DOI: 10.19650/j.cnki.cjsi.J2412990

基于循环跨视图转换和多状态特征融合的鸟瞰图生成方法*

刘明杰,何峥言,陈俊生,刘 平,朴昌浩
(重庆邮电大学自动化学院 重庆 400065)

摘 要:针对多数基于多视角透视图的鸟瞰图(BEV)生成算法难以实现对语义不一致多状态关联特征的提取,以及模型性能与复杂度的平衡等问题,提出一种基于轻量级 Transformer 的 BEV 生成模型。该模型采用端到端的单阶段训练策略,通过建立交通场景中动态车辆和静态道路信息的关联,滤除生成视图中的噪声。基于此,一方面设计面向多尺度特征的 Transformer 循环跨视图转换模块,通过注意力机制实现对输入的位置编码和表征学习,捕捉特征序列中不同位置的依赖关系,提升 BEV 特征的鲁棒性;另一方面设计面向语义不一致的多状态 BEV 特征融合模块,提取静态道路和动态车辆的关联信息,提升生成 BEV 视图的精度。在 NuScenes 数据集上进行实验,结果表明,方法在确保低模型复杂度的前提下,达到了先进的 BEV 视图生成性能。动态车辆和静态道路的语义分割精度分别达到了 43.2% 和 82.0%。

关键词: 视图转换;轻量化 Transformer 模型;鸟瞰图;透视图
中图分类号: TP391.4 TH865 **文献标识码**: A **国家标准学科分类代码**: 520.20

Bird's eye view generation based on recurrent cross-view transformation and multi-state feature fusion

Liu Mingjie, He Zhengyan, Chen Junsheng, Liu Ping, Piao Changhao
(*School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*)

Abstract: To address semantic inconsistency in multi-state associated feature extraction and balancing model performance with complexity in most multiple perspective view-based bird's eye view (BEV) generation method, a light-weight Transformer-based BEV generation model is proposed. The method utilizes an end-to-end one-stage training strategy to establish a mutual association between dynamic vehicle and static road information in traffic scenes, effectively filtering out noise in the generated BEV. A Transformer-based recurrent cross-view transformation module for multi-scale features is introduced to perform image encoding and representation learning. This module improves the robustness of the extracted BEV features by capturing the location-dependent relationships in the perspective view (PV) feature sequence. Additionally, a multi-state BEV feature fusion module is designed to address semantic inconsistencies, extracting correlated information between dynamic vehicles and static roads, thus enhancing the performance of the generated BEVs. Experiments on the NuScenes dataset show that this method achieves advanced BEV generation performance with low model complexity, achieving 43.2% and 82.0% semantic segmentation accuracy for dynamic vehicles and static roads, respectively.

Keywords: map-view transition; light-weight Transformer; bird's eye view; perspective view

0 引 言

鸟瞰图(bird's eye-view, BEV)因包含目标及周围环境详细的三维位置和规模信息,在自动驾驶领域得到了广泛关注[1]。BEV 视图一方面可以利用类似激光雷达等具有深度信息的传感器获得[2];另一方面也可以通过融合多视角车载相机感知图像的方式生成[3]。其中,基

收稿日期:2024-06-25 Received Date: 2024-06-25
*基金项目:国家重点研发计划项目(2022YFE0101000)、重庆市技术创新与应用发展专项重大项目(CSTB2023TIAD-STX0035)、重庆市教委科学技术研究项目(KJQN202200630)资助

于多视角车载相机的 BEV 生成技术具有成本低、实时性高等优点,成为了当前获取交通场景 BEV 视图的主流方式[4-6]。然而,车载相机所成图像是以二维特征表征物体三维信息的透视图(perspective view, PV)。因此,基于多视角车载相机的 BEV 生成技术的关键是建立 PV 视图到 BEV 视图特征的映射关系。

PV 到 BEV 的视图转换方法可以分为基于几何的方法和基于网络模型的方法。基于几何的方法是利用逆透视映射赋予图像深度信息,并针对不同视图之间的几何关系进行建模[7];基于网络模型的方法则是以数据驱动的方式,隐式地利用相机几何信息进行视图转换,神经网络充当 PV 和 BEV 之间的映射函数[8]。基于网络模型的方法因不依赖目标的 3D 先验和深度信息,且视图转换精度相对较高,成为了当前 BEV 视图生成的主流技术。

基于网络模型的视图转换方法又可细分为基于多层感知机(multi-layer perceptron, MLP)的视图转换和基于 Transformer 的视图转换。其中,基于 MLP 的方法是通过学习一个复杂映射函数,将输入映射到具有不同模态、维度或表示的输出上。Pan 等[9]利用两层 MLP 将 PV 特征扁平化处理,并学习特征图中像素位置之间的关系,基于该种关系将 PV 特征重塑为 BEV 特征。Lu 等[10]利用带 MLP 的编码器-解码器结构,将 PV 视图直接转换到二维 BEV 视图的笛卡尔坐标系中。编码器将 PV 特征扁平化处理,解码器通过上采样将其重建为 BEV 特征。Li 等[11]提出在 PV 向 BEV 转换基础上,再次使用 MLP 将 BEV 逆投影回 PV 以检查特征是否被正确映射。

相较于 MLP,基于 Transformer 的视图转换一方面能够依据输入动态调整模型对不同区域的关注程度,更好地适应不同场景变化;另一方面可以采用逆向方式搜索 BEV 下可学习查询所对应的 PV 特征,以更好地捕捉输入图像的关键信息。因此,基于 Transformer 生成的 BEV 特征具有更强的环境表征能力和泛化能力。文献[12]利用稀疏的 BEV 查询设计 3D 环境感知方法,通过稀疏查询预测出 3D 参考点,并利用几何信息将参考点投影到图像平面。Zhou 等[13]设计基于相机感知的交叉注意力模块,隐式地学习 PV 到 BEV 的几何变换。Li 等[14]提出基于时态结构的 BEVFormer 模型,通过预设网格状 BEV 查询与空间域和时间域进行信息交互,并将其聚合为统一的 BEV 特征。BEVFormer 在多个基准上达到了先进的性能,但其参数量和计算开销相对较大。轻量化的 Transformer 模型具有相对较小的参数量和计算开销。但是,该类 BEV 生成模型为提高视图的转换精度,会针对不同状态目标要素进行独立的特征提取。这种特征提取方式因丢失不同状态目标要素之间的关联信息,使得 BEV 视图的转换精度提升有限。

