DOI: 10.19650/j.cnki.cjsi.J2513715

基于要素信息补全的自动驾驶复杂场景语义理解*

赵树恩[1],袁 亮[1],赵东宇[2]
(1. 重庆交通大学机电与车辆工程学院 重庆 400074; 2. 四川大学计算机学院 成都 610065)

摘 要:针对自动驾驶复杂交通场景精准感知与理解过程中路侧设施及交通参与者二维视觉图像几何特征信息不全、场景语义信息缺乏等问题,构建一种基于要素信息补全的自动驾驶复杂场景语义理解模型。首先,运用稠密连接网络(DenseNet)提取视觉图像多尺度二维特征,通过特征视线投影模块(FLoSP)将体素逆向映射至三维空间,采用维度分解残差(DDR)模块构建 3D UNet,提取场景目标三维特征,实现单帧视觉图像二维特征向三维特征的转换,再在 3D UNet 编码器与解码器之间引入三维上下文先验层(3D CRP),并通过空洞空间金字塔池化(ASPP)与 Softmax 层输出场景语义补全结果,以增强语义补全模型的空间语义理解能力。同时,运用图像描述生成技术,构建基于改进 VGG-16 编码器和长短时记忆网络(LSTM)解码器的上下文语义嵌入场景理解语言描述模型,其中改进 VGG-16 编码器将不同尺度的交通场景特征进行融合与拼接,并通过投影矩阵输入到 LSTM 解码器,建立场景目标图像与谓词关系的语义表示,进而自动生成目标检测结果及自动驾驶决策规划建议自然语言描述。最后,运用 Semantic KITTI 数据集及实车实验,对所提出的复杂场景语义理解算法进行验证。结果表明,该算法相较于 JS3C-Net 算法平均交并比(mIoU)相对提升了 11.27%,通过语义补全实现了自动驾驶复杂场景的准确感知与语义理解,为自动驾驶决策规划提供可靠依据。

关键词: 自动驾驶;场景语义补全;图像描述;场景语义理解
中图分类号: TH89 TP751 **文献标识码**: A **国家标准学科分类代码**: 510.40

Semantic understanding of complex scenarios in autonomou driving based on element information completion

Zhao Shuen[1], Yuan Liang[1], Zhao Dongyu[2]
(*1. College of Mechatronics and Vehicle Engineering, Chongqing Jiaotong University, Chongqing 400074, China*;
2. College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: To address the challenges of incomplete geometric feature information of two-dimensional visual images of roadside facilities and traffic participants, as well as the lack of scene semantic information inaccurate perception and understanding of complex traffic scenarios for autonomous driving, a semantic understanding model for complex autonomous driving scenarios based on element information completion is proposed. Firstly, a dense connection network (DenseNet) is utilized to extract multi-scale 2D features from visual images. Then, the feature line-of-sight projection (FLoSP) module is used to inverse-map voxels to 3D space. A dimension decomposition residual (DDR) module is utilized to construct a 3D UNet, extracting 3D features of scene objects and enabling the transformation of single-frame 2D visual image features into 3D features. Additionally, a contextual residual prior (3D CRP) layer is introduced between the 3D UNet encoder and decoder. Atrous spatial pyramid pooling (ASPP) and Softmax layers are used to output scene semantic completion results, thereby enhancing the spatial semantic understanding capability of the model. Meanwhile, image caption generation technology is utilized to formulate a context-aware semantic embedding scene understanding language description model based on an improved VGG-16 encoder and a long short-term memory (LSTM) decoder. The improved VGG-16 encoder integrates and concatenates features of traffic scenes at different scales and inputs them into the LSTM decoder via a projection matrix, establishing

收稿日期:2025-01-22 Received Date: 2025-01-22
*基金项目:国家自然科学基金项目(52072054)、重庆市自然科学基金创新发展联合基金项目(CSTB2024NSCQ-LZX0105)、重庆交通大学自然科学类揭榜挂帅项目(XJ2023)资助

a semantic representation between scene object images and predicate relations, and automatically generating natural language descriptions of object detection results and autonomous driving decision-making suggestions. Finally, the proposed complex scene semantic understanding algorithm is validated on the Semantic KITTI dataset and through real vehicle experiments. Compared with the JS3C-Net algorithm, the results show that the proposed algorithm achieves a relative improvement of 11.27% in mean intersection over union (mIoU), realizes accurate perception and semantic understanding of complex scenarios in autonomous driving through semantic completion, and provides a reliable basis for autonomous driving decision-making and planning.

Keywords: autonomous driving; semantic scene completion; image description; scene semantic understanding

0 引 言

自动驾驶车辆面临城市道路、高速公路、村镇路段等多样化场景的动态切换,其环境要素呈现高维度、异构性和时空关联性特征。因此,交通场景精准感知与理解是自动驾驶安全可靠决策规划的基础和关键[1]。随着自动驾驶技术向高阶智能化演进,复杂行车场景的语义理解已成为环境感知的研究热点。现有研究通过多传感器融合和深度学习模型取得了一定进展,但在要素信息缺失、传感器噪声干扰等现实场景下,语义理解的准确性和鲁棒性仍有待提高。

