arXiv:2304.00525v1 [cs.CV] 2 Apr 2023

一次训练，多种部署：基于极坐标的自适应鸟瞰感知在自动驾驶中的应用

Huitong Yang[1], Xuyang Bai[2],Xinge Zhu[3⋆]，以及 Yuexin Ma[1⋆]

摘要—当前车载芯片的算力通常各不相同，这意味着要将同一种基于学习的算法适配到不同芯片上，需要进行多次训练，耗费大量计算资源。对于大型模型的三维感知方法而言，情况会更加严峻。以往以视觉为中心的三维感知方法使用固定分辨率的规则网格特征图进行训练，因此无法适配其他网格尺度，限制了更广泛的部署。本文在从图像构建鸟瞰图特征图时采用极坐标表示，以实现一次训练、多种部署的目标。具体而言，极坐标空间中沿射线分布的特征可以轻松地自适应采样，并投影到任意分辨率的笛卡尔空间特征中。为进一步提升适应能力，我们让多尺度上下文信息相互交互，以增强特征表示。在大规模自动驾驶数据集上的实验表明，凭借一次训练、多种部署这一优势，我们的方法优于其他方法。

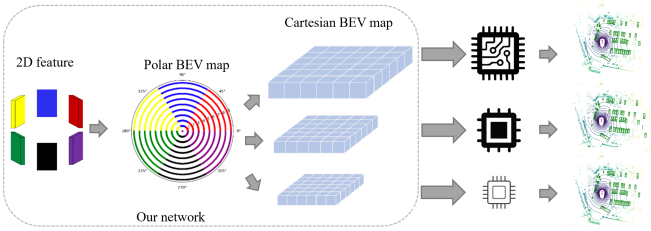

图 1：得益于基于极坐标的特征图表示具有网格无关采样特性，我们的网络能够在部署时根据实际计算预算，为三维感知提供多种笛卡尔鸟瞰图网格分辨率。

一、引言

准确的三维感知对于自动驾驶汽车理解周围场景的复杂性，并做出安全、有效的驾驶决策至关重要。大量研究致力于这一任务，其中以视觉为中心的方法因计算成本较低而日益普及。尽管这类基于学习的方法已取得令人瞩目的性能，但通常是在预先设定的特征图分辨率下训练的，因此只能部署到具备相应计算资源的有限车载芯片上，无法适用于计算预算任意变化的多种部署场景。采用不同分辨率重新训练模型会额外消耗大量计算资源。此外，芯片上运行的任务并非只有感知，其他任务同样会占用计算资源。动态分配资源能够提高芯片的服务效率，但这也要求感知算法具备动态调整能力。因此，能够自适应调整计算需求，并部署到计算性能各异的芯片上的以视觉为中心的方法至关重要。

当前主流的以视觉为中心的三维感知方法，通常会将透视视图（PV）图像特征图投影到鸟瞰图（BEV）特征图上，再附加检测头或分割头，以完成特定的3D感知任务。然而，这类方法在笛卡尔坐标系中构建规则网格表示的 BEV 特征图，并在训练期间为网格指定固定的物理尺度对应关系，因此难以适配其他分辨率。直接对 BEV 特征图进行下采样或上采样以满足不同的计算需求，会导致三维感知性能大幅下降。实际上，与规则网格表示相比，BEV 上的极坐标表示更为合理，因为根据相机成像的缩短效应，每条极坐标射线在物理上都可以对应图像像素的一列 [34, 19] 。更重要的是，沿每条射线的特征不受网格影响，可以灵活地投影到分辨率不同的多种笛卡尔特征图上，如图1所示。也就是说，即使网络是在将极坐标特征转换为128×128 笛卡尔特征图的条件下训练的，预训练的极坐标特征也能轻松适配分辨率为 256×256 或 64×64 的新笛卡尔特征图，以满足不同的计算预算分配需求。据我们所知，我们是首个利用基于极坐标的 BEV 表示这一特性，实现一次训练多种部署目标的研究团队。

本文基于极坐标鸟瞰图特征表示，提出一种新颖的、以视觉为中心的自动驾驶三维检测方法。该方法只需一次训练，即可灵活部署多次，以适应不同的计算预算。整个流水线包含三个主要模块，如图 2 所示，分别是列式 PV-to-BEV 变换器模块、与网格无关的极坐标到笛卡尔坐标采样模块，以及多尺度特征

⋆ 通讯作者[1]Huitong Yang 和 Yuexin Ma 任职于上海科技大学信息科学与技术学院，中国上海 201210。mayuexin@shanghaitech.edu.cn[2]Xuyang Bai 任职于香港科技大学计算机科学与工程系。xbaiad@connect.ust.hk[3]XingeZhu 任职于香港中文大学信息工程系。zx018@ie.cuhk.edu.hk

交互模块。第一个模块通过多头注意力和深度分布估计，将 PV 图像特征投影为 BEV 极坐标特征。对于第二个模块，极坐标系中沿射线分布的特征可以自适应采样，并投影为笛卡尔空间中具有不同网格尺寸的特征。因此，当计算预算有限时，我们的网络可以自然地将极坐标特征转换为较小的笛卡尔特征图，从而减少计算资源占用，同时保持高质量的性能。为进一步提升适应能力，我们设计了多尺度特征交互模块，通过 Transformer 机制增强特征表示。最后，我们接入一个检测头进行三维目标定位。我们在大规模自动驾驶数据集 nuScenes [3]，上评估了该方法，并在一次训练多种部署的条件下取得了当前最先进的性能。此外，我们还进行了消融研究，以验证网络中重要模块的有效性。