针对上述问题,本文提出一种端到端的基于轻量级 Transformer 的 BEV 生成模型。首先,利用 EfficientNet[15]作为特征提取器,提取同一时刻不同视角 PV 视图的多尺度特征;其次,设计面向多尺度特征的 Transformer 循环跨视图转换模块(Transformer recurrent cross-view transformation module, TRCTM),通过嵌入位置编码的 BEV 查询,将多尺度 PV 视图特征分别聚合至静态道路和动态车辆的 BEV 视图中;最后,设计面向语义不一致的多状态 BEV 特征融合模块(multi-state BEV feature fusion module, MSFM),通过提取静态道路和动态车辆特征的关联信息,实现两种状态的 BEV 特征融合。在 NuScenes[16]数据集上对比本文方法与其他先进的 BEV 转换方法,验证了本文方法能够在确保模型具有相对较低的参数量和计算复杂度的同时,有效提高 BEV 视图的生成精度。

1 端到端的基于轻量化 Transformer 的 BEV 生成模型架构

1.1 网络结构

本文提出的基于轻量化 Transformer 的 BEV 生成模型(light-weight Transformer-based perspective view to bird's eye view, LT-P2B)架构如图 1 所示。该模型结构包括编码(特征提取)、视图转换、多状态特征融合和解码 4 个主要部分。

首先,将同一时刻不同视角的 PV 视图合成一个维度为 $N_{\text{view}} \times H \times W \times C$ 的批量,其中 N_{view} 为车载相机的个数,H、W 和 C 分别为图像的高度、宽度和深度;以上述批量为输入,EfficientNet 为特征提取器,提取同一时刻不同视角 PV 视图的多尺度特征 $\boldsymbol{F} = \{\boldsymbol{F}_k\}_{k=1}^{N_{\text{view}}}$,其中 $\boldsymbol{F}_k$ 表示第 k 个 PV 视图的多尺度特征;然后,分别构建面向静态道路和动态车辆的 BEV 查询向量 Q_r 和 Q_v,通过面向多尺度特征的 Transformer 循环跨视图注意力模块,查询 PV 视图多尺度特征 $\boldsymbol{F}$ 中语义和位置信息,得到独立的静态道路和动态车辆 BEV 特征 $\boldsymbol{B}_r$ 和 $\boldsymbol{B}_v$;进一步,设计面向语义不一致的多状态 BEV 特征融合模块,建立静态道路和动态车辆的 BEV 特征关联,实现两种状态 BEV 视图特征的空间和语义融合;最后,经解码层细化输出更加精确的 BEV 视图。

1.2 面向多尺度特征的 Transformer 循环跨视图转换模块

视图转换的目的是实现 PV 视图像素坐标到 BEV 视图世界坐标的映射转换。对于任意 BEV 视图的世界坐标点,利用余弦相似度即可建立该点与 PV 视图像素坐标点之间的几何关系。但该方式仅能够建立点与点之间的映射,缺乏特征序列间的匹配,进而导致视图转换过程

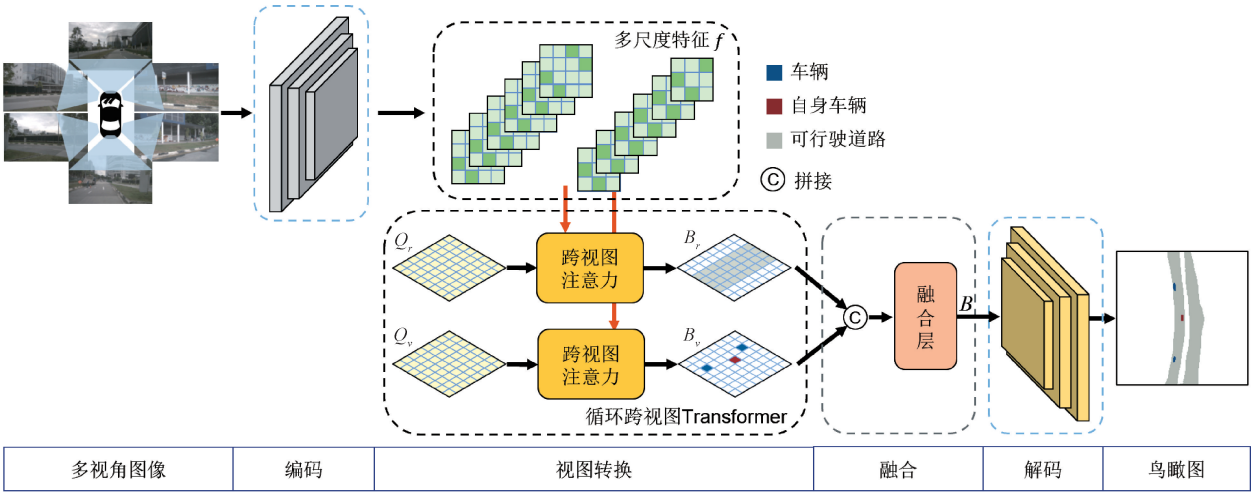

图 1　LT-P2B 方法整体架构

Fig. 1　Overall architecture of LT-P2B method

中缺乏语义信息的对应。Transformer 利用注意力机制实现对输入的编码和表征学习,通过捕捉特征序列中不同位置的依赖关系[17],可同时实现语义和位置信息的映射。因此,Transformer 模型十分适用于多视角下 PV 视图到 BEV 视图的转换。然而,Transformer 的排列不变性导致其无法学习 PV 与 BEV 之间的位置对应关系。利用车载相机本身的几何信息可以有效解决该问题[18]。基于此,本文提出面向多尺度特征的 Transformer 循环跨视图转换模块,结构如图 2 所示。

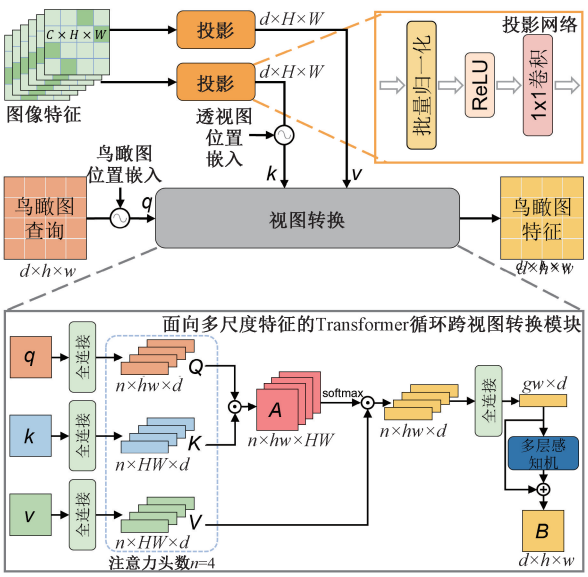

图 2　面向多尺度特征的 Transformer 循环跨视图转换模块

Fig. 2　Transformer recurrent cross-view transformation module for multi-scale feature

