doi:10.19562/j.chinasae.qcgc.2025.ep.001

PolarDet:基于位置与语义信息加权的极坐标BEV端到端3D目标检测算法*

时培成[1],戈润帅[1],Chadia Chakir[1],董心龙[1],梁涛年[2],杨爱喜[3]
(1. 安徽工程大学机械与汽车工程学院,芜湖 241000; 2. 奇瑞新能源汽车股份有限公司,芜湖 241000;
3. 浙江大学工程师学院,杭州 310015)

[摘要] 传统的笛卡尔坐标系下的3D目标检测方法因车载相机的固定楔形成像几何,导致相机图像编码时在一定程度上忽视了目标在不同视角下的对称性和连续性。鉴于此,本文提出一种基于位置与语义信息加权的极坐标BEV端到端3D目标检测方法—PolarDet。该方法通过极坐标查询与预定义的极坐标网格生成极坐标下的BEV位置与语义信息,再与前一帧的BEV信息进行特征交互以融入时间信息;在输出最终检测结果时,PolarDet再对位置与语义信息进行加权求和,以提高信息的利用效率,使网络能够达到更高的检测精度。本文在具有挑战性的BEV目标检测nuScenes数据集上进行了广泛的实验,结果表明,PolarDet最优模型的mAP(平均精度)达到0.469,NDS(nuScenes检测得分)达到0.56,显著优于基于笛卡尔坐标的BEV目标检测方法。

关键词:极坐标;BEV目标检测;位置和语义信息;特征加权;跨平面编码器

PolarDet: An End-to-End 3D Object Detection Algorithm in Polar Coordinates Based on Position and Semantic Information Weighting

Shi Peicheng[1], Ge Runshuai[1], Chakir Chadia[1], Dong Xinlong[1], Liang Taonian[2] & Yang Aixi[3]
1. *School of Mechanical and Automotive Engineering, Anhui Polytechnic University, Wuhu* 241000;
2. *Chery New Energy Automobile Co., Ltd., Wuhu* 241000;
3. *Polytechnic Institute, Zhejiang University, Hangzhou* 310015

[**Abstract**] Traditional 3D object detection methods in Cartesian coordinate systems often overlook the symmetry and continuity of the target from different perspectives to some extent during camera image encoding due to the fixed wedge-shaped imaging geometry of in-vehicle cameras. To address this, in this paper, PolarDet, an innovative end-to-end 3D object detection method in polar coordinates based on position and semantic information weighting is proposed. This method generates BEV (Bird's Eye View) position and semantic information in polar coordinates through polar coordinate queries and predefined polar grid, which then interacts with the BEV information from the previous frame to incorporate temporal information. When outputting the final detection results, PolarDet performs a weighted sum of position and semantic information to enhance information utilization efficiency, allowing the network to achieve higher detection accuracy. Extensive experiments on the challenging BEV object detection nuScenes dataset show that the optimal model of PolarDet achieves a mAP (mean average precision) of 0.469 and an NDS (nuScenes detection score) of 0.56, significantly outperforming Cartesian coordinate-based BEV detection methods.

Keywords: polar coordinates; BEV object detection; position and semantic information; feature weighting; cross-plane encoder

* 中央引导地方科技专项-长三角科技创新共同体联合攻关计划项目(2023CSJGG1600)、安徽省自然科学基金面上项目(2208085MF173)和芜湖市"赤铸之光"重大科技项目(2023zd01,2023zd03)资助。
原稿收到日期为 2024 年 08 月 07 日,修改稿收到日期为 2024 年 09 月 27 日。
通信作者:时培成,教授,博士,E-mail:shipeicheng@ahpu.edu.cn。

前言

鸟瞰图(bird's eye view,BEV)感知技术通过整合多传感器数据,将这些数据统一映射至单一视图,提供一个由上而下观察的俯视视角,有效解决了传统视图中物体被遮挡以及物体尺度不一致的问题。此集成视图不仅为后续的规划与控制模块提供了清晰的信息基础,还显著降低了系统部署的难度,对于自动驾驶技术的落地应用具有重要价值。在此技术背景下,相机因其低廉的成本以及其能够捕捉丰富纹理信息的优势,受到了学术界的广泛关注,因此基于相机的BEV感知技术[1-2]近几年得到了迅速发展。

在早期研究[3-5]中,基于相机的BEV感知任务通常依赖于物体的深度信息,因此须针对图像信息进行深度估计。这种方法不仅对深度估计的准确性要求高,而且计算成本也较高。为克服这些挑战,近年来研究转向采用Transformer[6]模型。Transformer模型最初用于自然语言处理,能够高度并行化且擅长捕捉长距离依赖关系,因此Transformer模型被引入到计算机视觉领域,并取得了一些显著成果[7-11]。在BEV目标检测领域中,一些方法[1,12-13]通过在笛卡尔坐标系中引入Transformer模型的注意力机制进行3D目标检测,并结合体素化栅格和双线性插值等技术生成BEV视图,巧妙地规避了对深度的直接估计,简化了传统检测流程,实现了高效的目标检测。然而,这些方法[1,12-13]绝大部分基于笛卡尔坐标系,其利用信息的方式并不符合传感器的工作模式。事实上,在极坐标下每个像素的位置均能用距离和角度来表示,与相机捕获图像的方式非常相似,而每个车载相机以径向轴成像几何感知世界,这导致在笛卡尔坐标系下感知世界相比于极坐标更加复杂。因此,本文基于以前的工作[1-2],提出了一种在极坐标系下进行端到端BEV目标检测的算法模型——PolarDet。

此外,在过往的研究[1-2,14]中,基于相机的BEV目标检测方法往往直接依赖于单一特征图来进行分类预测和边界框与速度等回归预测,这种做法忽视了RGB图像无法直接表示深度的本质特性。由于3D目标检测的最终任务须估计物体的3D位置以进行边界框预测,即使许多基于Transformer的方案不直接预测深度信息,也会使提取到的丰富图像语义特征中混杂大量位置信息。这种方法在训练过程中不仅存在收敛困难的问题,也会导致模型在最后检测阶段辨别信息的负担过重,使误差风险增加。鉴于此,本文提出了一种位置信息与语义信息加权方法,强调不同信息在不同任务中的作用大小,使模型在目标检测过程中有针对性地利用不同信息。如图1所示,PolarFormer在如图场景中出现了多处误检,而使用位置与语义信息加权策略的PolarDet则解决了该问题。

