点, 蓝色为预测正确点. 为了突出本文方法的优越性, 基于 DGCNN 的点云-图像双模态自监督学习方法 CrossPoint[13]和基于 Transformer 的点云-图像-文本三模态自监督学习方法 ReCon[35]作为参照对比方法. 在所展示的 4 种人工形状中, 本文方法的预测错误红色点最为稀少, 足以证明本文方法具有良好的部件分割能力.

表 7 ShapeNetPart 数据集上各种方法的平均部件分割对比结果 (%)

方法	期刊或会议/年份	骨干网络	类别mIoU	实例mIoU
PointNet[3]	CVPR/2017	—	80.4	83.7
PointNet++[4]	NIPS/2017	—	81.9	85.1
DGCNN[5]	TOG/2019	—	82.3	85.2
Jigsaw3D[9]	NIPS/2019	DGCNN	—	85.3
PointContrast[11]	ECCV/2020	MinkNet[53]+UNet[54]	—	85.1
PointMLP[43]	ICLR/2022	—	84.6	86.1

表 7 ShapeNetPart 数据集上各种方法的平均部件分割对比结果 (%) (续)

方法	期刊或会议/年份	骨干网络	类别mIoU	实例mIoU
CrossPoint[13]	CVPR/2022	DGCNN	—	85.5
Point-BERT[17]	CVPR/2022	Transformer	84.1	85.6
Point-MAE[19]	ECCV/2022	Transformer	—	86.1
P2P-SFPN[29]	NIPS/2022	CN[55]+SFPN[56]	82.5	85.7
P2P-UPer[29]	NIPS/2022	CN+UPerNet[57]	84.1	86.5
ACT[30]	ICLR/2023	Transformer	84.7	86.1
ReCon[35]	ICML/2023	Transformer	84.8	86.4
本文方法	—	Transformer	**85.1**	**86.6**

表 8 ShapeNetPart 数据集上各种方法的单个类别部件分割对比结果 (%)

方法	aero plane	bag	cap	car	chair	ear phone	guitar	knife	lamp	laptop	motor bike	mug	pistol	rocket	skate board	table
PointNet[3]	83.4	78.7	82.5	74.9	89.6	73.0	91.5	85.9	80.8	95.3	65.2	93.0	81.2	57.9	72.8	80.6
PointNet++[4]	82.4	79.0	87.7	77.3	90.8	71.8	91.0	85.9	83.7	95.3	71.6	94.1	81.3	58.7	76.4	82.6
DGCNN[5]	84.0	83.4	86.7	77.8	90.6	74.7	91.2	87.5	82.8	95.7	66.3	94.9	81.1	63.5	74.5	82.6
Point-BERT[17]	84.3	84.8	88.0	79.8	91.0	81.7	91.6	87.9	85.2	95.6	75.6	94.7	84.3	63.4	76.3	81.5
PointMLP [43]	83.5	83.4	87.5	80.5	90.3	78.2	**92.2**	88.1	82.6	96.2	**77.5**	**95.8**	**85.4**	64.6	**83.3**	**84.3**
P2P[29]	84.3	85.1	88.3	80.4	91.6	80.8	92.1	87.9	85.6	95.9	76.1	94.2	82.4	62.7	74.7	83.7
本文方法	**85.3**	**85.3**	**89.1**	**81.5**	**91.7**	**81.9**	92.1	**88.2**	**86.2**	**96.5**	75.6	95.1	84.7	**66.7**	80.1	81.9