假设编码后输出的某一尺度 PV 视图特征为 $\boldsymbol{f}^i \in \boldsymbol{R}^{N_{\text{view}}\times H^i\times W^i\times C^i}$,其中 i 表示 PV 视图特征的第 i 个尺度大小。利用车载相机的参数信息,将 $\boldsymbol{f}^i$ 中像素点集合 $P^i=\{(m^i,n^i)\mid m^i=1,2,\cdots,H^i;n^i=1,2,\cdots,W^i\}$ 投影至世界坐标系中对应的位置,形成像素坐标映射至世界坐标系中的位置集合 D^i:

$$D^i=\boldsymbol{R}^{-1}\boldsymbol{E}^{-1}P^i \tag{1}$$

式中:$\boldsymbol{E}\in\boldsymbol{R}^{3\times3}$,$\boldsymbol{R}\in\boldsymbol{R}^{3\times3}$,分别表示车载相机的内部参数和外部参数。同时,以车辆自身为原点建立 BEV 视图三维坐标网格 $\boldsymbol{g}\in\boldsymbol{R}^{N_{\text{view}}\times H^g\times W^g\times3}$,用于表示自车周围特定范围特征。构建方法如下:1)依据预设 BEV 查询尺寸 (H^g,W^g) 构建以自车位置中心,坐标原点位于自车左下方的三维网格 $\boldsymbol{g}'\in\boldsymbol{R}^{N_{\text{view}}\times H^g\times W^g\times3}$。其中每个网格坐标点包含 (x,y,z)3 个元素,$x\in\boldsymbol{R}^{W^g}$ 和 $y\in\boldsymbol{R}^{H^g}$ 分别对应 $[0,W^g]$ 和 $[0,H^g]$ 的等间隔坐标,由于不涉及目标的高度特征,z 的值始终为 1;2)依据构建的 BEV 网格尺寸 (H^g,W^g),以及该网格需要表征的真实世界中场景尺寸 (H^w,W^w),生成变换矩阵 $\boldsymbol{M}_V\in\boldsymbol{R}^{3\times3}$,该矩阵包含世界坐标映射至 BEV 三维网格的元素缩放比例,以及坐标原点移至网格中心(自车位置)的元素平移距离。具体公式如下:

$$\boldsymbol{g}=\boldsymbol{M}_V^{-1}\boldsymbol{g}'$$
$$\boldsymbol{M}_V=\begin{bmatrix}0 & -W^B/W^w & W^B/2\\ -H^B/H^w & 0 & H^B/2\\ 0 & 0 & 1\end{bmatrix} \tag{2}$$

式中:$-H^B/H^w$ 和 $-W^B/W^w$ 分别表示世界坐标系中元素的高度和宽度映射至网格坐标系的缩放比例;$W^B/2$ 和 $H^B/2$ 分别表示坐标原点移至网格中心的元素高度和宽度平移距离。基于此,利用共享权重的 1×1 卷积分别对

$\boldsymbol{g}$ 和 D^i 中的每个元素进行编码,生成 BEV 视图下的查询位置编码 $\boldsymbol{\varphi} \in \boldsymbol{R}^{N_{\text{view}} \times H^g \times W^g \times C^g}$,以及 PV 视图下的位置编码 $\boldsymbol{\delta}^i \in \boldsymbol{R}^{N_{\text{view}} \times H^i \times W^i \times C_g}$。进一步,将 $\boldsymbol{\delta}^i$ 嵌入对应的 $\boldsymbol{f}^i$,得到具有位置编码信息的 PV 视图特征 $\boldsymbol{\psi}^i$。利用 3 个并行的全连接层作为可学习权重 $\boldsymbol{W}^{Q^i}$、$\boldsymbol{W}^{K^i}$ 和 $\boldsymbol{W}^{V^i}$ 分别与 $\boldsymbol{\varphi}^i$、$\boldsymbol{\psi}^i$ 和 $\boldsymbol{f}^i$ 相乘:

$$\boldsymbol{Q}^i = \boldsymbol{W}^{Q^i}\boldsymbol{\varphi}^i, \boldsymbol{W}^Q \in \boldsymbol{R}^{C^g \times n_h d} \tag{3}$$

$$\boldsymbol{K}^i = \boldsymbol{W}^{K^i}\boldsymbol{\psi}_k^i, \boldsymbol{W}^K \in \boldsymbol{R}^{C^g \times n_h d} \tag{4}$$

$$\boldsymbol{V}^i = \boldsymbol{W}^{V^i}\boldsymbol{f}_k^i, \boldsymbol{W}^V \in \boldsymbol{R}^{C^g \times n_h d} \tag{5}$$

式中:n_h、d 分别表示本文采用的注意力头数以及每个注意力头的维度。为方便对所有视角的 PV 视图特征进行查询,需要对 $\boldsymbol{Q}^i$、$\boldsymbol{K}^i$ 和 $\boldsymbol{V}^i$ 进行矩阵变换,得到 $\boldsymbol{Q}^i \in \boldsymbol{R}^{n_h \times N_{\text{view}} \times H^g \cdot W^g \times d}$、$\boldsymbol{K}^i \in \boldsymbol{R}^{n_h \times N_{\text{view}} \times H^g \cdot W^g \times d}$ 和 $\boldsymbol{V}^i \in \boldsymbol{R}^{n_h \times N_{\text{view}} \cdot H^i \cdot W^i \times d}$。然后,对 $\boldsymbol{Q}^i$ 和 $\boldsymbol{K}^i$ 进行矩阵乘法,得到两者相关性矩阵 $\boldsymbol{W}_a{}^i$:

$$\boldsymbol{W}_a{}^i = \boldsymbol{Q} \otimes \boldsymbol{K}^{\mathrm{T}} \tag{6}$$

式中:$\boldsymbol{W}_a{}^i \in \boldsymbol{R}^{n_h \times H^g \cdot W^g \times N_{\text{view}} \cdot H^i \cdot W^i}$ 表示 BEV 特征查询与 PV 特征中每个元素之间的相关程度。利用 Softmax 函数将 $\boldsymbol{W}_a{}^i$ 中的权重元素归一化至(0,1)内,并将其与 $\boldsymbol{V}^i$ 相乘,实现 BEV 特征中每个元素对 PV 视图特征中最强相关元素的查询:

$$\boldsymbol{A} = \mathrm{softmax}(\boldsymbol{W}_a) \odot \boldsymbol{V} \tag{7}$$

最后,通过全连接层与 MLP 网络的残差连接得到当前尺度的 BEV 特征输出 $\boldsymbol{B}^i$:

$$\boldsymbol{B}^i = L(\boldsymbol{A}) + MLP(L(\boldsymbol{A})) \tag{8}$$

式中:$L(\cdot)$ 表示全连接层;$MLP(\cdot) = L(GELU(L(\cdot)))$,$GELU(\cdot)$ 表示 GELU 激活函数。

