Masked Autoencoders for Point Cloud Self-supervised Learning 的复现和改进

摘要

随着基于深度学习的点云理解发展，MAE 预训练方法备受关注，但跨模态方法计算成本高，实际应用更倾向单模态方法。本文深入分析的 Point-MAE 方案，由点云掩蔽和嵌入模块及自动编码器组成，解决了主干架构、位置信息泄漏和信息密度等问题，其简单架构性能优异且启发了多模态统一架构。同时，本文在 Point-MAE 基础上创新，提出新随机掩盖策略、预训练加入旋转操作、改用全局和局部掩蔽重建框架并引入局部增强模块，提升了点云数据特征提取和泛化能力。

关键词：3D 点云；自监督学习；掩蔽自编码器；Point-MAE；深度学习

1 引言

点云作为三维物体的有效表示，因其丰富的几何、形状和结构细节而被广泛应用于自动驾驶、机器人等广泛的应用领域。近年来，随着基于深度学习的点云理解的快速发展 [9]，基于掩蔽自编码器 (MAE) 的预训练方法受到了广泛关注 [3]，旨在从大量未标记点云中学习潜在的 3D 表示，可分为两类，即单模态和跨模态方法 [6]。其中，跨模态 MAE 方法利用其他模态的见解，通过获取整体 3D 表示取得了显著的性能。然而，这些方法严重依赖于从大量成对图像或文本中迁移知识，而这些知识在实践中往往是不可用的。具体而言，它们利用预先训练的图像或语言模型来提取跨模态知识，并使用投影或知识提炼等技术进行跨模态知识迁移。这种复杂的操作需要大量的计算成本，因此阻碍了它们在实践中的应用。与单模态相比，跨模态方法在 ScanObjectNN 上获得了 5% 的性能提升 [10]，同时需要 5 倍的预训练参数。

因此，由于其简单性和效率，实际应用中更倾向于仅以点云为输入的单模态方法。在这种背景下，本文通过研究一种掩蔽自编码器方案（Point-MAE）来进一步解决这些问题。如图 1 所示，Point-MAE 主要由点云掩蔽和嵌入模块以及自动编码器组成。输入点云被分成不规则的点斑块，这些点斑块以高比率随机掩蔽以减少数据冗余。然后，自动编码器从未掩蔽的点斑块中学习高级潜在特征，旨在在坐标空间中重建掩蔽的点斑块。同时，自动编码器的主干完全由标准 Transformer 块构建，采用非对称编码器-解码器结构 [7]。编码器仅处理未掩蔽的点斑块。然后，将编码标记和掩蔽标记作为输入，带有简单预测头的轻量级解码器重建掩蔽的点斑块。与从编码器输入处理掩码标记相比，将掩码标记转移到轻量级解码器可以节省大量计算资源，更重要的是，避免位置信息的早期泄露。Point-MAE 的关键技术总结如下：

- 提出了一种用于点云自监督学习的新型掩蔽自编码器方案，解决了包括主干架构、位置信息的早期泄漏和点云的信息密度等关键问题。
- 通过多个实验表明完全基于标准 Transformers 的简单架构可以超越监督学习中的专用 Transformer 模型。
- 从多模态学习的角度来看，Point-MAE 的提出启发了语言和特别是图像的统一架构，例如掩蔽自编码器，当配备特定于模态的嵌入模块和特定于任务的输出头时，也适用于点云。

