arXiv:2304.02867v2 [cs.CV] 3 Mar 2024

体素还是柱体：探索三维目标检测的高效点云表示

Yuhao Huang、Sanping Zhou、Junjie Zhang、Jinpeng Dong、Nanning Zheng*西安交通大学人机混合增强智能国家重点实验室、视觉信息与应用国家工程研究中心、人工智能与机器人研究所{hyh950623@stu, spzhou@mail, hooz1009@stu,djp1235a@stu, nnzheng@mail}.xjtu.edu.cn

摘要

高效表示点云是基于激光雷达的三维目标检测的基础。尽管近期基于网格的检测器通常将点云编码为体素或柱体，但这两种方法之间的差异仍未得到充分探索。本文量化了当前编码范式之间的差异，并指出其垂直方向学习能力有限。为解决这些局限，我们提出了一种混合体素-柱体融合网络（VPF），协同结合体素和柱体各自的优势。具体而言，我们首先设计了一种稀疏体素-柱体编码器，分别通过三维和二维稀疏卷积将点云编码为体素特征和柱体特征；随后提出稀疏融合层（SFL），促进稀疏体素特征与柱体特征之间的双向交互。我们提出的高效全稀疏方法可无缝集成至密集和稀疏检测器中。借助这一强大而简洁的框架，VPF取得了有竞争力的性能，并在nuScenes和Waymo开放数据集上实现实时推理速度。代码将发布于 https://github.com/HuangYuhao-0623/VPF。

引言

基于激光雷达的三维目标检测方法已广泛应用于自动驾驶和机器人导航系统，因为激光雷达传感器采集的点云能够明确反映几何信息，且很少受到天气条件的影响。与二维图像不同，点云稀疏且分布不均匀，这给三维目标检测带来了两项挑战：构建高效的检测网络，以及从分布各异的点中建立稳健的目标表征。

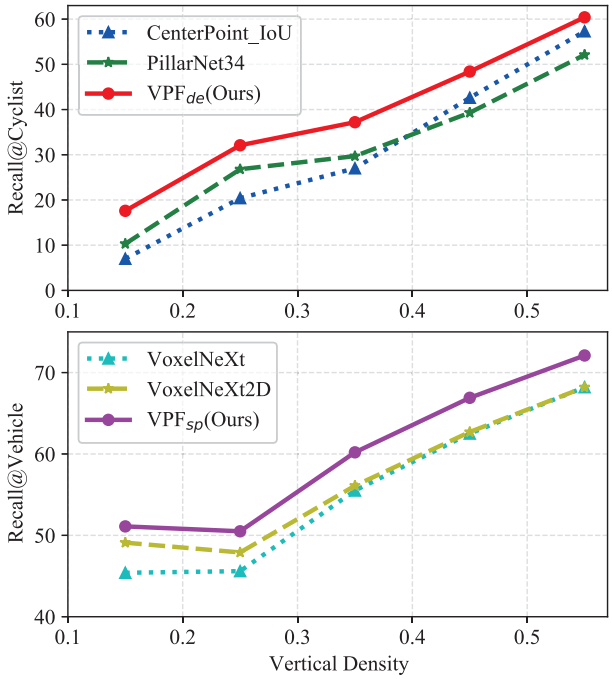


图 1：召回率与垂直密度对比。对于密集和稀疏检测器（Yin、Zhou 和 Krahenbuhl 2021；Guangsheng Shi 2022；Chen 等，2023年），基于柱体的表示在垂直密度较低时召回率更高，而基于体素的表示则往往在高密度场景中表现出色。值得注意的是，我们的混合表示在不同场景下均能带来稳定的改进。

点云表征有多种范式。基于距离视图的方法（Meyer等人，2019；Bewley等人，2021；Fan等人，2021）将点云转换为紧凑的2.5D距离图像，并采用经过深入研究的二维检测器（Ren等人，2015；Lin等人，2017a）预测三维框。这类方法效率较高，但可能会扭曲三维点云的几何结构，并引发尺度方差问题（Li等人，2019）。另一种方法是基于点的检测器（Qi等人，2019；Shi、Wang和Li，2018；Yang等人2020；Zhang 等，2022年）直接利用 PointNet 系列提取逐点特征（Qi 等，2017a,b），并受益于灵活的感受野。然而，这些方法受限于耗时的球形查询和激进的下采样策略。最后，基于网格的方法（Yin、Zhou 和 Krahenbuhl 2021；Hu 等，2022年；Guangsheng Shi 2022；Shi 等，2020a；He 等，2022年）将点云量化为规则的体素或柱体，并在二维鸟瞰图（BEV）中生成边界框，从而在性能和效率之间实现了显著的平衡。

尽管体素和柱都是广泛使用的表示-

*通讯作者。版权所有 © 2024，人工智能促进协会（www.aaai.org）。保留所有权利。

在基于网格的方法中，尽管存在一些表示方式，其间的差异却很少被讨论。鉴于它们在垂直方向上的行为不同，我们在Waymo开放数据集（Sun 等，2020）上开展探索性实验，以分析由此产生的影响。我们首先根据真实标注（GT）的垂直点分布进行划分：对于每个真实标注框，我们沿垂直方向将其均匀划分为 10 个区间，然后将垂直密度计算为 $S_Z = N_{bin}/10$，其中 N_{bin} 表示非空区间的数量。接下来，我们统计若干基于网格的检测器的召回率与垂直密度曲线[1]（Yin、Zhou 和 Krahenbuhl，2021；Guangsheng Shi，2022；Chen 等，2023）。如图 1 所示，基于体素的检测器（即 CenterPoint IoU 和 VoxelNeXt）在垂直密度较高时表现始终更好，而基于柱体的方法（即 PillarNet 和 VoxelNeXt2D）则在密度较低的情况下表现出色。

这一观察突显了当前体素和基于柱体的方法在垂直方向上表示能力的有限性。具体而言，采用体素的检测器通常使用 3D 子流形卷积（Graham 和vanderMaaten 2017），该卷积将特征扩散限制在从非空体素到空体素的范围内，从而限制了感受野。相反，基于柱体的方法将点云离散化为垂直体积，每个体积都编码其在特定X-Y坐标中的所有邻近点。尽管这种范式获得了完整范围的垂直感受野，但其在捕捉细粒度特征方面面临挑战，并且容易造成显著的信息损失，尤其是在处理点密度较高的区域时。鉴于体素和柱体的局限性，我们受到启发，致力于开发一种**协同结合二者的方法，以获得更加鲁棒的表示，尤其是在垂直方向上，同时保持计算效率**。

本文提出了体素-柱体融合（VPF），这是一种混合式点云表示，旨在协同发挥体素和柱体各自的优势。我们首先设计了一种稀疏体素-柱体编码器，将点云划分为体素和柱体，随后通过 3D 和 2D 稀疏卷积对这些稀疏体积特征进行编码。为了扩大体素的垂直感受野，并丰富柱体中的细粒度信息，我们提出稀疏融合层（SFL），以建立体素与柱体之间的双向交互。具体而言，SFL 沿垂直方向聚合体素特征，并将柱体特征广播到对应的垂直体素中。接着，它将聚合和广播后的特征与原始柱体特征和体素特征相结合。此外，我们的方法计算高效且完全稀疏，可无缝集成到密集检测器和稀疏检测器中（Yin、Zhou 和 Krahenbuhl，2021；Guangsheng Shi，2022；Chen 等，2023）。我们还在 nuScenes（Caesar 等，2020）和 Waymo开放数据集（Sun 等，2020）上进行了全面实验。结果表明，我们的方法以实时推理速度达到了最先进的性能，并提升了体素和柱体的垂直表示，如图 1 所示。

[1]我们采用单阶段体素编码的 CenterPoint，其检测头与 PillarNet 相同，详见消融研究。

我们的贡献总结如下：

- 我们指出了当前基于网格的方法在垂直表示学习方面的局限性，并引入一种混合点云表示，以协同利用体素和柱体各自的优势。
- 我们提出了稀疏融合层，通过促进体素-柱体双向交互，扩大体素的垂直感受野，并丰富柱体中的细粒度信息。
- 我们基于所提出的混合表示设计了密集和稀疏检测器。通过在大规模数据集上进行全面实验，我们验证了体素-柱体融合的重要性和实用性。