为实现不同尺度特征的融合,本文将得到的当前尺度 BEV 特征 $\boldsymbol{B}^i$ 作为下一尺度 BEV 查询,循环上述过程得到最后的 BEV 特征 $\boldsymbol{B}$。此外,为实现不同状态的 BEV 特征鲁棒性表征及模型的轻量化设计,本文在该阶段针对静态道路和动态车辆分别构建 BEV 查询,最后得到两种不同状态的 BEV 特征图谱。

1.3 面向语义不一致的多状态 BEV 视图特征融合模块

为实现视图转换 Transformer 模型的轻量化设计,在 PV 到 BEV 的跨视图转换阶段,本文针对静态道路和动态车辆分别提取 BEV 特征。为实现交通场景下高精度 BEV 视图的构建,需要对静态道路特征和动态车辆特征进行融合。然而,两种状态特征之间存在着较大的语义不一致性,简单的特征级联或元素相加的融合方式会影响特征之间的关联性。为解决该问题,本文提出针对语义不一致多状态 BEV 视图特征融合模块,结构如图 3 所示。

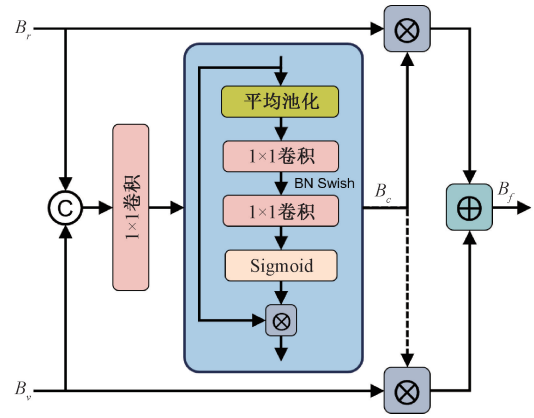

图 3 面向语义不一致的多状态 BEV 视图特征融合模块

Fig. 3 Multi-state BEV feature fusion module targeting semantic inconsistency

首先,将面向多尺度特征的 Transformer 循环跨视图转换模块分别输出的静态道路 BEV 特征 $\boldsymbol{B}_r \in \boldsymbol{R}^{H^g \times W^g \times C^g}$ 与动态车辆 BEV 特征 $\boldsymbol{B}_v \in \boldsymbol{R}^{H^g \times W^g \times C^g}$ 进行级联,形成 $\boldsymbol{B}_c \in \boldsymbol{R}^{H^g \times W^g \times 2C^g}$;然后,采用 1 × 1 卷积对 $\boldsymbol{B}_c$ 进行降维,得到 $\boldsymbol{B}_d \in \boldsymbol{R}^{H^g \times W^g \times C^g}$,其中 C^g 为每个通道的初始化学习权重;进一步,利用通道注意力输出 $\boldsymbol{B}_r$ 和 $\boldsymbol{B}_v$ 的融合权重 $\boldsymbol{B}' \in \boldsymbol{R}^{H^g \times W^g \times C^g}$,该权重通过网络的反馈传播实现自适应学习;最后,分别将 $\boldsymbol{B}_r$、$\boldsymbol{B}_v$ 与 $\boldsymbol{B}'$、$1 - \boldsymbol{B}'$ 进行逐元素相乘后再相加,得到融合后的 BEV 特征图谱。融合过程为:

$$\begin{aligned} \boldsymbol{B}' &= \mathrm{Conv}_{1\times1}((\mathrm{Conv}_{1\times1}(\mathrm{Concat}(\boldsymbol{B}_r, \boldsymbol{B}_v))) \times \boldsymbol{B}_d) \\ \boldsymbol{B}_f &= \boldsymbol{B}' \otimes \boldsymbol{B}_r \oplus (1 - \boldsymbol{B}') \otimes \boldsymbol{B}_v \end{aligned} \tag{9}$$

由于静态道路特征和动态车辆特征侧重的目标不同,采用 1×1 卷积和不同的融合权重能够对 BEV 视图特征进行重新排列,实现语义不一致的静态特征和动态特征融合精度提升。

基于面向多尺度特征的 Transformer 循环跨视图转换模块和面向语义不一致的多状态 BEV 视图特征融合模块的设计,本文采用端到端单阶段的模型训练策略,通过建立动态车辆和静态道路的特征关联,滤除生成 BEV 视图中的噪声,进一步提升 BEV 视图精度。

2 算法实现

结合上述模型网络结构及子模块描述,本文给出端到端的基于轻量化 Transformer 的多状态特征融合 BEV 生成算法实现过程,算法实现伪代码如下。

端到端的基于轻量化 Transformer 的多状态特征融合鸟瞰图生成算法

Input: 多视角图像 $\{\boldsymbol{I}_n\}_{n=1}^{6}$, 相机内外参 $\{\boldsymbol{E}_n\}_{n=1}^{6}$, $\{\boldsymbol{R}_n\}_{n=1}^{6}$, 图像尺度数量 N_{scale}

Output: 鸟瞰图 $\boldsymbol{B}$

1. 初始化:鸟瞰图查询 $\boldsymbol{Q}=\{\boldsymbol{Q}_r,\boldsymbol{Q}_v\}$,鸟瞰图三维坐标网格 $\boldsymbol{g}\in[0,1]$,$E_p(\boldsymbol{R},\boldsymbol{E},\boldsymbol{f})$ 为透视图位置嵌入,如式(1),$E_b(\boldsymbol{R},\boldsymbol{g})$ 为鸟瞰图位置嵌入。
   //使用 EfficientNet 提取多视角图像特征
2. $\boldsymbol{f}=\{\boldsymbol{f}_i\}_{i=1}^{N_{\text{scale}}}\leftarrow\text{EfficientNet}(\boldsymbol{I}_n)$
   //对多尺度特征进行循环视图转换
3. **for** $i=1,2,\cdots,N_{\text{scale}}$ **do**
   //对图像特征 $\boldsymbol{f}$ 进行位置嵌入,并投影得到 $\boldsymbol{K}$
4. $\quad\boldsymbol{K}\leftarrow\text{project}(\boldsymbol{f}_i+E_p(\boldsymbol{R}_n,\boldsymbol{E}_n,\boldsymbol{f}_i))$
   //对图像特征 $\boldsymbol{f}$ 投影得到 $\boldsymbol{V}$
5. $\quad\boldsymbol{V}\leftarrow\text{project}(\boldsymbol{f}_i)$
   //分别对静态道路和动态车辆进行视图转换
6. $\quad$**for** $j=r,v$ **do**
   //对查询进行鸟瞰图位置嵌入
7. $\quad\quad\boldsymbol{Q}_j\leftarrow\boldsymbol{Q}_j+E_b(\boldsymbol{R},\boldsymbol{g})$
   //计算注意力权重并得到新的查询
8. $\quad\quad\boldsymbol{A}_j\leftarrow\text{softmax}(\boldsymbol{Q}_j\boldsymbol{K}^T)\boldsymbol{V}$
9. $\quad\quad\boldsymbol{Q}_j\leftarrow\text{Linear}(\boldsymbol{A}_j)+\text{MLP}(\text{Linear}(\boldsymbol{A}_j))$
10. $\quad$**end for**
11. **end for**
   //对 $\boldsymbol{Q}_r,\boldsymbol{Q}_v$ 进行特征融合,如式(8)
12. $\boldsymbol{B}\leftarrow\text{Fusion}(\boldsymbol{Q}_r,\boldsymbol{Q}_v)$
   //对融合后的鸟瞰图特征 $\boldsymbol{B}$ 通过解码器上采样得到鸟瞰图
13. $\boldsymbol{B}\leftarrow\text{Decoder}(\boldsymbol{B})$

