自动化学报
ACTA AUTOMATICA SINICA

基于渐进式适配器的动态点云高效迁移学习方法

孙一丁 祝继华 王耀南 程浩 卢超翼 陈林

Efficient Transfer Learning for Dynamic Point Clouds With Progressive Adapters

SUN Yi-Ding, ZHU Ji-Hua, WANG Yao-Nan, CHENG Hao-Zhe, LU Chao-Yi, CHEN Lin

在线阅读 View online: https://doi.org/10.16383/j.aas.c250666

基于渐进式适配器的动态点云高效迁移学习方法

孙一丁[1, 2] 祝继华[1, 2] 王耀南[3, 4] 程浩喆[1] 卢超翼[1] 陈 林[1]

摘 要 动态点云视频时空耦合复杂, 现有端到端方法仅支持特定数据集与任务验证, 跨场景、跨任务泛化能力不足, 需重复训练. 为能够挖掘可迁移的通用先验, 本文将研究视角重新聚焦于静态点云基础模型. 首先分析现有迁移学习方法存在的两大掣肘, 即密集计算与冗余设计. 为实现轻量、紧凑的目标, 提出一种基于渐进式适配器的跨模态高效迁移学习方法. 该方法通过无重叠滑动窗口注意力, 将自注意力复杂度由二次阶降为线性阶. 其次引入洗牌希尔伯特-Z 序双向扫描曲线, 将动态点云视频约束为兼容 Mamba 的特征序列, 并以渐进式门控融合实现基础模型与新增适配器的高效协同. 整体方法不改变基础模型权重, 不依赖外部辅助设计. 最后在 MSR-Action3D、HOI4D 和 SHREC′17 上的实验结果表明, 仅微调 1.8% 参数即可获得超越现有基线模型的性能, 验证了该方法的优越性.

关键词 点云视频; 迁移学习; 混合架构; 注意力机制; 高效适配器

引用格式 孙一丁, 祝继华, 王耀南, 程浩喆, 卢超翼, 陈林. 基于渐进式适配器的动态点云高效迁移学习方法. 自动化学报, 2026, **52**(8): 1710−1720

DOI 10.16383/j.aas.c250666 **CSTR** 32138.14.j.aas.c250666

Efficient Transfer Learning for Dynamic Point Clouds With Progressive Adapters

SUN Yi-Ding[1, 2] ZHU Ji-Hua[1, 2] WANG Yao-Nan[3, 4] CHENG Hao-Zhe[1] LU Chao-Yi[1] CHEN Lin[1]

Abstract Dynamic point cloud videos exhibit intricate spatio-temporal coupling. Current end-to-end methods are only evaluated on specific datasets and tasks, generalize poorly across scenes or tasks, and must be retrained from scratch. To uncover transferable universal priors, we refocus our investigation on static point cloud foundation models. In this paper, we first identify two bottlenecks in existing transfer learning methods: Heavy computation and redundant design. To obtain a compact and lightweight solution, we propose an efficient cross-modal transfer method that uses progressive adapters. It reduces self-attention from quadratic to linear complexity with non-overlapping sliding window attention. A shuffled Hilbert-Z bidirectional scan curve converts the dynamic point cloud video into a Mamba-compatible feature sequence; progressive gating fusion then lets the foundation model and the newly added adapter cooperate efficiently. The foundation model weights remain unchanged and no external auxiliary designs are required. Experiments on MSR-Action3D, HOI4D and SHREC′17 show that tuning only 1.8% of the parameters already surpasses existing baselines, confirming the superiority of our method.

Keywords point cloud video; transfer learning; hybrid architecture; attention mechanism; efficient adapter

Citation Sun Yi-Ding, Zhu Ji-Hua, Wang Yao-Nan, Cheng Hao-Zhe, Lu Chao-Yi, Chen Lin. Efficient transfer learning for dynamic point clouds with progressive adapters. *Acta Automatica Sinica*, 2026, **52**(8): 1710−1720

近年来, 随着机器人技术[1]、增强现实与虚拟现实[2]、智慧城市等领域的快速发展, 点云视频理解引起了广泛关注. 与二维图像和静态点云不同, 点云视频将三维空间与一维时间融合, 拥有更强大的表示潜力. 目前的点云视频表示学习方法主要分为有监督学习[3-4]和自监督学习[5-6]两大类. 尽管基于 Transformer 骨干的一系列方法[3]已在主流基准上取得优秀的表现, 但单纯依赖该架构进行 4D 时空特征抽取时, 仍面临长序列建模效率骤降的固有瓶颈. 同时, 此类方法普遍遵循“一任务一模型”的专用化训练协议, 致使网络权重无法跨任务复用. 鉴于动态点云视频数据规模远小于静态点云, 且采集与标注成本高昂, 若对每个下游任务均从头训练独立模型, 将带来难以承受的计算与人力开销. 因此, 迁移静

收稿日期 2025-11-24 录用日期 2026-02-13
Manuscript received November 24, 2025; accepted February 13, 2026
国家自然科学基金 (62125305), 陕西省杰出青年科学基金 (2025JC-JCQN-091), 陕西省技术创新引导计划 (基金) 资助项目 (2024QY-SZX-23) 资助
Supported by National Natural Science Foundation of China (62125305), the Natural Science Basis Research Plan in Shaanxi Province of China (2025JC-JCQN-091), and the Technology Innovation Leading Program of Shaanxi (2024QY-SZX-23)
本文责任编委 樊彬
Recommended by Associate Editor FAN Bin
1. 西安交通大学软件学院 西安 710049 2. 人机混合增强智能全国重点实验室 西安 710049 3. 湖南大学人工智能与机器人学院 长沙 410082 4. 机器人视觉感知与控制技术国家工程研究中心 长沙 410082
1. School of Software Engineering, Xi′an Jiaotong University, Xi′an 710049 2. State Key Laboratory of Human-machine Hybrid Augmented Intelligence, Xi′an 710049 3. School of Artificial Intelligence and Robotics, Hunan University, Changsha 410082 4. National Engineering Research Center for Robot Visual Perception and Control Technology, Changsha 410082

态点云基础模型所蕴含的丰富先验知识, 构建通用、高效的点云视频理解框架, 已成为该领域极具前景的研究方向. 然而, 由于点云视频与静态点云在时空密度、采样拓扑及动态演化等维度存在天然模态差异, 直接迁移静态基础模型至视频理解任务往往出现显著的性能退化[7–8]. PointCSA[9]采用几何约束时空适配器 GCTA (geometric constraint spatial-temporal adapter), 借助轻量级适配器 (adapter) 结构实现时空信息融合, 在多个下游数据集上取得优秀的表现. 但深入剖析其机制可以发现, 该方法仍受限于双重瓶颈: 一方面, GCTA 选择完全冻结并复用静态 Transformer 的全部权重, 对模态差异未做任何针对性适配, 这一方案是直观但次优的. 静态点云基础模型[10–11] 普遍采用自注意力机制, 其计算复杂度与词元 (token) 数量呈二次关系. 当输入由单帧扩展为长程点云视频时, 序列长度随帧数倍增, 导致内存占用与计算开销急剧放大, 成为影响迁移学习效率和可扩展性的首要障碍. 另一方面, PointCSA[9] 基于多层感知机架构 (multilayer perceptron, MLP) 采用 GCTA 适配器. 该架构天然缺少显式的时序建模能力, 必须额外引入位置编码和手工时序约束来完成时空理解. 适配器的整体结构冗余度高且对位置编码器设计高度敏感, 限制了在跨任务、跨场景下的泛化能力.