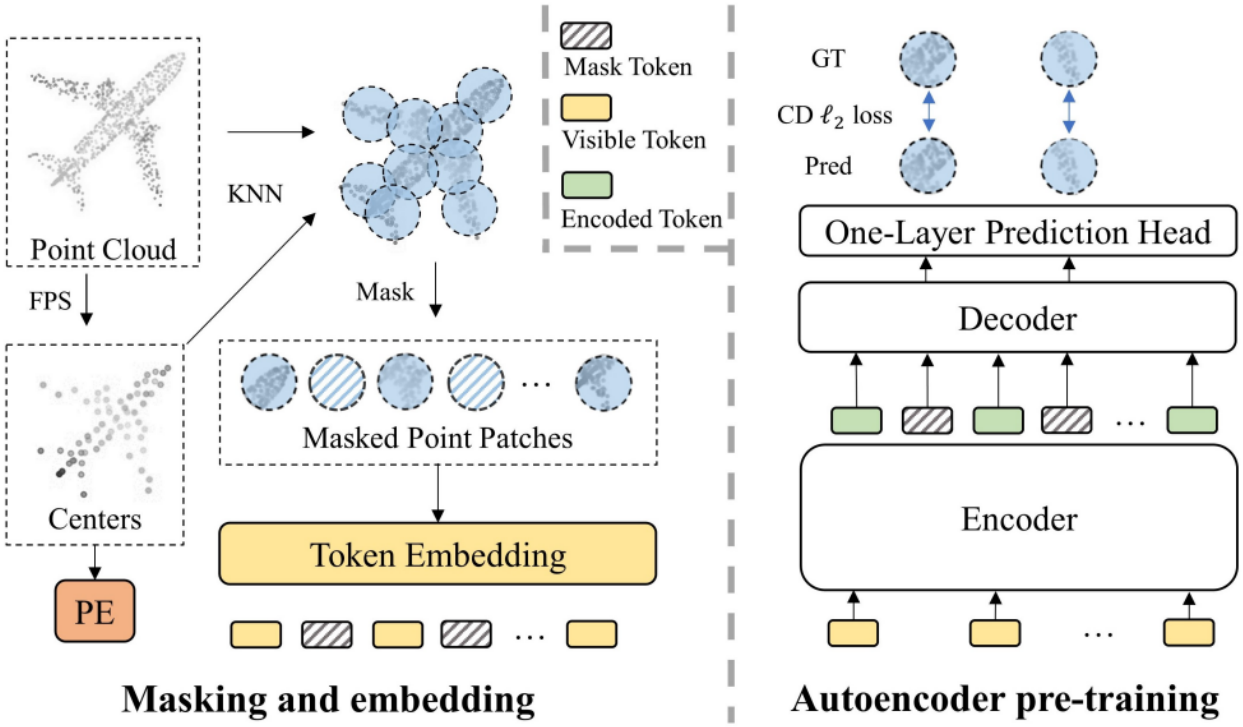

图 1. **Point-MAE 的整体方案图。**左侧展示了掩码和嵌入过程，输入云被分成点块，随机掩码然后嵌入。右侧展示了自动编码器预训练，编码器只处理可见的标记。掩码标记被添加到解码器的输入序列中以重建掩码点块。

2 相关工作

2.1 点云的自监督学习

点云的自监督学习已经被广泛探索和研究，并产生了许多出色的方法。这些方法使用精心设计的借口任务（即屏蔽输入标记并预先训练模型以预测原始图像）来训练模型，而不是使用标记数据。常见的借口任务包括点云重建工作，例如输入重建 [18]、局部到全局重建 [8]、异常部位修复和遮挡完成 [15]。此外，人们还探索了各种对比学习方法，通过区分预定义的正样本和负样本来学习点云表示。自监督学习还包括其他借口任务例如方向估计、混合 [13] 和解缠技术等。

最近，受到 BERT 模型和 MAE 模型的启发，许多基于 Transformer 的掩蔽点建模（MPM）方法被提出，并且在下游任务中显著优于其他方法。Point-BERT 模型将 BERT 预训练方案引入点云 [19]，其训练模型的任务是预测由 dVAE 生成的离散的 token [12]。Point-MAE 中提出了一种自监督的借口任务，基于已知 token 重建被遮蔽的输入点块。此外，Point-M2AE 采用了金字塔编码器和解码器设计 [21]，以分层方式进行被遮蔽的点云重建。另一方面，ReCon++ [11] 通过引入来自其他模态（如图像和自然语言）的预训练模型改进了 3D 掩蔽建模。而 Point-GPT 将点块按有序顺序排列 [2]，并将生成预训练 Transformer (GPT) 的概念扩展到点云。尽管取得了显著的成果，但这些 MPM 方法对旋转很敏感，限制了它们的可扩展性和通用性。