1)将同一时刻不同视角的 PV 视图合成维度为 $N_{\text{view}}\times H\times W\times C$ 的一个批量,并作为模型的输入。

2)利用 EfficientNet 提取两种尺度分别为 $N_{\text{view}}\times H^1\times W^1\times C^1$ 和 $N_{\text{view}}\times H^2\times W^2\times C^2$ 的 PV 视图特征 $\boldsymbol{f}^1$ 和 $\boldsymbol{f}^2$。

3)分别针对 $\boldsymbol{f}^1$ 和 $\boldsymbol{f}^2$ 进行投影和特征位置编码,得到面向多尺度特征的 Transformer 循环跨视图转换模块的输入 $\boldsymbol{V}$ 和 $\boldsymbol{K}$;同时,依据构建 BEV 视图的尺度,初始化静态道路和动态车辆的 BEV 查询 $\boldsymbol{Q}_r$ 和 $\boldsymbol{Q}_v$。

4)利用面向多尺度特征的 Transformer 循环跨视图转换模块,分别输出针对 $\boldsymbol{f}^1$ 的静态道路和动态车辆 BEV 视图特征。进一步,将其作为针对 $\boldsymbol{f}^2$ 的 BEV 查询,再次利用该转换模块输出多尺度融合的静态道路和动态车辆的 BEV 查询 $\boldsymbol{B}_r$ 和 $\boldsymbol{B}_v$。

5)将 $\boldsymbol{B}_r$ 和 $\boldsymbol{B}_v$ 输入面向语义不一致的多状态 BEV 视图特征融合模块,实现两种状态 BEV 视图特征的空间和语义融合。

6)通过解码层细化输出精确的 BEV 视图。

3　实验分析

3.1　实验环境与数据集

本文实验的训练和测试过程均基于 PyTorch 深度学习框架完成。实验硬件环境为:英特尔 i7-13700K @ 3.4 GHz CPU 和 NVIDIA RTX 4080 GPU,显存为 16 GB。每张输入图像尺寸调整为 224×480,采用随机旋转、随机缩放等方法进行数据增强。使用 focal 损失函数[19] 和 AdamW 优化器训练模型,批处理大小为 4,设置初始学习率和权重衰减分别为 4×10^{-3} 和 1×10^{-7},训练网络模型 30 个 epoch。反向传播时,分别计算针对静态道路和动态车辆的损失,最后通过损失权重计算总损失。计算方式如下:

$$\mathcal{L}=\lambda_1\mathcal{L}_r(\boldsymbol{B}_r)+\lambda_2\mathcal{L}_v(\boldsymbol{B}_v)\tag{10}$$

式中:λ_1 和 λ_2 分别表示静态道路和动态车辆的损失权重;$\mathcal{L}_r(\boldsymbol{B}_r)$ 和 $\mathcal{L}_v(\boldsymbol{B}_v)$ 分别表示静态道路和动态车辆的 BEV 特征与对应标签之间的 focal 损失。

模型的训练和测试在 NuScenes 数据集上进行。NuScenes 数据集包含了不同天气、时间、光照和交通条件下的 1 000 个复杂交通场景。每个场景持续 20 s,每 0.5 s 采集 1 次,共包含 40 帧数据。每帧数据中记录了由 6 个车载相机采集的车辆周围 360°的 PV 视图,以及校准后的相机内参 E 和外参 (R,t) 等数据。数据集利用地图信息、激光雷达和车辆姿态等数据,生成分辨率为 200×200 的占用掩模(occupancy mask)作为真值(ground truth)标签,其中对静态道路和动态车辆进行了标注。

3.2　评价指标

复杂交通场景下的 PV 到 BEV 视图的转换实际涉及不同交通要素的语义分割。因此,可采用 BEV 视图下语义分割的性能评价指标评判转换生成 BEV 视图的精度。

本文通过计算模型预测输出和真值标签之间的交并比(intersection over union, IoU)评估模型性能。IoU 是表征算法的轮廓描述能力常用评价指标,被定义为预测和标签之间交集和并集的面积之比,公式为:

$$IoU=\frac{TP}{TP+FP+FN}\tag{11}$$

式中:TP、FP 和 FN 分别表示真阳性、假阳性和假阴性的数量。为验证单阶段端到端模型训练的有效性,本文另外对模型参数量、运算量和推理时间进行评估。

3.3　对比实验分析

为验证本文方法的有效性,将本文方法 LT-P2V 与

几种先进的 BEV 视图生成方法 VPN[9]、CVT[13]、BEVFormer[14]、PON[20]、OFT[21]、FIERY[22]、M²BEV[23]、MSTFB[24]以及 PointBEV[25]进行性能对比实验。本文从动态车辆与静态道路的语义分割精度、模型参数量、运算量和推理时间等方面进行对比,结果如表 1 所示。从表 1 可以看出,LT-P2V 能够在实现模型轻量化的前提下,提升 BEV 视图的生成精度。在语义分割性能方面,本文针对生成 BEV 视图下的动态车辆和静态道路的分割精度分别达到了 43.2% 和 82.0%。相较于性能先进的 BEVFormer 模型,本文方法的精度提升了 0.1% 和 0.3%,但是 BEVFormer 模型的参数量和运算复杂度分别是本文方法的 5.7 倍和 3 倍。在模型轻量化方面,本文方法的模型参数量和运算量分别为 12×10^6 和 12 GB。相较于 MSTFB 模型,本文方法与其在模型参数量和运算量上性能相近,但本文方法在生成 BEV 视图下的动态车辆和静态道路语义分割的精度分别提升了 4.0% 和 5.2%。在模型效率方面,本文方法的推理时间为 16 ms,仅为 BEVFormer 的 0.06 倍。