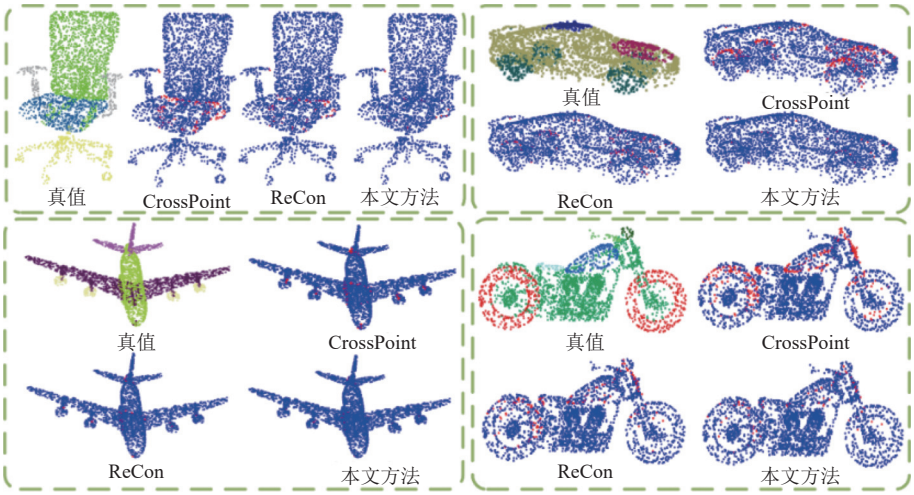


图 5 ShapeNetPart 数据集上 3 种方法的部件分割可视化

3.6 消融学习和超参数验证

(1) 子模型消融学习

为验证本文方法中各子模型的有效性以及对点云识别结果的贡献度, 本节选用真实世界点云数据集 ScanObjectNN 完成消融学习实验. 实验结果 (OA) 如表 9 所示. 其中, “√”表示添加, “—”表示删除. 本文方法所涉及的子模型分为 3 种: (1) “坏教师”模型; (2) 辅助点云生成模型; (3) 多模态教师模型. 在消融“坏教师”模型和多模态教师模型实验中, 骨干网络的正向拟合重建损失函数 $\mathcal{L}_{\mathrm{MR}}$ 和跨模态特征不变性对齐损失函数 $\mathcal{L}_s$ 保留以确保网络正常学习. 从表 9 中可以得出, 3 种子模型依次嵌入进骨干网络后, 均对分类结果产生显著贡献. 当“坏教师”模型

单独嵌入执行时, 在 3 种拆分数据集上分别取得 95.79%, 93.84% 和 91.77% 的精度结果. 相比于辅助点云生成模型 (95.64%/93.71%/91.65%) 和多模态教师模型 (95.54%/93.69%/91.42%) 的分类结果, “坏教师”模型的分类贡献增益最大. 其原因在于该模型在反向噪声拟合中提升了效率和对噪声的抗性. 因此, 预训练模型能较好地屏蔽真实世界噪声对表示学习的不良影响, 这对 ScanObjectNN 数据集尤为重要. 之后, 3 种模型依次组合以完成二阶消融学习测试. 当“坏教师”模型和辅助点云生成模型进行组合时, 分类结果为 96.15%, 94.02% 和 92.26%. 相比其他两种模型的结合, 展示出强劲的分类表现. 综合表 9 中单模型和两两结合的模型在真实世界数据集上的分类效果, 本文得出以下结论: (1) 本文所提出的 3 种子模型均对骨干网络的分类性能产生积极作用. (2) 逆密度尺度指导下的“坏教师”模型, 及其与基于 StyleGAN 的辅助点云生成模型的结合模型的效果提升最为显著. 经过分析, 其原因在于“坏教师”模型通过反向拟合策略在初始阶段给定与源局部特征分布相似的噪声分布. 经过逐步双向加速重建拟合, 最终所捕获的特征表示具有强鲁棒性和判别力. 加入辅助点云重新判别后, 掩码重建所生成的噪声点或异常点也被丢弃. 因此, 在面对包含噪声的 ScanObjectNN 数据集时, 两种模型分别从特征层级和重建层级加强了模型表示力, 因此得到优良的点云分类性能.

(2) 逆密度尺度指导下的“坏教师”模型消融学习

在本节, 逆密度尺度指导下的“坏教师”模型中 3 种子方法依次嵌入以验证其有效性, 其中包括基于高斯核密度估计的逆密度重加权方法 (简称逆密度尺度指导)、“坏教师”掩码重建损失函数 $\mathcal{L}_{\text{B-MR}}$ 及特征拟合损失函数 $\mathcal{L}_{\text{FF}}$. 当逆密度尺度指导策略消融时, 中心点直接通过添加固定小范围噪声对其进行偏移. 经过多层感知机抽象特征后, 与局部邻域元素级相加并完成后续重建工作. 在 ScanObjectNN 数据集上的点云分类消融学习结果 (OA) 如表 10 所示. 从表 10 中可以得到, 联合基于高斯核密度估计的逆密度重加权方法和“坏教师”掩码重建损失函数的结合模型在 3 种拆分数据集上的分类精度分别达到 95.66%、93.78% 和 91.71%. 相比缺少逆密度尺度指导, 该联合模型作用效果突出. 而特征拟合损失函数 $\mathcal{L}_{\text{FF}}$ 的主要功能在于为掩码重建添加特征层级上的定向拟合. 该方法与“坏教师”掩码重建损失函数结合的消融结果为 95.62%、93.76% 和 91.69%. 与单“坏教师”掩码重建损失函数的消融结果相比, $\mathcal{L}_{\text{FF}}$ 的分类结果贡献度显著.