总之，我们的贡献如下：

- 我们提出了一种新颖的基于极坐标的BEV感知方法，训练完成后可适配不同计算预算，支持多种部署。
- 我们利用多尺度BEV特征之间的信息交互，增强特征表示能力，从而提升适应性。
- 在大规模自动驾驶数据集上进行三维检测时，我们的方法在利用未见分辨率的特征图进行推理方面，实现了最先进的泛化能力。

II. 相关工作

随着低成本自动驾驶技术的普及，以视觉为中心的BEV感知算法逐渐受到关注。我们的工作聚焦于这一前景广阔的技术，并进一步提出一种自适应部署方案，这对实际应用迫切且意义重大。我们从两个方面介绍相关工作：基于笛卡尔坐标的BEV感知方法和基于极坐标的方法。

A. 基于笛卡尔坐标系的 *BEV* 感知

当前的BEV感知方法 [23] 主要可分为四类：基于单应性的方法、基于深度的方法、基于MLP的方法和基于Transformer的方法。基于单应性的方法 [1, 46, 12, 4, 6, 20, 35, 29] 利用逆透视映射（IPM） [24] 建立PV视图与BEV视图之间的几何变换关系，从而支持探索许多下游视觉任务，例如三维目标检测、语义地图构建和运动规划。此外，基于深度的方法 [31, 9, 42, 28, 7, 2]显式或隐式地估计深度分布，并将其作为桥梁，从多视图图像特征生成BEV特征。LSS [27, 9]根据像素级深度分布构建相机视锥，并将其投影到世界空间。为获得更准确的深度估计，BEVDepth [14] 利用稀疏激光雷达点的深度值提供显式深度监督，从而显著提高三维检测的准确率。此外，一些方法 [8, 40,49, 25, 13, 30, 33] 利用MLP作为映射函数来进行视角变换。对于基于Transformer的方法，[17, 18, 32,26, 16]，它们通常通过位置编码设计一组BEV查询，然后利用BEV查询与图像特征之间的交叉注意力来实现视角变换。DETR3D [38]率先采用稀疏的三维可学习参考点作为非显式BEV查询，以关联来自多视图相机的二维特征。BEVFormer [15] 利用Transformer统一空间信息和时间信息，无需深度估计即可学习稠密BEV特征，并实现多任务学习。尽管上述算法表现出有竞争力的性能，但其特征图均是在笛卡尔坐标系中设计的，其中规则网格划分与物理世界之间具有固定的尺度对应关系。如果使用在某一尺度下训练的此类网络来推断其他尺度的输出，性能将大幅下降。

B. 基于极坐标的 *BEV* 感知

极坐标系是表示三维空间的另一种选择。它最早出现在基于激光雷达的三维感知方法中 [48,5] 。PolarNet [39] 将点云投影为极坐标下的 BEV 表示，并通过可学习网格单元对点进行处理，从而显著改善了空间点云分布中的长尾问题。Cylinder3D [47]提出了在圆柱坐标系中进行动态三维体素化，以适应室外激光雷达点云的密度变化特性。实际上，由于相机成像的缩短效应，极坐标表示也适用于基于相机的 BEV 感知任务。此外，基于极坐标的 BEV 表示 [11] 更适合表示自车环视相机所捕获的完整场景。遵循光线追踪原理，Ego3RT [21] 将稠密的极坐标 BEV 网格视作人眼，并利用相机参数构建三维到二维反向投影，以进行多任务学习。PolarBEV [19]建立了一种非均匀 BEV 网格表示，根据目标与相机之间的距离，通过角度特定和半径特定嵌入来考虑不同的目标尺度。我们的方法也利用 Transformer 将图像列投影为极坐标射线。然而，我们首次关注极坐标特征图的 BEV 网格采样不敏感特性，并利用这一特性，根据实际应用中的实时计算预算和推理速度实现多部署适配。

III. 方法

我们提出了一种新颖的以视觉为中心的基于极坐标的三维目标检测器，只需训练一次，即可根据计算预算部署为不同粒度的鸟瞰图特征图。如图 2 所示，我们的框架以多视角 RGB 图像为输入，并输出 3D边界框及其类别。按照 [9],环视图像首先由单个共享骨干网络编码，以获得多尺度二维表示 $\{F_i\}_{i=1}^{scale}$