(a) PolarFormer[2]结果　　(b) 真值标签　　(c) PolarDet结果

图1　PolarFormer[2]与PolarDet的检测结果对比(黄色实线圈出的是PolarFormer[2]的误检结果,黄色虚线圈出的是PolarDet在相同位置处的检测结果)

综上所述,这项工作的创新点可总结如下。

(1)极坐标下的端到端3D目标检测。本文首次提出了一种在极坐标系下直接对汽车不同视角的输入图像进行端到端3D目标检测的新模型——PolarDet。相比于以前的方法,PolarDet通过在极坐标下进行3D目标检测,降低了坐标转换的难度,能够更自然地处理来自不同视角的数据,有效提升了3D目标检测的准确率。

(2)基于注意力机制的多信息学习方法。本文设计了一种新颖的多信息学习方法。该方法通过在注意力机制中复制、拼接和切割极坐标查询,分别提取位置与语义信息。通过分别学习不同信息,提升对相机数据中丰富纹理细节的利用率,这种创新的方法能够大大提升3D目标检测模型对图像信息的理解。

(3)基于信息加权的3D目标检测。本文提出了一种根据不同任务对信息施加不同权重的方法。通过利用神经网络在训练过程中学习到的语义信息和目标的3D位置信息,在最终预测时根据任务类型赋予权重,使模型在目标检测的不同任务中有针对性地利用信息。

1 相关工作

1.1 基于笛卡尔坐标系的BEV目标检测技术

早期BEV目标检测主要基于深度信息进行透视视角到BEV视角的变换。LSS[3]在OFT[15]的基础上,使用多个相机确定透视射线每个点的特征以减少深度预测偏差。BEVDet[4]直接将密集预测的深度信息投影到BEV视角下提取特征。CaDDN[5]使用激光雷达数据监督深度估计。BEVDet4D[16]保留过去帧的BEV特征并融合当前帧的BEV特征,以减小速度的预测误差。

近年来的工作更多致力于使用注意力机制,DETR3D[13]用参考点附近的2D特征优化3D物体表征。PETR[12]则受DETR[8]的影响,对3D位置信息进行编码,直接在3D空间内更新目标查询。BEVFormer[1]利用时间自注意机制与空间交叉注意力机制实现3D目标检测与语义分割任务。BEVFormerv2[17]进一步引入了透视空间监督来增强BEVFormer[1]的性能。BEVDepth[18]引入外部参数与点云数据提高深度估计准确性。Sparse4D[19]通过采用多帧采样的方式实现时序融合。Sparse4D v2[20]在Sparse4D[19]的基础上改进了时间融合模块,以降低计算复杂度。

然而,这些工作均基于笛卡尔坐标系进行目标检测,未能遵循真实世界中传感器的工作模式。本文提出在极坐标下进行的BEV目标检测方案,以更好地利用传感器的工作模式,提高检测的准确性与计算效率。

1.2 基于极坐标系的BEV目标检测技术

近两年出现了一些基于极坐标的BEV目标检测方案,PolarDETR[14]遵循DETR[8]的做法,使用目标查询迭代更新极坐标下的BEV特征。PolarFormer[2]引入一种多尺度极坐标表征策略,以解决极坐标中极射线方向上目标尺度随距离变化的问题。

其中,PolarFormer[2]虽然解决了目标尺度随远近而变化的问题,但它与PolarDETR[14]仍然受到位置信息与语义信息混杂而加重网络负担的影响。

2 网络结构

2.1 网络框架

如图2所示,PolarDet首先通过骨干网络对汽车

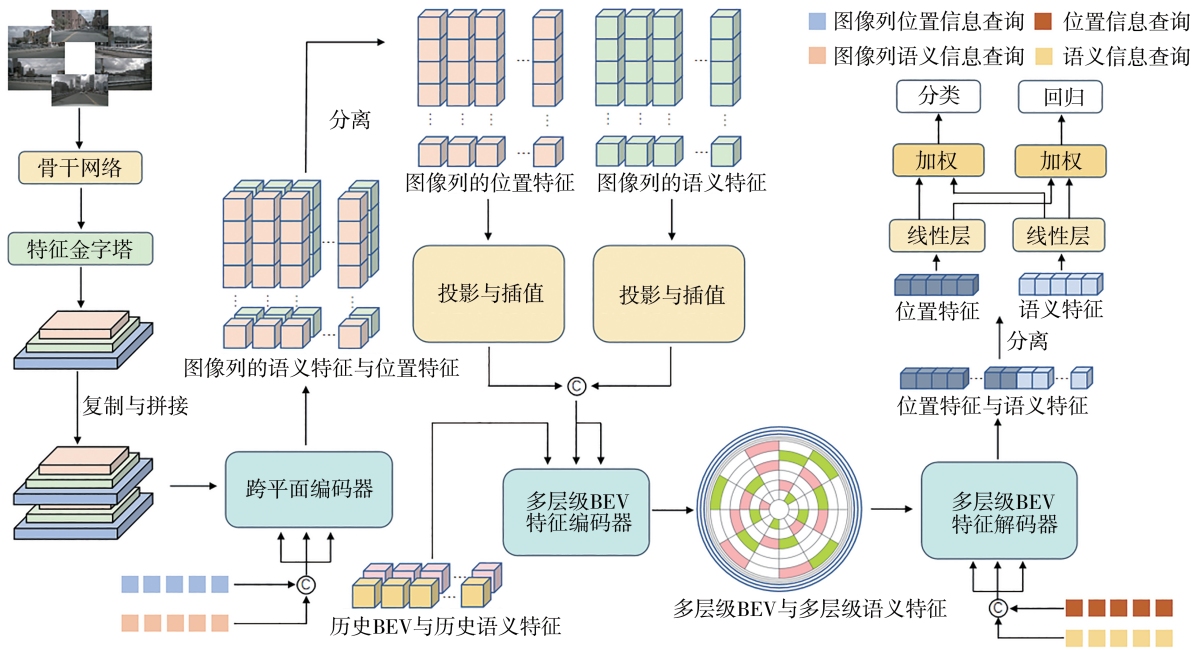
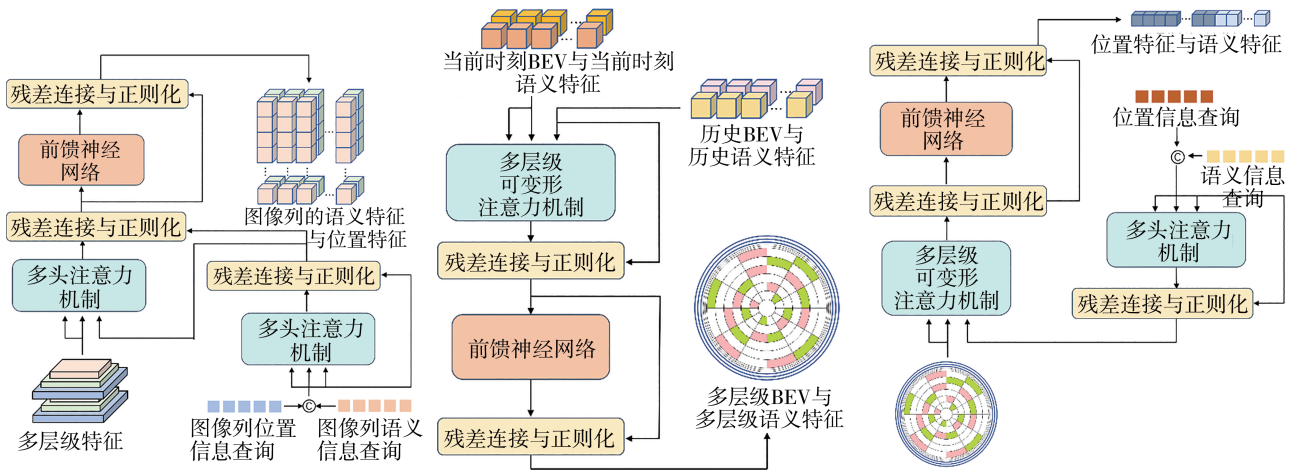