近年来,诸多学者集中于基于单目视觉的三维语义场景补全(semantic scene completion, SSC)研究。Xiao 等[2]提出了一种实例感知的单目语义场景补全框架,该框架利用两阶段区域 VQ-VAE 网络和可变形注意力机制优化体素查询,在 Semantic KITTI 数据集上取得了显著性能提升。Liu 等[3]设计语义引导的 SG-SSC 框架,引入语义融合模块,旨在补偿深度缺失问题。语义分割[4-5]在三维语义场景补全中不仅是基础的步骤,更是提升补全效果和质量的关键。Zhang 等[6]通过三分支非对称网络优化特征提取,Yu 等[7]结合上下文感知查询生成器和三维几何感知模块,显著提升语义分割精度。Luo 等[8]通过解耦网络结构与分辨率的直接关联,利用分辨率自适应特征和动态点采样策略,实现多分辨率场景的几何与语义重建。模型轻量化为实时应用和资源受限环境提供了高效且可扩展的解决方案。樊博等[9]提出融合注意力机制与重影特征映射的轻量级算法进行无人机视角下交通场景语义分割。侯志强等[10]提出一种基于跨层次聚合网络的实时城市街景语义分割算法,通过多尺度上下文信息提取和特征复用,提升分割准确性与实时性。开志强等[11]提出全局语义学习与显著目标感知框架,通过门控语义注意力机制和一致性损失约束,实现复杂遮挡场景下关键要素的纹理重建与语义恢复。靖永志等[12]通过轻量化的列向位置分类架构和多维度抗干扰机制,在复杂遮挡场景中实现毫米级要素间隙的鲁棒感知,为信息补全提供了高精度实时语义重建。基准数据集方面,Karangwa 等[13]系统评述了自动驾驶领域常用的 KITTI、nuScene、Waymo 等公开数据集,为复杂环境下的多任务感知研究提供了数据支撑。

自动驾驶环境感知不仅是对目标的精准检测,还应包含对复杂交通场景的理解及自然语言描述[14]。李国燕等[15]通过引入分割机制和空间关系机制,根据语义分割以及地理空间关系生成遥感图像的信息描述。Khurram 等[16]提出模块化深度学习的网络架构,生成复杂场景图像的自然语言描述。Pang 等[17]提出结合自然语言处理的图像理解和信息提取算法,通过深度学习实现图像和文本信息的深度融合,提高了图像描述生成准确性和效率。上述方法对于复杂驾驶场景的理解仍停留在目标级数据信息层面,缺乏语义级理解及自然语言描述,不能精确提取场景中的基建与参与者的几何特征以及语义信息。

针对复杂驾驶场景中缺乏语义级理解及自然语言描述的问题,该研究提出一种基于要素信息补全的场景语义理解模型,旨在对自动驾驶中复杂交通场景的几何特征与语义信息进行精准理解和自然语言描述。首先,运用稠密连接网络(densely connected networks, DenseNet)提取视觉图像的多尺度二维特征,通过特征视线投影模块(features line of sight projection, FLoSP)将体素逆向映射回三维空间转化为三维特征(three-dimensional features, F_{3D}),缓解传统方法中的投影模糊与尺度失配问题;接着,采用维度分解残差(dimension decomposition residual, DDR)模块提取 F_{3D} 中不同尺度的特征,构建 3D UNet 进行三维特征提取,并在其编解码器之间引入三维上下文先验层(three-dimensional context relation prior, 3D CRP)以增强模型的空间语义理解能力;然后,通过空洞空间金字塔池化(atrous spatial pyramid pooling, ASPP)与 Softmax 层输出场景语义补全结果;同时,基于改进的 VGG-16 网络提取语义补全后的场景要素特征,并通过长短时记忆网络(long short-term memory, LSTM)将语义补全结果转化为结构化自然语言描述,完成从几何感知到语义解释的端到端映射,实现对自动驾驶复杂场景的自然语言描述及决策规划建议。

本研究的贡献为:

1)设计了基于 FLoSP 和改进 3D UNet 网络的场景补全语义理解模型,实现了单帧视觉图像二维特征向三维

特征的转换,并通过在 3D UNet 编码器与解码器之间引入 3D CRP,以增强语义补全模型的空间语义理解能力,提升了实时性与鲁棒性。

2)构建了基于改进 VGG-16 的场景语义描述生成模型,并通过 LSTM 神经网络生成自然语言描述,实现对自动驾驶复杂场景的自然语言描述语句,为智能车辆驾驶辅助及无人驾驶决策规划提供了理论基础。

1　场景要素语义补全

场景要素语义补全在场景语义理解中包括目标几何级信息补全和语义级信息补全 2 个子任务[18]。无论是几何补全还是语义补全,均需对三维空间进行有效表征描述。体素作为基本单元,用于构建三维场景,通过二分类(空/实体)进行几何补全,随之进行语义分割[19],最终实现场景的语义理解。

为了有效表征复杂的驾驶场景,采用 DenseNet 网络提取图像特征,并通过 FLoSP 将体素逆向映射为三维特征,从而扩展场景特征维度。

1.1　场景二维特征提取

场景的二维特征提取是场景要素语义补全的重要步骤,场景要素语义补全模型采用 DenseNet 网络对 RGB 图像数据进行特征提取,并在 3 个维度上应用合理的模型缩放策略,图像分辨率设定为 600×600,通道维度倍率因子为 2,深度维度倍率因子为 3.1。DenseNet 的基线网络相较于其他扩展卷积神经网络,其密集连接机制通过跨层特征复用,可增强多尺度特征的连续性,缓解梯度消失问题,捕捉行驶场景中的细微变化。

DenseNet 基线网络由卷积层、池化层、4 个密集卷积模块、3 个过渡层和分类层构成,旨在从输入图像中提取多尺度特征。卷积层负责特征深度提取,而池化层则降低特征图分辨率以减轻计算负担。4 个密集卷积模块采用统一大小的卷积核,以提取不同尺度的特征,其间由 3 个过渡层通过 1×1 卷积和平均池化调整特征图大小,以保证计算效率。最终,分类层将特征映射至输出类别。表 1 列出了网络结构参数的设计。