表 9 ScanObjectNN 数据集上子模型的点云分类消融学习结果 (%)

“坏教师”模型	辅助点云生成模型	多模态教师模型	OBJ_BG	OBJ_ONLY	PB_T50_RS
—	—	—	95.35	93.63	91.26
√	—	—	95.79	93.84	91.77
—	√	—	95.64	93.71	91.65
—	—	√	95.54	93.69	91.42
√	√	—	96.15	94.02	92.26
√	—	√	96.02	93.90	92.11
—	√	√	95.98	93.88	91.91
√	√	√	**96.22**	**94.10**	**92.44**

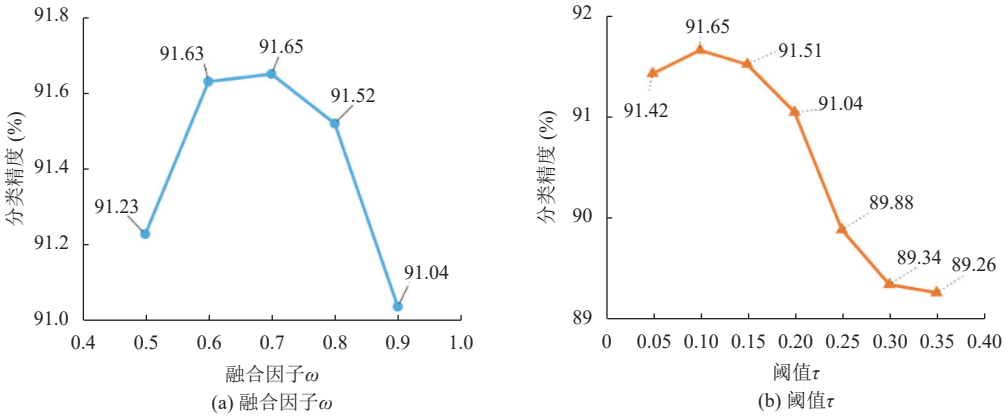
表 10 ScanObjectNN 数据集上逆密度尺度指导下的“坏教师”模型点云分类消融学习结果 (%)

逆密度尺度指导	$\mathcal{L}_{\text{B-MR}}$	$\mathcal{L}_{\text{FF}}$	OBJ_BG	OBJ_ONLY	PB_T50_RS
—	—	—	95.35	93.63	91.26
—	√	—	95.44	93.67	91.45
√	√	—	95.66	93.78	91.71
—	—	√	95.46	93.70	91.33
—	√	√	95.62	93.76	91.69
√	√	√	**95.79**	**93.84**	**91.77**

(3) 基于 StyleGAN 的辅助点云生成模型超参数验证

基于 StyleGAN 的辅助点云生成模型中涉及的超参数分别为: (1) 基于 StyleGAN 的融合因子 ω, 其功能在于有权地融合风格生成的点云结果和真值. (2) 基于 StyleGAN 的辅助点云生成结果与原始掩码重建点云的融合阈值 τ, 用于抵抗掩码重建所生成新噪声点的影响. 两种超参数验证实验均在 ScanObjectNN 中 PB_T50_RS 拆分部分上完成, 且仅在骨干网络基础上保留基于 StyleGAN 的辅助点云生成模型. 对于融合因子 ω, 本文取 [0.5,0.9] 范围内的 5 种不同值进行验证. 从图 6(a) 中可知, 当融合因子 $\omega = 0.7$ 时, 分类表现最佳. 其原因在于 StyleGAN 生成辅助点云过程中真实点云 $\boldsymbol{P}_n^{\text{real}}$ 担当生成风格引导, 保证生成结果能有助于服务原始重建结果. 因此, $\boldsymbol{P}_n^{\text{real}}$ 占相对主导地位时可获得良好的分类效果. 然而, 当生成结果与 $\boldsymbol{P}_n^{\text{real}}$ 几乎相同时, 即当融合因子 $\omega = 0.9$ 时, 辅助点云生成模型并未发挥作用, 分类结果最差. 不同阈值 τ 的分类结果如图 6(b) 所示. 当原始重建结果与辅助点云生成结果的余弦

距离衡量阈值 $\tau = 0.1$ 时, 分类结果最佳且为 91.65%. 随着阈值 τ 不断增大, 分类效果急剧下滑. 该结果显式地说明辅助点云与原始结果的相似性相对较大时, 基于 StyleGAN 的辅助点云生成模型将对分类结果产生贡献. 但阈值度量标准过于松弛将对下游分类任务产生不利影响.