2.2 自动编码器

自动编码器由编码器和解码器构成，具体而言，编码器负责将输入编码为高级潜在特征，而解码器将潜在特征进行解码，并且重建输入。而自动编码器的优化目标在于使重建的数据尽可能类似于原始输入，例如图像像素空间中的均方误差损失。值得注意的是，本文所提出的方法属于去噪的自动编码器类，主要思想是通过引入输入噪声来增强模型的鲁棒性。本文通过相同的原理，使用自动掩蔽操作引入输入噪声，随机掩蔽输入中的标记，然后通过应用自动编码器来预测与掩蔽标记相对应的图像区域，进一步地，将这种自动编码器应用到点云数据上。在计算机视觉中，MAE 和 SimMIM [17] 都提出了类似的掩蔽图像建模，随机屏蔽输入图像块，然后应用自动编码器预测在像素空间中被屏蔽的块。

2.3 Transformers

Transformers 通过自注意机制对输入的全局依赖关系进行建模 [14]，并在 NLP 领域中占据了主导地位 [1]。从第一个应用在计算机视觉的 Transformers 开始 [4]，Transformers 架构在计算机视觉领域就变得受欢迎起来了 [20]。然而，作为掩蔽自动编码器的骨干，Transformers 架构用于点云表示学习的发展较少并且值得进一步取扩展。PCT 模型设计了一个专用的输入嵌入层 [5]，并修改了 Transformer 层中的自注意机制。PointTransformer [22] 也修改了 Transformer 层，并在 Transformer 块之间使用了额外的聚合操作。而 Point-BERT 引入了标准的 Transformer 架构，但需要 DGCNN 来协助预训练 [16]。因此，与以前的工作不同，本文提出了一种完全基于标准 Transformers 的架构。

3 本文方法

本节将介绍掩码自动编码器方法的各个组成模块。

3.1 点云掩蔽和嵌入

与计算机视觉中的图像可以自然地划分为规则的块不同，点云由三维空间中无序的点组成。根据其特性，通过三个阶段处理输入点云：点块生成、掩蔽和嵌入。

点块的生成：通过最远点采样（FPS）和 K 最近邻（KNN）算法将输入点云划分为不规则点块（可能重叠）。形式上，给定一个包含 p 个点 $X^i \in \mathbb{R}^{p\times 3}$ 的输入点云，应用 FPS 为点块中的中心 CT 采样 n 个点。基于中心点，KNN 从输入中选择 k 个最近点作为相应点块 P,

计算公式总结如下，

$$CT = FPS(X^i), \quad CT \in \mathbb{R}^{n\times 3}; \tag{1}$$

$$P = KNN(X^i, CT), \quad P \in \mathbb{R}^{n\times k\times 3}. \tag{2}$$

在这种表达下，每个点块 P 中都由相对于中心点的归一化坐标表示。

掩码：考虑到点块与点块之间可能会出现重叠，因此，需要对点块分别进行掩蔽，以保持每个点块中的信息完整。掩蔽率为 m，掩蔽块集表示为 $P_{gt} \in \mathbb{R}^{mn\times k\times 3}$ ，在计算重建 损失时用作基本事实。同时注意到不同的掩蔽策略会得到不同的效果，而（$60\% - 80\%$）的随机掩蔽对最终得出的效果更好。

嵌入：对于每个被掩蔽的点块的嵌入，将其替换为一个共享加权的可学习掩码标记。同时，将全套掩码标记表示为 $T_m \in \mathbb{R}^{mn\times C}$ ，其中 C 是嵌入维度。传统方法上，会将可见的点块展平并用可训练的线性投影嵌入。然而，线性嵌入不符合置换不变性原则，本文方法采用更合理的嵌入方法。即通过一个轻量的 PointNet 后再进行嵌入，它主要由 MLP 和最大池化层组成。可见点块 $P_v \in \mathbb{R}^{(1-m)n\times k\times 3}$ 嵌入到可见标记 T_v 中，

$$T_v = PointNet(P_v), \quad T_v \in \mathbb{R}^{(1-m)n\times C}. \tag{3}$$

考虑到点块以归一化坐标表示，向嵌入标记提供中心的位置信息至关重要。而位置嵌入 (PE) 的一种简单方法是使用可学习的 MLP 将中心坐标映射到嵌入维度，所用的方法在自动编码器中分别对编码器和解码器使用两个单独的 PE。

