极坐标分层曼巴：使用点云作为自我中心序列进行流式 LiDAR 目标检测

Mellon M. Zhang **Glen Chou**[†] **Saibal Mukhopadhyay**[†]
Georgia Institute of Technology
{ meilongz,chou,smukhopadhyay6 } @gatech.edu [*]

Abstract

准确且高效的目标检测对于自动驾驶车辆至关重要，因为实时感知需要低延迟和高吞吐量。LiDAR 传感器提供了稳健的深度信息，但传统方法在单次扫描中处理完整的 360ř 扫描，导致显著的延迟。流式处理方法通过在本地极坐标系中顺序处理部分扫描来解决此问题，但它们依赖于与极几何不对齐的平移不变卷积——导致性能下降或需要复杂的畸变减缓。最近基于 Mamba 的状态空间模型（SSM）在 LiDAR 感知方面表现出潜力，但仅限于完整扫描设置，依赖于几何序列化和位置嵌入，这在内存中很密集且不适合流式处理。我们提出极坐标层次 Mamba（PHiM），这是一种为极坐标流式 LiDAR 设计的新型 SSM 架构。PHiM 使用局部双向 Mamba 块进行扇区内空间编码，并使用全局前向 Mamba 进行扇区间时间建模，采用畸变感知的、维度分解的操作来替代卷积和位置编码。在 Waymo Open Dataset 上，PHiM 在流式检测器中树立了新的技术水平，比之前的最佳性能高出 10 %，并在两倍吞吐量下与完整扫描基线相匹配。代码将在 https://github.com/meilongzhang/Polar-Hierarchical-Mamba 提供。

1 介绍

目标检测是自动驾驶汽车（AVs）的一项关键任务，不仅要求高精度，还要求在真实驾驶的不确定性情况下具有鲁棒性和效率。为了增强感知，大多数自动驾驶汽车引入了 LiDAR 传感器，在低光和远距离场景中表现优于其他类型的传感器。传统的 LiDAR 方法 [**? ? ? ? ? ? ?**] 在单次处理完整或聚合的点云扫描 [**? ? ?**] 时，导致了显著的延迟——大约在数百毫秒的数量级——从而使得实时检测具有挑战性。

为了解决这一问题，近期的研究探索了流式方法，这些方法在部分 LiDAR 扫描数据到达时顺序处理，从而避免了全局聚合 [**? ? ? ?**] 。这些方法通常直接在 LiDAR 传感器的原生极坐标系中运行，以节省计算和内存 [**?**] 。然而，尽管极坐标高效，它们引入了空间失真，违反了标准卷积骨干的假设 [**?**] 。先前的工作试图通过坐标变换或辅助模块事后减轻这些失真 [**? ? ?**] ，但通常导致模型复杂性增加或性能下降。

标准卷积与极坐标的不兼容性激发了对替代架构的探索。一个有前景的方向是 Mamba 状态空间模型 [**? ?**] ，它因在不依赖平移不变性的情况下表现出强大性能和接近线性的推理扩展而受到关注。最近的研究表明，Mamba 在各种流媒体模式 [**? ? ?**] 中的有效性。然而，流式激光雷达提出了一个独特的挑战：它将顺序结构与复杂的 3D 空间几何结合在一起，使得现有的 Mamba 设计不适用于纯粹的顺序数据。本质上，虽然 Mamba 擅长于对时间序列进行建模，但它捕捉空间结构，尤其是 3D 结构的能力仍然有限 [**?**] 。

最近的研究尝试将 Mamba 应用于 LiDAR 感知 [**? ?**] ，但仅限于全扫描设置，将 Mamba 视为类似于卷积或注意力的通用处理模块。为了增强空间意识，这些方法使用几何曲线如 Hilbert [**?**] 或 Z 序 [**?**] 序列化点云，保留局部结构但在推断时需要预计算的模式和额外的存储空间。它们还结合了手工制作的位置编码与可学习的映射，增加了计算和参数。然而，

[*†] Equal advising

Preprint. Under review.