在处理二维驾驶场景时,网络首先通过 7×7 卷积核提取特征,接着进行 3×3 最大池化以缩小特征图。随后,多个 MobileNetV2[20] 卷积模块(MobileNetV2 convolution, MBConv)被串联,其中包括 1×1 和 3×3 卷积核的组合,分别有 6、12、32 个模块。与原始 DenseNet 的 Bottleneck 层相比,MBConv 的线性瓶颈设计可避免 ReLU(rectified linear unit, ReLU)激活函数对低维特征的破坏,提升小目标检测鲁棒性。之后,1×1 卷积核进一步处理特征图,紧接着为 2×2 平均池化和 7×7 全局平均池化,最终通过全连接层输出车辆、行人、建筑物等目标的二维特征矩阵。

表 1　DenseNet 基线网络结构

Table 1　DenseNet baseline network architecture

网络层	输出维度	模型结构
卷积层	112×112	7×7 卷积核,步长 2
池化层	56×56	3×3 最大池化,步长 2
密集卷积模块(1)	56×56	$\begin{bmatrix}1\times1\text{ 卷积核}\\3\times3\text{ 卷积核}\end{bmatrix}\times6$
过渡层(1)	56×56 28×28	1×1 卷积核 2×2 平均池化,步长 2
密集卷积模块(2)	28×28	$\begin{bmatrix}1\times1\text{ 卷积核}\\3\times3\text{ 卷积核}\end{bmatrix}\times12$
过渡层(2)	28×28 14×14	1×1 卷积核 2×2 平均池化,步长 2
密集卷积模块(3)	14×14	$\begin{bmatrix}1\times1\text{ 卷积核}\\3\times3\text{ 卷积核}\end{bmatrix}\times32$
过渡层(3)	14×14 7×7	1×1 卷积核 2×2 平均池化,步长 2
密集卷积模块(4)	7×7	$\begin{bmatrix}1\times1\text{ 卷积核}\\3\times3\text{ 卷积核}\end{bmatrix}\times32$
分类层	1×1	7×7 全局平均池化 10 分类,全连接,Softmax

1.2　场景要素特征投影

为准确获取场景要素的三维语义信息,需将二维多尺度特征转化为三维空间特征。由于单目二维特征的尺度限制,直接运用体素定义存在一定歧义。因此,采用 FLoSP 逆向投影策略[18],将二维多尺度特征转换为三维特征,增强二维图像与三维体素网格之间的特征关系。FLoSP 结构如图 1 所示。

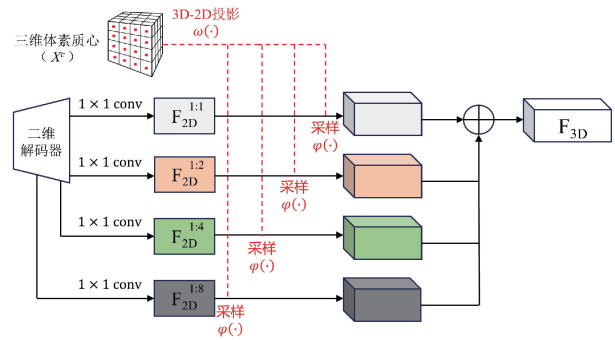
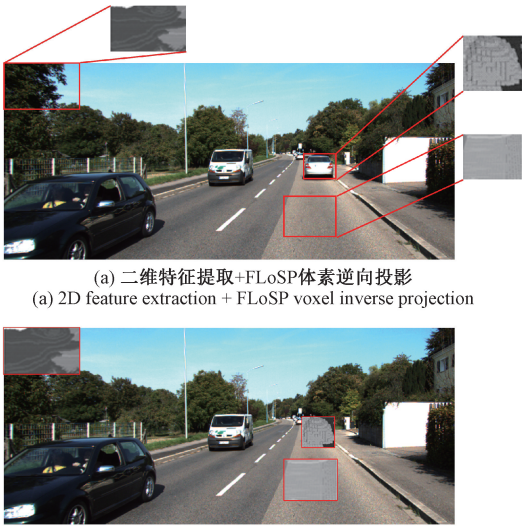


图 1　FLoSP 网络结构

Fig. 1　FLoSP network architecture

在 FLoSP 网络结构中,DenseNet 网络的多尺度输出特征 $F_{2D}^{1:s}(s\in\{1,2,4,8\})$ 经过 1×1 卷积处理后,被逆向映射回三维空间。具体而言,这一过程涉及将类别为 c 的

三维体素质心(x^c)投影到对应的二维特征图上,并对其进行填充。在投影过程中,任何落在图像边界之外的三维体素特征向量将被设置为 0,以避免边界效应对最终输出造成干扰。最终,通过对所有二维多尺度特征映射得到的三维体素特征进行逐元素加和,获得 F_{3D}。基于 FLoSP 的体素逆向投影过程如图 2 所示。



(a) 二维特征提取+FLoSP体素逆向投影
(a) 2D feature extraction + FLoSP voxel inverse projection

(b) 体素占据
(b) Voxel occupancy

图 2 FLoSP 体素逆向投影过程

Fig. 2 Reverse projection process of FLoSP voxels

在图 2(a)中,DenseNet 网络提取了二维特征,并通过 FLoSP 投影网络将体素映射至相应特征,生成了图 2(b)所示的体素占据状态。在此状态下,内部体素具有非零特征向量,代表其所包含的特征信息,但尚未包含具体的语义类别信息;而外部体素的特征向量则为 0,以标识其不包含有效特征。

DenseNet 网络提取图像中的多尺度像素特征,并将其与体素质心相对应。FLoSP 网络则将这些二维像素信息转化为三维特征,进而为后续 3D UNet 网络结构提供了补全几何形状和语义信息的能力。