3.2 自动编码器

自动编码器的主干基于 Transformer 而制定的，采用非对称编码器-解码器设计，而最后一层采用简单的预测头来实现目标重建。

编码-解码器：编码器由标准 Transformer 块组成，仅编码可见的标记 T_v 而不编码掩码的标记 T_m 。编码的标记表示为 T_e 。此外，位置嵌入被添加到每个 Transformer 块，提供位置信息。所使用的解码器与编码器类似，但包含的 Transformer 块较少，解码器将编码标记 T_e 和掩码标记 T_m 作为输入，将一整套位置嵌入添加到每个 Transformer 块，为所有标记提供位置信息。处理后，解码器仅输出解码的掩码标记 H_m ，并将其馈送到以下预测头。编码器-解码器结构公式表达为如下，

$$T_e = Encoder(T_v), \quad T_e \in \mathbb{R}^{(1-m)n\times C}; \tag{4}$$

$$H_m = Decoder(concat(T_e, T_m)), \quad H_m \in \mathbb{R}^{mn\times C}. \tag{5}$$

在这一结构中，可以将掩码标记移至轻量级解码器，而不是从编码器的输入中处理它们。这种设计有两个好处，首先，由于使用的是高掩码率，移位掩码标记会显著减少编码器的输入标记数量。因此，由于 Transformers 的二次复杂度，可以进一步节省计算资源。更重要的是，将掩码标记移至解码器可以避免位置信息过早泄露给编码器，从而使编码器更好地学习潜在特征。

预测头：作为主干的最后一层，预测头旨在重建坐标空间中的掩蔽点块。预测头从解码器获取输出 H_m，将其投影到一个向量，该向量的维数与点块中的坐标总数相同，然后进行重塑操作，获得预测的掩蔽点块 P_{pre}，

$$P_{pre} = Reshape(FC(H_m)), \quad P_{pre} \in \mathbb{R}^{mn\times k\times 3}. \tag{6}$$

3.3 重建损失函数定义

为了给自监督学习一个借口任务进行预训练，本文通过设定一个重建任务使得模型从点云中学习到更多高级潜在特征。重建目标是恢复每个掩蔽点块中点的坐标。给定预测点块 P_{pre} 和真实点 P_{gt}，我们通过 l_2 倒角距离计算重建损失，

$$L = \frac{1}{|P_{pre}|} \sum_{a \in P_{pre}} \min_{b \in P_{gt}} \|a - b\|_2^2 + \frac{1}{|P_{gt}|} \sum_{b \in P_{gt}} \min_{a \in P_{pre}} \|a - b\|_2^2 . \tag{7}$$

4 复现细节

4.1 与已有开源代码对比

我们的代码主要基于 Point-MAE 实现，并且做了相应的改进。

首先，由于点云数据信息密度不同，且分布不均匀，即如果信息密度较高的区域被掩盖，那么就难以恢复，因此我们提出一种更加合理的随机掩盖策略，在每个块中随机掩盖掉部分位于中部的点而非直接掩盖掉随机块。进一步地，点云数据对于旋转较为敏感，因此我们在自监督重建中加入了旋转等操作，促进模型更好地学习到点云数据潜在的旋转不变的特性。最后，我们在预训练时运用一种全局和局部掩蔽重建的方法来更好地获得点云数据潜在的特征，同时在编码器中引入局部增强模块增强数据特征感知能力。

4.2 实验环境搭建

本文实验均在 Ubuntu 20.04.1 LTS 系统下进行，使用 Python 3.8 作为编程语言，使用 Pytorch 1.11.0 作为深度学习框架。关于硬件环境，本文使用 Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz 和 NVIDIA GeForce GTX 3090 Ti GPU 进行实验。

4.3 创新点

对于复现的 Point-MAE 模型，本文对其进行了创新，创新点主要包括以下三点：

- 1. 采用新的随机掩盖策略，从每个块中随机掩盖掉位于中部的点。
- 2. 在预训练时加入了旋转操作，以促进模型更好地学习到点云数据潜在的旋转不变的特性。
- 3. 改用一种全局和局部掩蔽重建的框架，同时在编码器中引入局部增强模块增强数据特征感知能力。