表 1 不同方法性能对比

Table 1 Comparison of different method

方法	主干网络	动态车辆/%	静态道路/%	参数量/$\times10^6$	运算量/GB	推理时间/ms
VPN[9]	ResNet18	31.0	76.9	18	62	8
OFT[21]	ResNet18	30.1	71.7	22		48
PON[20]	ResNet50	24.7		38	90	20
BEVFormer[14]	ResNet101	43.1	80.7	68.1	36	259
M²BEV[23]	ResNeXt101		77.2	47.6		
CVT*[13]	EfficientNetB4	36.8	74.3	7(14)	7(14)	11
FIERY*[22]	EfficientNetB4	35.8		7(14)	15(30)	19
MSTFB[24]	EfficientNetB4	39.2	76.8	10.5	21.67	
PonitBEV[25]	EfficientNetB4	38.7		7.6	59.3	38.6
LT-P2B	EfficientNetB4	43.2	82.0	12	12	16

注:* 表示该模型需要经过两阶段运行,表中只展示模型针对动态车辆推理时的参数量和运算量;括号内数值表示模型针对动态车辆和静态道路推理时的参数量和运算量。

此外,为了直观感知本文方法的性能,本文方法、PointBEV 和 FIERY 生成的 BEV 视图的可视化结果如图 4 所示。其中,本文方法预测的 BEV 视图中可行驶道路和自身车辆已在图中标注,其余未标注的为预测出的其他车辆。在静态道路识别方面,对比本文方法推理结果与标签可以看出,本文方法在复杂道路场景中预测出的可行驶道路与原道路基本一致。在动态车辆识别方面,图中圆实线标记了本文方法相较于 PointBEV 和 FIERY 算法额外识别出的真阳性车辆,由此可以看出,本文方法能够实现更加精准的车辆分割,尤其是针对遮挡非常严重和距离相对较远的车辆。

3.4 消融实验

为了验证本文方法中各子模块(TRCTM 和 MSFM)及单阶段训练策略的有效性,在 NuScenes 数据集上进行了消融实验,结果如表 2 所示。从表 2 可知,采用单阶段训练策略分别对动态车辆和静态道路的分割精度提升了 4.9% 和 9.1%。由此可见,利用单阶段训练策略实现不同状态特征融合,能够显著提升生成 BEV 视图的精度。在单阶段训练策略的基础上,本文继续探究 TRCTM 和 MSFM 两个子模块的性能。从表 2 可以看出,分别采用 TRCTM 和 MSFM 能够显著提升生成 BEV 视图的精度。特别地,两个子模块的融合应用可以使生成的 BEV 视图精度达到最优,分别对动态车辆和静态道路的分割精度提升了 1.5% 和 2.1%。由此说明,TRCTM 能够有效利用注意力机制实现对输入图像的编码和表征学习,捕捉特征序列中不同位置的依赖关系,提升 BEV 视图特征的鲁棒性提取;MSFM 模块能够针对语义不一致的 BEV 特征进行融合,实现关联特征的增强,进而提升 BEV 视图的生成精度。

表 2 消融实验结果

Table 2 Results of ablation experiment (%)

单阶段训练	TRCTM	MSFM	动态车辆	静态道路
			36.8	70.0
√			41.7	79.1
√	√		42.2	81.5
√		√	42.3	80.2
√	√	√	43.2	82.0

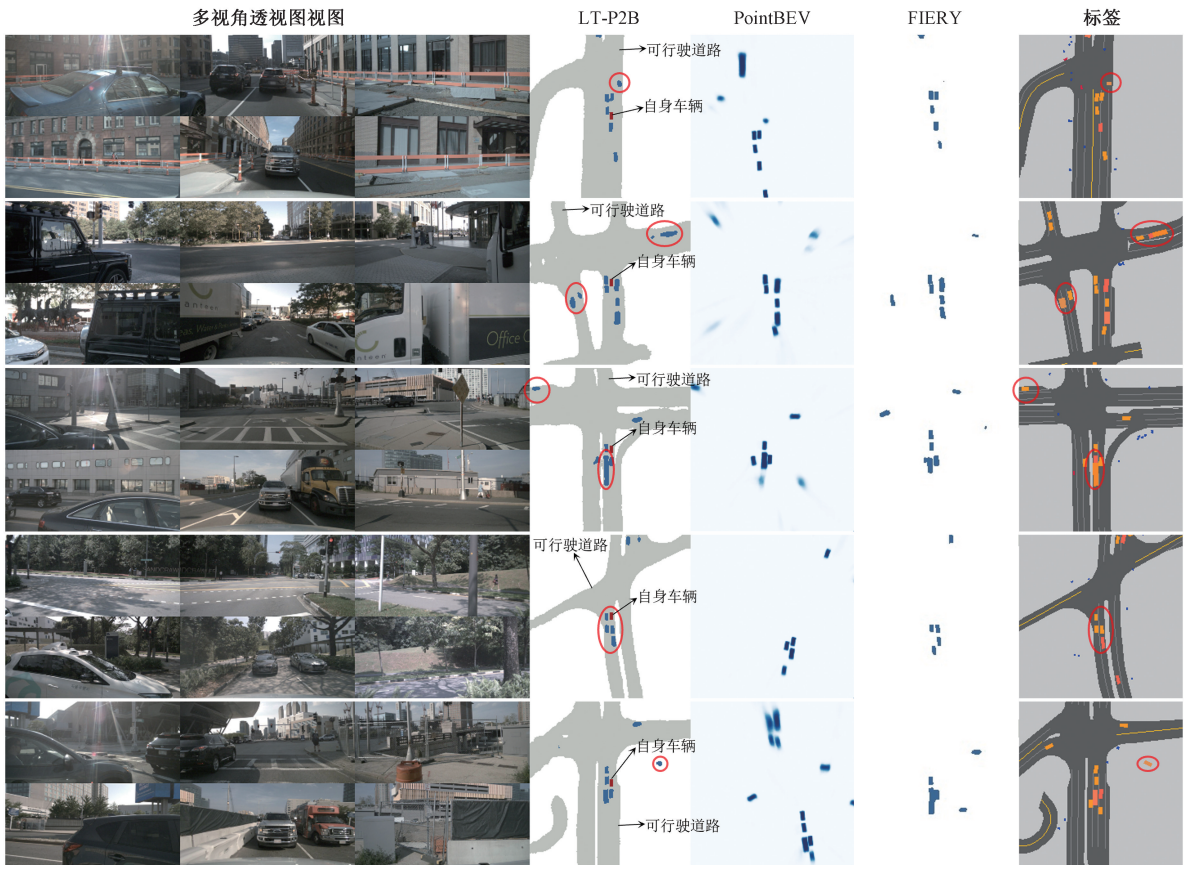

图 4　不同方法在 NuScenes 验证集中可视化结果

Fig. 4　Visualization results of different method on NuScenes

此外,为了研究 PV 视图下的特征嵌入对视图转换性能的影响。本文针对注意力机制中 K 的构建进行了消融实验,实验结果如表 3 所示。在 K 的构建过程中仅采用位置编码时,模型对静态道路和动态车辆的语义分割精度分别为 81.6% 和 40.6%。嵌入 PV 视图特征后,模型针对静态道路和动态车辆的语义分割精度分别提升了 0.4% 和 2.6%。由此可知,采用 PV 视图下的特征位置编码嵌入得到的 K,能够在生成注意力权重的同时,利用目标的外观和几何特征对 BEV 和 PV 视图特征的对应关系进行推理,提升目标的分割精度。