1.3 场景要素目标语义补全

1) 三维特征提取

为提取包含二维像素信息与三维体素信息的 F_{3D},设计了 3D UNet 对特征进行编解码,以输出准确场景的空间语义特征。该 3D UNet 包括编码器、三维上下文先验层、解码器及 ASPP。编码器的基本构件采用维度分解残差模块,以提取 F_{3D} 中不同尺度的特征。DDR 表达式如式(1)所示。

$$x_t = F^d(x_{t-1}, \{W_i\}) + x_{t-1} \tag{1}$$

式中:F^d 为学习的残差映射,并具有 d 维的空洞,缓解了梯度消失。

DDR 的结构如图 3 所示,设置输入 x_{t-1} 通道数为 c,运用残差连接方法最终提取 F_{3D} 输出 x_t,激活函数运用 relu 函数。

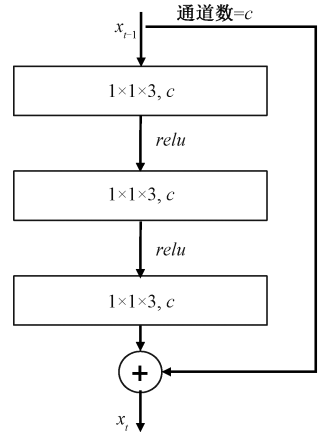


图 3 维度分解残差模块结构

Fig. 3 Dimension decomposition residual module architecture

将不同尺度的 DDR 模块组合,构建 3D UNet 编码结构,以提取驾驶场景的三维特征。F_{3D} 特征通过 DenseNet 网络的二维解码器与 FLoSP 层连接得到。针对公路场景多尺度目标并存的特性,设计渐进式 8 级降尺度采样 DDR 模块,以提取 F_{3D} 的多级特征,并通过两个 3×3 卷积核、膨胀率为 2 的逆向卷积层构成升尺度解码器,以调整特征大小。最后,ASPP 作为 Head 层,收集多尺度特征以输出场景体素的语义信息。ASPP 通过 1×1 卷积、不同膨胀率的卷积和最大池化层处理深层特征,进而堆叠这些特征以构建丰富的空间信息。3D UNet 结构如图 4 所示。

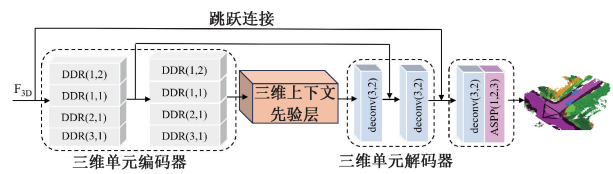


图 4 3D UNet 网络结构

Fig. 4 3D UNet network architecture

由于 F_{3D} 是将体素质心映射到二维特征上的融合结果,直接解码可能无法充分表达复杂场景的真实语义。因此,为增强模型的空间语义理解能力,在 3D UNet 的编码器与解码器之间引入了三维上下文先验层,以提升模型的空间语义表征能力。其结构如图 5 所示。

以三维张量($\boldsymbol{L},\boldsymbol{W},\boldsymbol{D}$)为输入,首先利用 3D ASPP 卷积扩大网络感受野,然后利用 1×1 卷积和 sigmoid 激活函数将体素图像生成大小为 ($\boldsymbol{L},\boldsymbol{W},\boldsymbol{D}$) × ($\boldsymbol{L},\boldsymbol{W},\boldsymbol{D}$)/$S^3$ 的 M 组关联矩阵 $\hat{\boldsymbol{A}}'_m$, $s \in \{1,2,4,8\}$, $m \in M$,并运用加权

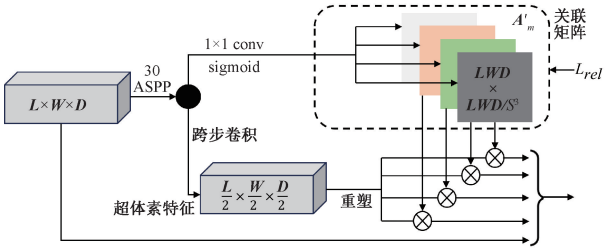


图 5　三维上下文先验层结构

Fig. 5　3D contextual prior layer architecture

的多标签二值分类交叉熵损失函数 L_{rel} 在真实关联矩阵 $\boldsymbol{A}'_m$ 监督下进行训练。

$$\begin{cases} L_{rel} = -\sum_{m\in M,i}[(1-\boldsymbol{A}^i_m)\log(1-\hat{\boldsymbol{A}}^i_m) + \\ \quad w_m\boldsymbol{A}^i_m\log\hat{\boldsymbol{A}}^i_m] \\ w_m = \dfrac{\sum_i(1-\boldsymbol{A}^i_m)}{\sum_i\boldsymbol{A}^i_m} \end{cases} \tag{2}$$

式中:w_m 为预测值与真实值之间的权重;i 为关联矩阵中对所有元素的索引。最后将关联矩阵与重塑后的超体素特征相乘输出全局特征。

2) 损失函数设计

设计多种损失函数,通过迭代优化以提升场景语义补全模型的性能。

(1)场景类别关联损失

模型引入场景类别关联损失 L_s 增强网络对全局场景几何与语义信息的理解。对于三维体素 x^c,其类别标记为 c,计算模型预测的准确率 P_c、召回率 R_c 和特异性 S_c。P_c 和 R_c 有助于区分同类体素,而 S_c 则用于识别非 c 类别的体素。具体的 P_c、R_c、S_c 计算方法为:

$$\begin{cases} P_c(\hat{p},p) = \log\dfrac{\sum_i\hat{p}_{i,c}[[p_i=c]]}{\sum_i\hat{p}_{i,c}} \\ R_c(\hat{p},p) = \log\dfrac{\sum_i\hat{p}_{i,c}[[p_i=c]]}{\sum_i[[p_i=c]]} \\ S_c(\hat{p},p) = \log\dfrac{\sum_i(1-\hat{p}_{i,c})(1-[[p_i=c]])}{\sum_i(1-[[p_i=c]])} \end{cases} \tag{3}$$