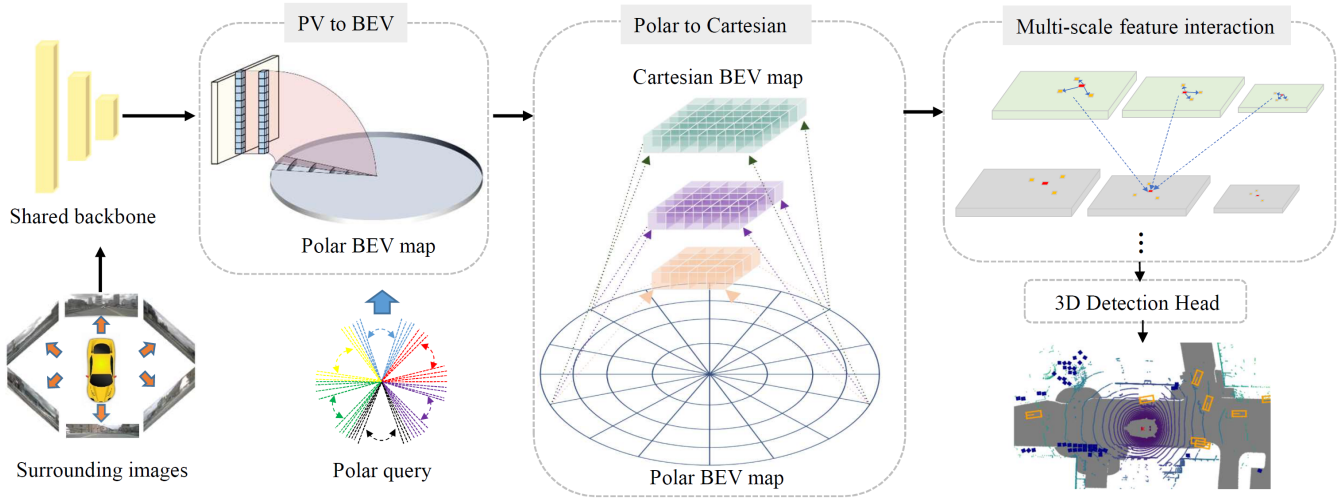

图 2：我们方法的整体流程。我们利用标准的 2D骨干网络提取多视角特征。随后，列式 PV-to-BEV 变换器模块通过交叉注意力机制，将图像特征图中的列投影到相应的极坐标射线上。接着，我们提出一种网格无关的极坐标到笛卡尔坐标采样策略，将极坐标BEV地图转换为不同粒度的笛卡尔 BEV 网格，并可根据实际自动驾驶场景中的计算预算，在部署时进行调整。此外，我们还提出一种多尺度鸟瞰图特征交互编码器，以关联多尺度笛卡尔BEV地图中的上下文信息。最后，添加一个检测头，用于预测目标的三维边界框和类别。

尺寸为 $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C$，随后使用 FPNLSS 将其统一为单一分辨率的二维特征图。为了从图像中执行三维感知任务，主流方法通常先将 PV特征图投影到 BEV空间，这种表示可直接用于许多下游实际应用，例如行为预测、运动规划等。然而，由于这些方法通常在训练期间采用具有预定义网格尺寸的规则网格笛卡尔BEV表示，因此难以在推理时适配不同尺度，这限制了它们在训练过程中计算预算各异的自动驾驶场景中的应用。为缓解这一问题，并设计出能够在不同计算预算下灵活部署的检测器，我们采用极坐标表示，它依据成像原理以更自然的方式表示物理世界。更重要的是，沿每条射线的特征不受网格影响，可用于投影到多个不同分辨率的笛卡尔 BEV特征图。通过这种方式，我们的方法无需针对不同计算预算下的部署进行多次训练，从而节省大量时间和资源；这对于依赖有限车载芯片的应用至关重要。为实现这一目标，我们提出了三个新颖模块，包括列式 PV-to-BEV 变换器模块、网格无关的极坐标到笛卡尔坐标采样模块，以及多尺度特征交互模块，后文将对此进行介绍。

A. 列式 *PV-to-BEV* 变换器模块

视觉中心感知算法中的一个关键步骤是将视角从PV变换到BEV，因为BEV是对世界的自然表示，且便于融合。主流方法可分为两类：基于几何的变换和基于网络的变换 [23]。前者充分利用相机的物理原理，以可解释的方式转换视角，但可能受到深度预测不准确的影响。后者以数据驱动方式进行视角变换，而不显式利用相机投影模型。在我们的工作中，我们通过将单图像视角变换器 [34]扩展到多视角场景，构建了一个基于网络的视角变换模块。如图2所示，我们首先假设图像中的垂直扫描线与从相机中心出发、位于BEV平面上的射线之间存在一一对应关系。随后，视角变换可表述为一组一维序列到序列转换问题，并由一个变换器模块建模。这种按列处理的变换器模块避免了二维图像特征图与BEV查询之间的密集注意力计算，转而仅依赖一维序列到序列转换，从而形成一种节省内存且数据高效的架构。

具体而言，我们首先在极化鸟瞰平面上初始化一组位置嵌入（称为极坐标查询）。接下来，根据相机几何建立的每个图像列与极坐标射线之间的映射关系，通过对该列上的图像特征与该射线上的极坐标查询执行交叉注意力，将图像特征从图像列传递到其对应的射线。此过程可视为将语义透视视图特征可学习地分配到鸟瞰图平面中沿极坐标射线排列的位置槽上。与单张图像的情况不同，在 [34], 中，多张环视图像通常彼此重叠；重叠区域的多视角观测对许多检测目标至关重要，例如在单张图像中被截断的大型物体。