表 3　特征嵌入对性能的影响

Table 3　The effect of feature embedding for performance

(%)

位置编码	特征嵌入	动态车辆	静态道路
√		40.6	81.6
√	√	43.2	82.0

3.5　鲁棒性分析

为验证提出方法的鲁棒性,分析对比了本文算法与 CVT、PointBEV、FIERY 在远距离车辆感知和车辆被遮挡情况下的性能表现,结果分别如图 5、6 所示。

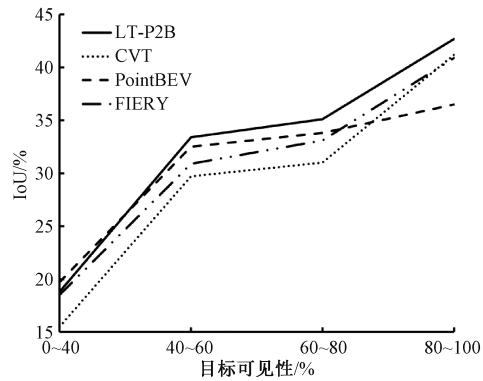

图 5　不同可见性车辆的 IoU 指标

Fig. 5　The IoU of different visibility vehicles

在车辆遮挡方面,根据数据标签将目标的可见性分为 4 个等级:车辆严重遮挡(0%~40%)、车辆遮挡(40%~60%)、车辆可见(60%~80%)和车辆高度可见(80%~100%)。从图 5 可以看出,在不同可见性条件下,本文方法的 IoU 指标均高于其他方法。以 CVT 为例,在车辆严

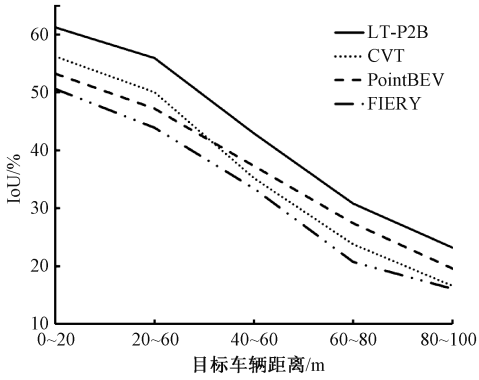

图 6 不同距离车辆的 IoU 指标

Fig. 6 The IoU of different distances vehicles

重遮挡时,本文方法的 IoU 指标为 18.8%,高于 CVT 的 15.2%。在车辆高度可见时,本文方法的 IoU 达到了 42.7%,比 CVT 高了 1.2%。

在远距离车辆感知方面,根据车辆与自车距离,将其分为 5 种不同距离情况:近距离(0~20 m)、相对近距离(20~40 m)、相对远距离(40~60 m)、较远距离(60~80 m)和远距离(80~100 m)。在近距离情况下,本文方法针对车辆的 IoU 指标为 61.2%。虽然识别精度随着距离逐渐下降,但其性能仍然优于其他模型。尤其在远距离车辆检测方法,本文方法高出次优的 PointBEV 大约 4%。

此外,为了研究损失权重 λ_1 和 λ_2 对模型性能的影响,本文设置多组 λ_1 和 λ_2 进行实验,实验结果如表 4 所示。当静态道路损失所占比重增大时,模型对动态车辆的识别精度从 26.2% 下降到 6%,且对静态道路的识别精度没有提升;在动态车辆损失所占比重逐渐增大时,针对动态车辆的识别精度从 26.2% 提升至 43.2%,且对静态道路的识别精度几乎没有影响。综上可知,λ_1 和 λ_2 的比值对静态道路识别精度影响较小。静态道路损失占比的增加并不会使其精度提升,而动态车辆的识别精度会随着其损失所占比重增加出现明显提升。本文取 $\lambda_1 = 1, \lambda_2 = 10$ 进行总损失的计算。相比于 λ_1/λ_2 为 1 和 0.2 时,虽然针对静态道路的识别精度分别下降了 0.3% 和 1.2%,但针对动态车辆的识别精度分别提升了 17% 和 25.8%。

表 4 参数 λ_1 和 λ_2 对性能的影响

Table 4 The effect of λ_1 and λ_2 for performance

λ_1	λ_1	动态车辆/%	静态道路/%
1	1	26.2	82.3
5	1	17.4	79.7
10	1	6	81
1	5	32.8	83.2
1	10	43.2	82.0

4 结 论

为有效平衡 BEV 视图生成模型的复杂度和视图生成精度,本文提出一种端到端的基于轻量级 Transformer 的 BEV 生成模型。本文首先设计面向多尺度特征的 Transformer 循环跨视图转换模块,利用注意力机制实现对输入图像的编码和表征学习,捕捉特征序列中不同位置的依赖关系,提升 BEV 视图特征的鲁棒性提取;然后利用面向语义不一致多状态 BEV 视图特征融合模块能够针对语义不一致的 BEV 特征进行融合,实现关联特征的增强,提升 BEV 视图的生成精度。在 NuScenes 数据集上进行的实验表明,本文方法能够在提高 BEV 视图生成精度的同时,有效降低模型参数量和计算复杂度。

参考文献

[1] NG M H, RADIA K, CHEN J F, et al. Bev-seg: Bird's eye view semantic segmentation using geometry and semantic point cloud [J]. ArXiv preprint arXiv: 2006.11436, 2020.

[2] 金宇锋, 陶重犇. 基于 Transformer 的融合信息增强 3D 目标检测算法[J]. 仪器仪表学报, 2023, 44(12): 297-306.
JING Y F, TAO CH B. Fusion information enhanced method based on transformer for 3D object detection[J]. Chinese Journal of Scientific Instrument, 2023, 44(12): 297-306.

[3] VORA S, LANG A H, HELOU B, et al. Pointpainting: Sequential fusion for 3d object detection [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 4604-4612.

