# 混合 Transformer-Mamba 模型用于 3D 语义分割

Xinyu Wang, Jinghua Hou, Zhe Liu, Yingying Zhu*

Abstract— 基于 Transformer 的方法通过其强大的注意力机制在 3D 语义分割中展示了卓越的能力，但其平方复杂度限制了其在大规模点云中对长程依赖关系的建模。虽然最近基于 Mamba 的方法提供了线性复杂度的高效处理，但在提取 3D 特征时，它们在特征表示方面显得局促。然而，有效结合这些互补优势在该领域仍然是一个未解决的挑战。在本文中，我们提出了 HybridTM ，这是第一个将 Transformer 和 Mamba 结合的 3D 语义分割混合架构。此外，我们提出了内层混合策略，在更细的粒度上结合了注意力和 Mamba，能够同时捕捉长程依赖和细粒度的局部特征。大量实验展示了我们的方法 HybridTM 在多种室内和室外数据集上的有效性和泛化能力。此外，我们的方法 HybridTM 在 ScanNet、ScanNet200 和 nuScenes 基准上实现了最先进的性能。代码将在 https://github.com/deepinact/HybridTM 上提供。

## I. 介绍

点云的语义分割已经成为 3D 视觉中的一个关键任务，使得系统能够为三维空间中的每个点分配语义标签。这一能力对于诸如自动驾驶 [1], [2], [3], [4] 、机器人导航 [5], [6] 和 3D 场景理解 [7] 等应用而言是基础性的。

最近，基于 transformer 的方法凭借其强大的注意力机制在 3D 语义分割中展示了非凡的能力。然而，由于计算复杂度的平方增长，在点云中为 transformers [8] 实现大感受野是不切实际的。为了解决这一限制，许多研究 [9], [10], [11], [12], [13], [14], [15], [16], [17] 将点云划分为较小的组，并在这些组内应用自注意力。尽管这一策略降低了计算成本，但其本质上限制了对长距离依赖关系的建模能力，从而导致分割性能的不理想。

基于状态空间模型（SSM），Mamba [18] 可以以线性计算复杂度有效地处理大规模点云。尽管 Mamba 在各种二维视觉任务中展示了令人印象深刻的性能 [19], [20], [21], [22], [23] ，但基于 Mamba 的方法 [24], [25], [26] 在提取三维特征时常常难以处理特征表示，特别是在捕捉细粒度局部特征方面，这对于精确分割很重要。

虽然最近的研究探讨了结合注意力机制和 Mamba 的混合架构在 NLP [30] 和 2D 视觉 [31], [32], [33] 任务中的应用，但 3D 语义分割面临新的挑战。与密集的 2D 图像不同，点云本质上是稀疏和不规则的。直接将现有的 2D 视觉 [31], [32], [33] 混合策略应用于序列化的稀疏体素会导致特征退化，从而限制其在 3D 语义分割中的有效性。为了应对这些挑战，我们提出了 HybridTM ，这是一种集成 Transformer 和 Mamba 的用于 3D 语义分割的首个混合架构。此外，我们提出了内部层混合策略，可以通过在更细粒度上结合这些运算符，同时保持捕获细粒度特征和建模长距离依赖的能力。如图 1 所示， HybridTM 在室内和室外数据集上的表现优于现有基于 transformers、3D 稀疏卷积和 Mamba 的方法。

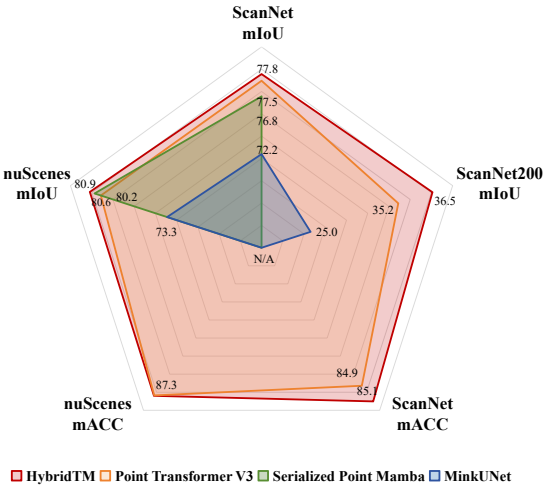


Fig. 1. 对比现有的具有不同算子（例如，Transformer [8] 、3D 稀疏卷积 [27] 和 Mamba [18] ）的代表性 3D 语义分割方法在 ScanNet [28] 、nuScenes [29] 和 ScanNet200 [28] 数据集上的表现。 HybridTM 在室内和室外数据集上均取得了优越的性能。

总而言之，我们的贡献如下。

- 我们提出了 HybridTM ，这是第一个集成 Transformer 和 Mamba 的 3D 语义分割混合架构。HybridTM 能够通过提出的内层混合策略高效捕捉细粒度的局部特征和建模长距离关系。
- 我们在几个室内和室外数据集上进行了实验，以验证 HybridTM 的有效性和泛化能力。 HybridTM 在 Scannet [28] 、Scannet200 [28] 和 nuScenes [29] 上实现了最先进的（SOTA）性能。