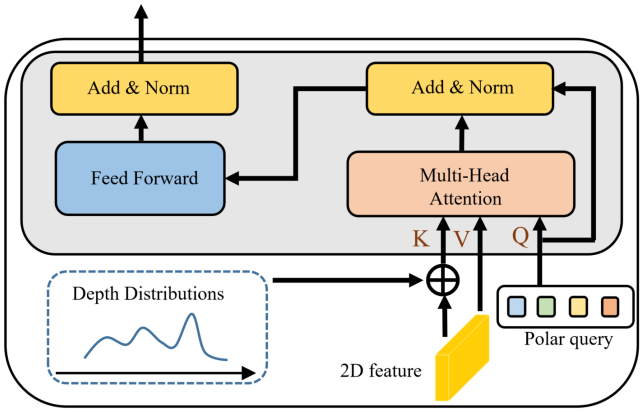

图 3：列式 PV-to-BEV 变换器模块的详细网络结构。

因此，为了增强重叠区域的信息，我们统计投影到同一射线上的图像列数量，并自适应地分配更高的注意力权重，使我们的列式交叉注意力模块能够更好地处理来自重叠区域的信息。

为进一步增强变换器模块的几何推理能力，我们将深度信息注入视角变换过程，如图 3 所示。我们首先按照[14] 预测每个图像像素的深度分布，然后使用多层感知机网络将深度分布嵌入到位置嵌入中，再将其与二维特征相加，为 PV-to-BEV 变换提供深度引导。尽管我们的列式 PV-to-BEV 变换模块在视角变换时不一定需要逐像素深度，但我们仍通过实验发现，深度信息对于 Transformer 的几何推理至关重要。

B. 网格无关的极坐标到笛卡尔坐标采样模块

为了能够根据实际计算预算在部署时调整感知粒度，我们提出了一种网格无关采样模块，使网络能够适应不同的鸟瞰图特征分辨率。具体而言，我们首先在多视角感知范围$[-51.2m, 51.2m]$ 内生成一个具有相应分辨率的笛卡尔网格 $M \in R^{H\times W\times 2} = (w_i,\ h_i)$，其中 $H \times W$ 表示鸟瞰图特征图的空间分辨率。在极坐标到笛卡尔坐标的特征变换过程中，将利用该网格对极坐标特征图进行采样，从而得到笛卡尔鸟瞰图特征图。极坐标系与笛卡尔坐标系之间的转换计算如下：

$$\phi^r = \arctan(\frac{w_i}{h_i}), \sigma^r = \sqrt{({w_i}^2 + {h_i}^2)}, \tag{1}$$

其中，(ϕ^r, σ^r) 是极坐标特征图的坐标。为覆盖 $[-51.2m, 51.2m]$ 的感知范围，我们将极坐标系的范围设为 $\sigma^r \in [0m, 72m]$，$\phi^r \in [0, 2\pi]$，并对极坐标到笛卡尔坐标变换进行归一化。最后，我们将归一化后的极坐标堆叠起来，得到极坐标到笛卡尔坐标变换的鸟瞰图网格：

$$\begin{aligned} \phi^r &= (\phi^r - P^r_{\min})/(P^r_{\max} - P^r_{\min}), \\ \sigma^r &= (\sigma^r - \sigma^r_{\min})/(\sigma^r_{\max} - \sigma^r_{\min}), \\ P &= \text{Stack}([\phi^r, \sigma^r)] \in R^{H\times W\times 2}. \end{aligned} \tag{2}$$

随后，我们的网格无关采样器可通过双线性插值运算，将极坐标BEV地图转换为具有不同分辨率的笛卡尔 BEV 网格。网格无关采样器生成的多分辨率BEV特征图，使单一网络能够在推理过程中适配不同粒度的特征图，从而支持采用不同计算预算进行部署。

C. 多尺度特征交互模块

为捕获多尺度上下文信息，并增强不同分辨率的多个BEV特征图的表征能力，我们提出了多尺度BEV交互编码器（MBIE），用于从采样得到的笛卡尔 BEV 特征图中提取不同粒度的上下文信息。遵循 [11]，，我们通过堆叠多个可变形注意力层来构建MBIE模块，以融合不同尺度的信息。形式上，可变形注意力算子定义如下：

$$\begin{aligned} &\text{MSBEVDeformAttn}(b_q, \hat{r_q}, \{b^s\}_{s=1}^S) = \\ &\sum_h^H W_h[\sum_{s=1}^S \sum_{r=1}^R A_{hsqr} {W_h}' b^s(\sigma_s(\hat{r_q}) + \Delta f_{hsqr})], \end{aligned} \tag{3}$$

其中，q 是BEV查询特征b_q中的一个查询元素，$\hat{r_q} \in [0,1]^2$ 表示参考点的归一化坐标，b^s 是多尺度BEV特征，$\sigma_s(*)$是重缩放算子，用于将 r_q 转换到相应尺度；Δf_{hsqr} 是偏移量，有助于查询特征与来自所有层级的相邻上下文BEV信息进行交互。MBIE在相邻像素之间以及多个笛卡尔 BEV 特征图之间执行信息交换，从而增强每个特征图的多尺度上下文，并成为泛化到不同BEV特征分辨率的关键步骤。

在 MBIE 模块之后，采用基于可变形卷积的特征金字塔网络模块 [10]，将不同分辨率的鸟瞰图特征图融合至目标分辨率。由于可以手动控制特征金字塔网络中的下采样或上采样倍率，因此便于在推理过程中调整输出特征的分辨率。推理策略的更多细节将在第 IV-C 节中说明。

给定由前序阶段生成的稠密鸟瞰图特征图，鉴于其简洁性和广泛应用，我们采用 CenterPoint 的第一阶段作为检测头。遵循[41]，，我们采用热力图损失进行分类监督，并采用 L1回归损失进行边界框监督。