相关工作

基于网格的三维目标检测

基于网格的方法主要采用体素或柱体表示。基于体素的检测器将点云划分为三维体素，并采用三维卷积进行逐体素特征提取。开创性工作 VoxelNet（Zhou 和 Tuzel，2018）用体素特征编码层取代手工设计的表示，从而实现端到端训练。随后，SECOND（Yan、Mao 和 Li，2018）引入三维稀疏卷积（Graham，2015；Graham 和 van der Maaten，2017），以避免对空体素进行冗余计算。两阶段检测器（Deng 等人，2021；Yin、Zhou 和 Krahenbuhl，2021）采用了二维检测中的由粗到细模式（Ren 等人，2015；Lin 等人，2017a）。例如，Voxel R-CNN（Deng 等人，2021）提出体素 RoI 池化，从体素中提取 RoI 特征。相比之下，单阶段方法（Zheng 等人，2021；Hu 等人，2022）直接在一个阶段生成边界框。一些近期工作（He 等人，2022；Wang 等，2023；Yang 等人，2023）开始在三维目标检测中探索强大的 Transformer 架构（Vaswani 等人，2017；Dosovitskiy 等人，2020；Kirillov 等人，2020），而 FSD（Fan 等人，2022b）和 VoxelNeXt（Chen 等人，2023）则探索全稀疏检测框架。对于基于柱体的方法（Yang、Luo 和 Urtasun，2018；Lang 等人，2019；Guangsheng Shi，2022），点云被编码为二维体（柱体）。PointPillars（Lang 等人，2019）将点云转换为柱体，并采用 PointNet（Qi 等人，2017a）进行逐柱体特征提取。PillarNet（Guangsheng Shi，2022）提出二维稀疏骨干网络以实现高效的柱体编码，在性能和推理速度之间取得了良好的平衡。随后，SST（Fan 等人，2022a）引入单步幅 Transformer，以提升具有挑战性的行人类别上的性能。尽管体素或柱体表示得到了广泛应用，但很少有人讨论二者的差异。

多源特征融合

提出了多源融合方法（Yang 等人，2019；Pang、Morris 和 Radha，2020；Zhou 等，2020a），以结合来自不同来源的不同信息（Zhou 等，2023）。例如，提出了点-体素联合表示（Yang 等人，2019；Shi 等，2020b、2020a）

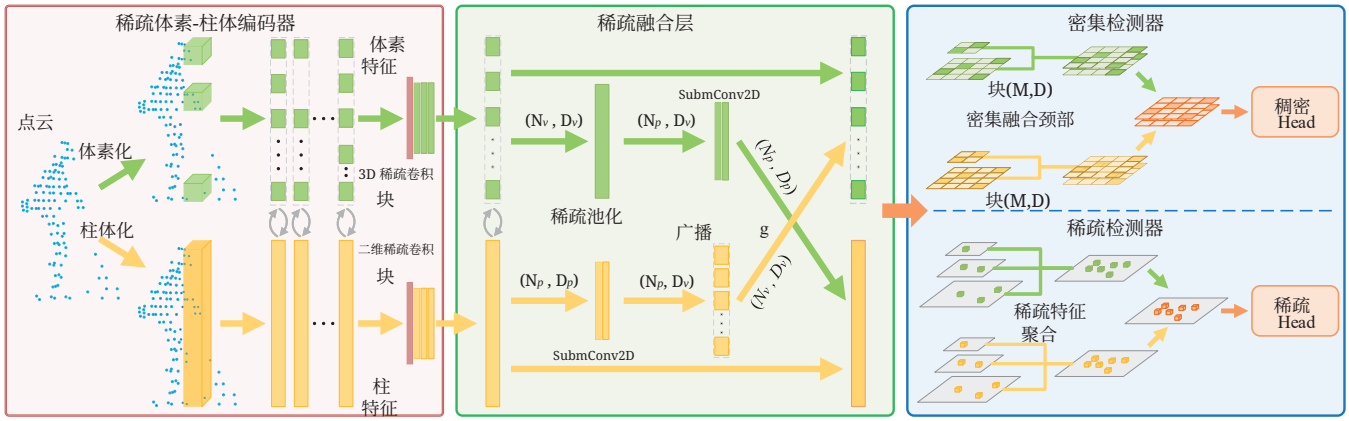


图 2：VPF 的框架。点云首先由稀疏体素-柱体编码器处理，提取相关联的稀疏体素特征和柱体特征。随后的稀疏融合层促进双向交互，从这两种稀疏特征中捕获补充信息。这些组件共同构成一个混合骨干网络，可与密集和稀疏检测器集成。

将灵活的感受野与高效的特征学习方案相结合。此外，多视图融合（Chen 等，2016；Zhou 等，2020b；Wang 等，2020；Xin 等，2019）是另一种补充信息的范式。MVF（Zhou 等，2020b）利用鸟瞰图和透视视图中的互补信息，而 Pillar-OD （Wang 等，2020）则在鸟瞰图和圆柱视图中应用柱体编码器。近年来，多模态融合（Liang 等，2019；Pang、Morris 和 Radha，2020；Li 等，2022年）取得了显著进展。CLOCS（Pang、Morris 和 Radha，2020）提出了一种后期融合策略，以利用二维和三维预测之间的几何与语义一致性。DeepFusion（Li 等，2022年）提出了 InverseAug 和 LearnableAlign，以在后期对齐多模态特征。尽管多源融合提供了来自不同来源的额外信息，但也带来了一个问题。由于不同来源通常存在差异，有效对齐多源特征变得颇具挑战且复杂。在我们的工作中，我们为体素-柱体融合设计了一种简洁而高效的结构。

方法论

为增强点云的垂直表示能力，我们提出体素-柱体融合（VPF），这是一种同时利用二维和三维体积信息的混合点云表示。如图 2 所示，首先将点云量化为体素和柱体，两者在 X-Y 平面上的分辨率相同。随后，我们采用包含四个中间步骤的稀疏体素-柱体编码器，每个步骤均包含一个稀疏卷积块，分别使用三维和二维稀疏卷积提取体素特征和柱体特征。为丰富柱体分支的局部上下文，并将垂直语义融入体素分支，我们在每个步骤的末尾部署稀疏融合层（SFL），使体素特征与柱体特征能够双向交互。最后，我们设计了稠密和稀疏检测器，即 VPF_{de} 和 VPF_{sp}，它们均配备了本文提出的混合表示骨干网络。

稀疏体素-柱体编码器

一致体素-柱体编码。 尽管体素和柱体表示点云的方式不同，但在 X-Y 平面分辨率相同的情况下，特定 X-Y 坐标下垂直方向上汇集的体素包含与对应柱体相同的点云。我们利用这一特性构建一致的体素-柱体编码器。

给定一个点云 $\mathcal{S} = \{s_i \in \mathbb{R}^4\}_{i=1}^{N}$，其中 N 是点的数量。我们首先将三维空间分别划分为体素和柱体，空间分辨率分别为 $L \times W \times H$ 和$L \times W$。接下来，采用动态体素化（Zhou 等，2020b）来生成初始稀疏体素和柱体特征。在体素分支中，$\mathcal{S}$ 基于预定义体素大小进行量化。该过程会生成点到体素索引，表示为 $\mathcal{V} = \{v_i = (l_i,\ w_i,\ h_i) \in \mathbb{N}^3\}_{i=1}^{N}$。随后，通过稀疏平均池化（Fey 和 Lenssen，2019）计算每个非空体素中所有点的均值，得到初始体素特征。在柱体分支中，由于体素和柱体在鸟瞰图（BEV）中的一致性，移除垂直索引后即可得到点到柱体索引 $\mathcal{P} = \{p_i = (l_i,\ w_i) \in \mathbb{N}^2\}_{i=1}^{N}$。最后，我们将 PointNet（Lang 等，2019）与稀疏最大池化结合，生成初始柱体特征。

稀疏卷积块。我们通过双分支稀疏卷积（卷积）块提取稀疏体素和柱体特征，这些块同时包含三维和二维稀疏卷积。与之前的单分支框架（Yan、Mao 和 Li2018；GuangshengShi2022）类似，对于每个块，体素分支由一个三维常规稀疏卷积组成，后接多个三维子流形稀疏卷积层，而柱体分支则由其二维等效结构组成。常规稀疏卷积用于空间下采样，而子流形稀疏卷积则确保输出位置与输入位置相同，以优化效率。如图3 所示，对于二维和三维常规稀疏卷积，我们在水平维度上统一卷积核大小、步幅和填充操作。

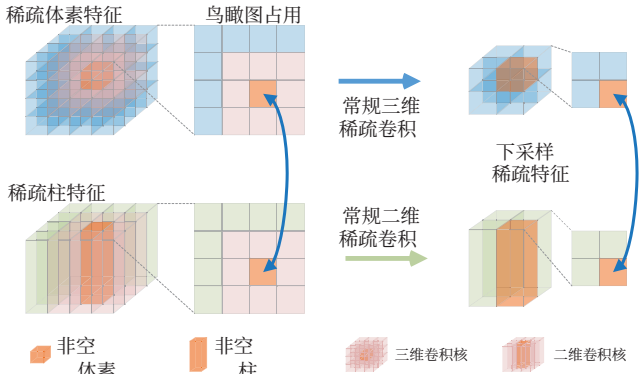


图 3：一致的体素-柱体下采样过程。在下采样过程中，通过使 X-Y 维度上的二维和三维常规稀疏卷积具有相同的卷积核大小、步幅和填充操作，稀疏体素和柱体特征得以保持一致的鸟瞰图占用情况。