[4] JIANG Y Q, ZHANG L, MIAO ZH W, et al. Polarformer: Multi-camera 3d object detection with polar transformer[C]. AAAI Conference on Artificial Intelligence, 2023, 37(1): 1042-1050.

[5] CHEN SH Y, CHENG T H, WANG X G, et al. Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer[J]. ArXiv preprint arXiv:2206.04584, 2022.

[6] LIU ZH J, TANG H T, AMINI A, et al. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation[C]. IEEE International Conference on Robotics and Automation, 2023: 2774-2781.

[7] ABBAS S A, ZISSERMAN A. A geometric approach to obtain a bird's eye view from an image [C]. IEEE

Conference on International Conference on Computer Vision, 2019: 4095-4104.

[8] PHILION J, FIDLER S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D [C]. European Conference on Computer Vision, 2020: 194-210.

[9] PAN B W, SUN J K, LEUNG H Y T, et al. Cross-view semantic segmentation for sensing surroundings [J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4867-4873.

[10] LU CH Y, VAN DE MOLENGRAFT M J G, DUBBELMAN G. Monocular semantic occupancy grid mapping with convolutional variational encoder-decoder networks[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 445-452.

[11] LI Q, WANG Y, WANG Y L, et al. Hdmapnet: An online HD map construction and evaluation framework[C]. International Conference on Robotics and Automation, 2022: 4628-4634.

[12] LIU Y F, YAN J J, JIA F, et al. Petrv2: A unified frame-work for 3D perception from multi-camera images[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 3262-3272.

[13] ZHOU B, KRÄHENBÜHL P. Cross-view transformers for real-time map-view semantic segmentation [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2022: 13760-13769.

[14] LI ZH Q, WANG W H, LI H Y, et al. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers [C]. European Confer-ence on Computer Vision, 2022: 1-18.

[15] TAN M X, LE Q V. Efficientnet: Rethinking model scaling for convolutional neural networks [C]. IEEE International Conference on Machine Learning, 2019: 6105-6114.

[16] CAESAR H, BANKITI V, LANG A H, et al. Nuscenes: A multimodal dataset for autonomous driving[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 11621-11631.

[17] 李冰锋, 刘帅, 杨艺. 基于改进的 Transformer 细粒度图像识别算法研究[J]. 电子测量技术, 2024, 47(2): 114-120.
LI B F, LIU SH, YANG Y. Research on improved Transformer fine-grained image recognition algorithm[J]. Electronic Measurement Technology, 2024, 47(2): 114-120.

[18] 吴军, 袁少博, 祝玉恒, 等. 采用自适应背景聚类的激光雷达与相机外参标定优化方法[J]. 仪器仪表学报, 2023, 44(2): 230-237.
WU J, YUAN SH B, ZHU Y H, et al. Optimization method for external parameters calibration of lidar and camera using adaptive background clustering [J]. Chinese Journal of Scientific Instrument, 2023, 44(2): 230-237.

[19] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2980-2988.

[20] RODDICK T, CIPOLLA R. Predicting semantic map representations from images using pyramid occupancy networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11138-11147.

[21] RODDICK T, KENDALL A, CIPOLLA R. Orthographic feature transform for monocular 3D object detection[J]. ArXiv preprint arXiv: 1811.08188, 2018.

[22] HU A, MUREZ Z, MOHAN N, et al. Fiery: Future instance prediction in bird' s-eye view from surround monocular cameras [C]. IEEE Conference on International Conference on Computer Vision, 2021: 15273-15282.

[23] XIE EN Z, YU ZH D, ZHOU D Q, et al. M^2bev: Multi-camera joint 3D detection and segmentation with unified birds-eye view representation[J]. ArXiv preprint arXiv: 2204.05088, 2022.

[24] 冯霞, 陈爽, 卢敏, 等. 融合多源时空信息鸟瞰图的未来实例分割预测[J/OL]. 吉林大学学报(工学版), 1-11[2024-10-23].
FENG X, CHEN SH, LU M, et al. Future instance segmentation prediction based on bird's eye view of multi-source spatiotemporal information fusion[J/OL]. Journal of Jilin University (Engineering and Technology Edition), 1-11[2024-10-23].

[25] CHAMBON L, ZABLOCKI E, CHEN M, et al. Pointbev: A sparse approach for bev predictions [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2024: 15195-15204.

作者简介

刘明杰,2019 年于韩国仁荷大学获得博士学位,现为重庆邮电大学副教授,主要研究方向为智能网联汽车环境智能感知、多源信息融合、机器学习。

E-mail:liumj@ cqupt. edu. cn

Liu Mingjie received his Ph. D. degree from Inha University in 2019. Now he is an associate professor at Chongqing University of Posts and Telecommunications. His main research interests include environment perception for internet of vehicles, multi-information fusion, and machine learning.

何峥言,2022 年于重庆邮电大学获得学士学位,现为重庆邮电大学硕士研究生,主要研究方向为智能网联汽车环境智能感知。

E-mail:S210331004@ stu. cqupt. edu. cn

He Zhengyan received his B. Sc. degree from Chongqing University of Posts and Telecommunications in 2022. Now he is a M. Sc. candidate at Chongqing University of Posts and Telecommunications. His main research interests include environment perception for internet of vehicles.

陈俊生,2019 年于重庆大学获得博士学位,现为重庆邮电大学讲师,主要研究方向为智能网联汽车环境智能感知、能源设备状态智能诊断。

E-mail:chenjunsheng@ cqupt. edu. cn

Chen Junsheng received his Ph. D. degree from Chongqing University in 2019. Now he is a lecturer at Chongqing University of Posts and Telecommunications. His main research interests include environment perception for internet of vehicles, intelligent diagnosis of energy equipment status.

刘平,2017 年于浙江大学获得博士学位,现为重庆邮电大学副教授,主要研究方向为智能网联汽车环境智能感知及控制、轨迹优化。

E-mail:liuping_cqupt@ cqupt. edu. cn

Liu Ping received his Ph. D. degree from Zhejiang University in 2017. Now he is an associate professor at Chongqing University of Posts and Telecommunications. His main research interests include environment perception & control for internet of vehicles, and trajectory optimization.

朴昌浩(通信作者),2001 年于西安交通大学获得学士学位,2006 年于韩国仁荷大学获得博士学位,现为重庆邮电大学教授,主要研究方向为自动驾驶、智能网联汽车。

E-mail:piaoch@ cqupt. edu. cn

Piao Changhao (Corresponding author) received his B. Sc. degree from Xi′an Jiaotong University in 2001, received his Ph. D. degree from Inha University in 2006. Now he is a professor at Chongqing University of Posts and Telecom-munications. His main research interests include autonomous driving and internet of vehicles.