表 I：本表展示了在固定训练分辨率下，不同输出分辨率对应的结果，这也符合我们的初衷，即一次训练多种部署。- 表示该方法没有提供此数值。

方法	BEV 网格尺寸	输出分辨率	训练头	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVDet	0.4 米	128×128	128×128	0.392	0.312	0.691	**0.272**	0.523	0.909	0.247
	0.256 米	200×200		0.324	0.244	0.888	0.308	0.606	0.958	0.211
	0.2 米	256×256		0.275	0.201	1.004	0.328	0.702	1.010	0.221
BEVerse	0.4 米	128×128	128×128	-	0.321	0.682	0.278	**0.464**	**0.327**	**0.189**
	0.256 米	200×200		-	0.275	0.804	0.318	0.543	**0.517**	**0.195**
	0.2 米	256×256		-	0.230	0.898	0.337	0.615	**0.662**	**0.194**
Ours	0.4 米	128×128	128×128	**0.398**	**0.321**	**0.669**	0.275	0.494	0.956	0.231
	0.256 米	200×200		**0.378**	**0.305**	**0.710**	**0.280**	**0.519**	1.024	0.241
	0.2 米	256×256		**0.346**	**0.278**	**0.789**	**0.300**	**0.568**	0.991	0.252

表 II：更具挑战性的设定结果 性的设定。可以发现，这些输出分辨率都是完全未见过的，并且灵活 le.

方法	BEV 网格尺寸	输出分辨率	训练头	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	平均速度误差↓	mAAE↓
BEVDet	0.8 米	64×64	128×128	0.181	0.092	1.078	0.505	0.781	0.877	0.486
	0.53 米	96×96		0.336	0.231	0.860	0.301	0.575	0.842	0.217
	0.4 米	128×128		0.392	0.312	0.691	**0.272**	0.523	0.909	0.247
BEVerse	0.8 米	64×64	128×128	-	0.107	**1.054**	0.365	0.641	**0.510**	**0.251**
	0.53 米	96×96		-	0.254	0.778	0.296	0.509	**0.385**	**0.187**
	0.4 米	128×128		-	0.321	0.682	0.278	**0.464**	**0.327**	**0.189**
Ours	0.8 米	64×64	128×128	**0.290**	**0.143**	1.057	**0.288**	**0.589**	0.713	0.217
	0.53 米	96×96		**0.370**	**0.280**	**0.749**	**0.279**	**0.504**	0.951	0.223
	0.4 米	128×128		**0.398**	**0.321**	**0.669**	0.275	0.494	0.956	0.231

IV. 实验

本节首先介绍数据集、评估指标和实现细节。随后，我们详细比较了不同基线方法的泛化能力。此外，我们在 nuScenes 数据集上使用多种方法评估我们的方法，以展示其有效性，并通过大量消融实验验证各个提出的组件。

A. 数据集与指标

我们在大规模数据集 nuScenes [3]，上评估了所提出的方法。该数据集包含 1000 个场景，每个场景持续 20 秒，每帧包含 6 个分辨率为 1600×900 的环视相机。对于三维目标检测，我们主要使用以下指标评估性能：平均精度均值 (mAP)、平均平移误差 (ATE)、平均尺度误差 (ASE)、平均方向误差 (AOE)、平均速度误差 (AVE)、平均属性误差 (AAE)，以及 nuScenes 检测得分 (NDS)，其计算方式为 (NDS $= \frac{1}{10}[5 * \text{mAP} + \sum_{mTP\in \text{TP}}(1 - \min(1, \text{mTP}))]$)。

表 III：所提出模块的消融实验。CPBT 表示按列处理的 PV到BEV变换器模块。MBIE 表示多尺度特征交互模块。

CPBT	MBIE	NDS↑	mAP↑
		0.384	0.294
✓		0.391	0.307
✓	✓	**0.398**	**0.321**

B. 实现细节

为公平比较，我们遵循现有工作 [14, 15, 42]，采用预训练的 Swin-Transformer 作为二维骨干网络。我们的感知范围为 $[-51.2m, 51.2m]$，沿 X 轴和 Y 轴。考虑到多视图相机的视图对称性，半径和方位角设为 (144,360)，表示极坐标感知范围。输入的多视图