维度（即，L 和 W 维度）。这确保体素和柱体特征扩展到相同的 X-Y 平面位置。以这种方式，非空体素和柱体在鸟瞰图中具有一致的占用情况，即，每个非空柱体在相同的X-Y 坐标处都有数-个对应的非空体素（≥ 1）。这种一致性有助于融合稀疏体素和柱体特征。

总体而言，稀疏体素-柱体编码器包含 4 个中间步骤，其中依次部署稀疏卷积块，以生成下采样尺寸为 $1\times$，$2\times$，$4\times$，和 $8\times$ 的稀疏体素和柱体特征。

稀疏融合层

横向连接通常用于融合不同层级的语义信息（Lin 等，2017a）或双分支网络（Fe- ichtenhofer等，2018）。由于稀疏体素和柱体特征分别通过3D和二维稀疏卷积提取，我们引入稀疏融合层（SFL），为稀疏体素和柱体特征之间的逐元素融合建立双向横向连接。

稀疏池化与广播。 在融合前，我们对稀疏特征执行稀疏池化和广播操作，以匹配其大小和维度。将稀疏体素特征表示为体素索引 $\{v_i = (l_i, w_i, h_i) \in \mathbb{N}^3\}_{i=1}^{N_v}$ 及其对应的特征 $\{f_i^v \in \mathbb{R}^{D_v}\}_{i=1}^{N_v}$，其中$N_v$ 和 D_v 分别表示非空体素的数量和体素特征维度。同样，稀疏柱索引和特征分别表示为 $\{p_j = (l_j, w_j) \in \mathbb{N}^2\}_{j=1}^{N_p}$ 和 $\{f_j^p \in \mathbb{R}^{D_p}\}_{j=1}^{N_p}$。由于许多体素对应一个柱，$N_v$ 大于 N_p。

在确保鸟瞰图中占用情况一致的前提下，对于 X-Y 坐标处的非空柱，我们通过将体素索引的水平向量 (l_i, w_i) 与柱索引 $p_j = (l_j, w_j)$ 匹配，找到对应的体素，并构建 $N_v \times N_p$ 体素-柱索引矩阵 C，其中每个元素 $c_{ij}, i = 1, \ldots, N_v, j = 1, \ldots, N_p$ 定义如下：

$$c_{ij} = \begin{cases} 1, & l_i = l_j, w_i = w_j \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

因此，体素到柱的稀疏特征池化及其逆操作——特征广播，定义如下：

$$\text{Pool: } f_j^{v\to p} = \text{Pool}\{f_i^v | c_{ij} = 1, \forall i\}, \tag{2}$$

$$\text{BroadCast: } f_i^{p\to v} = \{f_j^p | c_{ij} = 1, \forall j\}, \tag{3}$$

其中 Pool$\{\cdot\}$ 表示逐元素最大池化，柱特征 f_j^p 通过广播以相同方式投影到非空体素上。

稀疏体素-柱体融合。 除了稀疏池化和广播之外，我们还进一步引入了稀疏融合层（SFL），如图 2 所示。给定一对稀疏体素特征和柱体特征，SFL 首先计算体素-柱体索引矩阵 C。随后，稀疏特征进行双向交互。在体素到柱体的连接中，体素特征通过稀疏最大池化进行垂直聚合，详见公式 2，从而得到形状为 $(N_p,\ D_v)$ 的逐柱特征。这些逐柱特征在空间上与原始柱体特征保持一致，但特征维度不同。接下来，应用 2D 子流形卷积，生成维度为 $(N_p,\ D_p)$ 的逐柱池化特征 $\{f_j^{v\to p} \in \mathbb{R}^{D_p}\}_{j=1}^{N_p}$。对于柱体到体素分支，首先通过 2D 子流形卷积变换柱体特征，然后按照公式 3 进行广播，形成形状为 $(N_v,\ D_v)$ 的逐体素特征 $\{f_i^{p\to v} \in \mathbb{R}^{D_v}\}_{i=1}^{N_v}$。值得注意的是，我们采用这种非对称结构以提升效率。鉴于柱体特征的维度通常是体素特征维度的数倍，直接使用 3D 子流形卷积将体素特征转换为柱体维度并不经济。随后，我们进行逐元素求和以聚合特征，

$$\begin{aligned} f_i^v &= f_i^v + f_i^{p\to v}, i = 1, ..., N_v, \\ f_j^p &= f_j^p + f_j^{v\to p}, j = 1, ..., N_p. \end{aligned} \tag{4}$$

SFL 被置于稀疏体素-柱体编码器的每个中间步骤之后，以促进体素与柱体之间的多层级交互。我们强调 SFL 的两项主要优势。首先，SFL通过保持稀疏体素-柱体编码器中的水平一致性，为特征对齐提供了一种更直接的方法。其次，由于其具有动态性，它避免了体素和柱体的随机丢弃或填充，从而有助于过渡到全稀疏检测框架。

检测框架

我们的方法是一种完全稀疏的骨干网络，可以无缝集成到不同类型的检测器中。我们分别将VPF$_{\text{de}}$ 和 VPF$_{\text{sp}}$ 引入为密集检测器和稀疏检测器的变体。

VPF$_{\text{de}}$. 对于密集的基于网格的方法，稀疏体素或柱体特征首先会转换为密集鸟瞰图特征图，然后由检测头处理。本文构建了一种名为 VPF$_{\text{de}}$ 的密集检测器。给定所提出骨干网络提取的稀疏特征，我们首先设计了密集融合颈部（DFN），用于融合体素分支和柱体分支的密集特征图。DFN 遵循常见的分层结构（Lin et al. 2017a；Yan, Mao, and Li 2018）进行多尺度特征聚合。如图 2 所示，我们分别在体素/柱体分支中应用卷积块 Block(M, D)，以提取具有 $8\times$ 和 $16\times$ 下采样尺寸的密集特征，其中 M 和 D 分别表示卷积层数和输出维度。接下来，通过逐元素求和融合相同尺度的密集体素和柱体特征。最后，我们通过上采样和拼接来组合不同尺度的特征（Yan, Mao, and Li 2018）。对于检测头，我们采用常用的基于中心点的检测头（Yin, Zhou, and Krahenbuhl 2021），并结合交并比感知校正（Hu et al. 2022），将回归精度融入分类分数。最终预测分数由校正函数（Hu et al. 2022）计算,

$$S_{\text{pred}} = S_{\text{cls}}^{1-\alpha} * \text{IoU}_{\text{pred}}^{\alpha}, \tag{5}$$

其中，S_{cls} 是分类得分，IoU^{pred} 是IoU预测，α 是用于平衡二者的超参数。

VPF$_{\text{sp}}$。 为避免冗余计算（Chen 等，2023年；Fan 等，2022a）并支持远距离检测（Fan 等，2022b），研究者提出了稀疏检测框架。由于所提出的体素-柱体编码器和 SFL 均采用完全稀疏架构，我们可以轻松地将混合表示骨干网络整合到现有的稀疏检测器中，因此提出了 VPF$_{\text{sp}}$。如图 2 所示，我们在体素分支和柱体分支中应用了额外的下采样层（Chen 等，2023年），以获得 $16\times$ 和 $32\times$ 下采样稀疏特征。这些多尺度稀疏特征通过各自分支中的稀疏高度压缩进行聚合，然后通过逐元素求和进行融合。最终，稀疏检测头根据鸟瞰图空间中的稀疏柱体特征预测目标，框回归和 IoU 预测范式与基于中心的方法及公式 5 相同。

训练损失。 在损失函数中，我们分别使用焦点损失（Lin 等，2017b）和 L1 损失作为分类损失和边界框回归损失，分别记为 $\mathcal{L}_{\text{cls}}$ 和 $\mathcal{L}_{\text{reg}}$。IoU 头由 L1 损失监督，并编码为（$2 *$交并比 -0.5） $\in[-1,1]$。我们还使用距离交并比（DIoU）损失（Zheng 等，2020）来进一步优化物体中心回归,

$$\mathcal{L}_{\text{diou}} = 1 - \text{IoU}(b, b^{gt}) + \frac{c^2(b, b^{gt})}{d^2}, \tag{6}$$

其中，交并比(b, b^{gt})表示预测框 b 与对应真实框 b^{gt}之间的三维交并比，c 表示 b 和 b^{gt}的中心偏移，d 表示覆盖 b 和 b^{gt}的最小包围长方体的对角线距离。

结合 IoU预测损失 $\mathcal{L}_{\text{iou}}$ 和 DIoU 损失 $\mathcal{L}_{\text{diou}}$，整体损失表示为:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{iou}} + \gamma(\mathcal{L}_{\text{diou}} + \mathcal{L}_{\text{reg}}), \tag{7}$$