(a) 融合因子ω (b) 阈值τ

图 6 不同融合因子 ω 和阈值 τ 对分类结果的影响

(4) 多模态教师模型消融学习

多模态教师模型主要包括图像模态和文本模态. 发挥关键作用的 3 种损失函数分别为: (1) 跨模态不变性特征对齐损失函数 $\mathcal{L}_s$. (2) 模态内变量正则化损失函数 $\mathcal{L}_v$. (3) 协方差正则化损失函数 $\mathcal{L}_c$. 基于此, 本节以 ScanObjectNN 数据集中 OBJ_BG 拆分部分数据集为验证基础, 对两种模态中的 3 个关键损失函数进行消融学习测试. 其中, 实验实施细节为保留基于掩码重建表示学习过程, 并在两种模态上分别依次加入 3 种损失函数得到分类精度以验证有效性, 其结果如表 11 所示. 当图像和文本模态首先与三维点云模态表示通过 $\mathcal{L}_s$ 进行特征对齐后, 分类精度以 0.4% 和 1.0% 开始增长. 在此基础上, 为防止信息崩塌, $\mathcal{L}_v$ 和 $\mathcal{L}_c$ 依次嵌入进骨干网络后, 均产生分类贡献. 综合数据结果可得出, 文本模态的分类表现更加突出, 并且模态内变量正则化损失函数 $\mathcal{L}_v$ 发挥重要作用. 其原因在于文本数据包含具体类型信息, 关键性特征描述易被提取. 此外, $\mathcal{L}_v$ 是防止信息崩溃的关键, 而协方差正则化损失函数 $\mathcal{L}_c$ 用以解除关联性. 因此, $\mathcal{L}_v$ 在两种模态下的平均贡献度更高.

表 11 ScanObjectNN 数据集上多模态教师模型点云分类消融学习结果 (%)

图像模态			文本模态			ScanObjectNN
$\mathcal{L}_s$	$\mathcal{L}_v$	$\mathcal{L}_c$	$\mathcal{L}_s$	$\mathcal{L}_v$	$\mathcal{L}_c$	OBJ_BG (OA)
—	—	—	—	—	—	95.35
√	—	—	—	—	—	95.39
√	√	—	—	—	—	95.41
√	—	√	—	—	—	95.40
√	√	√	—	—	—	95.46
—	—	—	√	—	—	95.45
—	—	—	√	√	—	95.50
—	—	—	√	—	√	95.48
—	—	—	√	√	√	95.52
√	√	√	√	√	√	**95.54**

(5) 文本模态消融学习

为了探究文本模态对预训练模型理解三维物体的贡献度, 本文基于零样本分类任务, 跟随 ReCon 的参数设置, 采用逐个合并文本嵌入的方式对 ModelNet 数据集完成分类实验. 其中, 文本被分类为 3 种类型: 前缀、类别以及后缀. 前缀包含“A model of”“A point cloud of”“A 3D rendered model of”等. 类别具体描述点云形状, 包括 Chair 和

Airplane 等. 后缀包含“with white background”“with white context.”等. 上述文本描述最终将被合并成一句完整的信息作为零样本分类任务的输入. 表 12 展示了文本模态消融学习结果. 从表中结果可以看出, “A rendered image of”+类别的效果最佳. 其原因在于零样本分类任务以图像和文本作为输入, 针对图像的准确文本描述将更有助于模型学习. 然而, 固定描述性语言的信息丰富度和贡献性有限, 并且文本信息模糊可能导致表示学习模型吸纳信息疲软等不利影响.

表 12 ModelNet40 数据集上文本模态点云零样本分类消融学习结果 (%)

前缀+类别	OA	类别+后缀	OA
“ ”+类别	60.22	类别+“ ”	60.22
“A”+类别	62.90	类别+“.”	56.23
“A model of”+类别	56.32	类别+“with white background.”	63.89
“A model of a”+类别	58.65	类别+“with white context.”	63.98
“An image of”+类别	62.38	—	—
“An image of a”+类别	62.22	—	—
“A 3D model of”+类别	63.02	—	—
“A 3D model of a”+类别	32.15	—	—
“A rendered model of”+类别	63.31	—	—
“A rendered model of a”+类别	62.98	—	—
“A point cloud of”+类别	60.11	—	—
“A point cloud of a”+类别	60.85	—	—
“A point cloud model of”+类别	63.55	—	—
“A point cloud model of a”+类别	62.59	—	—
“A 3D rendered model of”+类别	62.26	—	—
“A 3D rendered model of a”+类别	62.95	—	—
“A rendered image of”+类别	66.38	—	—
“A rendered image of a”+类别	62.39	—	—
“A 3D rendered image of”+类别	65.21	—	—
“A 3D rendered image of a”+类别	64.05	—	—

(6) 模型复杂度

3D 预训练模型的复杂度常通过模型空间和时间消耗量级来衡量. 为了评价所提出方法的模型复杂度, 本文将在相同运行条件下通过 4 种不同类型的指标测试现有方法及本文方法. 指标包括参数量、显存消耗、单轮运行时间以及 FLOPs. 此外, 待对比的方法包括 Point-MAE[19]、Point-M2AE[20]、I2P-MAE[28]、ACT[30]以及 ReCon[35]. 在 ModelNet40 数据集上微调测试模型复杂度对比结果如表 13 所示. 空间复杂度上, 联合掩码重建和对比学习的 ReCon 和本文方法参数量较大, 但显存消耗较小, 源于优良的参数共享机制. 时间复杂度上, Point-MAE、ACT、ReCon 和本文方法的单轮微调时间和 FLOPs 相似并且明显小于 Point-M2AE 和 I2P-MAE. 其原因在于两种方法在多尺度特征构建和双向映射上时间消耗较大. 综上, 本文在拥有最佳微调效果的情况下仍然具有相对较好的模型复杂度.