本文克服了上述两项局限并提出 PointM4 方法. 该方法主要包含两部分操作: 针对注意力机制时间复杂度高的问题, 提出一种无重叠滑动窗口注意力机制. 该机制可以精准地控制注意力的感受野并大幅降低计算负担. 针对适配器设计冗余的问题, 提出一种基于 Mamba 架构[12] 的全新适配器. 与 PointCSA[9] 中的多层感知机架构相比, 本文方法通过选择性状态空间将序列记忆直接融入模型, 在线性复杂度下实现长程建模. 适配器整体设计简洁, 不依赖多余的外部组件. 同时, 通过洗牌时空扫描曲线和渐进式适配方法, 本文弥补 Mamba 架构在全局建模上的先天不足, 并避免静态点云基础模型在跨模态迁移学习中的信息冲突. 如图 1 所示, 本文方法仅用 1.8% 的参数即可在线性复杂度下开展实验. 实验结果表明, 该方法在 3D 动作识别、4D 动作分割、4D 语义分割、手势识别任务上分别取得了+1.4%、+8.9%、+2.7% 和 +1.1% 的性能提升. 本文也对提出的各类算法和策略进行了消融实验, 保证了模型的可靠性. 本文主要贡献如下: 1) 分析现有动态点云视频迁移学习方法的不足之处并提供针对性的改进; 2) 设计无重叠滑动窗口注意力, 线性化时间复杂度, 以提升模型局部感知能力; 3) 提出洗牌时空扫描曲线与渐进式适配协同的 Mamba 适配器, 在保持模型紧凑的前提下提升跨域泛化能力; 4) 本文方法在多个下游数据集上达到最先进的性能, 在 MSR-Action3D 数据集上相较于直接迁移精度平均提升了 15.0%.

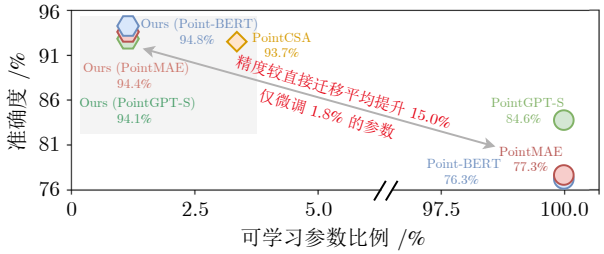

图 1　本文与现有方法在性能与可学习参数比例上的对比

Fig. 1　Comparison between our method and existing methods on performance and learnable parameter ratio

1　相关工作

1.1　静态点云理解

静态点云以三维空间离散采样刻画物体几何与表面属性, 是自动驾驶目标检测、环境建模与定位建图的核心感知媒介. Qi 等[13] 提出直接作用于无序点集的深度网络 PointNet, 以最大池化对称函数实现置换不变性, 为点云分类与分割提供统一框架. 随着点云采集设备的普及, 获取大量未标注的点云数据不再困难. 为能够充分利用这些数据, 点云自监督预训练方法应运而生. 这些方法主要通过设计精巧的代理任务来充分学习点云潜在的语义信息, 并在下游数据集上进行具体任务的适配与微调. 现有的预训练任务主要分为两类: 一种是对比式任务[14–15], 另一种是生成式任务[10, 16]. 对比式任务通过构建正负样本对, 在特征空间将相似度高的样本拉近, 相似度低的样本推远, 从而使模型能够区分不同的点云实例. 例如, Xie 等[14] 提出几何增强方法来生成正负样本对; Afham 等[15] 采用模态间和模态内的对比学习. 另一方面, 生成式任务要求模型通过重建被掩码的输入数据来学习结构化知识, 鼓励模型恢复局部关联. Pang 等[10] 首先将该思想应用于三维点云领域; Zhang 等[16] 进一步使用层次化 Transformer 并设计相应的掩码策略. 然而, 这些预训练模型虽然性能强大, 但是它们仍受限于专用的下游任务, 如静态点云分类、分割等. 本文旨在通过插入适配器, 将这些模型迁移到点云视频理解领域来扩展模型的实用性.

1.2　动态点云理解

动态点云理解作为静态点云的延伸, 由于其在

具身机器人、自动驾驶领域的广泛应用, 近年来吸引了学术界的关注. 动态点云视频拥有帧内无序与帧间有序的双重特性, 因此需要更复杂的方法来聚合和理解时空数据. 当前该领域的方法主要分为基于体素的方法和基于点的方法: 基于体素的方法由 Choy 等[17] 提出, 通过对原始点云进行体素化, 从 4D 体素中提取特征; 基于点的方法适用性更高并集成了多种最新架构; 例如, Fan 等[3] 利用 Transformer 架构来增强模型对时空相关性的捕捉; Deng 等[18] 首次将视觉–语言模型知识迁移至 4D 点云视频, 通过跨模态对比学习弥补点云细节缺失, 实现 RGB-点云协同动作识别; Fan 等[19] 提出点时空 Transformer, 通过全局关联点集并解耦编码时空结构, 无需显式跟踪即可精准建模点云视频动态; Liu 等[20] 则开发出一种基于状态空间模型的 4D 骨干网络. 尽管这些方法有效, 但在实际部署过程中往往无法承受庞大的计算开销. 本文方法通过复用预训练好的 3D 基础模型来高效地学习动态点云数据, 能够在减少训练时间的同时, 达到与最新方法媲美的性能.

1.3 Mamba 与 Transformer 混合架构

Mamba 架构由于其线性复杂度和强大的信息聚合能力, 近年来成为表示学习领域的研究重点之一. 考虑到 Mamba 架构的高效性, 许多方法尝试将 Mamba 与其他架构结合. 例如, 对于视觉任务, Wang 等[21] 同时将 Transformer 和 Mamba 架构结合, 并用于静态点云表示学习. 然而, 目前尚未有专用于点云视频理解的混合架构方法. 本文首次尝试在不改变原始 Transformer 预训练权重的前提下, 将 Mamba 作为适配器融入现有模型中, 辅以特定的扫描和融合策略, 最大化地发挥 Mamba 架构的优势, 填补该领域的空白.

2 方法架构

本文方法旨在利用静态点云预训练模型所蕴含的丰富空间先验, 并将其迁移至动态点云理解任务, 以提升表征效率与泛化能力. 第 2.1 节介绍了 3D 基础模型的预处理操作及其基于 Transformer 架构的前向过程; 为了降低计算复杂度, 提升模型运行效率, 第 2.2 节介绍了基于无重叠滑动窗口的注意力机制; 为了使 Mamba 模块捕捉长距离的时空动态, 第 2.3 节提出专用于动态点云理解的 Mamba 适配器, 其主要包含两阶段的适配过程, 能够在逐层对特征进行精调的同时, 将时间复杂度控制在线性级别. 方法的总体架构如图 2 所示.