其中，γ是类似于（Guangsheng Shi 2022年）中使用的损失权重。

实验

数据集

Waymo开放数据集 (WOD)（Sun 等，2020）由798个训练序列、202个验证序列和150个测试序列组成，共包含20万个带标注帧。检测范围覆盖 $[-75m, -75m, 75m, 75m]$的区域，评估指标为平均精度（平均精度）和按航向加权的平均精度（APH）。测试样本被分为两个难度级别：对于内部点超过5个的目标，属于第1级；对于至少包含1个点的目标，属于第2级。

nuScenes数据集 （Caesar 等，2020）包含 700 个训练序列、150 个验证序列和 150 个测试序列，其中有 3.4 万个带标注的关键帧和 10 个类别。评估指标为 nuScenes 检测得分（NDS），它是平均精度均值和其他真阳性指标的加权和。

实现细节

网络架构。 我们的 VPF 构建于稀疏骨干网络之上，该网络由稀疏体素-柱编码器（包含 4 个卷积块）和稀疏融合层（SFL）组成。在 VPF$_{\text{de}}$ 模型中，每个卷积块内体素分支和柱体分支的特征维度分别设为 $[16, 32, 64, 64]$ 和 $[32, 64, 128, 256]$,。而在 VPF$_{\text{sp}}$ 模型中，这些维度分别为$[16, 32, 64, 128]$ 和 $[32, 64, 128, 256]$。

此外，对于应用于密集检测器的密集融合颈部（DFN），卷积层数量 M 和输出维度 D 分别设为 5 和 128。对于稀疏检测器，额外下采样层（即，$16\times$ 和 $32\times$ 个卷积块）中的特征维度在体素分支和柱体分支中均设为 $[128, 128]$, $[256, 256]$。请注意，在最终的体素-柱特征融合过程中，体素特征会上采样至 $256D$。

模型设置。 在数据预处理过程中，点云被划分为体素和柱体，WOD的尺寸为 $[0.1m, 0.1m, 0.15m]$, $[0.1m, 0.1m]$，nuScenes数据集的尺寸为$[0.075m, 0.075m, 0.2m]$, $[0.075m, 0.075m]$ 。我们将WOD中Vehicle、行人和Cyclist的交并比阈值 (0.8,0.55,0.55) 以及公式5中的α 分别设为 (0.68,0.71,0.65)。在nuScenes中，所有类别的交并比和 α 分别设为0.2和0.5。

训练详情。 所有模型均使用 Adam 优化器，在 4 块 NVIDIA 3090 GPU 上训练。nuScenes 和 Waymo 数据集的学习率分别设为 1e-3 和 3e-3。我们采用常见的数据增强策略（Shi 等人，2022年）和 Wang 等人（2023年）提出的真值采样衰减策略。对于 WOD，我们以 16 的批量大小分别训练 VPF$_{\text{de}}$ 和 VPF$_{\text{sp}}$，训练轮数分别为 15 和 12。对于 nuScenes，我们使用相同的批量大小训练模型 20 轮。

与最先进方法的比较

Waymo开放数据集。 我们在 WOD验证 集上，将 VPF 的不同变体与现有单帧方法进行了比较。如表 1 所示，在单阶段框架中，我们的方法

方法	阶段	第2级 平均精度均值/平均精度航向均值	车辆（第1级/第2级） 平均精度 APH 平均精度 APH				行人（第1级/第2级） 平均精度 APH 平均精度 APH				骑行者（第1级/第2级） 平均精度 APH 平均精度 APH			
SECOND (2018)	单一	61.0/57.2	72.3	71.7	63.9	63.3	68.7	58.2	60.7	51.3	60.6	59.3	58.3	57.0
MVF (2020b)	单一	-	62.9	-	-	-	65.3	-	-	-	-	-	-	-
PointPillars (2019)	单一	62.8/57.8	72.1	71.5	63.6	63.1	70.6	56.7	62.8	50.3	64.4	62.3	61.9	59.9
Pillar-OD（2020）	单一	-	69.8	-	-	-	72.5	-	-	-	-	-	-	-
AFDetV2-Lite（2022年）	单一	71.0/68.8	77.6	77.1	69.7	69.2	80.2	74.6	72.2	67.0	73.7	72.7	71.0	70.1
IA-SSD（2022年）	单一	62.3/58.1	70.5	69.7	61.6	61.0	69.4	58.5	60.3	50.7	67.7	65.3	65.0	62.7
CenterFormer（2022年）	单一	71.1/68.9	75.0	74.4	69.9	69.4	78.6	73.0	73.6	68.3	72.3	71.3	69.8	68.8
CenterPoint† (2021)	单一	71.7/69.4	77.7	77.1	69.7	69.2	80.9	75.1	72.6	67.2	75.6	74.4	72.8	71.7
PillarNet（2022年）	单一	71.0/68.5	79.1	78.6	70.9	70.5	80.6	74.0	72.3	66.2	72.3	71.2	69.7	68.7
VoxelNeXt2D（2023年）	单一	70.9/68.2	77.9	77.5	69.7	69.2	80.2	73.5	72.2	65.9	73.3	72.2	70.7	69.6
VoxelNeXt$_{K3}$ (2023年)	单一	72.2/70.1	78.2	77.7	69.9	69.4	81.5	76.3	73.5	68.6	76.1	74.9	73.3	72.2
GD-MAE (2023年)	单一	72.9/70.4	79.4	78.9	70.9	70.5	82.2	75.9	74.8	68.8	75.8	74.8	73.0	72.0
DSVT（2023年）	单一	73.2/71.0	79.3	78.8	70.9	70.5	82.8	77.0	**75.2**	69.8	76.4	75.4	73.6	72.7
VPF$_{\text{sp}}$-Lite	单一	73.1/70.9	79.1	78.6	71.1	70.6	82.1	76.7	74.5	69.4	76.5	75.5	73.6	72.7
VPF$_{\text{de}}$-Lite	单一	73.6/71.4	80.2	79.7	71.9	71.5	82.5	76.9	74.8	69.4	77.1	76.0	74.2	73.2
VPF$_{\text{sp}}$	单一	**73.6/71.6**	**80.2**	**79.8**	**71.7**	**71.3**	**82.5**	**77.5**	74.9	**70.1**	**77.2**	**76.2**	**74.3**	**73.3**
VPF$_{\text{de}}$	单一	**73.9/71.7**	**80.5**	**80.0**	**72.3**	**71.9**	**82.8**	**77.3**	**75.1**	**69.9**	**77.2**	**76.0**	**74.3**	**73.2**
Voxel RCNN（2021）	Two	-	75.6	-	66.6	-	-	-	-	-	-	-	-	-
PartA^2 （2020b）	Two	66.9/63.8	77.1	76.5	68.5	68.0	75.2	66.9	66.2	58.6	68.6	67.4	66.1	64.9
激光雷达 R-CNN（2021）	Two	65.8/61.3	76.0	75.5	68.3	67.9	71.2	58.7	63.1	51.7	68.6	66.9	66.1	64.4
RSN（2021）	Two	-	75.1	74.6	66.0	65.5	77.8	72.7	68.3	63.7	-	-	-	-
PV-RCNN (2020a)	Two	66.8/63.3	77.5	76.9	69.0	68.4	75.0	65.6	66.0	57.6	67.8	66.4	65.4	64.0
SST TS 1f (2022a)	Two	-	76.2	75.8	68.0	67.6	81.4	74.0	72.8	65.9	-	-	-	-
PV-RCNN++ (2022年)	Two	71.7/69.5	79.3	78.8	70.6	70.2	81.3	76.3	73.2	68.0	73.7	72.7	71.2	70.2
FSD$_{spconv}$ (2022b)	Two	72.9/70.8	79.2	78.8	70.5	70.1	82.6	77.3	73.9	69.1	77.1	76.0	74.4	73.3
GD-MAE（2023年）	Two	74.1/71.6	80.2	79.8	72.4	72.0	83.1	76.7	75.5	69.4	77.2	76.2	74.4	73.4
DSVT（2023年）	Two	74.3/72.1	80.2	79.7	72.0	71.6	83.7	78.0	76.1	70.7	**77.8**	**76.8**	**74.9**	**73.9**
VPF$_{\text{sp}}$-TS	Two	**74.7/72.5**	**81.4**	**80.9**	**73.1**	**72.7**	**83.7**	**78.5**	**76.2**	**71.3**	77.3	76.3	74.4	73.5
VPF$_{\text{de}}$-TS	Two	**75.1/72.9**	**81.4**	**80.9**	**73.6**	**73.2**	**84.0**	**78.6**	**76.6**	**71.3**	**77.8**	**76.7**	**75.2**	**74.1**