图2 PolarDet网络结构

N个不同视角相机拍摄的图片$I=\{I_i\in \mathrm{R}^{3\times H\times W},\ i=1,2,3,\cdots,N\}$进行2D特征提取,再使用特征金字塔[21]网络在2D特征中提取多层级特征;在跨平面编码器处理后,特征通过投影与插值得到极坐标下的BEV特征;随后多层级BEV特征编码器将历史BEV信息融合到当前帧的极坐标BEV特征中;接下来在多层级BEV特征解码器中使用信息查询得到目标信息;最后,将特征切割,得到极坐标BEV位置与语义信息,通过后续网络得到分类与回归的结果。

2.2　跨平面编码器与插值采样

如图3(a)所示,跨平面编码器将特征金字塔[21]输出的多层级特征转为极坐标射线,根据相机坐标系中坐标$(x,\ y,\ z)$,用以下公式得到BEV视角下的极坐标$(\varphi,\ \rho)$:



(a) 跨平面编码器网络结构　(b) 多层级BEV特征编码器结构　(c) 多层级BEV特征解码器结构

图3　PolarDet的跨平面编码器、多层级BEV特征编码器与多层级BEV特征解码器结构

$$\varphi=\arctan\frac{x}{z}\tag{1}$$

$$\rho=\sqrt{x^2+z^2}\tag{2}$$

式中:x代表相机像素中的水平位置;y代表相机像素中的竖直位置;z代表深度;φ代表与极点连线相对于极轴的旋转角度;ρ代表与极点的距离。

PolarDet采用Transformer解码器获取列像素特征,多层级特征须进行一次复制拼接再进入跨平面编码器,通过特征的复制拼接与查询的拼接,可以利用分块矩阵的乘法,分别学习同一个图像特征内部的不同信息。

图像特征在该阶段首先经历的处理过程如下:

$$\boldsymbol{p}_{n,u,w}=MultiHead\left(\dot{\boldsymbol{p}}_{n,u,w},\boldsymbol{f}_{n,u,w},\boldsymbol{f}_{n,u,w}\right)=Concat(\boldsymbol{head}_1,\boldsymbol{head}_2,\cdots,\boldsymbol{head}_h)\boldsymbol{W}_u^o\tag{3}$$

$$\boldsymbol{head}_i=Attention(\dot{\boldsymbol{p}}_{n,u,w}\boldsymbol{W}_{i,u}^Q,\boldsymbol{f}_{n,u,w}\boldsymbol{W}_{i,u}^K,\boldsymbol{f}_{n,u,w}\boldsymbol{W}_{i,u}^V)\tag{4}$$

式中:$\dot{\boldsymbol{p}}_{n,u,w}$代表极射线查询;$\boldsymbol{f}_{n,u,w}$代表第$u$个尺度中,第$n$个相机图片的第$w$列像素;$\boldsymbol{W}_u^O\in\mathbb{R}^{hd_{\mathrm{model}}\times d_k}$、$\boldsymbol{W}_u^Q\in\mathbb{R}^{d_{\mathrm{model}}\times d_q}$、$\boldsymbol{W}_u^K\in\mathbb{R}^{d_{\mathrm{model}}\times d_k}$、$\boldsymbol{W}_u^V\in\mathbb{R}^{d_{\mathrm{model}}\times d_v}$代表注意力机制权值矩阵,其中$d_q=d_k=d_v=d_{\mathrm{model}}/h$。

特征的复制拼接与查询的拼接公式如下:

$$\dot{\boldsymbol{p}}_{n,u,w}=Concat(\dot{\boldsymbol{p}}_{n,u,w}^l,\dot{\boldsymbol{p}}_{n,u,w}^t)\tag{5}$$

$$\boldsymbol{f}_{n,u,w}=Concat(\dot{\boldsymbol{f}}_{n,u,w},\dot{\boldsymbol{f}}_{n,u,w})\tag{6}$$

式中:$\dot{\boldsymbol{p}}_{n,u,w}^l$代表位置信息极射线;$\dot{\boldsymbol{p}}_{n,u,w}^t$代表语义信息极射线;$\dot{\boldsymbol{f}}_{n,u,w}$代表第$u$个尺度中,第$n$个相机图片的第$w$列像素。

随后将$\boldsymbol{p}_{n,u,w}$沿方位角展开,得到单个相机语义信息与位置信息的极射线:

$$\boldsymbol{P}_{n,u}=Stack(\left[\boldsymbol{p}_{n,u,1},\boldsymbol{p}_{n,u,2},\cdots,\boldsymbol{p}_{n,u,w_u}\right],dim=1)\in\mathbb{R}^{R_u\times W_u\times C}\tag{7}$$