2.1 3D 基础模型概览

给定一个点云 $P \in \mathbf{R}^{N\times 3}$, N 表示点云中蕴含的点数. 以 PointMAE[10] 模型为例, 首先通过最远点采样和 KNN 算法将完整点云分为多个块. 接着使用轻量级 PointNet[13] 网络完成对点云初始嵌入的编码, 记为 $P^0 \in \mathbf{R}^{N\times K\times 3}$, 其中 K 代表 KNN 算法中的近邻数量. PointMAE 编码器由串联式的 PointTransformer 层组成, 在每一层中都会执行自注意力 (self-attention, Attention) 与前馈神经网络 (feed-forward network, FFN) 两个操作. 具体来说, 对于第 i 层的点云输入特征 P^i, 其前向过程如下所示:

$$P^{i+1} = \mathrm{FFN}\left(\mathrm{Attention}\left(P^i\right)\right) \tag{1}$$

然而, 静态点云基础模型本质上缺乏整合时间信息的能力. 当前主流的方法是通过插入额外的适配器来赋能静态基础模型并捕捉时序动态, 在保持基础模型冻结的前提下进行有监督微调. 本文架构在此范式下实现了两方面的创新: 一方面, 通过灵活的无重叠滑动窗口, 将原先一帧完整的点云在空

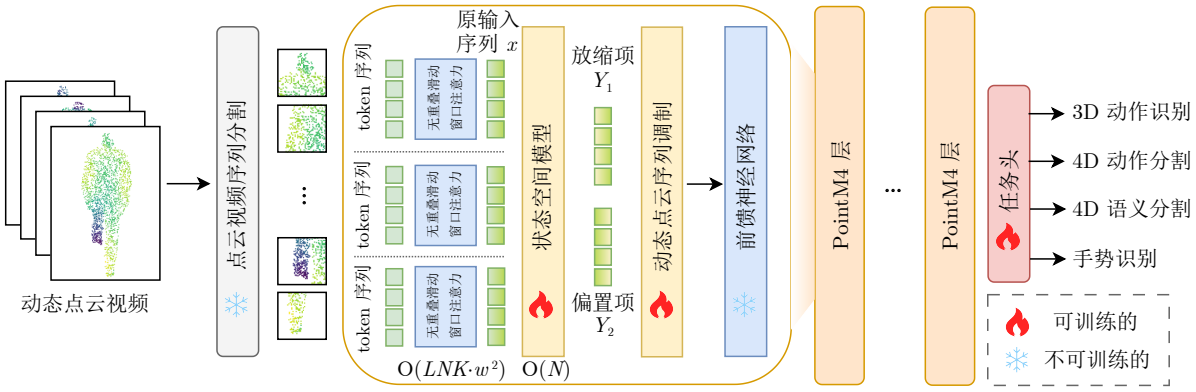

图 2 PointM4 模型结构示意图

Fig. 2 Schematic diagram of PointM4 model architecture

间上划分开, 将全局注意力范围降低到窗口级别, 且保持预训练权重不变. 利用这种方法, 本文在 Transformer 层内部降低了注意力的计算复杂度, 同时有效建模短期的空间关系, 实现精度与效率的权衡. 另一方面, 本文方法首次将 Mamba 架构用于动态点云迁移学习. 在保证线性复杂度的前提下, 仅插入少量可学习的参数, 即可实现静态 3D 基础模型的跨模态参数高效微调. 为使 Mamba 适配器的输出与基础模型相匹配, 本文使用了一种渐进式适配操作, 将原先完整的特征序列映射为两条序列, 以双向时空扫描后逐特征融合的形式捕捉 4D 理解所需的长程依赖.

2.2　基于无重叠滑动窗口的注意力机制

Transformer 的归纳偏置表明, 相对于基于 CNN[22] 和 Mamba[20] 的方法, 其必须从头学习空间关系和位置关系, 这导致标准的注意力机制往往仅关注全局词元的变化, 而忽略了局部的语义细节. 同时, 最近的一些工作已经开始关注全局自注意力带来的开销与其性能的平衡[20, 23], PointCSA[9] 适配器虽然能够逐帧地引入时序信息, 但是该方法仍然停留在全局注意力这一开销极大的机制下. 考虑到点云天然具有的无序性和非结构性, 本节的目标是在不改变输入序列形状的前提下, 更精确地控制注意力的感受野, 并相应地削减冗余计算. 本节提出一种基于滑动 3D 窗口的局部注意力机制. 具体来说, 一帧完整的点云被切分为多个窗口, 并在窗口内部应用局部注意力. 对于第 i 层的点云视频特征 PV^i, 使用不重叠的三维窗口将其切分, 窗口尺寸固定为 $w \times w \times 3$. 设点云视频共有 n 个窗口, 则有:

$$PV^i \to \left[S_1^i,\ S_2^i,\ \cdots,\ S_n^i\right] \tag{2}$$

式中, S_1, S_2, $\cdots$, S_n 是独立的 3D 窗口. 对于每一个窗口, 在窗口内部应用自注意力机制, 则有经过 Attention 处理后的窗口特征:

$$\hat{S}_n^i = \text{Attention}\left(S_n^i\right) \tag{3}$$

式中, 注意力权重与预训练好的编码器注意力层相同. 通过无重叠滑动窗口注意力, 使原本空间语义丰富的点云视频序列在每个窗口内部进行局部信息交换. 在窗口注意力操作后, 本节将每一个窗口提供的输出拼接 (concatenate) 起来, 以 $[\,\cdot\,]$ 表示, 从而获得可用于下一阶段处理的完整点云特征序列. 整个过程可以被表示为:

$$X^i = \left[\hat{S}_1^i,\ \hat{S}_2^i,\ \cdots,\ \hat{S}_n^i\right] \tag{4}$$

给定一个 $L \times N \times K \times 3$ 的点云视频序列, L 为点云视频帧数量, 如果按照现有的动态点云骨干[3] 进行时间–空间联合注意力的计算, 那么复杂度为 $\mathrm{O}(LNK)^2$; 按照 PointCSA[9] 进行逐帧空间注意力计算, 复杂度为 $\mathrm{O}\left((NK)^2L\right)$. 而本节提出的无重叠滑动窗口注意力, 则将其计算复杂度下降至 $\mathrm{O}(LNK \cdot w^2)$. 考虑到对于一个标准的输入点云视频序列, $\sqrt{NK}$ 通常会远大于 w 的值, 从而可以大幅降低计算负担, 在静态点云骨干中也可以实现近似线性复杂度的动态点云理解.

2.3　基于时空扫描与渐进式适配的 Mamba 适配器

通过第 2.2 节提出的方法, 可以获得动态点云视频中每一帧的空间特征; 第 2.3 节首次使用 Mamba 适配器来捕捉点云视频中完整的时空动态, 并持续保持线性复杂度. 综合分析 Mamba 适配器在其他领域的现有方法[21] 后发现, 在动态点云视频领域中想要成功应用仍需克服三个障碍: 1) 4D 动态点云展平为 1D 序列时, 时空邻接关系被割裂; 2) Mamba 的递归性质决定了其信息交互只能依赖于过去的隐藏状态 (hidden state), 而无法像 Transformer 一样实现全局建模; 3) 由于点云视频的特征序列相对于静态图像或者静态点云更加复杂, Mamba 架构中早期输入的隐藏状态对于序列的影响将迅速衰减, 长程依赖难以维持.

本文方法提出一种专为动态点云理解而定制的 Mamba 适配器. 与先前直接处理模型产生的特征不同, 为了解决障碍 1), 考虑到点云视频时间维度的连续性和空间维度的离散性, 本节通过一种“洗牌时空扫描曲线”来获取不同扫描曲线的重排变体. 通过希尔伯特曲线和 Z 序曲线产生的六种变体来随机扫描原始序列, 在基础模型冻结的前提下提供上下文关系的不同视角, 从而向静态点云基础模型中注入完整的时空信息. 为了缓解障碍 2) 与障碍 3), 对于任意的点云视频输入特征 PV^i, 提出一种可学习的双向空间状态模型 (bidirectional state space models, BSSM) 层, 通过并行运行两个递归过程, 并使用渐进式适配将时间尺度上不同帧的点云放缩到统一尺度, 从而在保证全局建模的同时弥补单向扫描中可能被忽略的局部关系.