式中:定义三维体素 x^c 类别为 c,p 为真实值;$\hat{P}$ 为 p 的预测概率;$[[\cdot]]$ 定义为若括号内的条件满足则为赋值为 1,否则为 0;P_i 是索引为 i 的体素对应的真实值;$\hat{P}_{i,c}$ 代表将第 i 个体素预测为 c 类别的概率。为增加损失函数在计算体素几何损失和语义损失时的通用性,通过最大化上述指标,定义场景类别关联损失 L_s 如式(4)所示。

$$L_s(\hat{p},p) = -\frac{1}{C}\sum_{c=1}^{C}(P_c(\hat{p},p) + R_c(\hat{p},p) + S_c(\hat{p},p)) \tag{4}$$

因此,场景类别关联损失中的语义损失 L_s^{sem} 如式(5)所示。

$$L_s^{sem} = L_s(\hat{y},y) \tag{5}$$

几何损失 L_s^{geo} 如式(6)所示。

$$L_s^{geo} = L_s(\hat{y}^{geo},y^{geo}) \tag{6}$$

式中:$\hat{y}$ 和 $\hat{y}^{geo}$ 分别为模型所预测的体素语义标签和几何标签,将这些预测值纳入场景类别关联损失中,可以计算出不同任务下预测与实际值之间的损失,从而提高模型在场景语义补全任务中的理解能力。

(2)截体比例损失

图像遮挡易造成被遮体素预测偏差,常误归为目标形状。为克服此难题,模型增加截体比例损失,以优化截体类别分布。如图 6 所示,将图像划分为 1×1 网格块,每个网格块对应一个截体,并将 L_f 损失应用于每一个局部截体。

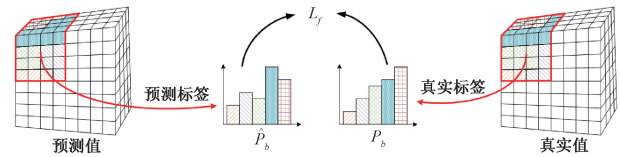


图 6　L_f 截体比例损失

Fig. 6　Loss of proportion of L_f truncates

通过计算预测值与真实值的 KL(Kullback-Leibler)散度来衡量信息损失,加强前后语义一致性,从而消除遮挡带来的语义歧义。L_f 计算如式(7)所示。

$$L_f = \sum_{b=1}^{l^2}D(P_b\|\hat{P}_b) = \sum_{b=1}^{l^2}\sum_{c\in C_b}P_b(c)\log\frac{P_b(c)}{\hat{P}_b(c)} \tag{7}$$

式中:P_b 为截体 b 中体素的真实类别分布;C_b 为截体 b 中的体素总类别;c 为一种体素类别,由于截体包括一些没有类别的小场景,无法定义 KL 散度,因此在计算 L_f 时计算了 C_b 中的 KL 散度,使得存在于截体 b 中的真实类别 C_b 能定义未知类别的小场景,因此取 $c\in C_b$;$P_b(c)$ 为 c 在截体 b 中所占的比例的真实值。在截体 b 中,P_b 与 $P_b(c)$ 分别为预测的体素类别分布与预测的 c 在 b 中的分布比例。然后,通过式(7)计算预测值与真实值之间的损失,逐步优化网络性能。

最终,在全局标准交叉熵损失函数 L_c 基础上,场景语义补全模型的损失函数对在三维上下文先验层的关系损失 L_{rel}、场景类别关联损失中的语义损失 L_s^{sem} 几何损失 L_s^{geo} 以及截体比例损失 L_f 进行求和,因此,综合损失函数 L_t 如式(8)所示。

$$L_t = L_c + L_{rel} + L_s^{sem} + L_s^{geo} + L_f \tag{8}$$

3) 实验验证

实验环境配置为:模型程序采用 Python 编写,并在 PyTorch 和 TensorFlow 1.4 环境中训练验证;硬件条件包

括 AMD Ryzen 7 5800X 8C 16 T CPU、Kingston 16GB 内存及 Leadtek Quadro RTX4000 16 G GPU。

使用 Semantic KITTI 数据集进行训练，KITTI 系列数据集进行测试，以验证模型的有效性。场景以 0.2 m 分辨率体素化，形成 256×256×32 的网格，标记 21 类目标，包括 19 类语义信息、1 类自由语义和 1 类未知语义。

训练时，将数据集划分为 3 834 帧训练集和 815 帧验证集。利用迁移学习优势，设置学习率为 0.000 1，*batch size*=4，*epoch*=30，采用 AdamW 优化器，截体比例损失的二维网格块大小设为 8。模型训练损失变化趋势如图 7 所示，补全模型根据全局综合损失 L_t，在预训练权重下端到端训练了 30 epochs，随着 epoch 增加，模型训练集及验证集的损失值均呈收敛趋势。

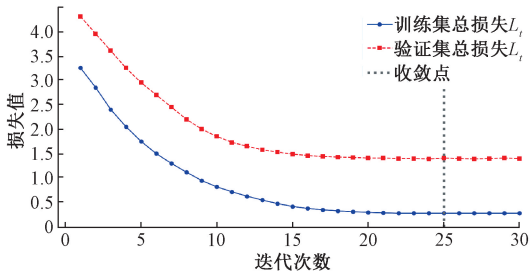


图 7 损失函数变化曲线

Fig. 7 Loss function change curves

当训练至第 25 epoch 时，验证集的损失值呈现收敛趋势，此时，训练集的损失值为 0.3 左右，验证集的损失值在 1.4 左右，表明模型在验证集上的预测结果能够有效接近真实值。