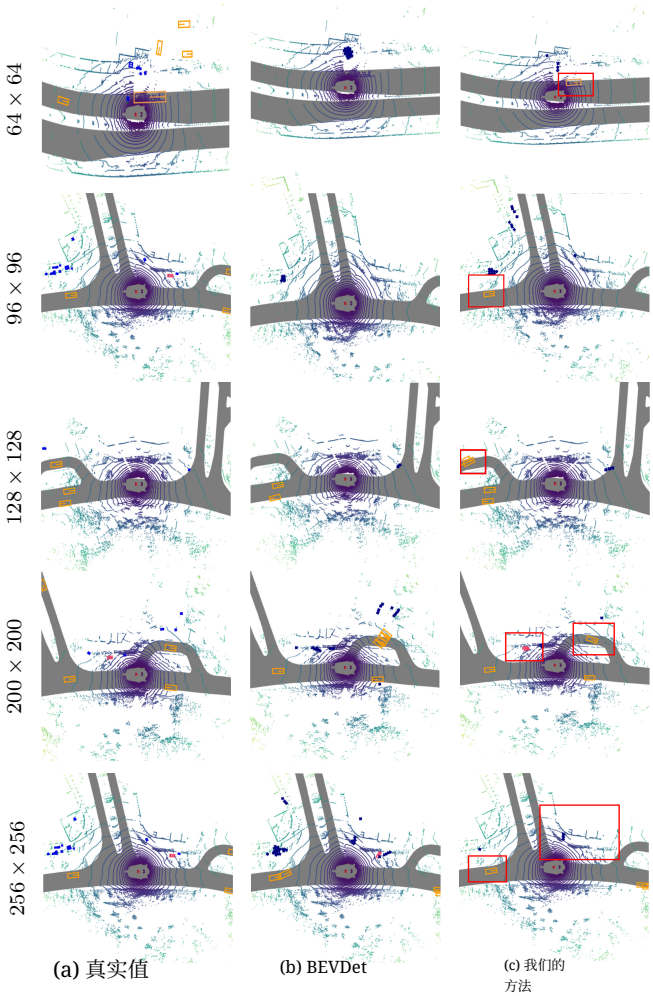

图 4：表 I 和 II 中实验结果的可视化。我们用蓝色矩形标出我们的方法预测得更准确的区域。

表 IV：在 nuScenes 验证集上对近期以视觉为中心的三维检测工作的定量评估。在所有方法中，DETR3D、PETR、BEVDet 和我们的方法均采用 CBGS [45] 进行训练。

方法	骨干网络	Size	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	平均速度误差↓	mAAE↓
CenterNet [44]	DLA	–	0.328	0.306	0.716	0.264	0.609	1.426	0.658
FCOS3D [36]	ResNet-50	1600×900	0.368	0.288	0.777	0.266	0.544	1.228	0.170
DETR3D [38]	ResNet-50	1600×900	0.373	0.302	0.811	0.282	0.493	0.979	0.212
PGD [37]	ResNet-50	1600×900	0.394	0.320	0.735	0.266	0.492	1.114	0.170
Ego3RT [22]	ResNet-50	1600×900	0.380	0.332	0.706	0.281	0.663	0.964	0.249
PETR [17]	ResNet-50	1056×384	0.381	0.313	0.768	0.278	0.564	0.923	0.225
BEVDet [9]	ResNet-50	704×256	0.379	0.298	0.725	0.279	0.589	0.860	0.245
BEVDet [9]	ResNet-101	704×256	0.381	0.302	0.722	0.269	0.543	0.900	0.269
BEVDet [9]	Swin-tiny	704×256	0.392	0.312	0.691	0.272	0.523	0.909	0.247
PersDet [43]	ResNet-50	704×256	0.389	0.319	0.676	0.284	0.589	0.924	0.229
CaDDN [28]	ResNet-50	704×256	0.370	0.294	0.702	0.283	0.579	0.988	0.222
Ours	Swin-tiny	704×256	**0.398**	**0.321**	**0.669**	**0.275**	**0.494**	**0.956**	**0.231**

图像缩放至 (256, 704)，用于原型验证。网络使用 CBGS [45] 进行训练，在 4 块 RTX 3090 GPU 上的总批量大小为 8。推理时，我们在固定感知范围下，将模型应用于不同的 BEV 网格分辨率，并将极坐标特征图采样为具有不同网格分辨率的矩形笛卡尔BEV特征图，以适应不同的计算预算。就推理速度而言，我们的方法在 64×64BEV分辨率下达到 9 FPS，在 256×256 BEV分辨率下达到 6.5 FPS。

C. 一次训练多种部署

本节展示了我们的方法在一次训练多种部署设置下的结果。具体而言，所提出的方法仅使用一个检测头（128×128特征图）进行训练，但能够在尺寸各异的多个特征图上进行推理，例如 200×200 和256×256，这也代表了不同的计算预算。得益于极坐标特征空间的分辨率不变特性，我们的方法可以通过不同程度的采样来适配不同的鸟瞰图网格尺寸。而对于 BEVDet [9]和 BEVerse [42]，等现有方法，它们无法直接应用于其他分辨率。为进行比较，我们对这些方法进行了修改：在鸟瞰图特征图上执行插值，同时保持检测头不变，从而使其适配不同的输出尺寸。其他一些相关的 BEV感知方法 [38, 17, 18], 并未构建显式的鸟瞰图特征图，也缺乏支持不同计算预算下多种部署的能力，因此这里不与它们进行比较。

如表 I 所示，所有方法均使用大小为 128×128 的特征图进行训练，并使用不同的特征图进行推理（200×200 和 256 ×256）。需要注意的是，输出特征大小不同意味着鸟瞰图网格的分辨率不同，也意味着计算预算不同。可以发现，与这些基线方法相比，我们的方法在几乎所有指标上都取得了稳定的性能提升。具体而言，在不同的输出特征图（200×200 和 256 ×256 ）上进行推理时，我们的方法展现出良好的泛化能力，性能仅略有下降，例如 NDS 分别下降 -2%（128×128 ⟶ 200×200）和 -5%（128×128 ⟶ 256×256）；而 BEVDet [9] 的性能则大幅下降，分别为 -7% 和 -12%。需要注意的是，BEVerse 在平均速度误差、mAAE 和 NDS 指标上表现更好，这是因为它利用了多帧图像作为输入。在我们的工作中，我们更关注不同输出分辨率之间的泛化能力从输出分辨率不同时进行推理所导致的 mAP 仅略有下降，可以清楚地看出这一点。