2.3.1　洗牌时空扫描曲线

如图 3 所示, 基于 Mamba 已集成的硬件感知算法, 对输入序列 X^i 分别执行前向 S6[12] 和后向 S6[12]. 前向 S6 从索引方向的第一个点开始进行递归扫描 (由深粉色表示), 后向 S6 从索引方向的最后一个点开始进行递归扫描 (由深绿色表示). 通过

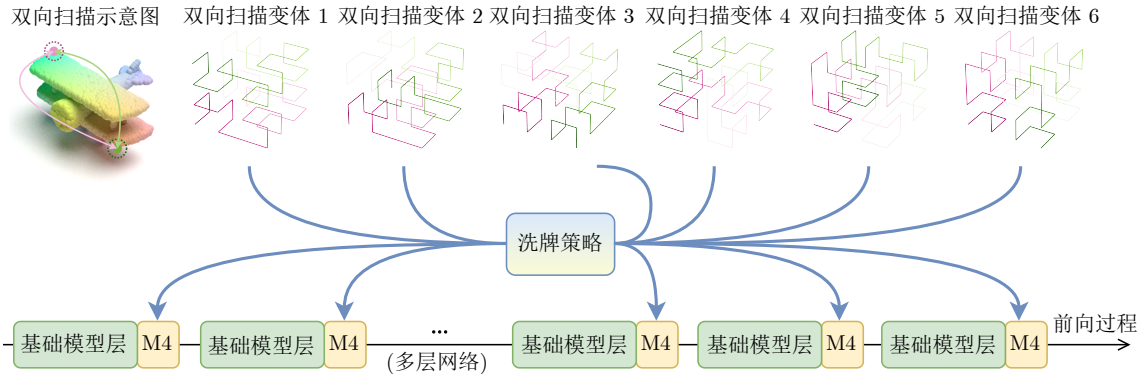

图 3 双向扫描与洗牌策略的示意图

Fig. 3 Schematic diagram of the bidirectional scan and shuffle strategy

这种方式, 输入序列可以实现双向的并行运行, 并扩展点云视频中每一个点的感受野, 缓解长程遗忘问题, 提升模型表现.

扫描曲线通过连续单参数映射将高维空间中的几何结构转化为一维序列. 当其用于点云视频理解时, 高维空间通常指点坐标所在的三维欧几里得空间与一维时序空间的组合. 典型的扫描曲线有逐轴曲线、交叉曲线、希尔伯特曲线和 Z 序曲线等. 先前工作 [20, 23] 已经验证希尔伯特曲线拥有良好的局部特性, 而 Z 序曲线则以高效率扫描著称. 因此, 考虑到点云视频的性质与高效迁移学习的目标, 本节期望以这两种曲线为基准来实现准确的因果推断. 这种简单的先进行空间排序再进行时间排序的方案是次优的. 这是由于在点云视频中, 一些在时空尺度上相距较远的点实际上语义相似度极高甚至完全相同, 而单一的扫描模式使得 Mamba 缺乏从多个视角捕捉时空关系的能力, 这将导致语义关联密切的点在序列扫描后比其他语义无关的点距离更远. 为此, 本节提出一种“洗牌时空扫描曲线”, 如图 3 所示, 本节将原先固定的希尔伯特曲线和 Z 序曲线以坐标轴为优先级构造出六种变体, 优先级顺序为 X-Y-Z-T, 其中 T 作为时间轴, 对变体构建不产生影响. 通过洗牌 (shuffle) 策略, 随机将两种曲线的不同变体分配给不同层的适配器, 使整体模型能够从多个视角推断时空关系, 隐式地提升 Mamba 长程依赖能力, 从而增强迁移学习的泛化能力.

2.3.2 渐进式动态点云序列适配

实验发现, 扫描完成后直接进行迭代会对模型性能造成负面影响. 这是由于在点云视频理解场景中, 时空信息粒度更细, 将整个输出序列全部压缩为单个张量供适配器处理会使模型产生信息冲突. 为了避免静态点云模型现有知识与适配器微调知识之间的混淆, 本节希望在保持线性复杂度的前提下能将其特征有效地融合在一起. 考虑到单个张量在实现全局调节中的局限性, 引入放缩项 Y_1 与偏置项 Y_2, 提出一种专用于动态点云序列的适配方案, 称为“渐进式动态点云序列适配”. 具体来说, 对于式 (4) 中得到的输出 X^i, 将投影层的通道数加倍. 之后, 按通道将输出拆分为两个序列 Y_1^i, Y_2^i, 则有:

$$Y_1^i,\ Y_2^i = \mathrm{BSSM}\left(X^i\right) \tag{5}$$

式中, Y_1^i, Y_2^i 与 X^i 的张量形状相同, 且拥有适配器所蕴含的动态点云信息, 可以充当引导条件, 使整体模型渐进式关注静态交互和宏观动态. 渐进式适配的公式如下:

$$g = \sigma\left(\mathrm{Linear}\left([X,\ Y_1,\ Y_2]\right)\right) \tag{6}$$

$$\mathrm{ProAda}\left(X,\ Y_1,\ Y_2\right) = g \odot \left(Y_1 \odot X + Y_2\right) + (1-g) \odot X \tag{7}$$

式中, σ 为 Sigmoid 函数, g 表示渐进式门控, Y_1 和 Y_2 表示放缩项和偏置项, ProAda 表示渐进式适配, $\odot$ 表示逐元素相乘 (element-wise multiplication). 适配后的特征序列随后将被送入冻结的点云基础模型前馈层进一步处理, 特征处理的公式如下:

$$PV^{i+1} = \mathrm{FFN}\left(\mathrm{ProAda}\left(X^i,\ Y_1^i,\ Y_2^i\right)\right) \tag{8}$$

式中, PV^{i+1} 指的是第 i 层基础模型与 Mamba 适配器串行处理后的特征, 并将被送入第 $i+1$ 层作为输入. 最终, 可以对最后一层的输出特征取平均, 以获得这个点云视频的最终表示:

$$\mathrm{Output} = \mathrm{Average}\left(PV^{Layer}\right) \tag{9}$$

式中, $Layer$ 指点云基础模型的层数. 在迁移学习过程中, 点云基础模型的所有参数都保持冻结, 只有新引入的 Mamba 适配器是可训练的. 整体的架构在端到端微调的基础上最大化保留了点云视频特

征, 避免与静态特征信息产生冲突.

3　实验结果与分析

为了证明 PointM4 的优越性, 本节分别在 3D 动作识别、4D 动作分割、4D 语义分割和手势识别任务上进行了实验. 为了与已有方法进行公平对比, 本文使用三个经典的点云基础模型 Point-BERT[24]、PointMAE[10]、PointGPT[11] 作为基线模型. 将 PointM4 应用于上述三个基线模型中并对这三个基线模型使用了完全相同的实验设置, 均在 ShapeNet[25] 数据集上进行 300 次迭代 (epoch) 的预训练. 在迁移学习过程中, 将静态点云基础模型的权重冻结, 仅更新 PointM4 模块和任务专用头的参数. 选择随机梯度下降作为优化器, 学习率的衰减因子为 0.1. 整个迁移学习共包含 50 次迭代. 所有的实验均在 NVIDIA RTX 4090 上进行.