表1：WOD 验证集上的单帧性能比较，不使用测试时增强或模型集成。 †：采用IoU预测设置的单阶段方法。不同阶段的前两名结果以粗体突出显示。

在所有单阶段方法中取得了出色的性能。具体而言，VPF$_{\text{sp}}$ 和 VPF$_{\text{de}}$ 均显著优于稠密或稀疏基线方法（即 Centerpoint、PillarNet 和 VoxelNeXt）。此外，我们的方法优于近期基于Transformer的 GD-MAE（Yang 等，2023）和 DSVT（Wang 等，2023），且训练计划（4 张 3090 GPU，相比之下需要 8 张 A100 GPU 或进行 MAE 预训练）和推理速度更为高效。我们还提供了框架的轻量版本 VPF$_{\text{sp}}$-Lite 和 VPF$_{\text{de}}$-Lite，通过减少卷积层和通道维度来降低计算预算。如图 4 所示，它们在运行时间分别为 60（±5）毫秒和 58（±4）毫秒的情况下，仍取得了有竞争力的结果。为与两阶段检测器进行比较，我们直接采用 CT3D（Sheng 等，2021）作为第二阶段，同样取得了有竞争力的性能。由于稀疏架构更难估计目标中心，VPF$_{\text{sp}}$ 在车辆类别上的表现不如 VPF$_{\text{de}}$；而 VPF$_{\text{sp}}$ 在远距离检测中计算效率更高。

nuScenes数据集。 我们在 nuScenes 测试 集上评估并比较了 VPF 与基于LiDAR的方法。如表 2 所示，我们的方法在平均精度均值和 NDS 指标上均超越了先前的方法，并取得了显著提升。更具体地说，VPF$_{\text{de}}$ 在单模型设置下实现了 67.0 的平均精度均值和 72.7 的 NDS，使用双重翻转测试时，NDS 提升至 73.8。此外，VPF 提高了性能-

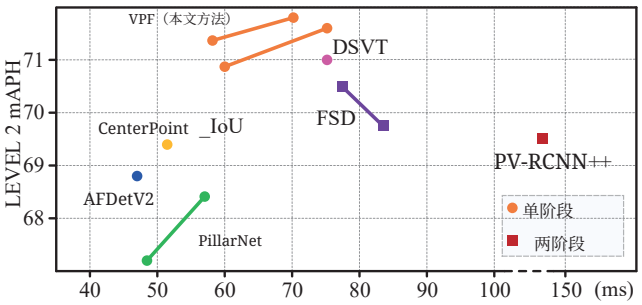


图4：WOD验证集上的性能与推理延迟。使用单一3090 GPU、批量大小为1进行测试。

性能，在行人、自行车等垂直敏感类别上大幅领先，同时在卡车等其他类别上保持了具有竞争力的结果。我们将其归因于稳健的垂直表示。鉴于行人等细长物体相比水平方向具有更丰富的垂直信息，我们的混合体素-柱体编码范式有助于为此类实体建立有效的表示。

消融研究

在本节中，我们在 WOD 验证 集上进行实验，以验证体素-柱混合表示的有效性。此外，我们还研究了不同点云表示和架构对 3D 检测的影响。所有

方法	NDS	mAP	Car	卡车	Bus	预告片	C.V.	Ped	Mot	Byc	T.C.	Bar
PointPillars（2019）	45.3	30.5	68.4	23.0	28.2	23.4	4.1	59.7	27.4	1.1	30.8	38.9
3DSSD (2020)	56.4	42.6	81.2	47.2	61.4	30.5	12.6	70.2	36.0	8.6	31.1	47.9
CBGS（2019）	63.3	52.8	81.1	48.5	54.9	42.9	10.5	80.1	51.5	22.3	70.9	65.7
HotSpotNet（2020b）	66.0	59.3	83.1	50.9	56.4	53.3	23.0	81.3	63.5	36.6	73.0	71.6
CVCNET (2020a)	66.6	58.2	82.6	49.5	59.4	51.1	16.2	83.0	61.8	38.8	69.7	69.7
CenterPoint (2021)	65.5	58.0	84.6	51.0	60.2	53.2	17.5	83.4	53.7	28.7	76.7	70.9
CenterPoint† (2021)	67.3	60.3	85.2	53.5	63.6	56.0	20.0	84.6	59.5	30.7	78.4	71.1
AFDetV2-Lite (2022年)	68.5	62.4	86.3	54.2	62.5	58.9	26.7	85.8	63.8	34.3	80.1	71.0
VISTA-OHS† (2022年)	69.8	63.0	84.4	55.1	63.7	54.2	25.1	84.6	70.0	45.4	78.5	71.1
Focals Conv（2022a）	70.0	63.8	86.7	56.3	**67.7**	59.5	23.8	87.5	64.5	36.3	81.4	74.1
LargeKernel3D（2022b）	70.5	65.3	85.9	55.3	66.2	60.2	26.8	85.6	72.5	46.6	80.0	74.3
PillarNet-34† (2022年)	71.4	66.0	**87.6**	57.5	63.6	63.1	27.9	87.3	70.1	42.3	83.3	77.2
VoxelNeXt† (2023年)	71.4	66.2	85.3	55.7	66.2	57.2	29.8	86.5	75.2	48.8	80.7	76.1
LinK（Lu 等，2023年）	71.0	66.3	86.1	55.7	65.7	62.1	30.9	85.8	73.5	47.5	80.4	75.5
VPFde	72.7	67.0	85.8	55.1	63.5	62.1	33.3	87.6	72.5	48.6	82.9	**78.2**
VPF的†	**73.8**	**68.6**	86.3	**56.8**	66.1	**64.5**	**34.6**	**88.3**	**75.8**	**51.9**	**84.2**	77.6

表 2：nuScenes 测试集上的三维目标检测性能比较。 † 表示使用了翻转测试。我们用粗体突出显示最佳结果，并用下划线标出非集成方法中的最佳结果。

体素	柱	DFN	SFL	Lat.	二级 APH 车辆	行人	骑行者	mAPH
✓				51ms	67.8	66.5	68.2	67.5
	✓			57ms	68.5	64.1	66.2	66.3
✓	✓			102毫秒	69.1	66.7	69.2	68.3
✓	✓	✓		69ms	69.2	66.0	68.9	68.0
✓	✓	✓	✓	75ms	**70.2**	**68.2**	**70.7**	**69.7**

表 3：混合编码范式的影响。CenterPoint 和 PillarNet 被用作基于体素/柱的基线。

Exp.	参数	二级 APH Veh.	行人	骑行者	mAPH
缩放版 PillarNet	92.4M	69.0	64.8	67.8	67.2
缩放版 CenterPoint	93.2M	69.2	67.0	70.1	68.8
缩放版 CenterPoint†	96.8M	69.2	67.2	70.5	69.0
体素-体素	67.1M	68.9	66.2	69.5	68.2
柱-柱	111.1M	69.3	64.5	67.1	66.9
VPF_{de}	92.4M	**70.2**	**68.2**	**70.7**	**69.7**

表 4：架构与可扩展性的影响。 †：采用垂直残差进行特征聚合。

模型均在完整训练集上训练 7 个周期，并根据第2级难度下的 APH 进行比较。

体素-柱体编码范式的影响。我们采用 VPF_{de} 作为框架，以阐释混合表示的影响。对于基于体素/柱体的基线方法，我们采用单阶段 Centerpoint（Yin、Zhou 和 Krahenbuhl，2021）和 PillarNet（Guangsheng Shi，2022）。为确保公平比较，我们按照（Guangsheng Shi，2022）的方法为 CenterPoint 添加 IoU 预测。首先，我们将单一表示与朴素模型集成进行比较：如 3rd 行所示，通过 NMS 合并 CenterPoint 和 PillarNet 的预测结果。集成策略始终能够提升检测精度。随后，如 4th 行所示，我们采用所提出的体素-柱体编码器，并结合密集融合颈部（DFN）。这种后期融合

Exp.	连接	二级 APH Veh.	行人	骑行者	mAPH
1	体素到柱	70.0	67.3	70.4	69.2
2	柱到体素	69.4	66.3	68.7	68.1
3	双向	**70.2**	**68.2**	**70.7**	**69.7**

表 5：不同交互范式的消融实验。

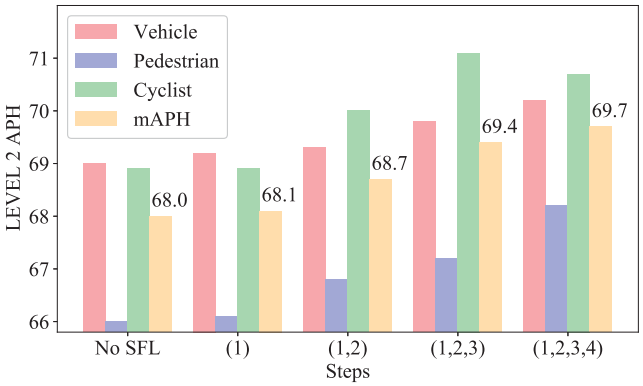


图5：SFL部署步骤的消融实验。

该策略由于体素和柱体的统一编码过程，在取得相当结果的同时更加高效。最后，我们部署稀疏融合层（SFL），在延迟略微增加的情况下获得了 1.7mAPH 的提升，这证明了稀疏主干网络中体素与柱体交互的重要性。