## II. 相关工作

### A. 三维语义分割中的 Transformer

Transformer 在许多视觉任务中被证明是一种成功的架构。因此，许多工作尝试在 3D 语义分割任务中采用注意力操作符。点云 Transformer (PCT) [34] 是开创性的工作，它用偏移注意力替换了传统的自注意力，以更好地捕捉点云中的几何关系。然而，Transformers 的二次计算复杂性使得直接对整个点云应用注意力操作符变得不切实际。为了解决这个问题，一些方法 [9], [10], [11], [12], [13], [14], [15], [16], [17], [35] 将点云划分为小群体，

This work was supported by the NSFC (Grant 62225603).
This work was also supported by the NSFC (Grant U2341227).
Xinyu Wang and Jinghua Hou contributed equally to this work.
*Corresponding author: Yingying Zhu.
Xinyu Wang is with School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), Wuhan 430074, China (e-mail: xinyuwang@hust.edu.cn)
Jinghua Hou is with School of Electronic Information and Communications, Huazhong University of Science and Technology (HUST), Wuhan 430074, China (e-mail: jhhou@hust.edu.cn)
Zhe Liu is with the Department of Computer Science of The University of Hong Kong (HKU), Pokfulam 999077, Hong Kong (e-mail: zheliu12@hku.hk)
Yingying Zhu is with School of Computer Science and Technology, Huazhong University of Science and Technology (HUST), Wuhan 430074, China (e-mail: yyzhu@hust.edu.cn)

www.xueshuxiangzi.com

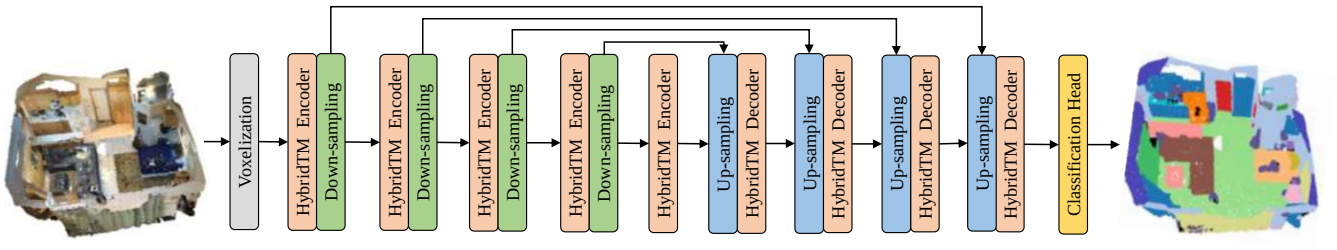


Fig. 2. HybridTM 的示意图，由 HybridTM 编码器、 HybridTM 解码器、下采样、上采样和一个分类头组成。 HybridTM 将输入的点云体素化，并采用类似 UNet 的架构来提取多尺度特征。最后，提取的特征被输入到分类头中用于 3D 语义分割。

从而降低计算成本。例如，Point Transformer [9] 通过向量注意力改进了自注意力机制，带来了更强的特征表示。同样，OctFormer [12] 使用八叉树对点云进行划分，并实施膨胀八叉树注意力以减少计算成本。尽管有这些创新，现有的方法往往在有效建模 3D 语义分割任务中固有的长距离依赖性方面表现不佳。Point Transformer v3 [14] 通过扩大接受域和序列化邻域映射实现了更好的性能，从而提高了性能。然而，它仍然无法充分抓住实现最佳 3D 语义分割所需的广泛长距离依赖性。

### B. 3D 视觉中的 Mamba

由于 Mamba 的线性复杂度 [18], [36] ，越来越多的人对将这种架构改造成各种视觉任务中替代 transformers 表现出兴趣 [19], [20], [21], [22], [23], [37], [38], [39] 。在 2D 视觉中的最初探索，特别是 Vision Mamba [19] 和 VMamba [20] ，通过有效地将 2D 图像序列化进行 Mamba 处理，展示了很有希望的结果。

在这些成功之后，研究人员开始研究 Mamba 在 3D 视觉任务中的潜力 [25], [40], [26], [24] 。点云 Mamba [25] 开创了 Mamba 在处理点云应用中的先河，展示了其在捕捉复杂几何关系方面的能力。LION-Mamba [40] 进一步将这一方法扩展到基于 LiDAR 的 3D 目标检测，通过建模长程依赖关系实现了最先进的性能。在 3D 语义分割领域，Serialized Point Mamba [26] 引入了一种线性复杂性的方法，试图在局部特征提取与全局上下文建模之间取得平衡。

然而，尽管这些技术取得了进展，纯粹基于 Mamba 的方法在不同视觉任务中仍难以始终如一地达到最先进的性能。这一限制主要源于 Mamba 在特征表示方面相对于基于 transformer 的架构的固有弱点，这表明需要更复杂的架构设计。