下一步将极射线融合,得到全局坐标系下的BEV特征图。在圆柱坐标上生成一组3D点$\mathcal{G}^{\mathrm{P}}=\{(\rho_i^{(P)},\phi_j^{(P)},Z_k^{(P)})|i=1,\cdots,\mathcal{R}_u;\ j=1,\cdots,\mathcal{N}_u;\ k=1,\cdots,\mathcal{Z}_u\}$,其中:$\mathcal{R}_u$代表圆周上生成的点数;$\mathcal{N}_u$代表一条极射线上生成的点数;$\mathcal{Z}_u$代表在圆柱坐标范围内,沿圆柱高方向上生成的点数。

下列公式对圆柱点进行极坐标与笛卡尔坐标的转换:

$$x_{i,j}=\rho_i^{(P)}\sin\ \phi_j^{(P)}\tag{8}$$

$$y_{i,j}=\rho_i^{(P)}\cos\ \phi_j^{(P)}\tag{9}$$

$$z_k^{(P)} = z_k^{(P)} \tag{10}$$

接下来使用相机参数进行视角变换,用s代表缩放因子,得到极射线的索引:

$$\begin{pmatrix} sx_{i,j,k,n}^{(I)} \\ sy_{i,j,k,n}^{(I)} \\ s \\ 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Pi}_n & 0 \\ 0 & 1 \end{pmatrix} \boldsymbol{E}_n \begin{pmatrix} x_{i,j} \\ y_{i,j} \\ z_k^{(P)} \\ 1 \end{pmatrix} \tag{11}$$

再使用下列公式得到同一世界坐标系下的极坐标BEV特征图:

$$G_u\left(\rho_i^{(P)}, \phi_j^{(P)}\right) = \frac{1}{\sum_{n=1}^{N}\sum_{k=1}^{Z}\lambda_n\left(\rho_i^{(P)}, \phi_j^{(P)}, z_k^{(P)}\right)} \cdot \sum_{n=1}^{N}\sum_{k=1}^{Z}\lambda_n\left(\rho_i^{(P)}, \phi_j^{(P)}, z_k^{(P)}\right) \cdot B(\boldsymbol{P}_{n,u}, (\bar{x}_{i,j,k,n}^{(I)}, \bar{r}_{i,j,n})) \tag{12}$$

式中:$\lambda_n(\rho_i^{(P)}, \phi_j^{(P)}, z_k^{(P)})$是二进制加权因子;$B(P_{n,u}, (x, y))$表示极射线$\boldsymbol{P}_{n,u}$在坐标$(x, y)$处双线性插值采样;$(\bar{x}_{i,j,k,n}, \bar{r}_{i,j,n})$表示已经归一化的极坐标角度索引与半径索引。

2.3 多层级BEV特征编码器

如图3(b)所示,多层级BEV特征编码器使用可变形注意力[7]机制融入时间信息,并进行更细致的特征提取。该机制在特征中学习参考点,使用双线性插值得到局部区域特征,相比于多头注意力机制,该机制对小目标的检测能力更强。通过融入时间信息,PolarDet对被遮挡物体以及运动物体的检测能力也更加出色。

用$\{G_u\}_{u=1}^{U}$表示当前时刻的BEV语义信息与位置信息,可变形注意力机制处理过程如下:

$$MSDeformAttn\left(\{\boldsymbol{G}_u\}_{u=1}^{U}, \boldsymbol{BEV}_{prev}\right) = \sum_{m=1}^{M} W_m \left[\sum_{u=1}^{U}\sum_{k=1}^{K}\boldsymbol{A}_{muqk}\boldsymbol{W}'_m \cdot sample(\boldsymbol{BEV}_{prev}, \zeta_u\left(x_q\right) + \Delta x_{muqk})\right] \tag{13}$$

式中:$sample(\boldsymbol{BEV}_{prev}, \zeta_u\left(x_q\right) + \Delta x_{muqk})$表示利用偏移$\Delta x_{muqk}$与初始位置$\zeta_u\left(x_q\right)$在$\boldsymbol{BEV}_{prev}$上进行采样;$m$与$k$表示注意力头与采样点的索引;$\boldsymbol{BEV}_{prev}$表示上一帧的BEV;$\boldsymbol{W}_m$与$\boldsymbol{W}'_m$都是可学习的权重矩阵;$\boldsymbol{A}_{muqk}$表示第$m$个注意力头的第$u$个尺度上第$k$个点的注意力权重。

2.4 多层级BEV特征解码器与检测输出

如图3(c)所示,多层级BEV特征解码器中信息查询进行多头注意力机制避免学习相同特征,随后与上游特征通过多层级可变形注意力机制[7]进行特征交互。

为了对检测结果进行加权,该模块后续还须对检测结果进行分割,假设输出特征为$\boldsymbol{p}\left[d_1, \cdots, d_n\right]$,$d_1, \cdots, d_n$代表$p$在各个维度上的元素个数,PolarDet进行如下处理:

$$\boldsymbol{p}_1 = \left[d_1, \cdots, d_{\frac{n}{2}}\right] \tag{14}$$

$$\boldsymbol{p}_2 = \left[d_{\frac{n}{2}}, \cdots, d_n\right] \tag{15}$$

式中:$\boldsymbol{p}_1$代表位置信息;$\boldsymbol{p}_2$代表语义信息。模块中使用线性层对$\boldsymbol{p}_1$与$\boldsymbol{p}_2$进行处理,并分别在分类与回归计算过程中进行加权:

$$\boldsymbol{Class} = \boldsymbol{A}_1^{\mathrm{C}}\boldsymbol{p}_1 + \boldsymbol{A}_2^{\mathrm{C}}\boldsymbol{p}_2 \tag{16}$$

$$\boldsymbol{Location} = \boldsymbol{A}_1^{L}\boldsymbol{p}_1 + \boldsymbol{A}_2^{L}\boldsymbol{p}_2 \tag{17}$$

通过上述步骤,得到目标分类的置信度$c \in \mathbb{R}^o$,目标的位置与速度信息$\theta_{\mathrm{ori}}, d_\rho, d_\phi, \theta_v$。$o$表示被分类的类别数;$\theta_{\mathrm{ori}}$代表边界框的偏航角;$d_\rho$表示沿极射线方向的位置;$d_\phi$表示极射线角度;$\theta_v$表示目标的绝对速度角。