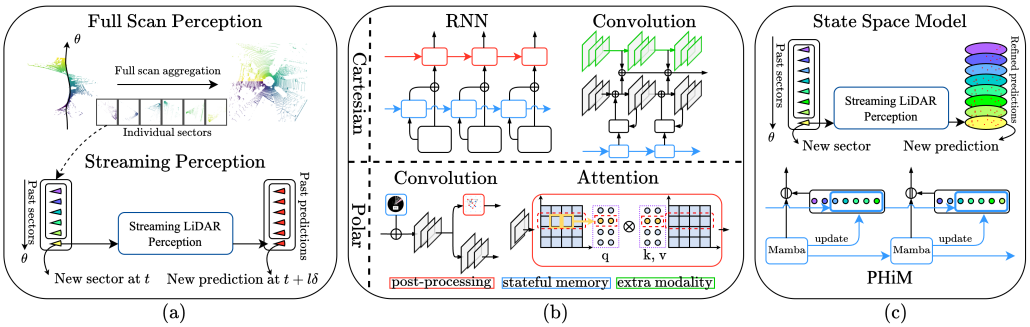


Figure 1: 与现有的流式工作比较。(a) 不同于从 LiDAR 传感器的全方位旋转中聚合点形成完整点云的全扫描方法，流式方法在数据收集时处理部分点云扇区。大多数方法基于这些部分输入扇区进行部分预测。(b) 以前的流式方法包括基于 RNN 或卷积操作于笛卡尔坐标系的模型（上），以及使用极坐标与卷积或注意力机制的模型（下）。这些方法依赖于有状态记忆、后处理或辅助模态的组合来对新旧扇区之间的时空相关性进行建模。(c) PHiM 直接将时空交互编码到状态空间模型（SSM）的隐藏状态中，同时存储扇区级特征。这使得从单个扇区进行全场景预测成为可能，并且不需要额外的模态、先验上下文填充或后处理。

这些方法在流模式或极坐标设置中转移效果不佳：几何曲线假设正方形网格，这与部分扇形的矩形形状不匹配，而位置编码依赖于平移不变窗口 [**?**]，这在极坐标空间中会失效。

总体而言，Mamba 架构在流式 LiDAR 中具有很大的潜力，但现有基于 Mamba 的模型却有所欠缺——要么完全忽略空间结构，要么依赖于在极坐标中表现不佳且代价高昂的偏差重的启发式方法。在本文中，我们介绍了极坐标分层 Mamba（PHiM），这是一种弥补这一空白的全新架构，具有通用且有效的时空设计。PHiM 结合了一个分层结构，可以捕获本地扇区内的细节和全球扇区间的上下文，并使用维度分解卷积来增强空间感知，同时避免极坐标失真。此外，一个后骨干特征缓冲区使得可以在不依赖序列化曲线或位置编码的情况下进行连续的预测优化。该设计支持快速的流水线推理，并在流式模型中实现了最新的性能表现。

我们的主要贡献如下：

- 我们提出了极地层次曼巴（Polar Hierarchical Mamba，PHiM），这是第一个基于状态空间模型（SSM）的架构，专门为使用极坐标的流媒体 LiDAR 感知而设计。与之前基于卷积的流媒体架构不同，我们摒弃了平移不变的处理，以实现最先进的检测性能。
- 我们设计了一种新颖的基于状态空间模型的 PHiM 模块，该模块能够实现快速的局部和全局特征学习以及流式推理。我们的设计有效地捕捉了局部扇区内的空间细节和全球扇区间的时空信息，并通过方位角使用轻量化的串行化方式，而不是昂贵的几何启发式方法，与全扫描 Mamba 架构相比，这样的设计减少了峰值内存使用。
- 我们引入了维度分解卷积（DDCs），以提高 Mamba 的空间局部性感知，同时避免极性扭曲的 (r, θ) 平面。DDCs 有效地在更具平移不变性的维度上提取局部扇区特征，以更轻量的方法取代带有归纳偏差和较多参数的位置信嵌入方案。结合我们的 PHiM 块架构，我们的贡献使模型参数相比全扫描 Mamba 架构减少了近 2 倍。
- 我们在 Waymo Open 数据集上对我们的方法进行了全面的比较，比较对象为现有的全扫描和流式检测器，展示了流式检测方面的新领先性能。我们的方法在对比全扫描方法（使用笛卡尔坐标）时也表现出具有竞争力的性能，且端到端延迟减半。