表 13 ModelNet40 数据集上模型复杂度测试结果

方法	空间复杂度		时间复杂度	
	参数量 (M)	显存消耗 (Mib)	单轮运行时间 (s)	FLOPs (M)
Point-MAE[19]	22.1	23 038	49	76 761.32
Point-M2AE[20]	12.8	14 821	78	149 390.23
I2P-MAE[28]	12.8	18 849	118	355 708.04
ACT[30]	22.1	11 145	46	76 761.32
ReCon[35]	43.6	11 064	38	85 037.04
本文方法	43.6	11 037	40	85 128.32

3.7 鲁棒性测试

逆密度尺度指导下的“坏教师”模型和基于 StyleGAN 的辅助点云生成模型分别以反向拟合策略和辅助点云融合策略来改善不同类型和尺度下点云噪声对表示学习的不良影响. 此外, 在丢失部分点云时, 模型的分类表现也是衡量稳健性的重要指标. 因此, 本节以输入大小为 1 024 个点的 ModelNet40 数据集为基础, 分别添加等级范围为 [0, 0.1] 的随机点云噪声、随机丢失比率范围为 [0, 0.8] 的点云以及空洞丢失比率范围为 [0, 0.5] 的点云作为输入以测试预训练模型的点云分类表现. 为更好地对比观察, 本文选用点云-图像-文本三模态现有先进方法 ReCon[35]的预训练模型在同等实验实施设定下完成点云分类任务, 其结果如图 7 所示. 首先, 图 7(a) 展示不同噪声等级下的分类结果. 可以观察得到, 本文方法的分类精度一直处于领先位置. 原因归结于本文方法中所提出的两种子模型均对噪声影响具有良好的抗性. 此外, 图 7(b) 展示不同随机丢失率下的分类结果. 当一定比率的原始点被随机丢失后, 本文方法也能以较为平稳的趋势保持良好的分类表现. 此外, 在图 7(b) 中, 比率为 0.8 时本文方法仍然能具有良好的性能. 因此, 本文认为按一定比率随机丢弃点的方法保留了三维几何结构信息, 致使模型仍然可以学习并利用上下文. 为了进一步验证本文方法在空洞点云上的性能, 按 10%–50% 的比率掏空某一局部区域并将剩余部分输入进微调模型. 结果如图 7(c) 所示, 整体下降幅度与随机丢弃法相似, 但同一比率对比下可以看出, 空洞丢弃导致模型识别率下降水平更大, 同样验证了随机丢弃能保留几何信息的论断. 此外, 在相同丢失比率下, 本文方法也能相较于先进方法 ReCon 具有更高的分类结果. 综上, 本文方法在面对不同程度和不同类型的数据干扰时, 具有良好的鲁棒性.