目标的位置与速度信息用来细化参考点(ρ, ϕ, z),该模块使用目标的位置与速度信息生成参考点的偏移,得到θ_{ori}的正交分量θ_ϕ与θ_ρ、v的正交分量v_ϕ与v_ρ:

$$\bar{\theta}_{\mathrm{ori}} = \theta_{\mathrm{ori}} - \phi \tag{18}$$

$$\theta_\phi = \sin\bar{\theta}_{\mathrm{ori}} \tag{19}$$

$$\theta_\rho = \cos\bar{\theta}_{\mathrm{ori}} \tag{20}$$

$$\bar{\theta}_v = \theta_v - \phi \tag{21}$$

$$v_\phi = v_{\mathrm{abs}}\sin\bar{\theta}_v \tag{22}$$

$$v_\rho = v_{\mathrm{abs}}\cos\bar{\theta}_v \tag{23}$$

式中v_{abs}代表目标的绝对速度。最后使用焦点损失与L1损失用来衡量分类与回归的误差。

3 实验

3.1 数据集

本节在nuScenes[22]数据集上对PolarDet进行了全面测试,nuScenes[22]数据集包含在汽车周围构成360°全景环绕的6个相机在每个场景下以每0.5 s拍摄一张图片的频率持续采样20 s的1 000个场景,其中700个场景被划分为训练集,150个场景被划分为验证集,150个场景被划分为测试集。

3.2 评价指标

本节采用nuScenes[22]数据集的评价指标,包括nuScenes检测得分(NDS)和平均精度(mAP)。mAP是根据鸟瞰图中心距离D={0.5,1,2,4} m和类别集

C的匹配阈值计算的平均值,NDS是mAP与其他物体属性检测结果(如平均平移误差ATE、平均尺度误差ASE、平均方向误差AOE、平均速度误差AVE和平均属性误差AAE)的加权组合。

3.3　实验设置与训练

PolarDet采用ResNet101-DCN[26]作为骨干网络,使用预训练的FCOS3D[27]作为骨干网络的检查点,其余数据随机初始化。特征尺度与跨平面编码器数量设置为3,多尺度极坐标BEV特征图的半径与分辨率分别设置为(64, 256)、(32, 128)、(16, 64),设置6个多层级BEV特征编码器层与6个解码器层。

PolarDet使用AdamW[28]优化器训练,权重衰减为0.075,初始学习率为2×10^{-5},并采用余弦退火策略。模型在一个NVIDIA A40 GPU进行了24轮的训练,6个摄像头的总批次大小为6,采用同步归一化。

4　实验结果与分析

4.1　与先进模型的对比

如表1所示,本节在nuScenes[22]验证集上将PolarDet与目前最先进的模型进行比较。由于PolarDet包含时间信息,为了公平比较,对于PolarFormer[2]与PolarDETR[14]的结果,本节采用融入时间信息的PolarFormer-T[2]与PolarDETR-T[14]。结果显示PolarDet在NDS和mAP等多项指标上均取得了最佳性能。

表1　PolarDet与先进网络的对比

方法	坐标系	骨干网络	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
BEVFormer[1]	笛卡尔	ResNet101	0.445	0.535	0.631	0.257	0.405	0.435	0.143
BEVDet-Beta[4]	笛卡尔	VovNet	0.422	0.482	**0.529**	0.236	0.396	0.979	0.152
SpatialDETR[23]	笛卡尔	ResNet101	0.424	0.486	0.613	0.253	0.402	0.857	0.131
PETR[12]	笛卡尔	ResNet101	0.366	0.441	0.717	0.267	0.412	0.834	0.190
DETR3D[13]	笛卡尔	ResNet101	0.349	0.434	0.716	0.268	0.379	0.842	0.200
PETRv2[24]	笛卡尔	ResNet101	0.421	0.524	0.681	0.267	0.357	**0.377**	0.186
ORA3D[25]	笛卡尔	VovNet	0.423	0.489	0.595	0.254	0.392	0.851	0.128
PolarFormer-T[2]	极坐标	ResNet101	0.457	0.543	0.612	0.257	0.392	0.467	0.129
PolarDETR-T[14]	极坐标	ResNet101	0.383	0.488	0.707	0.269	0.344	0.518	0.196
PolarDet	极坐标	ResNet101	**0.469**	**0.560**	0.593	**0.218**	**0.313**	0.470	**0.113**

4.2　实验定性分析

为了进一步证明PolarDet的有效性,本节使用训练完成的网络对nuScenes[22]数据集内包含的复杂道路、夜晚场景进行了检测结果可视化。如图4(a)所示,即使在夜晚,对于远处、被截断、模糊物体,PolarDet均能准确地检测到,证明了PolarDet融入时序特征方法的有效性。从BEV图中可以看出,PolarDet生成的检测框都十分精准,证明了本文提出的位置信息与语义信息加权方法的有效性。图4(c)展示了夜晚场景,即使在有大量行人和光污染的交叉路口,PolarDet仍表现出卓越性能,能够在密集人群检测中取得良好的检测效果。

如图4(b)所示,在复杂街道场景中,PolarDet依然表现出色,特别是在边界框的精准度上性能卓越。实验证明,PolarDet的检测方案能够充分利用图像信息,提升检测精度并增强网络的鲁棒性。

4.3　消融实验

4.3.1　极坐标系与笛卡尔坐标系对比

为更进一步证明在极坐标下进行3D目标检测相比于笛卡尔坐标系的优越性,本节设置了多组对比实验,与PolarFormer[2]在笛卡尔坐标系和极坐标系下进行了全面对比。

表2展示了PolarDet网络在不同配置下的性能,包括是否采用多层级特征、是否使用极坐标进行特征采样、是否使用极坐标表示的参数进行预测。其中,“s”代表未使用多层级特征;“CC”代表在特征采样和结果检测中均采用笛卡尔坐标系;“PC”代表在特征采样时采用极坐标系,但在预测中使用笛卡尔坐标系。

由表2可见,全局使用极坐标系的PolarDet相比于全局使用笛卡尔坐标系,mAP提升了大约3.5%,NDS提升了大约1.8%。

4.3.2　采用信息加权方法与融入时间信息的有效性

本文选取的基线模型是PolarFormer[2],其特征维度是256维,由于PolarDet须同时学习位置信息与语义信息,使用了512维特征,因此无法直接对比。为了排除特征维度对模型性能的影响,如表3所示,