为验证场景语义理解模型在复杂场景中的理解能力，选择 Semantic KITTI 数据集中的相关场景进行实验。场景要素语义补全模型输出语义补全结果部分示例如图 8 所示。

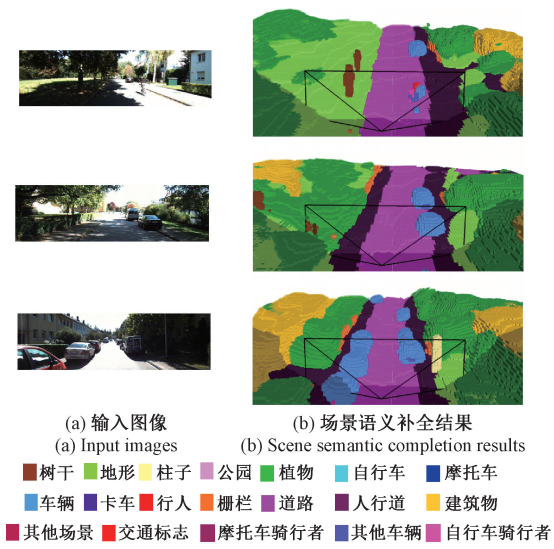


图 8 场景语义补全结果

Fig. 8 Scene semantic completion results

图 8(a)为数据集摄像头捕获的行驶场景，图 8(b)为语义补全模型对该场景进行的体素化语义补全结果。实验结果表明，通过模型训练，三维场景语义补全模型能够通过视觉二维图像输入，有效捕捉复杂场景中的可行驶道路、非行驶道路、路边植物、建筑物、车辆及道路标志等交通要素。同时，该模型能够有效预测视野以外的要素形状，补全结构化道路中交通要素的完整三维几何信息与语义信息。

2 场景语义理解描述

采用图像描述生成技术，以编码器-解码器架构融合计算机视觉与自然语言处理技术，将交通场景转化为易懂的语言描述。场景语义理解描述模型如图 9 所示，编码器采用改良 VGG-16 网络提取特征，解码器采用 LSTM 网络来生成自然语言描述和驾驶安全建议。

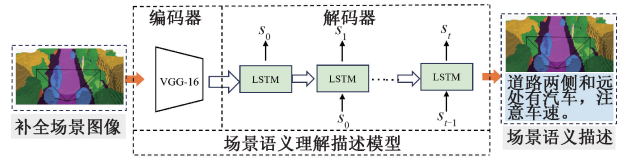


图 9 场景语义理解描述模型

Fig. 9 Scene semantic understanding description model

2.1 基于改进 VGG-16 的编码器

为详细描述交通场景图像特征，改进 VGG-16 网络结构，通过多层次特征融合与拼接，将深层次信息与低层次信息结合，有效解决因物体距离不同导致的尺度影响问题，分层呈现交通场景内容，包括汽车、自行车和信号灯等。改进后的 VGG-16 网络结构如图 10 所示，包含 16 个可训练层，其中 13 个卷积层，3 个全连接层，5 个最大池化层，使用 ReLU 作为激活函数。

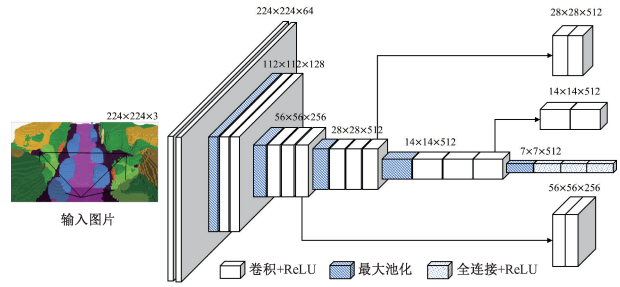


图 10 改进的 VGG-16 网络结构

Fig. 10 Diagram of the improved VGG-16 network architecture

经改进的 VGG-16 网络模型所提取的交通场景图像特征如式(9)所示。

$$\begin{cases} f_1 = F(f_{7\times7\times512}) \\ f_2 = F(f_{14\times14\times512}) \\ f_3 = F(f_{28\times28\times512}) \\ f_4 = F(f_{56\times56\times256}) \\ f_t = concat(f_1, f_2, f_3, f_4) \end{cases} \tag{9}$$

式中:F 为尺度归一化操作;f_t 表示交通场景特征;$f_{7\times7\times512}$、$f_{14\times14\times512}$、$f_{28\times28\times512}$、$f_{56\times56\times512}$ 分别表示卷积网络中尺寸分别为 7×7×512、14×14×512、28×28×512、56×56×256 的卷积层输出结果;$concat(\cdot)$表示特征拼接操作。

提出的编码器通过投影矩阵将输入图像的特征变换到解码器的输入形式,如式(10)所示。

$$\boldsymbol{v} = [\boldsymbol{VGG}_{\theta_v}(I)] \cdot \boldsymbol{W}_I + \boldsymbol{b}_I \tag{10}$$

式中:I 为输入图像;$\boldsymbol{VGG}_{\theta_v}(\boldsymbol{I})$ 为网络提取的图像特征向量;θ_v 为网络预训练参数;$\boldsymbol{W}_I$ 为投影矩阵;$\boldsymbol{b}_I$ 表示偏置向量。