3.1　3D 动作识别

本节首先在 MSR-Action3D[26] 数据集上评估了本文方法, 以展示其在 3D 动作识别领域中的有效性. MSR-Action3D[26] 数据集由 567 个深度视频组成, 包含 20 种不同的动作类别. 每一个视频平均包含 40 帧. 按照 P4Transformer[3] 的方案将该数据集划分为 270 个训练视频和 297 个测试视频. 为了评估提出的模块在不同时间分辨率下的表现, 本节分别以 12 帧、16 帧、24 帧、32 帧进行了实验, 并在每帧中采样 2 048 个点. 以 2 帧为单位选择锚帧并从中采样 64 个锚点. 采用与 P4Transformer[3] 相同的设置来获取点云视频的特征词元.

表 1 展示了 PointM4 模块迁移学习的结果 (在跨模态迁移学习中的 Point-BERT + M4、PointMAE + M4、PointGPT-S + M4 方法). 与专用于 4D 点云视频理解的 P4Transformer[3] 相比, PointM4 在冻结的静态点云模型上取得了全方位的性能提升. 特别是, 随着帧序列长度的增加, P4Transformer[3] 的结果提升受限, 证明 Transformer 架构在长序列 4D 视频建模方面存在缺陷. 相反, PointM4 的识别准确率随着序列长度的增加稳步提高. 这一结果表明, 将静态点云基础模型的知识通过适配器迁移到动态点云领域是可行的. 与此同时, 与现有的适配器迁移学习方法[9] 相比, 本文方案展示了 Mamba 作为跨模态迁移学习的适配器, 在解决长程依赖和全局感知问题中, 拥有很大的发展潜力.

3.2　4D 动作分割

本节进一步在 HOI4D[27] 数据集上对动作分割任务进行了实验. 与 3D 动作识别不同, 4D 动作分割的目标是对点云视频中的每一帧都分配动作标签. 按照 HOI4D[27] 的官方操作, 将该数据集划分为

表 1　MSR-Action3D 数据集动作识别准确率 (%)

Table 1　Action recognition accuracy on the MSR-Action3D dataset (%)

训练方式	方法	12 帧	16 帧	24 帧	32 帧
有监督训练	MeteorNet[28]	86.53	88.21	88.50	—
	PSTNet[4]	87.88	89.90	91.20	—
	P4Transformer[3]	87.54	89.56	90.94	87.93
	Kinet[29]	88.53	91.92	93.27	—
	PPTr[30]	89.89	90.31	92.33	—
	LeaF[31]	—	91.50	93.84	—
	PST-Transformer[19]	88.15	91.98	93.73	—
	X4D-SceneFormer[32]	—	92.56	93.90	—
	Mamba4D[20]	—	—	93.38	93.10
自监督训练 (端到端微调)	CPR[5]	91.00	92.15	93.03	—
	C2P[33]	—	91.89	94.76	—
	PointCMP[34]	91.58	92.26	93.27	—
	PointCPSC[35]	90.24	92.26	92.68	—
	MaST-Pre[6]	—	—	94.08	—
跨模态迁移学习 (3D 预训练模型适配)	PointCSA[9]	92.04	93.73	95.12	95.47
	Point-BERT + M4	$92.33_{(+0.29)}$	$\mathbf{94.77}_{(+1.04)}$	$95.47_{(+0.35)}$	$96.17_{(+0.70)}$
	PointMAE + M4	$93.37_{(+1.33)}$	$94.08_{(+0.35)}$	$95.12_{(+0.00)}$	$\mathbf{96.86}_{(+1.39)}$
	PointGPT-S + M4	$\mathbf{93.72}_{(+1.68)}$	$94.43_{(+0.70)}$	$\mathbf{95.62}_{(+0.50)}$	$96.52_{(+1.05)}$

2971 个训练场景和 892 个测试场景. 每一个视频场景都由 150 帧组成, 每帧包含 2 048 个点. 该数据集共包含 579 450 帧, 每一帧都标注了 19 类细粒度动作标签的具体类别. 在评估过程中使用帧级准确率、分段编辑距离和带重叠阈值 $k\%$ 的分段 F1 分数 (F1@k) 作为指标.

表 2 展示了 PointM4 在 HOI4D[27] 数据集上的结果 (P4Trans + M4 方法, P4Trans、M4 分别为 P4Transformer、PointM4 的缩写). 与当前 P4Trans 相比, P4Trans + M4 在准确率、编辑距离和 F1@50 的分数上分别提高了 8.9%、1.6 和 13.5%. 本文将这一结果归因于双向时空扫描与渐进式适配策略, 其证明了 PointM4 在潜在空间中学习到了丰富的语义信息以区分高层语义和低层的几何表示, 满足了细粒度理解的需求. 为进一步彰显 PointM4 在时序解析上的优势, 本节与基线模型进行了可视化对比. 如图 4 所示, 基线模型因过度依赖局部表征变化, 在运动连续区域中出现严重过分割, 与文献 [32] 的观察一致. PointM4 凭借双向 Mamba 适配器, 动作边界定位更准确, 并显著抑制了过分割现象, 在保证效率的同时提升了分割精度.

表 2 HOI4D 数据集动作分割准确率
Table 2 Action segmentation accuracy on the HOI4D dataset

训练方式	方法	准确率 (%)	编辑距离	F1@50 (%)
有监督训练	P4Trans[3]	71.2	73.1	58.2
	PPTr[30]	77.4	80.1	69.5
	Mamba4D[20]	85.5	91.3	85.5
自监督训练 (端到端微调)	P4Trans + C2P[33]	73.5	76.8	62.4
	PPTr + C2P[33]	81.1	84.0	74.1
	X4D[32]	84.1	91.1	84.8
	CrossVideo[36]	83.7	86.0	76.0
跨模态迁移学习 (3D 预训练模型适配)	P4Trans + M4	80.1$_{(+8.9)}$	74.7$_{(+1.6)}$	71.7$_{(+13.5)}$

3.3 4D 语义分割

为验证本文方法在细粒度任务上的泛化能力, 本节在 HOI4D[27] 数据集上继续开展了 4D 语义分割实验. 考虑到 GPU 内存限制与长序列时序建模需求, 在微调与测试时采用与基线一致的 3 帧序列. 评估指标为 39 类标签的平均交并比 (mIoU). 表 3 结果显示, 本文方法取得了 42.8% 的 mIoU, 显著优于基线模型[3, 30] 及其改进方案. 这一提升表明, 本文的跨模态迁移范式不仅适用于动作识别等粗粒度任务, 更能通过渐进式适配器将基础模型的强空间先验与 Mamba 的线性时序建模能力深度融合, 从而在 4D 语义分割这类对逐点分类精度要求极高的细粒度任务中, 有效提取高表征能力特征, 缓解因帧数受限导致的信息丢失问题.