模型规模和架构的影响。由于我们的模型具有不同的规模和双分支流程，我们在表4中对模型规模和架构的有效性进行了消融实验。首先，在保持鸟瞰图地图分辨率不变的同时，通过综合增加层数和特征维度，我们构建了扩展版的 PillarNet 和 Center-Point。

保持不变，这使得它们在参数规模上可与VPF相比较。如 $1^{st} - 2^{nd}$ 行所示，我们的混合体素-柱体编码范式明显优于单一表征的对应方案。随后，我们在 3rd 行中探究垂直特征聚合的有效性。对于体素骨干网络中的每个阶段，我们引入了一个垂直残差。该残差通过稀疏池化沿垂直方向聚合体素特征来构建。聚合得到的特征作为残差，并与原始体素特征相加。该垂直残差提升了细长物体（即行人和骑行者）的检测效果，但与我们的方法相比仍存在较大差距。此外，我们研究了不同的双分支流水线，以探究更宽或多分支网络的影响（Chen等，2017；Xie等，2017）。表4的 $4^{th} - 5^{th}$ 行中仅使用体素或仅使用柱的框架结果表明，尽管多分支架构能够提升检测性能，但我们工作的改进主要来自体素与柱之间的协同作用。

稀疏融合层。 我们从两个角度研究 SFL 的影响。首先，我们验证 SFL 的双向连接。鉴于其体素到柱和柱到体素的连接，我们在表 5 中比较了交互范式的性能。结果表明，每个单独的连接都能带来明显提升，而采用双向连接可取得最佳结果。其次，我们在图 5 中分析了为体素-柱编码器部署 SFL 的中间步骤。随着每一步中模块的堆叠，性能不断提升。由于 SFL 放大了柱体分支的局部上下文，并将垂直语义注入体素分支，因此它主要作用于稀疏骨干网络的后续步骤，有助于在稀疏卷积过程中丰富垂直信息。

结论

受体素和柱体编码范式的差异与限制启发，我们提出了 VPF，这是一种结合两者优势的混合检测框架。VPF 使用稀疏体素-柱体编码器进行一致的双分支特征提取。此外，我们的稀疏融合层促进稀疏体素特征与柱体特征之间的双向交互，共同增强垂直表示学习。值得注意的是，我们提出的组件可无缝集成到密集和稀疏检测器中，实现出色的性能和实时推理速度。

局限性。 VPF依赖于鸟瞰图中一致的占用情况来进行逐元素融合，因此要求对体素和柱体进行等水平划分。未来，我们将探索异步的体素-柱体融合策略（例如，采用不同分辨率或多步幅融合），以实现更灵活的点云表示学习。

致谢

本研究得到国家自然科学基金（编号：62088102）的资助。我们感谢四位匿名审稿人提出的建设性意见。

参考文献

Bewley, A.; Sun, P.; Mensink, T.; Anguelov, D.; and Sminchisescu, C. 2021. 用于尺度不变三维目标检测的范围条件膨胀卷积。发表于*CoRL*。

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan,Y.;Baldan, G.; and Beijbom,O. 2020. nuScenes：用于自动驾驶的多模态数据集。载于*CVPR*。

Chen, Q.; Sun, L.; Cheung, E.; and Yuille, A. 2020a. 通过混合圆柱-球面体素化实现 3D 目标检测中的跨视角一致性。发表于*Neurips*。

Chen, Q.; Sun, L.; Wang, Z.; Jia, K.; and Yuille, A. 2020b. 目标即热点：通过热点触发实现无锚框三维目标检测的方法。发表于*ECCV*。Chen, X.; Ma, H.; Wan, J.; Li, B.; and Tian, X. 2016. 面向自动驾驶的多视角三维目标检测网络。发表于*CVPR*。Chen, Y.; Li, J.; Xiao, H.; Jin, X.; Yan, S.; and Feng, J. 2017. 双路径网络。发表于*NeurIPS*。Chen, Y.; Li, Y.; Zhang, X.; Sun, J.; and Jia, J. 2022a. 用于三维目标检测的焦点稀疏卷积网络。发表于*CVPR*。Chen, Y.; Liu, J.; Qi, X.; Zhang, X.; Sun, J.; and Jia, J. 2022b. 扩展三维CNN中的卷积核。*arXiv*预印本。Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; and Jia, J. 2023. VoxelNeXt：用于三维目标检测与跟踪的全稀疏VoxelNet。发表于*CVPR*。Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; and Li, H. 2021. Voxel R-CNN：迈向高性能体素三维目标检测。发表于*AAAI*。Deng, S.; Liang, Z.; Sun, L.; and Jia, K. 2022. Vista：通过双重跨视角空间注意力提升三维目标检测性能。发表于*CVPR*。Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. 一张图像胜过16×16个词：用于大规模图像识别的Transformer。发表于*ICLR*。Fan, L.; Pang, Z.; Zhang, T.; Wang, Y.-X.; Zhao, H.; Wang, F.; Wang, N.; and Zhang, Z. 2022a. 采用稀疏Transformer的单一特征步幅三维目标检测器。发表于*CVPR*。Fan, L.; Wang, F.; Wang, N.; and Zhang, Z. 2022b. 全稀疏三维目标检测。发表于*NeurIPS*。Fan, L.; Xiong, X.; Wang, F.; long Wang, N.; and Zhang, Z. 2021. RangeDet：为基于激光雷达的三维目标检测中的距离视图正名。发表于*ICCV*。Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2018. 用于视频识别的慢速-快速网络。发表于*ICCV*。Fey, M.; and Lenssen, J. E. 2019. 使用PyTorch Geometric实现快速图表示学习。发表于*ICLRW*。Graham, B. 2015. 稀疏三维卷积神经网络。发表于*BMVC*。Graham, B.; and van der Maaten, L. 2017. 子流形稀疏卷积网络。*arXiv*预印本。