最近的一些努力开始探索结合 Transformer 和 Mamba 的混合结构，以利用每个组件的优势并提高整体性能。在自然语言处理任务中，Jamba 交错使用了 Transformer 和 Mamba 层，实现了较低的计算成本和更好的性能。在二维视觉任务中，最近的方法通过外层混合策略将 Transformer 层与 Mamba 层结合。这种策略通过捕获短程和长程依赖关系来有效增强特征表示，适用于包括通用计算机视觉、骨干预训练和图像生成的各种应用。然而，将这种混合策略扩展到三维语义分割任务面临显著挑战，导致次优性能。因此，在本文中，我们提出了 HybridTM ，这是第一个通过我们提出的内层混合策略有效结合 Transformer 和 Mamba 的混合架构，在三维语义分割中实现了卓越的性能。

## III. 方法

在这项工作中，我们提出了 HybridTM ，这是一种用于 3D 语义分割的首个混合架构，通过我们提出的内部层混合策略，将 Transformer [8] 和 Mamba [18] 层协同结合。我们的架构能够同时提取精细的局部特征并建模长距离依赖，从而实现更准确的 3D 语义分割。如图 2 所示， HybridTM 遵循类似于 UNet 的设计，其编码器和解码器建立在我们的混合层之上。该流程首先将输入的点云体素化，通过混合层处理以提取 3D 特征，并最终通过分类头生成每个点的语义预测。在后续章节中，我们详细介绍了混合层的关键组件。

### A. 混合层

混合层结合了注意力和 Mamba 算子，捕捉细粒度和全局特征。如图 3 所示，我们首先将输入体素分成几个等大小的小组，以便注意力层提取细粒度特征。然后我们将体素分成几个等大小的大组，以便后续的 Mamba 层提取全局特征。此外，我们使用 xCPE [14] 进一步提高性能。接下来，我们介绍混合层的每一步。

1) *注意力层*：注意力层用于在局部区域提取细粒度特征。给定混合层输入体素的特征 $F \in \mathbb{R}^{N\times C}$ 和组大小 $L$ 。我们首先通过空间填充曲线 [14] 将 $F$ 划分为不重叠的 3D 等大小小组 $F_g = \{F_i, i = 1,2,3...,\lceil \frac{N}{L} \rceil\}$ 。然后，我们应用注意力算子从这些组中提取特征。最后，我们将这些组恢复为输入体素的形状，以用于后续的 Mamba 层。此过程可以表示为：

$$\begin{aligned} F_g &= \text{Parition}(F, L), \\ F_g^{'} &= F_g + \text{MSA}(F_g), \\ F^{'} &= \text{Restore}(F_g^{'}, L), \end{aligned} \tag{1}$$

，其中 Parition 、Restore 和 *MSA* 分别表示分区、恢复和多头注意力算子。

2) *Mamba 层*：Mamba 层扩展了特征学习，以便在更大的感受野中捕捉远程特征交互。我们将增强的特征 $F^{'} \in \mathbb{R}^{N\times C}$ 分割成大小为 $K$ 的较大组，以进行全局上下文建模。此外，我们采用双向 Mamba 架构来捕捉双向的依赖性，增强模型理解远程关系的能力。经过此增强后，处理过的特征会恢复为与注意力层操作相同的形式。此过程可以表示为：

$$\begin{aligned} F^{'} &= \text{Parition}(F^{'}, K), \\ F_g^{''} &= F_g^{'} + \text{BiMamba}(F_g^{'}), \\ F^{''} &= \text{Restore}(F_g^{''}, K), \end{aligned} \tag{2}$$