5 实验结果分析

本部分对实验结果进行展示和分析，包括实验的复现结果以及添加创新模块的结果。

5.1 预训练阶段

我们在原先的 Point-MAE 模型上做了改进后的模型来做了相应的预训练。预训练是在 ShapeNet 数据集上进行的，包含 51,300 个 3D 模型一共 55 个类别。在预训练期间，我们应用了标准随机放缩和随机平移进行数据增强，在此基础上，我们加入了旋转操作，以加强模型更好地学习点云数据潜在的旋转不变性。注意到的是，在进行重建的预训练时，我们使用了 AdamW 优化器和余弦学习率衰减，并且在所提出的全局和局部掩蔽重建的框架上进行预训练。得到的重建结果如图 5.1 所示。

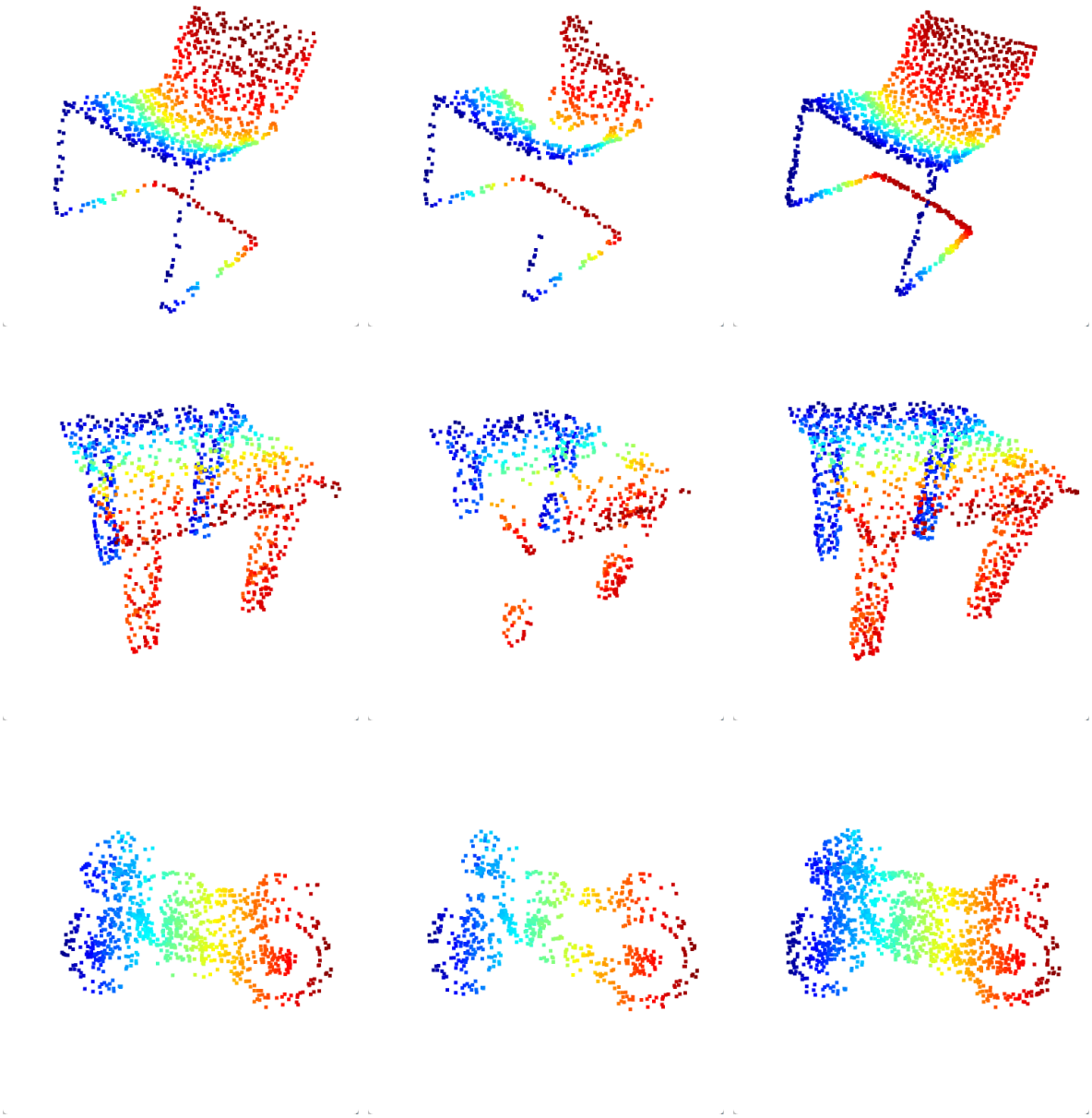

图 2. 原始 3D 图像　　图 3. 掩蔽 60% 的 3D 图像　　图 4. 重建后的 3D 图像

为了证明我们方法的有效性，我们在图 5.1 中可视化了 ShapeNet 验证集上的重建结果。该模型使用 60% 的掩蔽率进行预训练，但它能够重建具有不同掩蔽率的输入。由于我们的模型可以很好地学习高级潜在特征，因此可以提高模型的泛化能力。

5.2 下游任务

物体分类：为了进一步证明我们方法的有效性和优异性，我们在 ModelNet40 上评估了我们预先训练的模型以进行物体分类。ModelNet40 由 12,311 个干净的 3D CAD 模型组成，涵盖 40 个物体类别。将 ModelNet40 拆分为 9843 个训练集实例和 2468 个测试集实例。在训练期间，应用标准随机缩放和随机平移进行数据增强，同时，应用了所提出新的随机掩蔽策略。

实验结果如表 5.2 所示，用我们所提出的方法实现了 94.1% 的准确率，相较于 Point-MAE 的 93.2% 而言，提高了 0.9% 的准确率，提高了对 3D 图像的特征提取能力。值得注意的是，与其他自监督学习方法相比，我们的方法有着更优异的性能，这也可以验证出我们方法的有效性。

Self-supervised methods	Accuracy
OcCo	92.2%