我们还在图 4 中提供了可视化结果。从突出显示的区域可以明显看出，我们的方法具有更准确的检测能力，尤其是在适配其他特征图尺度时。这进一步证明了我们方法的优越性。

更具挑战性的部署 在表 II 中，我们在一种更具挑战性且灵活的设置下进行推理，其中使用了 64×64 和 96 ×96 输出特征图，训练头也为 128 × 128。值得注意的是，这些输出特征图具有灵活性，且在训练期间未曾出现。从结果中可以看出，与基线方法相比，我们的方法取得了最佳的泛化性能，性能下降幅度也最小，表明其在这种具有挑战性的设置下具有良好的泛化能力。

D. 消融研究与定性评估

为了验证网络中重要模块的有效性，我们进行了消融研究，结果如表 III 所示。我们可以看到，基于极坐标的鸟瞰图特征表示优于纯笛卡尔坐标特征表示，而多尺度特征交互过程进一步提升了性能。我们还在表 IV 中展示了以视觉为中心的BEV感知的定量评估结果。与其他方法相比，我们的方法表现出色且具有竞争力。

V. 结论

本文提出了一种面向自动驾驶的新型视觉中心鸟瞰图感知方法，可根据不同的计算预算灵活部署到多个平台。具体而言，我们利用极坐标表示的鸟瞰图特征图来解决这一问题，因为它具有良好的性质：沿射线的特征可以轻松映射为任意分辨率的笛卡尔网格表示的特征。我们还设计了一种有效的交互模块，通过多尺度特征交互进一步增强特征表示。大量实验表明，我们的方法仅需训练一次，便能在推断新特征图尺度时表现出优异的泛化能力，这对在多样化场景中实际部署具有重要意义。

VI. 致谢

本研究得到国家自然科学基金委员会（编号：62206173）、上海市扬帆计划（编号：22YF1428700）以及上海人本人工智能前沿科学中心（ShangHAI）的资助。

参考文献

[1] Syed Ammar Abbas 和 Andrew Zisserman。《一种从图像获取鸟瞰图的几何方法》。载于：*ICCV*。IEEE，2019年，第 4095–4104 页。[2] Adil Kaan Akan 和 Fatma G¨uney。《StretchBEV：在空间和时间维度上扩展未来实例预测》。载于：*arXiv*（2022年）。[3] Holger Caesar 等。《nuscenes：自动驾驶多模态数据集》。载于：*IEEE/CVF*计算机视觉与模式识别会议论文集。2020年，第 11621–11631 页。[4] Yigit Baran Can 等。《使用车载摄像头理解道路语义的鸟瞰图》。载于：*IEEE*机器人与自动化快报7.2（2022年），第 3302–3309 页。[5] Peishan Cong 等。《STCrowd：拥挤场景下行人感知的多模态数据集》。载于：*CVPR*。2022年6月，第 19608–19617 页。[6] Noa Garnett 等。《3d-lanenet：端到端三维多车道检测》。载于：*ICCV*。2019年，第 2921–2930 页。[7] Nikhil Gosala 和 Abhinav Valada。《使用单目正视图像进行鸟瞰图全景分割》。载于：*IEEE*机器人与自动化快报7.2（2022年），第 1968–1975 页。[8] Noureldin Hendy 等。《FISHING Net：网格中语义热力图的未来推断》。载于：*arXiv*abs/2006.09917（2020年）。[9] Junjie Huang 等。《BEVDet：鸟瞰视角下高性能多摄像头三维目标检测》。载于：*CoRR* abs/2112.11790（2021年）。[10] Shihua Huang 等。《FaPN：用于密集图像预测的特征对齐金字塔网络》。载于：*2021 IEEE/CVF*国际计算机视觉会议（*ICCV*）（2021年），第 844–853 页。[11] Yanqin Jiang 等。《PolarFormer：采用极坐标变换器的多摄像头三维目标检测》。载于：*arXiv*预印本 *arXiv:2206.15398*（2022年）。[12] Youngseok Kim 和 Dongsuk Kum。《基于逆透视映射图像的深度学习车辆位置与姿态估计》。载于：*2019 IEEE*智能车辆研讨会（*IV*）。IEEE，2019年，第 317–323 页。[13] Qi Li 等。《Hdmapnet：在线高清地图构建与评估框架》。载于：*arXiv*（2021年）。[14] Yinhao Li 等。《BEVDepth：为多视角三维目标检测获取可靠深度》。载于：*CoRR*（2022年）。[15] Zhiqi Li 等。《BEVFormer：通过时空变换器从多摄像头图像中学习鸟瞰视角表示》。[16] Zhiqi Li 等. “Panoptic SegFormer：借助 Transformer 深入探索全景分割”。载于：*CVPR*。2022年，第 1280–1289 页。[17] Yingfei Liu 等. “PETR：用于多视图三维目标检测的位置嵌入变换”。载于：*arXiv*(2022年)。[18] Yingfei Liu 等. “PETRv2：用于多摄像头图像三维感知的统一框架”。载于：*arXiv*(2022年)。[19] Zhi Liu 等. 基于极坐标栅格化与表面估计的视觉非均匀鸟瞰图表示学习。2022年。[20] Abdelhak Loukkal 等. “在扁平汽车之间驾驶：单目相机生成的鸟瞰图占用栅格，用于整体轨迹规划”。载于：冬季计算机视觉应用会议。2021年，第 51–60 页。[21] Jiachen Lu 等. “通过光线追踪学习自车三维表示”。载于：*arXiv* 预印本 *arXiv:2206.04042*(2022年)。[22] Jiachen Lu 等. “通过光线追踪学习自车三维表示”。载于：*arXiv*(2022年)。[23] Yuexin Ma 等. “以视觉为中心的BEV感知：综述”。载于：*arXiv* 预印本 *arXiv:2208.02797* (2022年)。[24] Hanspeter A Mallot 等. “逆透视映射简化光流计算与障碍物检测”。载于：生物控制论 (1991年)，第 177–185 页。[25] Bowen Pan 等. “用于感知周边环境的跨视角语义分割”。载于：*IEEE*机器人与自动化快报5.3 (2020年)，第 4867–4873 页。[26] Lang Peng 等. “BEVSegFormer：来自任意相机配置的鸟瞰图语义分割”。载于：*arXiv* (2022年)。[27] Jonah Philion 和 Sanja Fidler. “Lift, Splat, Shoot：通过隐式反投影至三维，对任意相机配置拍摄的图像进行编码”。载于：*ECCV*。Andrea Vedaldi 等编。施普林格，2020年，第 194–210 页。[28] Cody Reading 等. “用于单目三维目标检测的类别深度分布网络”。载于：*CVPR*。2021年，第 8555–8564 页。[29] Lennart Reiher、Bastian Lampe 和 Lutz Eckstein. “一种仿真到现实深度学习方法：将来自多台车载相机的图像转换为鸟瞰视角语义分割图像”。载于：智能交通系统会议。IEEE。2020年，第 1–7 页。[30] Thomas Roddick 和 Roberto Cipolla. “利用金字塔占用网络从图像预测语义地图表示”。载于：*CVPR*。2020年，第 11138–11147 页。[31] Thomas Roddick、Alex Kendall 和 Roberto Cipolla. “用于单目三维目标检测的正交特征变换”。载于：*arXiv* (2018年)。[32] Won Kyoung Roh 等. “ORA3D：重叠区域感知的多视角三维目标检测”。载于：*arXiv*abs/2207.00865 (2022年)。