2 相关工作

全扫描方法。 传统的 LiDAR 物体检测方法依赖于在处理之前聚合整个 360 度点云扫描，导致高延迟并限制了其在实时自动驾驶中的适用性。早期的方法如 VoxelNet [**?**] 对全场景点云进行体素化并应用密集的 3D 卷积。为了减少计算负载，方法如 PointPillar [**?**] 及其后续方法 [**? ?**] 将数据投影到二维鸟瞰图（BEV）并使用二维卷积。SECOND [**?**] 及其扩展方法引入了稀疏和子流形卷积 [**? ?**]，以利用 LiDAR 的固有稀疏性，大大提高了效率。

CenterPoint [**?**] 通过将物体建模为中心点进一步提高了检测精度，激发了许多后续方法 [**? ? ?**] 来优化空间定位。基于 Transformer 的架构 [**? ? ? ? ? ?**] 和最近的基于 Mamba 的模型 [**? ? ?**] 通过超越卷积技术获得了关注。例如，SWFormer [**?**] 在基于 pillar 的网格中引入了局部窗口注意力，而 DSVT [**?**] 则在体素化数据上应用了分组注意力。

尽管 Mamba 在序列建模中提供了有前景的可扩展性和性能，但现有对 LiDAR [**?** **?**] 的应用仍在完整扫描上运行，并将 Mamba 视为传统处理流水线中的通用块。这些工作使用 Hilbert [**?**] 或 Z 曲线 [**?**] 等几何启发式方法对 3D 点云进行序列化，并通过位置编码 [**?**] 增强以弥补空间感知的缺失。然而，这些手工启发式方法引入了不必要的归纳偏差，增加了内存使用量，并且不太适合极坐标流数据，因为后者是在狭窄的楔形区域而非完整矩形网格中观察场景。相反，我们的方法重新思考了 Mamba 的应用方式，通过本地建模以自我为中心的扇区序列，消除了对序列化、手工嵌入或完整扫描输入的需求。

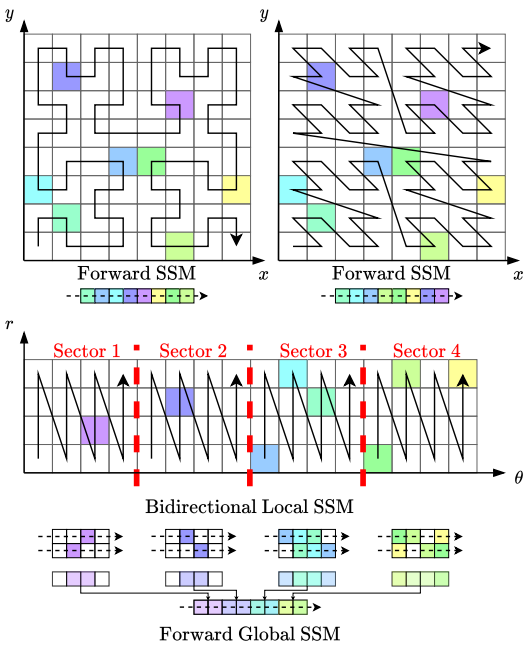


Figure 2: 与现有的 LiDAR Mamba 研究的比较。（上）之前基于 Mamba 的 LiDAR 模型处理完整的点云扫描，并使用昂贵的几何启发式方法，如 Hilbert（左）或 Z 曲线（右）来保持空间局部性。（下）我们的方法通过扇区到达的时间简单地进行序列化，以在传感器旋转过程中分散内存使用和计算，而不是在一个峰值中集中处理。

流式方法。 与一次处理所有聚合的激光雷达扫描的全扫描方法不同，流式方法在传感器发出时即对部分激光雷达扇区——完整 360ř 视野的角度切片进行操作。这使得低延迟、实时感知成为可能，因为模型可以在完整扫描完成之前开始处理。例如，Han [**?**] 引入了一种早期流式管道，该管道使用 LSTMs 和有状态非最大抑制 (NMS) 来保持传入扇区的时间上下文。STROBE [**?**] 通过多尺度聚合增强了空间记忆并结合高清地图，以恢复部分观测中缺失的上下文。