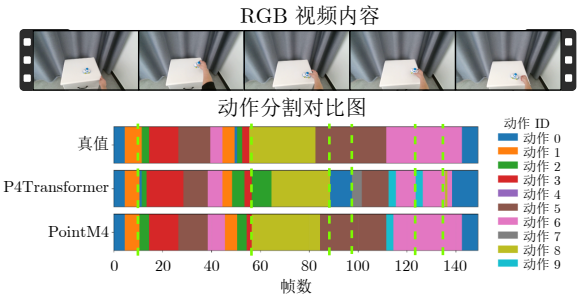

图 4 PointM4 与基线模型的动作分割可视化对比
Fig. 4 Visualization comparison of action segmentation between PointM4 and baseline models

表 3 HOI4D 数据集语义分割准确率
Table 3 Semantic segmentation accuracy on the HOI4D dataset

方法	输入帧数	mIoU (%)
P4Trans[3]	3	40.1
P4Trans + C2P[33]	3	41.4
P4Trans + CrossVideo[36]	3	42.1
PPTr[30]	3	41.0
PPTr + C2P[33]	3	42.3
P4Trans + M4	3	**42.8**$_{(+2.7)}$

3.4 手势识别

为验证模型的泛化能力, 本节在 SHREC′17[37] 数据集上进行手势识别的实验. SHREC′17[37] 由 2800 个视频组成, 包含 28 个类别. 按照 PointCSA[9] 中的方案, 将数据集中的 70% (1960 个视频) 划分为训练集, 其余 30% (840 个视频) 划分为测试集. 采用了与 MSR-Action3D[26] 中相同的架构和超参数来进行实验. 如表 4 所示, 与现有的自监督和有监督方案相比, 本文方案取得了更高的识别精度: 以 PointGPT[11] 为基础模型达到了 96.7% 准确率, 以 Point-BERT[24] 为基础模型达到了 96.4% 准确率, 以 PointMAE[10] 为基础模型达到了 96.3% 准确率. SHREC′17 实质上是一项噪声较多的细粒度任务, 需要准确识别手指和定位关节, 而这正是 3D 预训练模型的优势. 本文方法利用静态点云与动态点云之间的空间一致性, 成功将静态 3D 基础模型应用于多种下游任务.

3.5 消融实验

为了验证网络中不同模块的功能和有效性, 本

文对 PointM4 中的网络模块和策略在 MSR-Action3D[26] 数据集上进行了广泛的消融实验.

滑动窗口尺寸. 表 5 展示了不同窗口大小对模型性能及微调效率的影响. 实验表明, 无重叠滑动窗口操作可以在不改变注意力权重的前提下带来精度和速度的双重提升. 本文还有两点发现: 第一, 在整个点云上执行全局注意力的表现比窗口划分相对更差. 本文推测这是由于 PointMAE[10] 最初是在静态点云数据集上进行的预训练, 动态点云视频超出了其预训练域. 第二, 单纯增加窗口尺寸不能够给下游任务带来持续的性能提升. 这一现象印证了有关局部感受野的推测, 并证明了高频率小尺度注意力在动态点云迁移学习中的必要性.

Mamba 扫描策略. 本文的双向扫描策略与洗牌时空扫描策略可以引入随机性和多样性, 从而让每个 Mamba 模块从多个视角推断时空关系. 对六种变体及其洗牌顺序进行比较来验证该策略的有效性. 表 6 证明仅使用双向扫描策略即可实现 +1.9% 的性能提升, 这验证了通过全局上下文建模点云视频的积极影响. 此外, 洗牌策略比固定顺序在下游任务上拥有更好性能, 有效地打破了传统 Mamba 特征学习框架的局限性. 同时, 与其他现有点云排序方法[20, 23] 相比, 本文也观察到, 随着变体种类的增多, 模型的整体性能呈现上升趋势, 说明由希尔伯特曲线和 Z 序曲线产生的多样化变体对于提升静态点云模型理解复杂时空结构的能力确有裨益.

表 4　SHREC′17 数据集手势识别准确率 (%)
Table 4　Gesture recognition accuracy on the SHREC′17 dataset (%)

训练方式	方法	准确率
有监督训练	PLSTM-base[38]	87.6
	PLSTM-early[38]	93.5
	PLSTM-PSS[38]	93.1
	PLSTM-middle[38]	94.7
	PLSTM-late[38]	93.5
	Kinet[29]	95.2
自监督训练 (端到端微调)	PointCMP[34]	93.3
	MaST-Pre[6]	92.4
跨模态迁移学习 (3D 预训练模型适配)	Point-BERT + CSA[9]	96.2
	PointMAE + CSA[9]	95.2
	PointGPT-S + CSA[9]	96.5
	Point-BERT + M4	**96.4**$_{(+0.2)}$
	PointMAE + M4	**96.3**$_{(+1.1)}$
	PointGPT-S + M4	**96.7**$_{(+0.2)}$

表 5　滑动窗口尺寸对注意力效率以及模型性能的影响
Table 5　The impact of sliding window size on attention efficiency and model performance

模型	窗口尺寸 (点)	每秒帧数 (FPS)	注意力占比 (%)	准确率 (%)
A0	640 × 240 × 3	102.9	54.5	94.4
A1	16 × 16 × 3	110.7	36.9	95.4
A2	8 × 8 × 3	112.5	36.6	**96.8**
A3	4 × 4 × 3	113.2	36.5	96.2

表 6　Mamba 扫描策略及变体种类对模型性能的影响 (%)
Table 6　The impact of Mamba scanning strategy and variant types on model performance (%)

模型	双向扫描	洗牌策略	六种变体						准确率
			XYZ	*YXZ*	*XZY*	*ZXY*	*YZX*	*ZYX*	
B0	×	✓	✓	✓	✓	✓	✓	✓	94.9
B1	✓	×	✓	✓	✓	✓	✓	✓	94.2
B2	✓	✓	✓	✓	✓	✓	✓	✓	**96.8**
B3	✓	✓	✓	✓	×	×	✓	✓	95.8
B4	✓	✓	×	×	✓	✓	✓	✓	95.1
B5	✓	✓	✓	✓	×	×	×	×	94.9
B6	✓	✓	×	×	✓	✓	×	×	94.7
B7	✓	✓	×	×	×	×	✓	✓	94.7
B8	×	×	×	×	×	×	×	×	92.3

渐进式动态点云序列适配. 表 7 展示了适配器的潜在适配方法. 如第 2.3 节所述, 直接将 Mamba 模块插入到静态点云基础模型中的性能是次优的. 与此同时, 像相加和最大值这类简单适配方法的表现甚至比直接插入更差. 这表明输出序列导致的信息混淆必须通过精巧的特征融合方法来缓解. 渐进式适配方案通过类似“软门控”机制来避免过多地改变静态点云基础模型的特征分布, 从而使模型逐步拥有跨模态迁移学习能力. 为验证渐进式适配模块可以缓解静态基础模型与动态适配器之间的信息冲突, 本节分别提取 C1 与 C4 两种融合策略在 MSR-Action3D 数据集上进行了 t-SNE 特征可视化. 图 5 显示, 直接相加导致静态空间特征与动态时序特征分布存在一定程度的混叠, 表明两种异构信息相互干扰. 而 C4 模型利用渐进式适配, 避免分布混淆, 实现基础先验与时序补充的协同演化.

适配器与基础模型的级联顺序. 考虑到主流静态点云基础模型基本以 Transformer 架构为主, 以 PointMAE[10] 为例, 提供了几种架构设计方案. 为方便理解, 将 Transformer 层表示为 [T], Mamba 模块表示为 [M]. 从表 8 中发现 D1 模型所取得的性能较差, 其不符合适配器微调的设计理念与初衷. 本文选择的架构 D0 模型可以更好地平衡这两种架构, 使其在下游任务中展示出更好的协同能力.