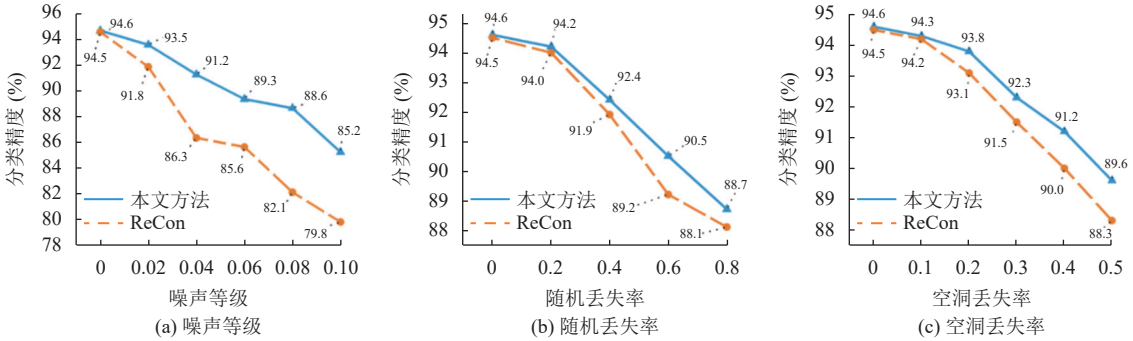


图 7 鲁棒性测试结果

4 总 结

针对现有多模态自监督点云表示学习方法中所存在的鲁棒性和泛化性弱、多模态特征难对齐及模态信息崩塌等缺陷, 本文提出基于双向拟合掩码重建的多模态自监督点云表示学习方法. 该方法包含逆密度尺度指导下的“坏教师”模型、基于 StyleGAN 的辅助点云生成模型及多模态教师模型, 分别通过双向拟合掩码重建、辅助点云替换及多模态特征对齐约束提升点云自监督表示学习能力. 在 ModelNet、ScanObjectNN 及 ShapeNet 数据集上, 所提出方法完成点云分类、线性支持向量机分类、小样本分类、零样本分类以及部件分割测试, 与现有先进方法对比具有优良效果. 此外, 消融学习等验证性实验结果证明, 所提出的每种子模型均对提升点云识别精度具有显著贡献. 最后, 鲁棒性测试结果验证本文方法对不同类型的数据干扰有良好的抵抗能力.

References:

[1] Zhu XL, Wang HC, You HM, Zhang WH, Zhang YY, Liu S, Chen JJ, Wang Z, Li KQ. Survey on testing of intelligent systems in autonomous vehicles. Ruan Jian Xue Bao/Journal of Software, 2021, 32(7): 2056–2077 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6266.htm [doi: 10.13328/j.cnki.jos.006266]

[2] Yan T, Gao HX, Zhang JF, Qian YH, Zhang LY. Grouping parallel lightweight real-time microscopic 3D shape reconstruction method. Ruan Jian Xue Bao/Journal of Software, 2024, 35(4): 1717–1731 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7013.htm [doi: 10.13328/j.cnki.jos.007013]

[3] Qi Charles R, Su H, Mo KC, Guibas LJ. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 652–660. [doi: 10.1109/CVPR.2017.16]

[4] Qi CR, Yi L, Su H, Guibas LJ. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5105–5114.

[5] Wang Y, Sun YB, Liu ZW, Sarma SE, Bronstein MM, Solomon JM. Dynamic graph CNN for learning on point clouds. ACM Trans. on Graphics, 2019, 38(5): 146. [doi: 10.1145/3326362]

[6] Cheng HZ, Lu J, Luo MX, Liu W, Zhang KB. PTANet: Triple attention network for point cloud semantic segmentation. Engineering Applications of Artificial Intelligence, 2021, 102: 104239. [doi: 10.1016/j.engappai.2021.104239]

[7] Cheng HZ, Zhu JH, Lu J, Han X. EDGCNet: Joint dynamic hyperbolic graph convolution and dual squeeze-and-attention for 3D point cloud segmentation. Expert Systems with Applications, 2024, 237: 121551. [doi: 10.1016/j.eswa.2023.121551]

[8] Lu J, Cheng HZ, Luo MX, Liu T, Zhang KB. PUConv: Upsampling convolutional network for point cloud semantic segmentation. Electronics Letters, 2020, 56(9): 435–438. [doi: 10.1049/el.2019.3705]

[9] Sauder J, Sievers B. Self-supervised deep learning on point clouds by reconstructing space. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 1161.

[10] Wang HC, Liu Q, Yue XY, Lasenby J, Kusner MJ. Unsupervised point cloud pre-training via occlusion completion. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 9782–9792. [doi: 10.1109/ICCV48922.2021.00964]

[11] Xie SN, Gu JT, Guo DM, Qi CR, Guibas L, Litany O. PointContrast: Unsupervised pre-training for 3D point cloud understanding. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 574–591. [doi: 10.1007/978-3-030-58580-8_34]

[12] Shi PC, Cheng HZ, Han X, Zhou YY, Zhu JH. DualGenerator: Information interaction-based generative network for point cloud completion. IEEE Robotics and Automation Letters, 2023, 8(10): 6627–6634. [doi: 10.1109/LRA.2023.3310406]

[13] Afham M, Dissanayake I, Dissanayake D, Dharmasiri A, Thilakarathna K, Rodrigo R. CrossPoint: Self-supervised cross-modal contrastive learning for 3D point cloud understanding. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 9902–9912. [doi: 10.1109/CVPR52688.2022.00967]

[14] Wu ZR, Song SR, Khosla A, Yu F, Zhang LG, Tang XO, Xiao JX. 3D ShapeNets: A deep representation for volumetric shapes. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1912–1920. [doi: 10.1109/CVPR.2015.7298801]

[15] Uy MA, Pham QH, Hua BS, Nguyen T, Yeung SK. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1588–1597. [doi: 10.1109/ICCV.2019.00167]

[16] Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang QX, Li ZM, Savarese S, Savva M, Song SR, Su H, Xiao JX, Yi L, Yu F. ShapeNet: An information-rich 3D model repository. arXiv:1512.03012, 2015.