为了进一步提高效率，基于极坐标的方法更贴近旋转激光雷达传感器的原始输出，并减少内存开销。PolarStream [**?**] 批评了将笛卡尔体素应用于极坐标数据的低效性，并提出了沿射程分层的卷积以及失真校正的双线性采样。PARTNER [**?**] 在此基础上扩展了交叉注意力和几何感知的自注意力，以更好地模拟极坐标空间中的空间关系。尽管有这些进步，这些方法仍然紧密耦合于卷积网络，并需要在事后进行修正，例如上下文填充或辅助模块，以缓解以自我为中心视角中固有的空间失真。

然而，这些流处理方法都未利用状态空间模型，也没有挑战诸如平移不变性这样的架构假设——这是一个不太适合极地数据的假设。在这项工作中，我们提出了极地层次贪吃蛇（PHiM），这是首个为流式 LiDAR 专门构建的基于贪吃蛇 (Mamba) 的架构。PHiM 直接对 LiDAR 扇区的自中心时间进程进行建模，使用维度分解卷积来增强空间感知，同时避免失真严重的平面。我们的方法避免了基于曲线的序列化、位置嵌入和空间校正模块，从而形成更简单、更快和更通用的架构。与现有流方法的比较在图 1 中显示，与现有 LiDAR 贪吃蛇方法的比较在图 2 中显示。

我们的方法受到自我中心的时空 LiDAR 感知中四个关键挑战的启发：我们的极坐标分层 Mamba (PHiM) (图 **??**)，通过 Mamba 进行时空学习 (Sec. 2.1)，采用维度分解卷积以避免变形严重的平面 (Sec. 2.2)，采用层次结构以实现多级特征提取 (Sec. 2.3)，以及极坐标到笛卡尔坐标的映射 (Sec. **??**)，结合扇区缓存 (Sec. **??**) 进行主干后细化，来应对这些挑战。

2.1 曼巴

流式 LiDAR 目标检测需要在部分点云区域上进行实时推理，这些点云区域是随着旋转 LiDAR 传感器扫描其周围环境而依次收集的。先前的方法通常对每个区域应用一个独立的模型，并依赖额外的机制来弥补部分视图中固有的有限上下文。例如，Han [**?**] 使用显式记忆模块来存储跨时间步的信息，STROBE [**?**] 整合辅助模态（例如，高清地图）以获得更丰富的场景理解，PolarStream [**?**] 则利用上下文填充来恢复相邻区域边界处的物体。相反，我们利用 Mamba 状态空间模型 [**?**] 来自然地将流式 LiDAR 建模为自我中心区域的时间序列。PHiM 并不依赖外部存储或上下文重构，而是通过 Mamba 的隐藏状态隐式地维护时间信息，使模型能够顺畅地跨区域边界前进相关特征。这提供了一种轻量但表现力丰富的时空建模

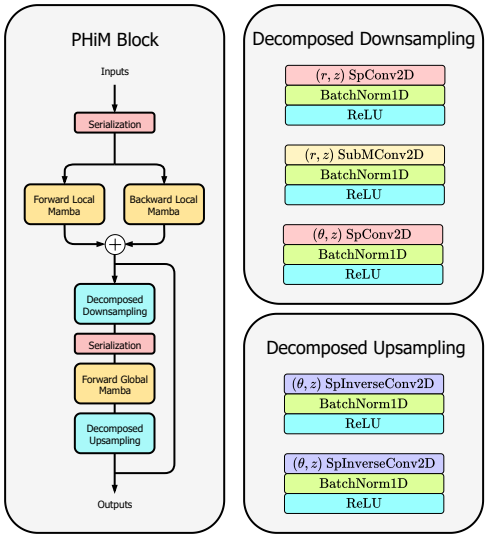
解决方案，无需辅助模块或手工制作的融合方案的负担。有关 Mamba 状态空间公式的详细解释，请参阅附录。

2.2 维度分解卷积

流式 LiDAR 方法通常采用极坐标表示以减少立方体体素化 [**?**] 的计算和内存开销，同时更好地与点云 [**?**] 的各向异性稀疏性对齐。然而，这种表示引入了显著的特征失真。如图 **??** 所示，相似物理尺寸的物体可能在极坐标视图中被严重扭曲。这种失真削弱了基于卷积的方法所依赖的平移不变性的假设，降低了卷积核的泛化能力。尽管如此，卷积仍然因其速度、效率及扩展有效感受野以建模远程依赖关系的能力——处理稀疏 LiDAR 数据的基本特性——而普遍存在。因此，以前基于极坐标的方法保留了卷积骨干网，但需要大量的后处理，例如参数化双线性采样 [**?**] 或关键点注意 [**?**] ，以补偿累积的失真。