，其中 BiMamba 表示双向 Mamba。

www.xueshuxiangzi.com

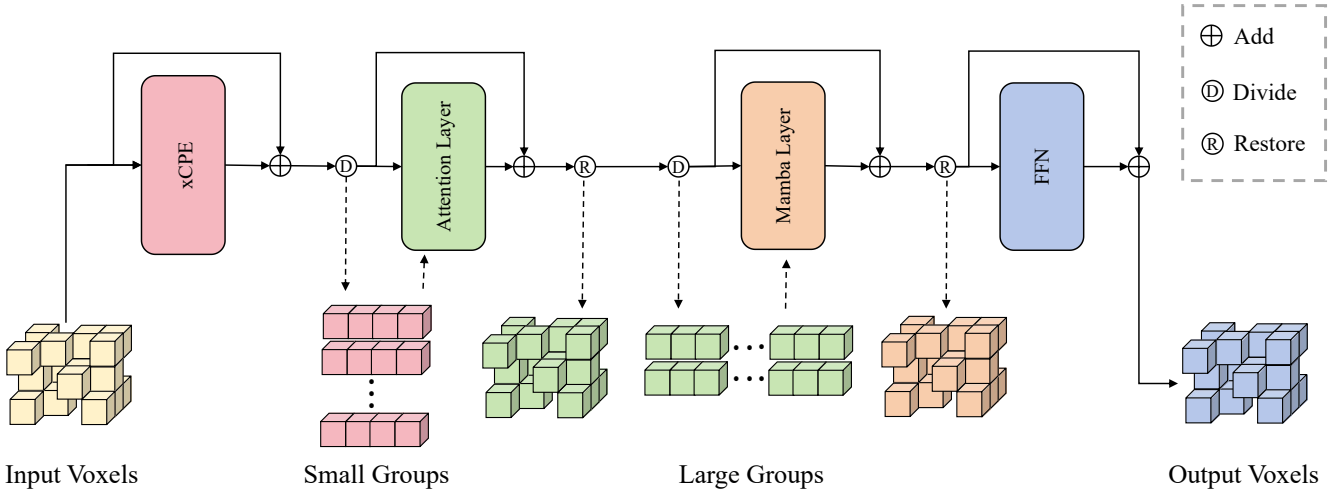


Fig. 3. 混合层的示意。混合层包含一个 xCPE、一个注意力层、一个 Mamba 层和一个 FFN 层。首先，我们将 xCPE 增强的体素划分为若干小组，以便注意力层提取细粒度特征。然后，我们将注意力层的输出体素恢复到原始形状。接着，我们将恢复的体素划分为若干大组，以便 Mamba 层提取全局特征。最后，我们采用 FFN 层增强融合特征以获得最终的输出体素。

3) 内层混合策略：我们的内层混合策略可以有效地将 Transformer 和 Mamba 结合用于 3D 语义分割。该设计基于两个观察结果：(1) 注意力操作可以有效地捕捉局部区域内体素之间的相对空间关系。(2) Mamba 能够以线性复杂度高效地建模长程依赖关系，但在提取 3D 特征时在精确特征表达上有所不足。因此，我们认为 Mamba 需要注意力产生的高质量局部特征来实现卓越的性能。

在每个混合层中，我们的策略是对注意力层和 Mamba 层进行排序，以最大化它们的互补优势。注意力层首先处理输入特征以提取丰富的局部特征，这些特征然后被送入 Mamba 层以实现高效的全局上下文建模。最终的前馈网络（FFN）实现了局部和全局信息的充分整合。通过 xCPE 模块进行的位置编码补充了这一过程，确保在特征提取过程中空间意识的保持。

该设计的序列确保 Mamba 层从注意力层接收到高质量的特征表示，而 FFN 则能够有效融合局部和全局特征。

### B. 不同策略的讨论

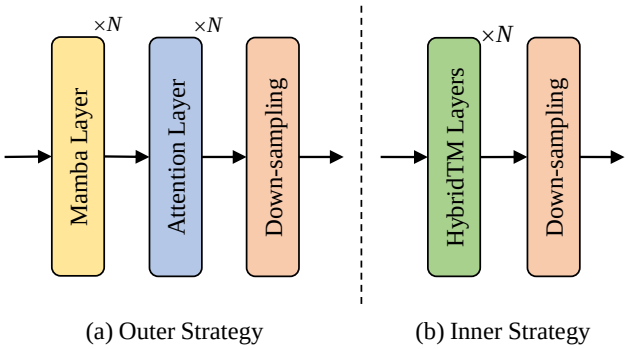


Fig. 4. 不同混合策略的比较。

与一般视觉任务相比，注意力和 Mamba 算子的整合在 3D 语义分割中提出了新的挑战。在 2D 视觉中，现有的方法 [31], [33], [32] 通常采用一种外部混合策略，即简单地将 Mamba 层分配给每个阶段的前 $N/2$ 层，而将注意力层分配给后 $N/2$ 层（图 4 (a)）。虽然这种方法对密集的 2D 图像效果良好，但由于输入数据特性的根本不同，它在 3D 语义分割中显得不够用。由于点云本质上的稀疏性和不规则性，必须通过填充曲线 [14] 或基于窗口的排序 [11] 将其序列化以进行进一步处理。虽然 Mamba 在一般视觉任务的序列建模中展示了卓越的能力，但它在有效捕捉这些序列化序列中的体素之间的相对空间关系方面表现不佳。这个问题严重阻碍了 Mamba 实现准确的分割性能。不像基于 Transformer 的架构，它们可以通过注意力算子直接模型化逐元素关系，仅仅采用 Mamba 层进行 3D 语义分割会导致特征退化。因此，我们认为整合注意力算子对于保持空间关系和捕捉细粒度的局部特征是必不可少的，同时让 Mamba 能够专注于建模远程依赖关系。

如图 4 所示，与外层策略（图 4 (a)）相比，我们的内层混合策略（图 4 (b)）在更细的粒度上整合了注意力层和 Mamba 层。通过允许注意力层首先保留相对的空间关系并提取细粒度特征，我们使得 Mamba 层能够基于细粒度的局部特征更好地建模全局依赖性。此外，我们在消融研究中凭经验证明了内层混合策略的有效性。

## IV. 实验

### A. 数据集

Scannet 数据集。ScanNet [28] 是一个大规模的三维室内场景数据集，由大约 1500 个在不同环境中拍摄的 RGB-D 视频序列组成。ScanNet 提供了广泛的标注，包括 20 类如家具和结构元素的三维物体实例标签和语义分割。该数据集是室内环境中三维语义分割的重要基准。