2.2　基于 LSTM 的解码器

在解码器阶段采用长短时记忆网络替代传统的循环神经网络。基于 LSTM 的解码器结构如图 11 所示。

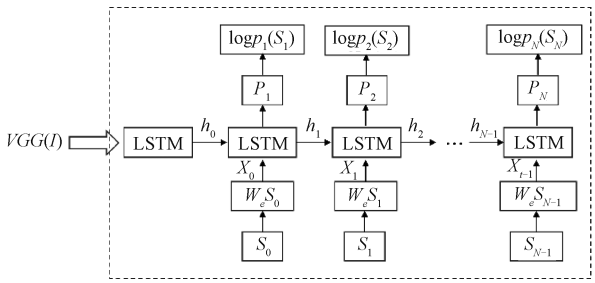


图 11　基于 LSTM 的解码器结构

Fig. 11　LSTM-based decoder architecture

针对给定的图像 I 和其对应的自然语言描述语句 $S=(S_0,S_1,\cdots,S_N)$,每一单元的输入可表示为单词向量和语句的嵌入矩阵,首先定义嵌入矩阵 x_t 为:

$$\boldsymbol{x}_t = \boldsymbol{W}_e\boldsymbol{S}_t,\ t \in \{0,\cdots,N-1\} \tag{11}$$

式中:$\boldsymbol{S}_t(t=0,1,\cdots,N-1)$表示第 t 个词的词向量。$\boldsymbol{S}_0$ 和 $\boldsymbol{S}_N$ 分别表示特殊的开始词和结束词,用于标识描述语句的开头和结尾;$\boldsymbol{W}_e$ 为词向量投影到嵌入空间的 $N_0\times h$ 维度的投影矩阵,N_0 为字典的尺寸;h 为嵌入空间的维度。

为了强化图像和语言之间的关联,语义理解模型将生成描述语句的问题建模为条件概率问题,即在给定图像 I 的条件下生成语句 S 的条件概率 $P(S\mid I)$。通过链式法则计算该条件概率的方法如式(12)所示。

$$P(S\mid I;\theta) = \prod_{t=0}^{N} P(\boldsymbol{S}_t \mid I, \boldsymbol{S}_0, \cdots, \boldsymbol{S}_{t-1};\theta) \tag{12}$$

式中:θ 为场景语义理解描述模型中需要训练的所有参数,包括 $\boldsymbol{W}_I$,$\boldsymbol{W}_e$,$\boldsymbol{b}_I$ 以及 LSTM 单元中的所有权重参数和偏置,模型通过训练 θ 来使 $P(S\mid I)$ 达到最大化。$P(I,S_0,\cdots,S_{T-1})$由 LSTM 建模。在 $t-1$ 时刻的 LSTM 单元的输出 h_t-1 传输给 t 时刻的 LSTM 单元,整体建模方法可表示为式(13)所示。

$$\begin{cases} \boldsymbol{x}_{-1} = \boldsymbol{v} \\ h_t = LSTM(\boldsymbol{x}_t, h_{t-1}), & t \in \{0,1,\cdots,N-1\} \\ \boldsymbol{y}_{t+1} = h_t \cdot \boldsymbol{W}_d + \boldsymbol{b}_d, & t \in \{0,1,\cdots,N-1\} \\ P_{t+1} = soft\max(y_{t+1}), & t \in \{0,\cdots,N-1\} \end{cases} \tag{13}$$

式中:表明 LSTM 中每个单元的隐层状态 h_t 被输入到 Softmax 功能块,其输出值代表每个单词的概率,从而形成单词的概率分布。图像特征通过投影矩阵生成语义生成网络的输入,即 $\boldsymbol{x}_{-1}$,后续每个 LSTM 单元的输入如式(11)所示。

在此过程中,损失函数定义如式(14)所示。

$$L_S(S,I;\theta) = -\sum^{N} \log p_t + \lambda_\theta \cdot \|\theta\|_2^2 \tag{14}$$

式中:$\lambda_\theta \cdot \|\theta\|_2^2$ 表示对参数的正则约束。

2.3　测试分析

为增强模型的泛化能力,通过多样化的数据集进行训练。基于 LaRA 数据集构建交通场景语义理解描述的数据集,标注样本包括 10 258 幅图像,若干样本数据如表 2 所示。

表 2　LaRA 数据集若干样本数据

Table 2　Several sample data of LaRA dataset

行驶图像	中文描述	行驶图像	中文描述
	道路前方有汽车,右侧远处有摩托车,注意减速。		道路两侧有汽车,道路前方有行人和公共汽车,注意车速。
	道路两侧有汽车和摩托车,注意车速。		道路两侧有行人、汽车和摩托车,前方有斑马线,注意行人。
	道路前方有斑马线,两侧有汽车,注意避让。		道路远处有行人、汽车和红色信号灯,请停车等待。

为确保评估的有效性,随机选取 1/10 作为测试集。语义理解模型通过预训练的 VGG-16 网络作为特征提取基础,并结合 LSTM 进行描述生成。为优化模型性能,采用自适应学习率策略结合随机梯度下降,初始学习率设置为 0.005,模型配置 512 维 LSTM 存储单元和 mini-batch 大小为 16,以提升模型对场景语义的把握及对图像内容的理解。

3 实验结果与分析

3.1 补全模型对比分析

为评估场景要素补全模型的性能,进行量化对比分析,以展示基于要素信息补全的自动驾驶复杂场景语义理解模型在几何与语义理解方面的优势。表 3 列出了所有语义类别的平均交并比(mean intersection over union, mIoU),以衡量模型的有效性。

表 3 不同算法在 KITTI 验证集中的性能表现

Table 3 Performance of different algorithms on KITTI validation set (%)