训练时长与内存占用. 本文比较了 PointM4 与

表 7 适配方法对模型性能的影响
Table 7 The impact of adaptation methods on model performance

模型	适配方法	具体操作	准确率 (%)
C0	跳过	X	92.4
C1	相加	$X + Y$	92.3
C2	最大值	$\max(X,\ Y)$	90.2
C3	拼接	$\text{Linear}([X,\ Y])$	95.5
C4	渐进式适配	$\text{ProAda}\,(X,\ Y_1,\ Y_2)$	**96.8**

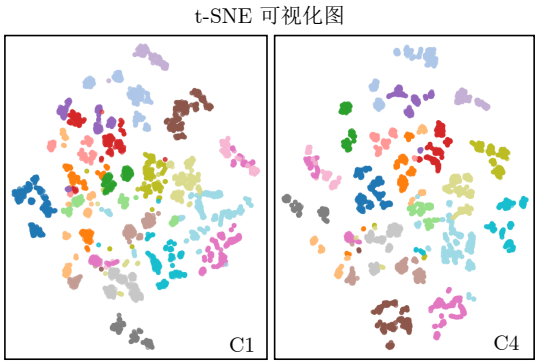

图 5 两种适配方法的 t-SNE 可视化
Fig. 5 t-SNE visualization of two adaptation methods

表 8 级联顺序对模型性能的影响
Table 8 The impact of cascade order on model performance

模型	是否适配	级联顺序	准确率 (%)
D0	✓	[TM] × 12	**96.8**
D1	×	[T] × 12 [M] × 12	90.6
D2	×	[T] × 6 [TMM] × 6	92.5
D3	✓	[TTMM] × 6	93.8

经典的动态点云迁移学习方案[9] 的训练时长与 GPU 内存占用. 如图 6 所示, 由于现有方案多数使用原生 Transformer 架构, 其内存的使用量会随着输入帧数的增加而迅速上升, 在 32 帧时会出现显存耗尽 (out of memory, OOM). 相比之下, 本文方法得益于 Mamba 的轻量化设计, 内存消耗更为缓慢, 同时也取得了更快的训练速度.

整体推理效率. 为验证所提方法在实际部署中的实时性, 表 9 对比了 PointM4 与 PointCSA 的端到端推理耗时分解. 虽然本文方法引入的窗口划分与六种扫描变体使预处理时间增加至 4.8 ms 左右, 但核心模型推理时间从 6.7 ms 降至 3.2 ~ 3.8 ms, 降幅超 43%, 这直接体现了线性复杂度对计算瓶颈的缓解. 更重要的是, 当输入帧数增至 36 帧时, PointCSA 因显存耗尽完全失效, 而 PointM4 总延迟仅 9.8 ms, 准确率进一步提升至 96.8%, 充分说明预处理开销是一次性常数代价, 而核心推理的可扩展性才是支撑长序列应用的关键.

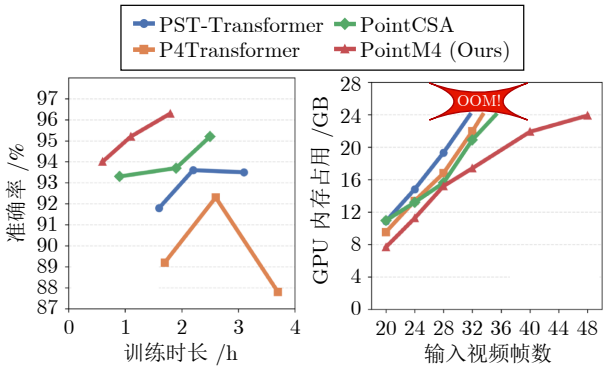

图 6 PointM4 与现有方法在训练时长与 GPU 占用上的对比
Fig. 6 Comparison of training time and GPU memory usage between PointM4 and existing methods

表 9 整体推理流程的时间分析
Table 9 Overall reasoning process time analysis

方法	帧数	预处理时间 (ms)	模型推理时间 (ms)	后处理时间 (ms)	准确率 (%)
PointCSA	24	3.3	6.7	0.9	95.1
	36	不适用 (OOM)			
PointM4	24	4.8	3.2	0.9	95.1
	36	5.0	3.8	1.0	96.8

4 结束语

本文提出一种基于渐进式 Mamba 适配器的动态点云迁移学习模型. 该模型通过无重叠滑动窗口注意力精准控制基础模型的感受野, 并将传统注意力机制的时间复杂度降低至线性级别. 考虑到点云视频的特殊性质, 本文引入一种拥有洗牌时空扫描曲线的双向 Mamba 适配器. 该适配器以线性复杂度为基础模型, 提供多视角的长程建模能力, 在多个下游任务中表现出优秀的性能. 尽管同步进行了一系列消融实验来验证本文方法的有效性和稳定性, 但目前该方法仍受限于固定的窗口尺寸, 对于跨数据集的泛化能力有待进一步验证. 未来的工作将探索实现自适应窗口以及更灵活协同方式的可能, 并尝试将该方法扩展到更具挑战性的室外运动场景中.

参考文献

1 Wang Yao-Nan, Hua He-An, Zhang Hui, Zhong Hang, Fan Ye-Xin, Liang Hong-Tao, et al. Performance function-guided deep reinforcement learning control for UAV swarm. *Acta Automatica Sinica*, 2025, **51**(5): 905−916
(王耀南, 华和安, 张辉, 钟杭, 樊叶心, 梁鸿涛, 等. 性能函数引导的无人机集群深度强化学习控制方法. 自动化学报, 2025, **51**(5): 905−916)

2 Tian Yong-Lin, Shen Yu, Li Qiang, Wang Fei-Yue. Parallel point clouds: Point clouds generation and 3D model evolution via virtual-real interaction. *Acta Automatica Sinica*, 2020, **46**(12): 2572−2582
(田永林, 沈宇, 李强, 王飞跃. 平行点云: 虚实互动的点云生成与三维模型进化方法. 自动化学报, 2020, **46**(12): 2572−2582)

3 Fan H H, Yang Y, Kankanhalli M. Point 4D transformer networks for spatio-temporal modeling in point cloud videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 14199−14208

4 Fan H H, Yu X, Ding Y H, Yang Y, Kankanhalli M S. PSTNet: Point spatio-temporal convolution on point cloud sequences. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2021.

5 Sheng X X, Shen Z Q, Xiao G. Contrastive predictive autoencoders for dynamic point cloud self-supervised learning. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI, 2023. Article No. 1101

6 Shen Z Q, Sheng X X, Fan H H, Wang L G, Guo Y L, Liu Q, et al. Masked spatio-temporal structure prediction for self-supervised learning on point cloud videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 16534−16543

7 Sun Y D, Cheng H Z, Lu C Y, Li Z Q, Wu M H, Lu H M, et al. HyperPoint: Multimodal 3D foundation model in hyperbolic space. *Pattern Recognition*, 2026, **173**: Article No. 112800

8 Sun Y D, Zhu J H, Cheng H Z, Lu C Y, Yang Z C, Chen L, et al. Align then adapt: Rethinking parameter-efficient transfer learning in 4D perception. arXiv preprint arXiv: 2602.23069, 2026.

9 Lv B X, Zha Y, Dai T, Yuerong X, Chen K, Xia S T. Adapting pre-trained 3D models for point cloud video understanding via cross-frame spatio-temporal perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 12413−12422

10 Pang Y T, Wang W X, Tay F E H, Liu W, Tian Y H, Yuan L. Masked autoencoders for point cloud self-supervised learning. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 604−621

11 Chen G Y, Wang M L, Yang Y, Yu K, Yuan L, Yue Y F. PointGPT: Auto-regressively generative pre-training from point clouds. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1291

12 Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv: 2312.00752, 2024.