ScanNet200。基于原始的 ScanNet 框架，ScanNet200 [28] 将语义标注的细粒度扩展到 200 个不同的对象类别。由于其细致的对象分类，该数据集保持了与 ScanNet 相同的场景分割，但提出了更具挑战性的语义分割任务。

nuScenes。nuScenes 数据集 [29] 包含在波士顿和新加坡捕获的 1000 个城市驾驶序列。每个序列大约 20 秒长，使用由 6 个摄像头、1 个激光雷达、5 个雷达单元以及

www.xueshuxiangzi.com

GPS/IMU 组成的传感器套件录制。数据集提供了 32 个与自动驾驶场景相关的类别的逐点语义分割标签，总计 1.1B 手动标注的激光雷达点。在我们的实验中，我们关注 16 个主要类别，遵循之前的研究 [14] 。

S3DIS 数据集。斯坦福大规模三维室内空间数据集 (S3DIS) [41] 包含了来自三个不同建筑中的六个大型室内区域的详细三维扫描。该数据集使用地面激光扫描仪收集，包含超过 2.15 亿个点，每个点都用 13 个语义类别之一进行了标注。

### B. 实现细节

我们使用一种 UNet 风格的架构实现了 HybridTM ，该架构具有非对称的编码器-解码器路径。编码器由五个阶段组成，包含 [2,2,2,6,2] 层，而解码器则包含四个阶段，含有 [2,2,2,2] 层。对于注意力机制，我们在所有编码器和解码器阶段中保持一致的 1024 组大小。同样地，Mamba 块 [18] 在整个网络中使用统一的 4096 组大小。

在训练时，我们采用 Point Transformer V3 [14] 中使用的损失函数和数据增强策略。我们使用 AdamW [42] 和 One-Cycle 学习率策略来优化网络。训练在 4 个 NVIDIA RTX 3090 GPU 上进行，批量大小为 12。训练的周期数根据数据集而有所不同：ScanNet 的两种变体为 800 个周期，S3DIS 为 3000 个周期，nuScenes 为 50 个周期。

### C. 主要结果

TABLE I

与最先进的方法在 ScanNet 验证集上的比较。

| Method | Present at | mIoU |
|---|---|---|
| PointNet++ [43] | NeurIPS 2017 | 53.5 |
| MinkowskiNet [44] | CVPR 2019 | 72.2 |
| O-CNN [45] | SIGGRAPH 2017 | 74.0 |
| ST [16] | CVPR 2022 | 74.3 |
| Point Transformer v2 [13] | NeurIPS 2022 | 75.4 |
| OctFormer [12] | SIGGRAPH 2023 | 74.5 |
| Swin3D [11] | arXiv 2023 | 75.5 |
| Point Transformer v3 [14] | CVPR 2024 | 77.5 |
| Pont Mamba [24] | arXiv 2024 | 75.7 |
| Serialized Point Mamba [26] | arXiv 2024 | 76.8 |
| Ours | – | 77.8 |

在 ScanNet 数据集上的结果。在表格 I 中，我们评估了 HybridTM 在 ScanNet 验证集上的性能，并将其与多个先进的（SOTA）方法进行了比较。 HybridTM 达到了 77.8 % mIoU，创造了新的 SOTA 结果。此外，HybridTM 比 Point Transformer v3 [14] 高出 0.3 % mIoU。同时，HybridTM 也超过了其他领先方法，如 Swin3D [11] 和 Serialized Point Mamba [26] ，分别高出 2.3 % 和 1.0 % mIoU。这些结果证明了 HybridTM 在推动 3D 语义分割边界方面的有效性。

在 Scannet200 数据集上的结果。为了验证具有更多类别的 HybridTM 。我们展示了 HybridTM 与几种 SOTA 方法在 ScanNet200 验证集上的比较。如表格 II 所示，HybridTM 显著优于之前的方法，并达到了新的 SOTA 结果（36.5 % mIoU）。具体来说， HybridTM 分别比 Point Transformer V3 和 OctFormer 高出 1.3 % mIoU 和 3.9 % mIoU。这些结果证明了 HybridTM 在更复杂的数据集上进行 3D 语义分割的有效性。

nuScenes 数据集上的结果。为了进一步验证户外数据集上的 HybridTM 。在表 III 中，我们展示了 HybridTM

TABLE II

在 ScanNet200 验证集上，我们的方法与最新方法的对比。

| Model | Present at | mIoU |
|---|---|---|
| MinkowskiNet [44] | CVPR 2019 | 25.0 |
| OctFormer [12] | SIGGRAPH 2023 | 32.6 |
| Point Transformer v2 [13] | NeurIPS 2022 | 30.2 |
| Point Transformer v3 [14] | CVPR 2024 | 35.2 |
| Ours | – | 36.5 |

TABLE III

我们的方法与最先进方法在 nuScenes 验证集上的比较

| Model | Present at | mIoU |
|---|---|---|
| MinkowskiNet [44] | CVPR 2019 | 73.3 |
| SPVNAS [46] | ECCV 2020 | 77.4 |
| Cylender3D [47] | CVPR 2021 | 76.1 |
| AF2S3Net [48] | CVPR 2021 | 62.2 |
| SphereFormer [17] | CVPR 2023 | 79.5 |
| Point Transformer v3 [14] | CVPR 2024 | 80.2 |
| Ours | – | 80.9 |