[17] Yu XM, Tang LL, Rao YM, Huang TJ, Zhou J, Lu JW. Point-BERT: Pre-training 3D point cloud Transformers with masked point modeling. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 19313–19322. [doi: 10.1109/CVPR52688.2022.01871]

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[19] Pang YT, Wang WX, Tay FEH, Liu W, Tian YH, Yuan L. Masked autoencoders for point cloud self-supervised learning. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 604–621. [doi: 10.1007/978-3-031-20086-1_35]

[20] Zhang RR, Guo ZY, Fang RY, Zhao B, Wang D, Qiao Y, Li HS, Gao P. Point-M2AE: Multi-scale masked autoencoders for hierarchical point cloud pre-training. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1962.

[21] Liu HT, Cai M, Lee YJ. Masked discrimination for self-supervised learning on point clouds. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 657–675. [doi: 10.1007/978-3-031-20086-1_38]

[22] Zhang ZW, Girdhar R, Joulin A, Misra I. Self-supervised pretraining of 3D features on any point-cloud. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 10252–10263. [doi: 10.1109/ICCV48922.2021.01009]

[23] Wang D, Yang ZX. Self-supervised point cloud understanding via mask Transformer and contrastive learning. IEEE Robotics and Automation Letters, 2023, 8(1): 184–191. [doi: 10.1109/LRA.2022.3224370]

[24] Mei GF, Huang XS, Liu J, Zhang J, Wu Q. Unsupervised point cloud pre-training via contrasting and clustering. In: Proc. of the 2022 IEEE Int'l Conf. on Image Processing. Bordeaux: IEEE, 2022. 66–70. [doi: 10.1109/ICIP46576.2022.9897388]

[25] Chen XL, He KM. Exploring simple siamese representation learning. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15750–15758. [doi: 10.1109/CVPR46437.2021.01549]

[26] Chen YJ, Nießner M, Dai A. 4DContrast: Contrastive learning with dynamic correspondences for 3D scene understanding. In: Proc. of

the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 543–560. [doi: 10.1007/978-3-031-19824-3_32]

[27] Jing LL, Zhang L, Tian YL. Self-supervised feature learning by cross-modality and cross-view correspondences. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Nashville: IEEE, 2021. 1581–1591. [doi: 10.1109/CVPRW53098.2021.00174]

[28] Zhang RR, Wang LH, Qiao Y, Gao P, Li HS. Learning 3D representations from 2D pre-trained models via image-to-point masked autoencoders. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 21769–21780. [doi: 10.1109/CVPR52729.2023.02085]

[29] Wang ZY, Yu XM, Rao YM, Zhou J, Lu JW. P2P: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans, 2022. 14388–14402.

[30] Dong RP, Qi ZK, Zhang LF, Zhang JB, Sun JJ, Ge Z, Yi L, Ma KS. Autoencoders as cross-modal teachers: Can pretrained 2D image Transformers help 3D representation learning? In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023.

[31] Zhang RR, Guo ZY, Zhang W, Li KC, Miao XP, Cui B, Qiao Y, Gao P, Li HS. PointCLIP: Point cloud understanding by CLIP. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8552–8562. [doi: 10.1109/CVPR52688.2022.00836]

[32] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.

[33] Zhu XY, Zhang RR, He BW, Guo ZY, Zeng ZY, Qin ZP, Zhang SH, Gao P. PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 2639–2650. [doi: 10.1109/ICCV51070.2023.00249]

[34] Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 159.

[35] Qi ZK, Dong RP, Fan GF, Ge Z, Zhang XY, Ma KS, Yi L. Contrast with reconstruct: Contrastive 3D representation learning guided by generative pretraining. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: JMLR.org, 2023. 1171.

[36] Chen HN, Zhu YY, Zhao JQ, Tian Q. 3D shape recognition based on multimodal relation modeling. Ruan Jian Xue Bao/Journal of Software, 2024, 35(5): 2208–2219 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7026.htm [doi: 10.13328/j.cnki.jos.007026]

[37] Xie CL, Wang CX, Zhang B, Yang H, Chen D, Wen F. Style-based point generator with adversarial rendering for point cloud completion. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 4619–4628. [doi: 10.1109/CVPR46437.2021.00459]

[38] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4401–4410. [doi: 10.1109/CVPR.2019.00453]

[39] Zhou LQ, Du YL, Wu JJ. 3D shape generation and completion through point-voxel diffusion. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 5826–5835. [doi: 10.1109/ICCV48922.2021.00577]

[40] Pan L, Chen XY, Cai ZG, Zhang JZ, Zhao HY, Yi S, Liu ZW. Variational relational point completion network. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8524–8533. [doi: 10.1109/CVPR46437.2021.00842]

[41] Bardes A, Ponce J, LeCun Y. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[42] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[43] Ma X, Qin C, You HX, Ran HX, Fu Y. Rethinking network design and local geometry in point cloud: A simple residual MLP framework. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[44] Qian GC, Li YC, Peng HW, Mai JJ, Hammoud H, Elhoseiny M, Ghanem B. PointNeXt: Revisiting PointNet++ with improved training and scaling strategies. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans, 2022. 23192–23204.