STRL	92.4%
IAE	92.9%
[ST]Transformer-OcCo	92.0%
[ST]Point-BERT	92.9%
[ST]Point-MAE	93.2%
Our method	**94.1%**

表 1. 不同模型在 ModelNet40 数据集上的分类效果

少样本学习：为了进一步验证方法的有效性，我们在 ModelNet40 上进行了小样本学习实验，采用 n 路，m 折的样本设置，其中 n 是从数据集中随机选择的类的数量，m 是每个类随机抽取的对象数量。我们使用上面提到的 $n \times m$ 个对象进行训练。在测试期间，我们从 n 个类中随机抽取 20 个未见过的对象进行评估。具体设置如表 5.2 所示，我们对每种设置进行 10 次独立实验，并报告平均准确率和标准差。

Methods	5-way,10-shot	5-way,20-shot	10-way,10-shot	10-way,20-shot
Point-BERT	95.0 ± 2.0	97.0 ± 1.0	91.0 ± 4.0	92.0 ± 3.0
Point-MAE	96.0 ± 2.0	98.0 ± 1.0	92.0 ± 3.0	95.0 ± 2.0
Our method	97.0 ± 1.0	98.0 ± 2.0	93.0 ± 3.0	96.0 ± 2.0

表 2. ModelNet40 上的少样本对象分类效果

少样本学习实验如表 5.2 所示，我们所提出来的方法有着更高的精度和更小的偏差，此外，尽管 Point-MAE 对比其他模型有着更好的性能，然而我们的方法有着比 Point-MAE 更高的精度和更小的标准差，进一步展示我们方法的有效性和优异性。

物体分割：另一方面，Point-MAE 在分割的下游任务中也表现优异，同样的，我们将我们提出的方法也在分割任务上进行实验。在 ShapeNetPart 数据集上做相应的实验，其中数据集包含 x16,881 个对象，涵盖 16 个类别，在初始化时，设置每个对象采样 2048 个点作为输入，从而产生 128 个点块。评估时使用每个类别的平均 IoU，即 mIoUc(%)。