TABLE IV

我们的方法与最新方法在 S3DIS 验证集上的比较。

| Model | Present at | Aera5 (mIoU) |
|---|---|---|
| MinkowskiNet [44] | CVPR 2019 | 65.4 |
| PointNeXt [49] | NeurIPS 2022 | 70.5 |
| Swin3D [11] | arXiv 2023 | 72.5 |
| Point Transformer v2 [13] | NeurIPS 2022 | 71.6 |
| Serialized Point Mamba [14] | arXiv 2024 | 70.6 |
| Ours | – | 72.1 |

与几个最先进方法在 nuScenes 验证集上的比较。 HybridTM 达到了 80.9 % mIoU，达到了新的 SOTA 结果。此外， HybridTM 分别比 Point Transformer v3 和 SphereFormer 高出 0.7 % 和 1.4 % mIoU。这些结果证明了 HybridTM 在处理户外大规模数据集的 3D 语义分割任务上的有效性。

S3DIS 数据集上的结果。我们在较小的数据集上评估了 HybridTM 的泛化能力。如表 IV 所示， HybridTM 取得了令人满意的性能（72.1 % mIoU），比 Serialized Point Mamba [26] 高出 1.5 % mIoU。这些结果表明了 HybridTM 的泛化能力。

我们在 ScanNet 验证集上进行了消融研究。训练设置与表 I 相同。

混合层中组件的有效性。我们进行了消融研究，以分析我们混合层架构中每个组件的贡献，结果如表 V 所示。仅使用注意力层（配置 I ）时， HybridTM 实现了 77.1 % mIoU。当仅使用 Mamba 层（配置 II ）时，性能下降到 76.9 % mIoU，这表明仅 Mamba 会导致特征退化。然而，利用 IL 策略集成两层（配置 III ）后，性能达到 77.8 % mIoU，显著超过单层变体（相比仅注意力层高 0.7 %，相比仅 Mamba 层高 0.9 %）。这些结果经验性地验证了我们混合架构的协同效益以及 IL 在结合注意力和 Mamba 操作符的互补优势方面的有效性。

不同混合策略的比较。为了验证我们的内层混合策略 (IL) 的有效性，我们对比了不同的混合策略。如表 VI 所

www.xueshuxiangzi.com

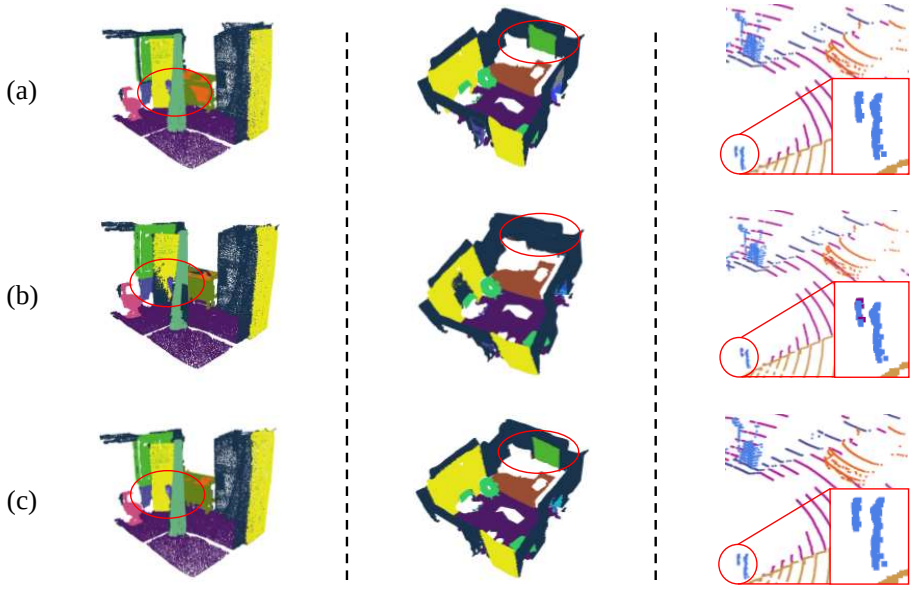


Fig. 5. 在 ScanNet 和 nuScenes 验证集上比较了 Point Transformer V3 (b) 和 HybridTM (c) 的表现。(a) 是地面实况。第一和第二列是 ScanNet 上的结果。最后一列是 nuScenes 上的结果。可以看出， HybridTM 能够取得比 Point Transformer V3 更好的结果，显示了 HybridTM 的优越性。

TABLE V
混合层中每一层的有效性。

| # | Attention Layer | Mamba Layer | mIoU |
|---|---|---|---|
| I | ✓ | | 77.1 |
| II | | ✓ | 76.9 |
| III | ✓ | ✓ | 77.8 |

TABLE VI
不同混合策略的比较。

| # | Strategy | mIoU |
|---|---|---|
| I | Outer Strategy [31] (Mamba before Attention) | 77.1 |
| II | Outer Strategy [31] (Mamba after Attention) | 77.4 |
| III | IL (Mamba before Attention) | 77.5 |
| IV | IL (Mamba after Attention) | 77.8 |