Guangsheng Shi, C. M., Ruifeng Li. 2022. PillarNet：实时高性能柱体三维目标检测。发表于*ECCV*。He, C.-H.; Li, R.; Li, S.; and Zhang, L. 2022. 体素集Transformer：一种从点云进行三维目标检测的集合到集合方法。发表于*CVPR*。Hu, Y.; Ding, Z.; Ge, R.; Shao, W.; Huang, L.; Li, K.; and Liu, Q. 2022. AFDetV2：重新思考点云目标检测是否需要第二阶段。发表于*AAAI*。Kirillov, A.; Usunier, N.; Carion, N.; Zagoruyko, S.; Syn- naeve, G.; Massa, F.; Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. 使用Transformer实现端到端目标检测。发表于*ECCV*。Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. PointPillars：用于点云目标检测的快速编码器。发表于*CVPR*。Li, Y.; Chen, Y.; Wang, N.; and Zhang, Z. 2019. 用于目标检测的尺度感知三叉网络。发表于*ICCV*。Li, Y.; Yu, A. W.; Meng, T.; Caine, B.; Ngiam, J.; Peng, D.; Shen, J.; Wu, B.; Lu, Y.; Zhou, D.; Le, Q. V.; Yuille, A.; and Tan, M. 2022. DeepFusion：用于多模态三维目标检测的激光雷达-相机深度融合。发表于*CVPR*。Li, Z.; Wang, F.; and Wang, N. 2021. LiDAR R-CNN：一种高效通用的三维目标检测器。发表于*CVPR*。Liang, M.; Yang, B.; Chen, Y.; Hu, R.; and Urtasun, R. 2019. 用于三维目标检测的多任务多传感器融合。发表于*CVPR*。Lin, T.-Y.; Doll´ar, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017a. 用于目标检测的特征金字塔网络。发表于*CVPR*。Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Doll´ar, P. 2017b. 面向密集目标检测的焦点损失。发表于*ICCV*。Lu, T.; Ding, X.; Liu, H.; Wu, G.; and Wang, L. 2023. LinK：用于基于激光雷达的三维感知的线性核。发表于*CVPR*。Meyer, G. P.; Laddha, A.; Kee, E.; Vallespi-Gonzalez, C.; and Wellington, C. K. 2019. Lasernet：一种用于自动驾驶的高效概率三维目标检测器。发表于*CVPR*。Pang, S.; Morris, D. H.; and Radha, H. 2020. CLOCs：用于三维目标检测的相机-激光雷达候选目标融合。发表于*IROS*。Qi, C.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. PointNet：基于点集的三维分类与分割深度学习。发表于*CVPR*。Qi, C.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++：度量空间中点集的深度层次特征学习。发表于*NeurIPS*。Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. 在点云中通过深度霍夫投票进行三维目标检测。发表于*ICCV*。Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN：借助区域提议网络实现实时目标检测。*IEEETPAMI*。Sheng, H.; Cai, S.; Liu, Y.; Deng, B.; Huang, J.; Hua, X.; and Zhao, M. 2021. 使用逐通道Transformer改进三维目标检测。发表于*ICCV*。Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020a. Pv-rcnn：用于三维目标检测的点-体素特征集抽象。载于*CVPR*。Shi, S.; Jiang, L.; Deng, J.; Wang, Z.; Guo, C.; Shi, J.; Wang, X.; and Li, H. 2022. PV-RCNN++：结合局部向量表示的点-体素特征集抽象，用于三维目标检测。*IJCV*。Shi, S.; Wang, X.; and Li, H. 2018. PointRCNN：从点云生成三维目标候选并进行检测。载于*CVPR*。Shi, S.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020b. 从点到部件：利用部件感知与部件聚合网络进行点云三维目标检测。*IEEETPAMI*。Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; 等。2020。面向自动驾驶感知的可扩展性：Waymo开放数据集。载于*CVPR*。Sun, P.; Wang, W.; Chai, Y.; Elsayed, G.; Bewley, A.; Zhang, X.; Sminchisescu, C.; and Anguelov, D. 2021. RSN：用于高效、精准激光雷达三维目标检测的距离稀疏网络。载于*CVPR*。Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. 注意力就是你所需要的一切。载于*NeurIPS*。Wang, H.; Shi, C.; Shi, S.; Lei, M.; Wang, S.; He, D.; Schiele, B.; and Wang, L. 2023. DSVT：具有旋转集合的动态稀疏体素变换器。载于*CVPR*。Wang, Y.; Fathi, A.; Kundu, A.; Ross, D. A.; Pantofaru, C.; Funkhouser, T.; and Solomon, J. 2020. 面向自动驾驶的基于柱体的目标检测。载于*ECCV*。Xie, S.; Girshick, R.; Doll´ar, P.; Tu, Z.; and He, K. 2017. 深度神经网络的聚合残差变换。载于*CVPR*。Xin, X.; Wang, J.; Xie, R.; Zhou, S.; Huang, W.; and Zheng, N. 2019. 利用多视图聚类进行半监督行人重识别。*PR*。Yan, Y.; Mao, Y.; and Li, B. 2018. SECOND：稀疏嵌入式卷积检测。*Sensors*。Yang, B.; Luo, W.; and Urtasun, R. 2018. PIXOR：基于点云的实时三维目标检测。载于*CVPR*。Yang, H.; He, T.; Liu, J.; Chen, H.; Wu, B.; Lin, B.; He, X.; and Ouyang, W. 2023. GD-MAE：用于激光雷达点云MAE预训练的生成式解码器。载于*CVPR*。Yang, Z.; Sun, Y.; Liu, S.; and Jia, J. 2020. 3DSSD：基于点的三维单阶段目标检测器。载于*CVPR*。Yang, Z.; Sun, Y.; Liu, S.; Shen, X.; and Jia, J. 2019. STD：由稀疏到稠密的点云三维目标检测器。载于*ICCV*。Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. 基于中心的三维目标检测与跟踪。载于*CVPR*。Yoo, J. H.; Kim, Y.; Kim, J.; and Choi, J. W. 2020. 3D-CVF：利用跨视角空间特征融合生成相机与激光雷达联合特征，以进行三维目标检测。载于*ECCV*。

Zhang, Y.; Hu, Q.; Xu, G.; Ma, Y.; Wan, J.; Guo, Y.; and Prediction, C. 2022. 并非所有点都相同：为三维激光雷达点云学习高效的基于点的检测器。载于*CVPR*。Zheng, W.; Tang, W.; Chen, S.; Jiang, L.; and Fu, C.-W. 2021. CIA-SSD：具有置信度和交并比感知能力的点云单阶段目标检测器。载于*AAAI*。Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU损失：更快、更好地学习边界框回归。载于*AAAI*。Zhou, S.; Wang, J.; Wang, L.; Wan, X.; Hui, S.; and Zheng, N. 2023. 用于网络集成的逆对抗多样性学习。*IEEETNNLS*。Zhou, S.; Wang, J.; Wang, L.; Zhang, J.; Wang, F.; Huang, D.; and Zheng, N. 2020a. 用于保边显著目标检测的层次化交互式细化网络。*IEEE TIP*。Zhou, Y.; Sun, P.; Zhang, Y.; Anguelov, D.; Gao, J.; Ouyang, T.; Guo, J.; Ngiam, J.; and Vasudevan, V. 2020b. 激光雷达点云三维目标检测的端到端多视图融合。载于*CoRL*。Zhou, Y.; and Tuzel, O. 2018. VoxelNet：基于点云的三维目标检测端到端学习。载于*CVPR*。Zhou, Z.; Zhao, X.; Wang, Y.; Wang, P.; and Foroosh, H. 2022. Centerformer：用于三维目标检测的基于中心的变换器。载于*ECCV*。Zhu, B.; Jiang, Z.; Zhou, X.; Li, Z.; and Yu, G. 2019. 面向点云三维目标检测的类别均衡分组与采样。*arXiv*预印本。

更多实现细节

网络结构。
我们的 VPF 构建于稀疏骨干网络之上，该网络由包含 4 个卷积块的稀疏体素-柱体编码器和稀疏融合层（SFL）组成。在 VPF_{de}模型中，每个卷积块中体素分支和柱体分支的特征维度分别设为 $[16, 32, 64, 64]$和 $[32, 64, 128, 256]$,。而在 VPF_{sp} 模型中，这些维度分别为 $[16, 32, 64, 128]$ 和 $[32, 64, 128, 256]$。图 6 展示了不同 VPF 变体的卷积块结构。

此外，对于应用于密集检测器的密集融合颈部（DFN），卷积层数量 M 和输出维度 D 分别设为 5 和 128。对于稀疏检测器，附加下采样层（即， $16\times$ 和 $32\times$ 个卷积块）的特征维度设为$[128, 128], [256, 256]$，分别用于体素分支和柱体分支。请注意，在最终的体素-柱特征融合过程中，体素特征会被上采样至 $256D$ 。

训练设置。
我们分别为 nuScenes 和 Waymo 数据集设置 1e-3 和 3e-3 的学习率。对于两阶段模型，我们遵循 DSVT（Wang 等，2023），固定检测模型的第一阶段，并对 CT3D（Sheng 等，2021）的精修阶段进行微调，训练 12 个周期。

实验结果

垂直密度分析
我们在图 7 中深入探讨体素与柱体之间的差异。具体而言，为了消除点的水平分布和真实标注中点数的影响，我们进一步根据真实标注中的点数及其水平占用率对物体进行划分。水平占用率通过 X 轴和 Y 轴方向密度的几何平均值量化，公式为 $\sqrt{S_X S_Y}$。作为对比，我们采用 CenterPoint IoU、PillarNet 和我们的 VPF_{de}作为对照组，以凸显传统基于体素/柱体的编码范式与我们的混合表示之间的差异。如图 7 所示，通过协同结合体素和柱体各自的优势，我们的方法在各种条件下均实现了稳定提升。

nuScenes验证 集上的性能
我们在表6中展示了nuScenes验证集的比较结果。VPF以较大优势超越了之前的单一体素/基于柱体的方法，尤其是在NDS方面。

更多消融实验
VPF_{sp}. 我们评估了 VPF_{sp} 在表 7 中各个组件的影响。请注意，原始VoxelNeXt 和 VoxelNeXt2D 均作为基于体素/柱的基线应用，且所有实验均使用相同的检测头。如 3rd 行所示，采用逐元素求和进行稀疏体素与柱特征融合的双分支体素-柱编码器，带来了1.9 和 3.7 的提升。

表 #3	体素太小	mAP	NDS
CenterPoint	0.75厘米	59.22	66.48
PillarNet		59.90	67.39
VoxelNeXt		60.53	66.65
VPF_{de}		**61.83**	**68.89**
$\text{VPF}^{\dagger}_{\text{de}}$		**64.68**	**71.05**

表 6：nuScenes验证集上的非集成对比。†：采用渐隐策略。

Exp.	设置	二级 APH			
		Veh.	Ped.	Cyc.	mAPH
1	体素	67.6	67.2	69.8	68.2
2	柱	67.5	63.7	68.1	66.4
3	不使用 SFL	69.5	68.5	72.3	70.1
4	使用 SFL	70.5	69.8	72.4	70.9