[45] Sanghi A. Info3D: Representation learning on 3D objects using mutual information maximization and contrastive learning. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 626–642. [doi: 10.1007/978-3-030-58526-6_37]

[46] Gadelha M, RoyChowdhury A, Sharma G, Kalogerakis E, Cao LL, Learned-Miller E, Wang R, Maji S. Label-efficient learning on point clouds using approximate convex decompositions. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 473–491. [doi: 10.1007/978-3-030-58607-2_28]

[47] Cheng HZ, Han X, Shi PC, Zhu JH, Li ZY. Multi-trusted cross-modal information bottleneck for 3D self-supervised representation learning. Knowledge-based Systems, 2024, 283: 111217. [doi: 10.1016/j.knosys.2023.111217]

[48] Wu Y, Liu JM, Gong MG, Gong PR, Fan XL, Qin AK, Miao QG, Ma WP. Self-supervised intra-modal and cross-modal contrastive learning for point cloud understanding. IEEE Trans. on Multimedia, 2024, 26: 1626–1638. [doi: 10.1109/TMM.2023.3284591]

[49] Anvekar T, Bazazian D. GPr-Net: Geometric prototypical network for point cloud few-shot learning. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Vancouver: IEEE, 2023. 4178–4187. [doi: 10.1109/CVPRW59228.2023.00440]

[50] Sharma C, Kaul M. Self-supervised few-shot learning on point clouds. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. 2020. 7212–7221.

[51] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4080–4090.

[52] Huang TY, Dong BW, Yang YH, Huang XS, Lau RWH, Ouyang WL, Zuo WM. CLIP2Point: Transfer CLIP to point cloud classification with image-depth pre-training. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 22157–22167. [doi: 10.1109/ICCV51070.2023.02025]

[53] Choy C, Gwak J, Savarese S. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3075–3084. [doi: 10.1109/CVPR.2019.00319]

[54] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[55] Liu Z, Mao HZ, Wu CY, Feichtenhofer C, Darrell T, Xie SN. A ConvNet for the 2020s. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11976–11986. [doi: 10.1109/CVPR52688.2022.01167]

[56] Kirillov A, Girshick R, He KM, Dollár P. Panoptic feature pyramid networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6399–6408. [doi: 10.1109/CVPR.2019.00656]

[57] Xiao TT, Liu YC, Zhou BL, Jiang YN, Sun J. Unified perceptual parsing for scene understanding. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 418–434. [doi: 10.1007/978-3-030-01228-1_26]

附中文参考文献:

[1] 朱向雷, 王海弛, 尤翰墨, 张蔚珩, 张颖异, 刘爽, 陈俊洁, 王赞, 李克秋. 自动驾驶智能系统测试研究综述. 软件学报, 2021, 32(7): 2056–2077. http://www.jos.org.cn/1000-9825/6266.htm [doi: 10.13328/j.cnki.jos.006266]

[2] 闫涛, 高浩轩, 张江峰, 钱宇华, 张临垣. 分组并行的轻量化实时微观三维形貌重建方法. 软件学报, 2024, 35(4): 1717–1731. http://www.jos.org.cn/1000-9825/7013.htm [doi: 10.13328/j.cnki.jos.007013]

[36] 陈浩楠, 朱映映, 赵骏骐, 田奇. 基于多模态关系建模的三维形状识别方法. 软件学报, 2024, 35(5): 2208–2219. http://www.jos.org.cn/1000-9825/7026.htm [doi: 10.13328/j.cnki.jos.007026]

程浩喆(1997－), 男, 博士生, 主要研究领域为深度学习, 三维计算机视觉.

胡乃文(2000－), 男, 硕士生, 主要研究领域为深度学习, 三维计算机视觉.

祝继华(1982－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机视觉, 机器学习.

谢奕凡(2001－), 男, 硕士生, 主要研究领域为深度学习, 三维计算机视觉.

史鹏程(1998－), 男, 硕士生, 主要研究领域为深度学习, 三维视计算机觉.

李仕奇(2000－), 男, 硕士生, 主要研究领域为深度学习, 三维计算机视觉.