示，MambaVision [31] ( I ) 中使用的外部策略 (Mamba 在注意力之前) 只能达到 77.1 % 的 mIoU。当我们将注意力层放在外部策略中 Mamba 之前 ( II ) 时， HybridTM 可以达到 77.4 %，比 ( I ) 高出 0.3 % 的 mIoU。这些结果表明注意力应该放在 Mamba 之前。当我们使用 IL (Mamba 在注意力之前) ( III ) 时， HybridTM 可以达到 77.5 % 的 mIoU，比 ( I ) 高出 0.4 % 的 mIoU。当我们使用 IL (Mamba 在注意力之后) ( IV ) 时， HybridTM 可以达到 77.8 % 的 mIoU，分别比 ( I ), II , ( III ) 高出 0.7 % , 0.4 % , 0.5 % 的 mIoU。这些结果证明了我们的 IL 与其他策略相比的优越性。

### D. 可视化

为了说明 HybridTM 的优越性，我们在图 5 中展示了 Point Transformer V3 [14] (b) 和 HybridTM (c) 在 ScanNet 和 nuScenes 验证集上的定性结果可视化。在第一和第二列中， HybridTM 能够对大型物体实现更精确的分割。在最后一列中，受益于所提出的 IL 混合策略，HybridTM 也能准确地分割小型物体。可视化结果表明，我们的方法不仅关注局部细节，还顾及更广泛的区域，从而实现更好的整体分割性能。

在本文中，我们提出了 HybridTM ，这是首个将 Transformer 和 Mamba 协同结合的 3D 语义分割混合架构。此外，我们提出了内层混合策略，能够有效整合细粒度空间特征提取和高效全局上下文建模，专为 3D 语义分割设计。我们的方法的有效性通过在具有挑战性的室内 (ScanNet, ScanNet200) 和室外 (nuScenes) 基准测试中的最先进性能得到验证，展示了我们的混合架构在 3D 语义分割中的优越性。

## References

[1] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, M. Jagersand, and H. Zhang, "A comparative study of real-time semantic segmentation for autonomous driving," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 587–597.

[2] R. Marcuzzi, L. Nunes, L. Wiesmann, J. Behley, and C. Stachniss, "Mask-based panoptic lidar segmentation for autonomous driving," IEEE Robotics and Automation Letters, vol. 8, no. 2, pp. 1141–1148, 2023.

[3] Y. Sun, W. Zuo, H. Huang, P. Cai, and M. Liu, "Pointmoseg: Sparse tensor-based end-to-end moving-obstacle segmentation in 3-d lidar point clouds for autonomous driving," IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 510–517, 2020.

[4] Z. Liu, T. Huang, B. Li, X. Chen, X. Wang, and X. Bai, "Epnet++: Cascade bi-directional fusion for multi-modal 3d object detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 7, pp. 8324–8341, 2022.

[5] W. Kim and J. Seok, "Indoor semantic segmentation for robot navigating on mobile," in 2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN). IEEE, 2018, pp. 22–25.

[6] J. Chen, Y. K. Cho, and Z. Kira, "Multi-view incremental segmentation of 3-d point clouds for mobile robots," IEEE Robotics and Automation Letters, vol. 4, no. 2, pp. 1240–1246, 2019.

www.xueshuxiangzi.com

[7] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in Porc. of IEEE Intl. Conf. on Computer Vision, 2019, pp. 9297–9307.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Proc. of Advances in Neural Information Processing Systems, vol. 30, 2017.

[9] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in Porc. of IEEE Intl. Conf. on Computer Vision, 2021, pp. 16 259–16 268.

[10] C. Zhang, H. Wan, X. Shen, and Z. Wu, "Patchformer: An efficient point transformer with patch attention," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, 2022, pp. 11 799–11 808.

[11] Y.-Q. Yang, Y.-X. Guo, J.-Y. Xiong, Y. Liu, H. Pan, P.-S. Wang, X. Tong, and B. Guo, "Swin3d: A pretrained transformer backbone for 3d indoor scene understanding," arXiv preprint arXiv:2304.06906, 2023.

[12] P.-S. Wang, "Octformer: Octree-based transformers for 3d point clouds," ACM Transactions ON Graphics, vol. 42, no. 4, pp. 1–11, 2023.

[13] X. Wu, Y. Lao, L. Jiang, X. Liu, and H. Zhao, "Point transformer v2: Grouped vector attention and partition-based pooling," Proc. of Advances in Neural Information Processing Systems, vol. 35, pp. 33 330–33 342, 2022.

[14] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao, "Point transformer v3: Simpler faster stronger," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, 2024, pp. 4840–4851.

[15] C. Park, Y. Jeong, M. Cho, and J. Park, "Fast point transformer," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, 2022, pp. 16 949–16 958.

[16] X. Lai, J. Liu, L. Jiang, L. Wang, H. Zhao, S. Liu, X. Qi, and J. Jia, "Stratified transformer for 3d point cloud segmentation," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, 2022, pp. 8500–8509.

[17] X. Lai, Y. Chen, F. Lu, J. Liu, and J. Jia, "Spherical transformer for lidar-based 3d recognition," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, 2023, pp. 17 545–17 555.

[18] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv preprint arXiv:2312.00752, 2023.