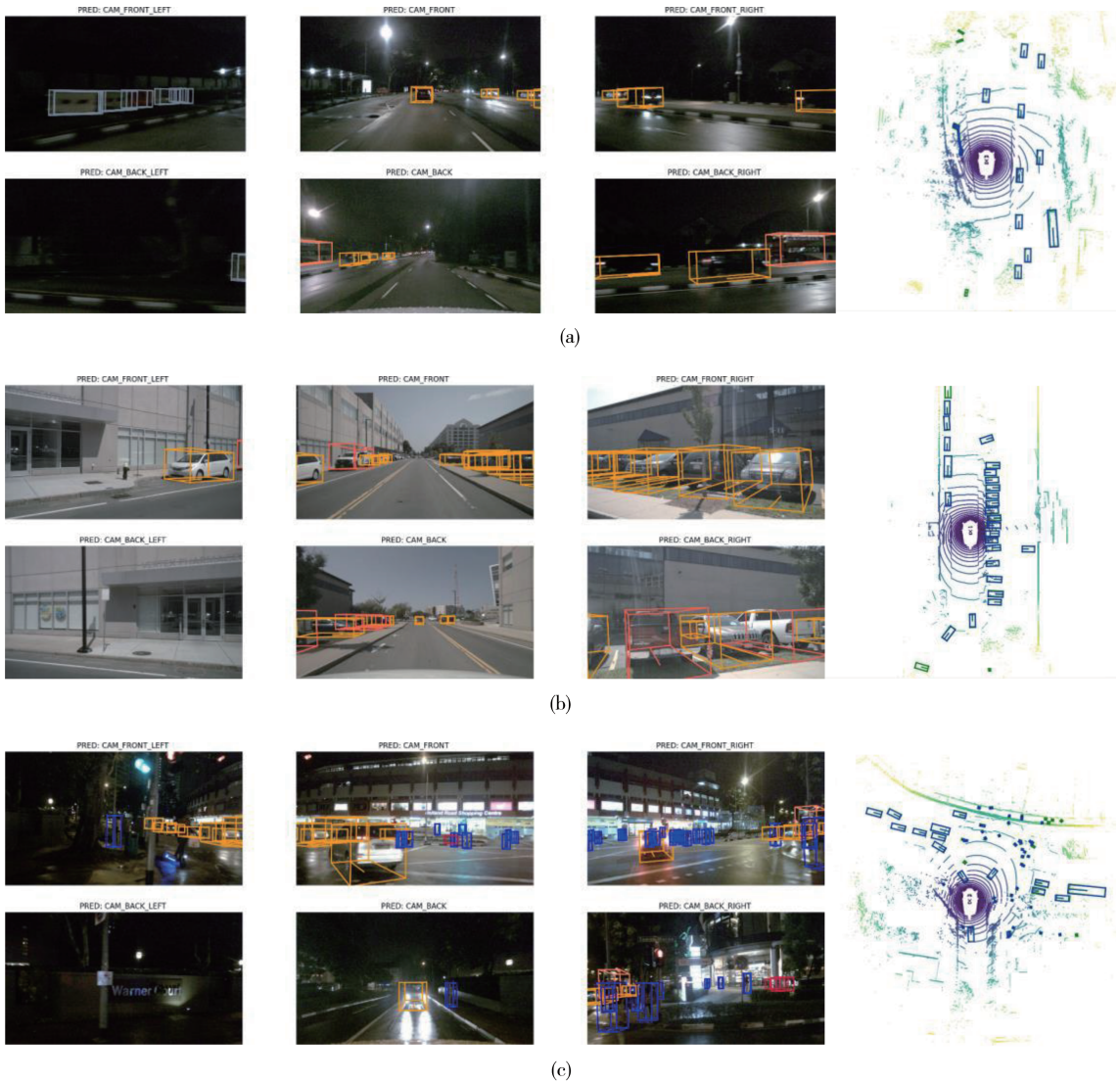


图4 不同场景下的PolarDet检测结果(为便于查看,BEV使用笛卡尔坐标系绘制;蓝色框代表检测结果,绿色框代表真实标签)

表2 PolarDet与PolarFormer检测结果对比

方法	是否使用多层级特征	特征所在坐标系	预测所在坐标系	mAP	NDS
PolarFormer-CC-s	否	笛卡尔	笛卡尔	0.381	0.449
PolarDet-CC-s	否	笛卡尔	笛卡尔	0.426	0.531
PolarFormer-PC-s	否	极坐标	笛卡尔	0.388	0.450
PolarDet-PC-s	否	极坐标	笛卡尔	0.430	0.535
PolarFormer-s	否	极坐标	极坐标	0.391	0.458
PolarDet-s	否	极坐标	极坐标	0.433	0.537
PolarFormer-CC	是	笛卡尔	笛卡尔	0.381	0.450
PolarDet-CC	是	笛卡尔	笛卡尔	0.453	0.550
PolarFormer-PC	是	极坐标	笛卡尔	0.385	0.455
PolarDet-PC	是	极坐标	笛卡尔	0.461	0.557
PolarFormer	是	极坐标	极坐标	0.396	0.458
PolarDet	是	极坐标	极坐标	0.469	0.560

本节设置了多组实验,为确保PolarDet的性能提升并非均来自特征维度的增加,部分实验将语义信息与位置信息的权重设计为0.5,确保模型仅增加了一倍维度而未对信息进行加权,表格中E表示权重均为0.5;NT表示没有输入时间信息,此时使用当前帧信息代替历史信息。

实验结果显示,PolarDet通过信息加权,mAP提升约3%,NDS提升约3.9%;PolarDet融入时间信息,mAP提升约7.3%,NDS提升约17%。

4.4 选择最优参数的正交实验

为探究不同信息权重对网络检测效果的影响,确定最优权重配置,本节采用正交实验法进行了一系列实验。

如图5所示,横坐标使用"(语义信息权重,位置

表3　PolarDet是否施加权重与是否使用时间信息的消融实验

方法	是否加权	维度数量	是否嵌入时间信息	mAP	NDS
PolarFormer	否	256	否	0.396	0.458
PolarDet-E-NT	否	512	否	0.402	0.460
PolarDet-E	否	512	是	0.455	0.539
PolarDet-NT	是	512	否	0.437	0.479
PolarDet	是	512	是	0.469	0.560