表 7：VPF_{sp} 的消融实验。

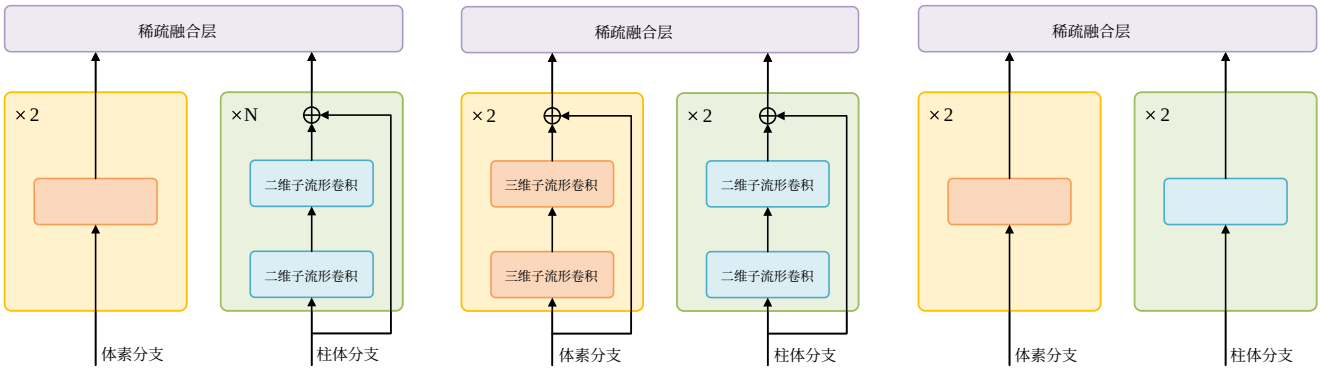
Exp.	设置	二级 APH			
		Veh.	Ped.	Cyc.	mAPH
1	仅体素	69.3	65.4	68.4	67.7
2	仅柱方式	69.7	67.4	70.3	69.1
3	求和	70.2	**68.2**	**70.7**	**69.7**
4	拼接	70.1	67.6	**71.1**	69.6
5	自适应（2020）	**70.3**	67.4	69.8	69.2

表 8：DFN 消融实验。在仅使用体素和仅使用柱的消融实验中，参数规模相当。

与单一表示方法相比，mAPH。此外，配备 SFL 后，VPF_{sp} 的性能显著超过基线方法。

密集融合颈部。 在表8中，我们分析了 DFN 的影响。我们首先研究，将单一体素或柱体分支作为密集颈部是否能够实现相似的性- 能。为此，我们在最终的 SFL 中保留单一体素或柱体分支之一，并按照（Yin、Zhou 和 Krahenbuhl 2021；GuangshengShi 2022）将密集颈部应用于孤立的密具体素或柱特征图。表8 的前三行验证了所提出的融合颈部是首选方案。此外，我们在 3rd-5th 行中探索了多种融合策略，包括逐元素求和、拼接和门控融合（Yoo 等人 2020）。尽管性能结果相近，我们仍因求和操作的简洁性而采用该方法。

定性结果
在图8中，我们展示了一些定性结果。前两行表明，VPF 能够准确识别分布紧密或部分被遮挡的物体。我们还在最后两行展示了一些失败案例，其中显示了前景点与背景点相互交错，或点极其稀疏的情况。据我们了解，



（a）VPF_{de}　　（b）VPF_{sp}　　（c）VPF 轻量版

图 6：VPF 各变体的模型架构。（a）、（b）和（c）分别展示了 VPFde、VPFsp 以及所有 VPF 轻量版模型所采用的双分支稀疏卷积块。

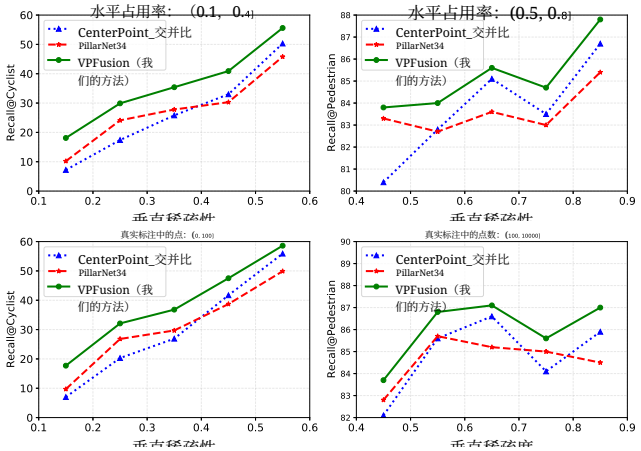


图 7：不同条件下召回率与垂直密度的对比。我们的方法在各种场景和垂直密度下均表现出更优的性能。

ing，整合图像中的纹理细节或累积过去帧可能有助于应对这些情况。

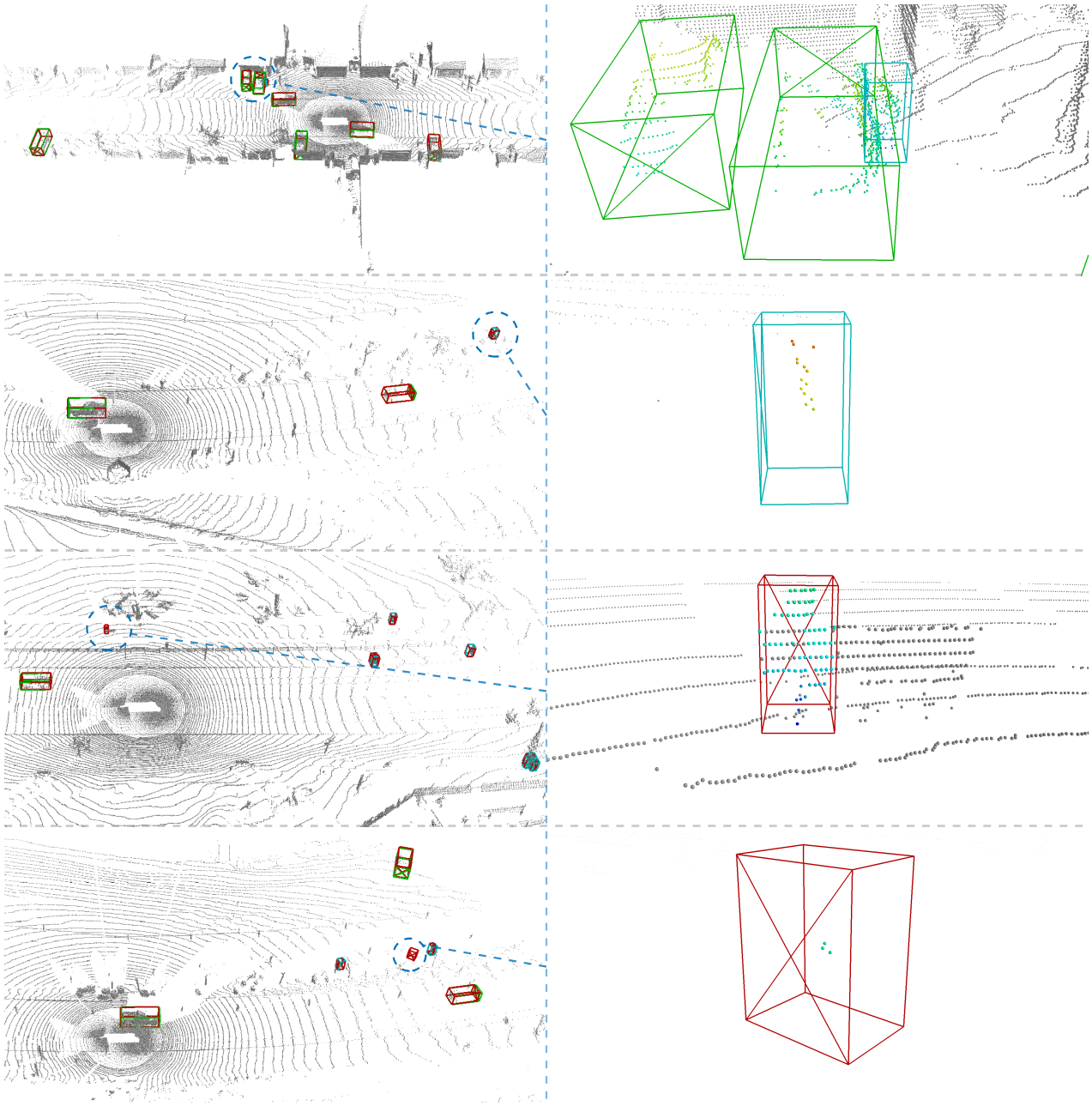


图 8：Waymo开放数据集 验证集上的可视化。车辆、行人和骑行者的预测边界框分别以绿色、青色和蓝色表示；真实标注以红色突出显示。第一列展示完整场景，第二列则详细展示特定的局部实例。