[19] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," in Proc. of Intl. Conf. on Machine Learning, 2024.

[20] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," Proc. of Advances in Neural Information Processing Systems, 2024.

[21] T. Huang, X. Pei, S. You, F. Wang, C. Qian, and C. Xu, "Localmamba: Visual state space model with windowed selective scan," arXiv preprint arXiv:2403.09338, 2024.

[22] X. Pei, T. Huang, and C. Xu, "Efficientvmamba: Atrous selective scan for light weight visual mamba," arXiv preprint arXiv:2403.09977, 2024.

[23] J. Ruan and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," arXiv preprint arXiv:2402.02491, 2024.

[24] J. Liu, R. Yu, Y. Wang, Y. Zheng, T. Deng, W. Ye, and H. Wang, "Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy," arXiv preprint arXiv:2403.06467, 2024.

[25] T. Zhang, X. Li, H. Yuan, S. Ji, and S. Yan, "Point cloud mamba: Point cloud learning via state space model," arXiv preprint arXiv:2403.00762, 2024.

[26] T. Wang, W. Wen, J. Zhai, K. Xu, and H. Luo, "Serialized point mamba: A serialized point cloud mamba segmentation model," arXiv preprint arXiv:2407.12319, 2024.

[27] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, 2018, pp. 9224–9232.

[28] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, 2017, pp. 5828–5839.

[29] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, 2020, pp. 11 621–11 631.

[30] O. Lieber, B. Lenz, H. Bata, G. Cohen, J. Osin, I. Dalmedigos, E. Safahi, S. Meirom, Y. Belinkov, S. Shalev-Shwartz et al., "Jamba: A hybrid transformer-mamba language model," arXiv preprint arXiv:2403.19887, 2024.

[31] A. Hatamizadeh and J. Kautz, "Mambavision: A hybrid mamba-transformer vision backbone," arXiv preprint arXiv:2407.08083, 2024.

[32] Y. Liu and L. Yi, "Map: Unleashing hybrid mamba-transformer vision backbone's potential with masked autoregressive pretraining," arXiv preprint arXiv:2410.00871, 2024.

[33] W. Chen, L. Niu, Z. Lu, F. Meng, and J. Zhou, "Maskmamba: A hybrid mamba-transformer model for masked image generation," arXiv preprint arXiv:2409.19937, 2024.

[34] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," Computational Visual Media, vol. 7, pp. 187–199, 2021.

[35] M. Zeller, J. Behley, M. Heidingsfeld, and C. Stachniss, "Gaussian radar transformer for semantic segmentation in noisy radar data," IEEE Robotics and Automation Letters, vol. 8, no. 1, pp. 344–351, 2022.

[36] T. Dao and A. Gu, "Transformers are ssms: Generalized models and efficient algorithms through structured state space duality," in Proc. of Intl. Conf. on Machine Learning, 2024.

[37] S. Hwang, A. Lahoti, T. Dao, and A. Gu, "Hydra: Bidirectional state space models through generalized matrix mixers," arXiv preprint arXiv:2407.09941, 2024.

[38] T. Chen, Z. Ye, Z. Tan, T. Gong, Y. Wu, Q. Chu, B. Liu, N. Yu, and J. Ye, "Mim-istd: Mamba-in-mamba for efficient infrared small target detection," IEEE Transactions on Geoscience and Remote Sensing, 2024.

[39] J. Wang, D. Huang, X. Guan, Z. Sun, T. Shen, F. Liu, and H. Cui, "Omega: Efficient occlusion-aware navigation for air-ground robot in dynamic environments via state space model," IEEE Robotics and Automation Letters, 2024.

[40] Z. Liu, J. Hou, X. Wang, X. Ye, J. Wang, H. Zhao, and X. Bai, "Lion: Linear group rnn for 3d object detection in point clouds," Proc. of Advances in Neural Information Processing Systems, 2024.

[41] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3d semantic parsing of large-scale indoor spaces," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, 2016, pp. 1534–1543.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. of Intl. Conf. on Learning Representations, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:53592270

[43] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," Proc. of Advances in Neural Information Processing Systems, vol. 30, 2017.

[44] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, 2019, pp. 3075–3084.

[45] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong, "O-cnn: Octree-based convolutional neural networks for 3d shape analysis," ACM Transactions ON Graphics, vol. 36, no. 4, pp. 1–11, 2017.

[46] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching efficient 3d architectures with sparse point-voxel convolution," in Proc. of European Conference on Computer Vision. Springer, 2020, pp. 685–702.

[47] X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, and D. Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar segmentation," in Proc. of IEEE Intl. Conf.


www.xueshuxiangzi.com

on Computer Vision and Pattern Recognition, 2021, pp. 9939–9948.

[48] R. Cheng, R. Razani, E. Taghavi, E. Li, and B. Liu, "2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network," in Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition, 2021, pp. 12 547–12 556.

[49] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, and B. Ghanem, "Pointnext: Revisiting pointnet++ with improved training and scaling strategies," Proc. of Advances in Neural Information Processing Systems, vol. 35, pp. 23 192–23 204, 2022.

www.xueshuxiangzi.com