信息权重)”来表示两种信息的权重。对于分类预测,本节选取了以下几组具有代表性的权重设置:(0.8, 0.2);(0.7, 0.3);(0.6, 0.4);(0.5, 0.5);(0.4, 0.6)。对于边界框与速度信息等回归预测,实验权重设置的范围为:(0.7, 0.3);(0.6, 0.4);(0.5, 0.5);(0.4, 0.6);(0.3, 0.7)。这些组合覆盖了从偏向语义信息到偏向位置信息的各种情况,使实验结果能够全面显示不同权重分配对网络性能的影响。

在实验中,采用标准的$L_{25}(5^2)$正交表,每次实验测试两个因子中的一个组合,确保每个组合都被充分测试。

本节进行了25次实验,图5将实验数据可视化成点线图,以直观展示不同权重配比下PolarDet的性能;表4列出了不同权重配比下的实验数据,以方便进行进一步分析。本节使用极差分析法来分析实验数据,从而选出最优参数,下列公式计算因子A与因子B中各水平的平均值:

$$\bar{A}_i = \frac{1}{n}\sum_j Y_{ij} \tag{24}$$

$$\bar{B}_i = \frac{1}{n}\sum_j Y_{ij} \tag{25}$$

式中:Y_{ij}表示第i次实验中的因子A和B的实验结果;n是重复次数。

计算得到因子A与因子B中各水平的平均值后,通过下列公式计算极差:

表4　PolarDet不同权重对性能的影响

语义信息-C	位置信息-C	语义信息-B	位置信息-B	mAP	NDS
0.8	0.2	0.5	0.5	0.461	0.550
0.8	0.2	0.4	0.6	0.466	0.558
0.8	0.2	0.3	0.7	0.463	0.550
0.8	0.2	0.6	0.4	0.453	0.536
0.8	0.2	0.7	0.3	0.441	0.526
0.7	0.3	0.5	0.5	0.463	0.557
0.7	0.3	0.4	0.6	**0.469**	**0.560**
0.7	0.3	0.3	0.7	0.464	0.552
0.7	0.3	0.6	0.4	0.454	0.533
0.7	0.3	0.7	0.3	0.444	0.529
0.6	0.4	0.5	0.5	0.460	0.553
0.6	0.4	0.4	0.6	0.464	0.556
0.6	0.4	0.3	0.7	0.463	0.555
0.6	0.4	0.6	0.4	0.452	0.540
0.6	0.4	0.7	0.3	0.440	0.523
0.5	0.5	0.5	0.5	0.455	0.539
0.5	0.5	0.4	0.6	0.458	0.548
0.5	0.5	0.3	0.7	0.456	0.541
0.5	0.5	0.6	0.4	0.443	0.530
0.5	0.5	0.7	0.3	0.440	0.520
0.4	0.6	0.5	0.5	0.442	0.528
0.4	0.6	0.4	0.6	0.445	0.530
0.4	0.6	0.3	0.7	0.436	0.527
0.4	0.6	0.6	0.4	0.423	0.518
0.4	0.6	0.7	0.3	0.421	0.509

注:C表示分类预测权重,B表示回归预测权重。

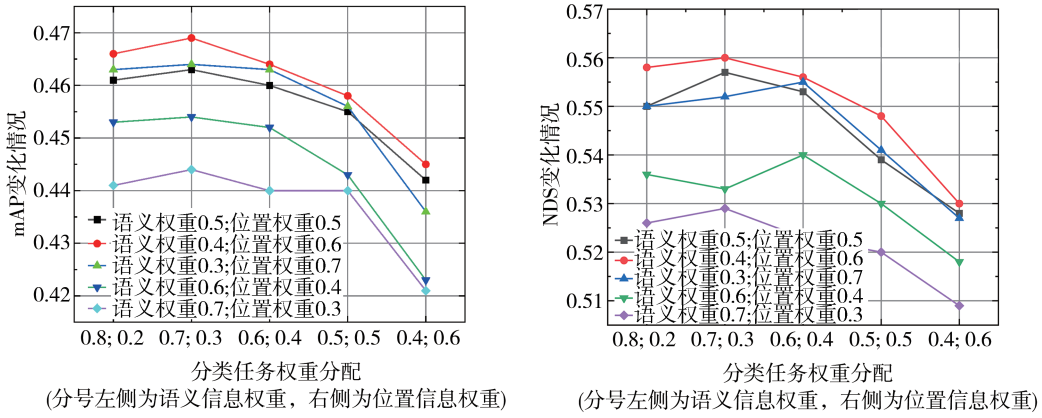


图5　PolarDet不同权重对性能的影响

$$R_A = \max(\bar{A}_i) - \min(\bar{A}_i) \tag{26}$$

$$R_B = \max(\bar{B}_i) - \min(\bar{B}_i) \tag{27}$$

本节使用NDS作为评价指标,通过式(24)与式(25)计算得到如表5所示的分类预测与回归预测中各种权重分配结果的NDS平均值。

表5 各种权重分配下的不同任务中的NDS平均值

权重分配	分类预测结果 NDS平均值	回归预测结果 NDS平均值
(0.8, 0.2)	0.545 4	
(0.7, 0.3)	0.546 2	0.521 4
(0.6, 0.4)	0.545 4	0.531 4
(0.5, 0.5)	0.535 6	0.545 4
(0.4, 0.6)	0.522 4	0.550 4
(0.3, 0.7)		0.545 0

随后使用式(26)与式(27)计算分类预测权重分配的极差与回归预测权重分配的极差,选取差值最大的因子对应的水平作为最优水平,得到最佳权重分配:对于分类预测,最佳权重分配为(0.7, 0.3),对于回归预测,最佳权重分配为(0.4, 0.6)。

5 结论

(1)本文提出了一种基于位置与语义信息加权的极坐标BEV端到端目标检测方法—PolarDet,其可以生成极坐标下的BEV位置与语义信息,降低坐标转换难度,可更自然地处理来自不同视角的数据,有效提升3D目标检测的准确率,结果表明其mAP达到0.469,NDS达到0.56,显著优于基于笛卡尔坐标的BEV目标检测方法。

(2)基于可变形注意力机制能够灵活关注不同局部区域的优势,PolarDet将上一帧的BEV信息融入当前帧,使其能够更准确地捕捉目标的运动轨迹与速度变化,结果表明融入上一帧时间信息后,可使PolarDet的mAP提升7.3%,NDS提升17%。

(3)在结果输出时,对位置与语义信息进行加权求和,可提高信息的利用效率,使网络能够达到更高的检测精度,并减少误检和漏检的情况,结果表明使用信息加权的PolarDet能够提升3%的mAP与3.9%的NDS。

参考文献

[1] LI Z, WANG W, LI H, et al. BEVFormer: learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers[J/OL]. arXiv, 2022[2023-10-17]. http://arxiv.org/abs/2203.17270.