13 Qi C R, Su H, Mo K C, Guibas L J. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 77−85

14 Xie S N, Gu J T, Guo D M, Qi C R, Guibas L, Litany O. PointContrast: Unsupervised pre-training for 3D point cloud understanding. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 574−591

15 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 9892−9902

16 Zhang R R, Guo Z Y, Fang R Y, Zhao B, Wang D, Qiao Y, et al. Point-M2AE: Multi-scale masked autoencoders for hierarchical point cloud pre-training. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1962

17 Choy C, Gwak J, Savarese S. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 3070−3079

18 Deng Z C, Li X T, Li X, Tong Y H, Zhao S, Liu M Y. VG4D: Vision-language model goes 4D video recognition. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 5014−5020

19 Fan H H, Yang Y, Kankanhalli M. Point spatio-temporal transformer networks for point cloud video modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(2): 2181−2192

20 Liu J M, Han J R, Liu L H, Aviles-Rivero A I, Jiang C K, Liu Z, et al. Mamba4D: Efficient 4D point cloud video understanding with disentangled spatial-temporal state space models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 17626−17636

21 Wang Z C, Chen Z H, Wu Y M, Zhao Z, Zhou L P, Xu D. PoinTramba: A hybrid Transformer-Mamba framework for point cloud analysis. arXiv preprint arXiv: 2405.15463, 2024.

22 Li Y Y, Bu R, Sun M C, Wu W, Di X H, Chen B Q. PointCNN: Convolution on *X*-transformed points. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc., 2018. 828−838

23 Han X, Tang Y, Wang Z X, Li X Z. MAMBA3D: Enhancing local features for 3D point cloud analysis via state space model. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024. 4995−5004

24 Yu X M, Tang L L, Rao Y M, Huang T J, Zhou J, Lu J W. 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. New Orleans, USA: IEEE, 2022. 19291−19300

25 Chang A X, Funkhouser T, Guibas L, Hanrahan P, Huang Q X, Li Z M, et al. ShapeNet: An information-rich 3D model repository. arXiv preprint arXiv: 1512.03012, 2015.

26 Li W Q, Zhang Z Y, Liu Z C. Action recognition based on a bag of 3D points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. San Francisco, USA: IEEE, 2010. 9−14

27 Liu Y Z, Liu Y, Jiang C, Lyu K B, Wan W K, Shen H, et al. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 20981−20990

28 Liu X Y, Yan M Y, Bohg J. MeteorNet: Deep learning on dynamic 3D point cloud sequences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 9245−9254

29 Zhong J X, Zhou K C, Hu Q Y, Wang B, Trigoni N, Markham A. No pain, big gain: Classify dynamic point cloud sequences with static models by fitting feature-level space-time surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 8500−8510

30 Wen H, Liu Y Z, Huang J W, Duan B, Yi L. Point primitive transformer for long-term 4D point cloud video understanding. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 19−35

31 Liu Y Z, Chen J Y, Zhang Z K, Huang J W, Yi L. LeaF: Learning frames for 4D point cloud sequence understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 604−613

32 Jing L L, Xue Y, Yan X, Zheng C D, Wang D, Zhang R M, et al. X4D-SceneFormer: Enhanced scene understanding on 4D point cloud videos through cross-modal knowledge transfer. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 2670−2678

33 Zhang Z Y, Dong Y H, Liu Y Z, Yi L. Complete-to-partial 4D distillation for self-supervised point cloud sequence representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 17661−17670

34 Shen Z Q, Sheng X X, Wang L G, Guo Y L, Liu Q, Zhou X. PointCMP: Contrastive mask prediction for self-supervised learning on point cloud videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 1212−1222

35 Sheng X X, Shen Z Q, Xiao G, Wang L G, Guo Y L, Fan H H. Point contrastive prediction with semantic clustering for self-supervised learning on point cloud videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 16469−16478

36 Liu Y Z, Chen C X, Wang Z F, Yi L. CrossVideo: Self-supervised cross-modal contrastive learning for point cloud video understanding. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 12436−12442

37 de Smedt Q, Wannous H, Vandeborre J P, Guerry J, le Saux B, Filliat D. 3D hand gesture recognition using a depth and skeletal dataset: SHREC′17 track. In: Proceedings of the Workshop on 3D Object Retrieval. Lyon, France: ACM, 2017. 33−38

38 Min Y C, Zhang Y X, Chai X J, Chen X L. An efficient PointLSTM for point clouds based gesture recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 5760−5769

孙一丁 西安交通大学软件学院硕士研究生. 2024 年获得西安工业大学计算机科学与技术学院学士学位. 主要研究方向为点云理解.
E-mail: sunyiding@stu.xjtu.edu.cn
(**SUN Yi-Ding** Master student at the School of Software Engineering, Xi′an Jiaotong University. He received his bachelor degree from the School of Computer Science and Technology, Xi′an Technological University in 2024. His main research interest is point cloud understanding.)

祝继华 西安交通大学软件学院教授. 2004 年获得中南大学自动化专业学士学位. 2011 年获得西安交通大学控制科学与工程专业博士学位. 主要研究方向为计算机视觉与机器学习. 本文通信作者.
E-mail: zhujh@xjtu.edu.cn
(**ZHU Ji-Hua** Professor at the School of Software Engineering, Xi′an Jiaotong University. He received his bachelor degree in automation from Central South University in 2004 and his Ph.D. degree in control science and engineering from Xi′an Jiaotong University in 2011. His research interests include computer vision and machine learning. Corresponding author of this paper.)

王耀南 中国工程院院士, 湖南大学人工智能与机器人学院教授. 1995 年获得湖南大学博士学位. 主要研究方向为机器人学, 智能控制和图像处理.
E-mail: yaonan@hnu.edu.cn
(**WANG Yao-Nan** Academician at Chinese Academy of Engineering, professor at the School of Artificial Intelligence and Robotics, Hunan University. He received his Ph.D. degree from Hunan University in 1995. His research interests include robotics, intelligent control, and image processing.)

程浩喆 西安交通大学软件学院博士研究生. 2022 年获得西安工程大学电子信息学院硕士学位. 主要研究方向为点云处理.
E-mail: chz97@stu.xjtu.edu.cn
(**CHENG Hao-Zhe** Ph.D. candidate at the School of Software Engineering, Xi′an Jiaotong University. He received his master degree from the School of Electronic Information, Xi′an Polytechnic University in 2022. His main research interest is point cloud processing.)

卢超翼 西安交通大学软件学院硕士研究生. 2024 年获得浙江科技大学计算机科学与技术学院软件工程专业学士学位. 主要研究方向为计算机视觉.
E-mail: chaoyi@stu.xjtu.edu.cn
(**LU Chao-Yi** Master student at the School of Software Engineering, Xi′an Jiaotong University. He received his bachelor degree in software engineering from the School of Computer Science and Technology, Zhejiang University of Science and Technology in 2024. His main research interest is computer vision.)

陈 林 西安交通大学软件学院助理教授. 2018 年获得西北师范大学电子信息工程专业学士学位. 2021 年获得中国科学院大学计算机技术专业硕士学位. 2025 年获得湖南大学控制科学与工程专业博士学位. 主要研究方向是多机器人系统、深度强化学习和视觉导航. E-mail: chenlin21@hnu.edu.cn
(**CHEN Lin** Assistant professor at the School of Software Engineering, Xi′an Jiaotong University. He received his bachelor degree in electronic information engineering from Northwest Normal University in 2018, his master degree in computer technology from the University of Chinese Academy of Sciences in 2021, and his Ph.D. degree in control science and engineering from Hunan University in 2025. His research interests include multi-robot systems, deep reinforcement learning, and visual navigation.)