Methods	$mIoU_I$	aero	bag	cap	car	chair	e-phone	guitar	knife
		lamp	laptop	motor	mug	pistol	rocket	s-board	table
Point-BERT	84.1	84.2	84.0	88.4	79.1	91.0	80.9	91.5	87.6
		84.6	95.6	75.2	94.7	84.1	61.9	75.1	79.8
Point-MAE	85.6	84.6	84.1	88.9	80.6	91.1	81.1	92.1	87.1
		84.7	96.1	75.6	95.1	84.4	62.4	75.6	80.5
Our method	**86.7**	85.1	85.7	89.1	80.1	92.4	81.0	92.8	88.6
		83.7	96.1	75.6	95.1	85.8	64.7	75.1	83.3

表 3. ShapeNetPart 数据集上的分割效果

如表 5.2 所示，我们将 Point-MAE 和我们提出的方法在 ShapeNetPart 数据集上进行实验，得到的相应的结果。所提出的方法实现了 86.7% 的 mIoU，比 Point-MAE 所展示的结果更好，高了 1.1%。此外，值得注意的是，相较于其他模型，所提出的方法在大部分类别上都有着较高的分割性能，只有在少部分（如汽车 car 类别）表现不如 Point-MAE。因此，也可以证明我们所提出的方法有着更优异的结果。

6 总结与展望

在本文中，我们介绍了点云自监督学习的新型掩蔽自动编码器方案，称为 Point-MAE。所用的 Point-MAE 简洁高效，仅根据点云的属性进行了少量修改。同时，Point-MAE 的有效性和高泛化能力在各种任务上得到了验证，包括对象分类、少样本学习和物体分割，并优于所有其他自监督学习方法。同时，本文在 Point-MAE 复现的基础上进行了创新，从随机掩盖策略出发，先从不同的点块进行相对随机的掩盖，此外，在预训练阶段，加入旋转操作，使得模型更好学习到点云数据潜在的旋转不变特性，在此基础上，我们改用了一种全局和局部掩蔽重建的框架，同时在编码器中引入局部增强模块以增强数据特征感知能力。实验结果表明，我们所做出来的创新方法可以有效提升点云数据的特征提取能力和泛化能力。另一方面，所用的方法的性能在 Point-MAE 的基础上得到了一定的提升，表现了方法的可行性和有效性。

相对的，方法也存在一定的不足，首先是在计算资源上，所用的双框架（全局和局部掩蔽）对计算资源要求更高，并且占用较多的计算内容和需要较长的训练时间。此外，需要进一步提升点云数据的特征提取能力，在做了相应的改进之后，模型的精度和 IoU 提升有限，需要进一步探索更优越的模型。未来，需要进一步地扩展模型框架，将全局信息和局部信息进行结合，不仅在预训练阶段，也需要在编码解码阶段对特征提取进行进一步加强，同时，加强模型在预训练阶段对点云数据其他潜在特征的理解能力。最终，在这些基础上，可以将模型应用在更为广阔的场景上（如医学图像等）。

参考文献

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Lan-

guage models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*, 2022.

[4] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021.

[6] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzhi Li, and Pheng-Ann Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. *arXiv preprint arXiv:2302.14007*, 2023.

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[8] Xinhai Liu, Zhizhong Han, Xin Wen, Yu-Shen Liu, and Matthias Zwicker. L2g auto-encoder: Understanding point clouds by local-to-global reconstruction with hierarchical self-attention. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 989–997, 2019.

[9] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[10] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, pages 28223–28243. PMLR, 2023.

[11] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision*, pages 214–238. Springer, 2025.

[12] Jason Tyler Rolfe. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*, 2016.

[13] Chao Sun, Zhedong Zheng, Xiaohan Wang, Mingliang Xu, and Yi Yang. Self-supervised point cloud representation learning via separating mixed shapes. *IEEE Transactions on Multimedia*, 25:6207–6218, 2022.

[14] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[15] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9782–9792, 2021.

[16] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019.

[17] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.

[18] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018.

[19] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022.

[20] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021.

[21] Zhiyuan Zhang, Binh-Son Hua, and Sai-Kit Yeung. Riconv++: Effective rotation invariant convolutions for 3d point clouds deep learning. *International Journal of Computer Vision*, 130(5):1228–1243, 2022.

[22] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.