[33] Avishkar Saha 等. “在鸟瞰视角车辆估计中实现时空聚合”。载于：*ICRA* (2021年)，第 5133–5139 页。[34] Avishkar Saha 等. “将图像转换为地图”。载于：*ICRA*。IEEE。2022年。[35] Sunando Sengupta 等. “从街景图像自动生成稠密视觉语义地图”。载于：*IROS*。IEEE。2012年，第 857–862 页。[36] Tai Wang 等. “FCOS3D：全卷积单阶段单目三维目标检测”。载于：*IEEE/CVF*国际计算机视觉会议研讨会，*ICCVW 2021*，蒙特利尔，*BC*，加拿大，*2021*年*10*月*11*日至*17*日。IEEE，2021年，第 913–922 页。[37] Tai Wang 等. “概率与几何深度：透视场景中的目标检测”。载于：机器人学习会议，*2021*年*11*月*8*日至*11*日，伦敦，英国。Aleksandra Faust、David Hsu 和 Gerhard Neumann 编。第 164 卷。机器学习研究论文集。PMLR，2021年，第 1475–1485 页。[38] Yue Wang 等. “DETR3D：通过三维到二维查询实现多视图图像三维目标检测”。载于：机器人学习会议，*2021*年*11*月*8*日至*11*日，伦敦，英国。Aleksandra Faust、David Hsu 和 Gerhard Neumann 编。PMLR，2021年，第 180–191 页。[39] Xiongwei Wu、Doyen Sahoo 和 Steven C. H. Hoi. “PolarNet：学习优化用于基于关键点的目标检测的极坐标关键点”。载于：第九届国际学习表征会议，*ICLR 2021*，线上会议，奥地利，*2021*年*5*月*3*日至*7*日。OpenReview.net，2021年。[40] Weixiang Yang 等. “专注地投射视角：通过跨视角变换进行单目道路场景布局估计”。载于：*CVPR*。2021年，第 15536–15545 页。[41] Tianwei Yin、Xingyi Zhou 和 Philipp Krahenbuhl. “基于中心的三维目标检测与跟踪”。载于：*IEEE/CVF*计算机视觉与模式识别会议论文集。2021年，第 11784–11793 页。[42] Yunpeng Zhang 等. “BEVerse：面向视觉中心自动驾驶的鸟瞰图统一感知与预测”。载于：*arXiv*(2022年)。[43] Hongyu Zhou 等. “PersDet：透视鸟瞰图中的单目三维检测”。载于：*CoRR*abs/2208.09394 (2022年)。[44] Xingyi Zhou、Dequan Wang 和 Philipp Kr¨ahenb¨uhl. “物体即点”。载于：*CoRR*abs/1904.07850 (2019年)。[45] Benjin Zhu 等. “用于点云三维目标检测的类别均衡分组与采样”。载于：*CoRR*abs/1908.09492 (2019年)。[46] Minghan Zhu 等. “ 利用单应性通过未标定交通摄像机进行单目三维车辆检测”。载于：*IROS*。2021年。[47] Xinge Zhu 等. “用于基于激光雷达感知的圆柱形与非对称三维卷积网络”。载于：*TPAMI* (2021年)。[48] XingeZhu 等人。“SSN：基于点云的多类目标检测形状特征网络”。发表于：*ECCV* （2020年）。[49] Jiayu Zou 等人。“HFT：通过混合特征变换提升透视表示”。发表于：*arXiv*（2022年）。