[2] JIANG Y, ZHANG L, MIAO Z, et al. PolarFormer: multi-camera 3D object detection with polar transformer[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(1): 1042-1050[2024-01-16]. https://ojs.aaai.org/index.php/AAAI/article/view/25185. DOI:10.1609/aaai.v37i1.25185.

[3] PHILION J, FIDLER S. Lift, Splat, Shoot: encoding images from arbitrary camera rigs by implicitly unprojecting to 3D[J/OL]. arXiv, 2020 [2023-10-25]. http://arxiv.org/abs/2008.05711.

[4] HUANG J, HUANG G, ZHU Z, et al. BEVDet: high-performance multi-camera 3D object detection in bird-eye-view[J/OL]. arXiv, 2022 [2023-10-17]. http://arxiv.org/abs/2112.11790.

[5] READING C, HARAKEH A, CHAE J, et al. Categorical depth distribution network for monocular 3D object detection[J/OL]. arXiv, 2021[2023-12-16]. http://arxiv.org/abs/2103.01100.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J/OL]. arXiv, 2023[2023-10-06]. http://arxiv.org/abs/1706.03762. DOI:10.48550/arXiv.1706.03762.

[7] ZHU X, SU W, LU L, et al. Deformable DETR: deformable transformers for end-to-end object detection[J/OL]. arXiv, 2021[2024-03-07]. http://arxiv.org/abs/2010.04159.

[8] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C/OL]//VEDALDI A, BISCHOF H, BROX T, et al. Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 213-229. DOI:10.1007/978-3-030-58452-8_13.

[9] BEAL J, KIM E, TZENG E, et al. Toward transformer-based object detection[J/OL]. arXiv, 2020[2023-11-20]. http://arxiv.org/abs/2012.09958. DOI:10.48550/arXiv.2012.09958.

[10] LIU S, LI F, ZHANG H, et al. DAB-DETR: dynamic anchor boxes are better queries for DETR[J/OL]. arXiv, 2022[2023-11-25]. http://arxiv.org/abs/2201.12329.

[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J]. 2020.DOI:10.48550/arXiv.2010.11929. .

[12] LIU Y, WANG T, ZHANG X, et al. PETR: position embedding transformation for multi-view 3D object detection[J/OL]. arXiv, 2022[2023-10-25]. http://arxiv.org/abs/2203.05625.

[13] WANG Y, GUIZILINI V, ZHANG T, et al. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries[J/OL]. arXiv, 2021[2024-01-16]. http://arxiv.org/abs/2110.06922.

[14] CHEN S, WANG X, CHENG T, et al. Polar parametrization for vision-based surround-view 3D detection[J/OL]. arXiv, 2022[2024-02-28]. http://arxiv.org/abs/2206.10965.

[15] RODDICK T, KENDALL A, CIPOLLA R. Orthographic feature transform for monocular 3D object detection[J/OL]. arXiv, 2018[2023-12-16]. http://arxiv.org/abs/1811.08188.

[16] HUANG J, HUANG G. BEVDet4D: exploit temporal cues in

multi-camera 3D object detection[J/OL]. arXiv, 2022[2024-07-09]. http://arxiv.org/abs/2203.17054.

[17] YANG C, CHEN Y, TIAN H, et al. BEVFormer v2: adapting modern image backbones to bird's-eye-view recognition via perspective supervision[J/OL]. arXiv, 2022[2024-04-04]. http://arxiv.org/abs/2211.10439.

[18] LI Y, GE Z, YU G, et al. BEVDepth: acquisition of reliable depth for multi-view 3D object detection[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(2): 1477-1485[2023-11-16]. https://ojs.aaai.org/index.php/AAAI/article/view/25233. DOI:10.1609/aaai.v37i2.25233.

[19] LIN X, LIN T, PEI Z, et al. Sparse4D: multi-view 3D object detection with sparse spatial-temporal fusion[J/OL]. arXiv, 2023[2024-04-11]. http://arxiv.org/abs/2211.10581.

[20] LIN X, LIN T, PEI Z, et al. Sparse4D v2: recurrent temporal fusion with sparse model[J/OL]. arXiv, 2023[2024-01-16]. http://arxiv.org/abs/2305.14018.

[21] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[J/OL]. arXiv, 2017[2023-11-15]. http://arxiv.org/abs/1612.03144.

[22] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: a multimodal dataset for autonomous driving[J/OL]. arXiv, 2020[2024-05-08]. http://arxiv. org/abs/1903.11027. DOI: 10.48550/arXiv.1903.11027.

[23] DOLL S, SCHULZ R, SCHNEIDER L, et al. SpatialDETR: robust scalable transformer-based 3D object detection from multi-view camera images with global cross-sensor attention[M/OL]// AVIDAN S, BROSTOW G, CISSÉ M, et al. Computer vision-ECCV 2022: Vol. 13699. Cham: Springer Nature Switzerland, 2022: 230-245 [2024-06-23]. https://link. springer. com/10.1007/978-3-031-19842-7_14. DOI: 10.1007/978-3-031-19842-7_14.

[24] LIU Y, YAN J, JIA F, et al. PETRv2: a unified framework for 3D perception from multi-camera images[C/OL]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 3239-3249[2024-02-27]. https://ieeexplore. ieee. org/document/10377268/. DOI: 10.1109/ICCV51070.2023.00302.

[25] ROH W, CHANG G, MOON S, et al. ORA3D: overlap region aware multi-view 3D object detection[J/OL]. arXiv, 2023[2024-06-23]. http://arxiv.org/abs/2207.00865.

[26] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[J/OL]. arXiv, 2015[2024-05-08]. http://arxiv.org/abs/1512.03385. DOI:10.48550/arXiv.1512.03385.

[27] WANG T, ZHU X, PANG J, et al. FCOS3D: fully convolutional one-stage monocular 3D object detection[J/OL]. arXiv, 2021[2023-10-17]. http://arxiv.org/abs/2104.10956.

[28] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[J/OL]. arXiv, 2019[2024-06-14]. http://arxiv.org/abs/1711.05101.