尽管利用无卷积的 Mamba 骨干网络以减少归纳偏差，我们仍保留卷积用于降采样，以捕获非空区域之间的长距连接。为减少失真，我们避免在失真严重的平面上进行卷积。我们通过笛卡尔空间和极坐标空间中点的成对增量 L2 距离来量化失真：$\Delta d = \|\boldsymbol{x}_i^{\text{cart}} - \boldsymbol{x}_j^{\text{cart}}\|_2 - \|\boldsymbol{x}_i^{\text{polar}} - \boldsymbol{x}_j^{\text{polar}}\|_2$ 。如图 **??** 所示，(r, θ) 平面占总失真的一半以上。因此，我们将 3D 卷积分解为两个 2D 卷积——在 (r, z) 和 (θ, z) 上。

我们在图 3a （右）中展示了分解的降采样和上采样策略。降采样使用步幅为 $(3, 3)$ 的稀疏二维卷积作用于 (z, r) ，然后是具有相同步幅的子流形二维卷积，接着是步幅为 $(1, 3)$ 的稀疏卷积作用于 (z, θ) 。上采样则涉及逆卷积，使用相同的步幅参数但顺序相反。第三轴被重塑为批处理维度，从而实现独立处理。这解耦了 r 和 θ 的缩放效果，使卷积核能够在具有较大平移不变性的平面中更好地泛化。



(a)

Input: Voxel features V, coordinates C
batch size B, spatial shape S
Initialize local feature tensor V_{local}
foreach *frustum* f **do**
  extract V_f from V
  $F_{\text{fw}} \leftarrow \text{fsm}(V_f)$
  $F_{\text{bw}} \leftarrow \text{bsm}(V_f^{\text{rev}})$
  $V_{\text{local}}[f] \leftarrow F_{\text{fw}} + F_{\text{bw}}^{\text{rev}}$
end
build SparseConvTensor $X \leftarrow (V_{\text{local}}, C)$
foreach *encoder* E **do**
  $X \leftarrow E(X)$
end
initialize global feature tensor V_{global}
merge partial sectors into same batch
foreach *batch* b **do**
  extract F_b from X
  $V_{\text{global}}[b] \leftarrow \text{fgm}(F_b)$
end
$X.\text{features} \leftarrow \text{normalize}(V_{\text{global}})$
foreach *decoder* D **do**
  $X \leftarrow D(X)$
end
$X.\text{features} \leftarrow X.\text{features} + V_{\text{local}}$
return $X.\text{features}, X.\text{indices}$

(b)

Figure 3: PHiM 块。(a) 序列化是根据方位角进行的，双向局部 SSM 通过元素加法进行聚合。(b) PHiM 块的前向传递。

2.3 PHiM 块

多分辨率特征学习对获取竞争性能至关重要，大多数最新的 LiDAR 目标检测方法利用某种形式的局部和全局特征学习，如分层降采样 [**? ?**] 、隐式窗口嵌入 [**?**] 或注意力机制 [**?**] 。受其他模式的分层架构的启发 [**? ?**] ，我们设计了一个流式分层架构，该架构首先使用双向 SSM 学习扇区内空间关系，使用维度分解卷积提取扇区特征，然后在时间维度使用前向 SSM 学习扇区间时空关系。通过这种方式，我们可以捕获高保真局部细节和高层次的全局场景信息，同时使新扇区能够从先前扇区学习时空上下文。通过利用传入 LiDAR 扇区的自然时间顺序，我们也为我们的 SSM 组件提供了一个自然的序列化方案，使几何基础的序列化技术 [**? ?**] 变得不必要。

图 3a 的左侧显示了 PHiM 块。稀疏输入特征首先按照方位角、径向距离和高度进行序列化——这些信息直接由极坐标中的体素索引给出。两个 Mamba 块随后分别在前向和后向方向上处理特征，并将它们的输出相加，以增强每个扇区内的局部空间表示。为了捕捉全局上下