类别	JS3C-Net	NDC-Scene	SurroundOcc	PanoSSC	本文方法
道路	50.69	**59.20**	56.90	56.36	57.07
人行道	23.67	28.24	**28.30**	26.40	24.19
公园	11.82	21.42	**30.20**	17.76	21.53
其他场景	0.08	1.67	**6.80**	0.88	1.76
建筑物	15.17	14.94	15.20	14.26	**15.29**
车辆	25.31	26.26	20.60	19.63	**26.36**
卡车	4.15	14.75	1.40	14.79	**14.86**
自行车	0.27	1.67	1.60	0.63	**2.73**
摩托车	0.00	2.37	1.20	0.36	**2.46**
其他车辆	5.86	7.73	4.40	6.22	**7.87**
植物	18.02	**19.09**	14.90	16.69	18.11
树干	**4.53**	3.51	3.40	1.83	3.58
地形	26.89	**31.04**	19.30	28.05	28.16
行人	0.67	**3.60**	1.40	0.87	1.58
自行车骑行者	1.47	2.74	2.00	0.00	**3.35**
摩托车骑行者	1.47	0.00	0.10	0.00	**1.58**
栅栏	3.97	6.65	**11.30**	5.72	6.98
柱子	3.67	**4.53**	3.90	1.94	4.11
交通标志	1.47	**2.73**	2.40	0.70	2.45
平均交并比	11.54	12.70	11.86	11.22	**12.84**

注:表中加粗表示同一类别物体语义补全性能表现最优结果,加下划线表示次优结果。

结果表明,相较于以图像为输入的其他算法(如 JS3C-Net[21]、NDC-Scene[22]、SurroundOcc[23] 和 PanoSSC[24]),该研究的模型在 Semantic KITTI 数据集上性能更优。具体而言,该方法相较于 JS3C-Net 平均交并比相对提升了 11.27%,明显优于 JS3C-Net、SurroundOcc 和 PanoSSC,且略高于 NDC-Scene,这反映出模型在几何特征捕捉上的进步。在建筑物、车辆、自行车、摩托车和骑行者等类别中,该模型表现尤为突出,显著提升了交通场景的语义理解能力。

3.2 消融研究

为评估场景语义理解模型各组件的可行性和有效性,对模型关键组件在 Semantic KITTI 数据集上进行了消融研究。为确保实验公平性,所有实验数据及参数设置均保持一致,表 4 列出了该模型在 KITTI 验证集中的消融研究结果。

表 4 场景语义理解模型在 KITTI 验证集中的消融对比

Table 4 Ablation comparison of scene semantic understanding models on KITTI validation set

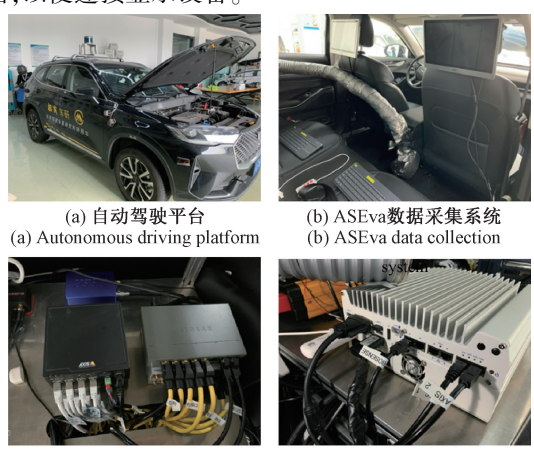
序号	FLoSP	3D UNet	L_s^{sem}	L_s^{geo}	平均交并比/%
1	√	√	√	√	12.84
2	×	√	√	√	5.96
3	√	×	√	√	11.75
4	√	√	×	√	10.43
5	√	√	√	×	12.54

结果显示,引入 FLoSP 模块后,通过二维图像与三维体素网格之间的映射,增加了多尺度特征信息,解决了网络初始输入特征几何信息不足的问题,mIoU 整体提升了 6.88%,这一改进显著提高了模型的场景语义理解精度。

与不引入 3D UNet 的对照网络相比,场景语义理解模型在 mIoU 上整体提升了 1.09%。消融实验证明,引入 3D UNet 网络对补全场景的空间语义特征的重要性。另外,两种损失函数的贡献与预期一致,分别实现了 2.41% 和 0.30% 的 mIoU 提升,从而进一步优化了模型性能。

3.3 实车实验设计

采用实车进行多样化且复杂的驾驶场景数据采集,以评估模型在实际场景中的表现。自动驾驶采集平台的整体配置如图 12 所示,自动驾驶数据采集平台以哈弗 H6 为主体,配备了摄像头等传感器,并将工控机置于后备箱,以便连接显示设备。



(a) 自动驾驶平台 (a) Autonomous driving platform

(b) ASEva数据采集系统 (b) ASEva data collection

(c) 驾驶数据采集设备 (c) Driving data collection device

图 12 自动驾驶采集平台

Fig. 12 Autonomous driving acquisition platform

在图 12 的实验中,车辆装备了 ASEva 数据采集系统、GPS 组合惯导、MIC-7700 工控机、AXIS BOX 及 AXIS 高清摄像头等设备。ASEva 系统负责在线标定传感器并输出结构化数据,Wire Shark 则用于抓取工控机数据。通过视频采帧技术,将视频流分解为 11 452 帧图像,依照 KITTI 配置文件处理实际场景数据。

图 13 展示了模型对采集路段的语义理解成果。模型将采集的实车场景数据作为输入,在多样化的行驶场景中精确捕捉布局,如十字路口和狭窄道路,并合理预测摄像头视野外的道路语义。尽管在交通密集场景中存在一定的语义连滞性,但实验在现有硬件条件下仍实现超过 10 Hz 的理解速率。

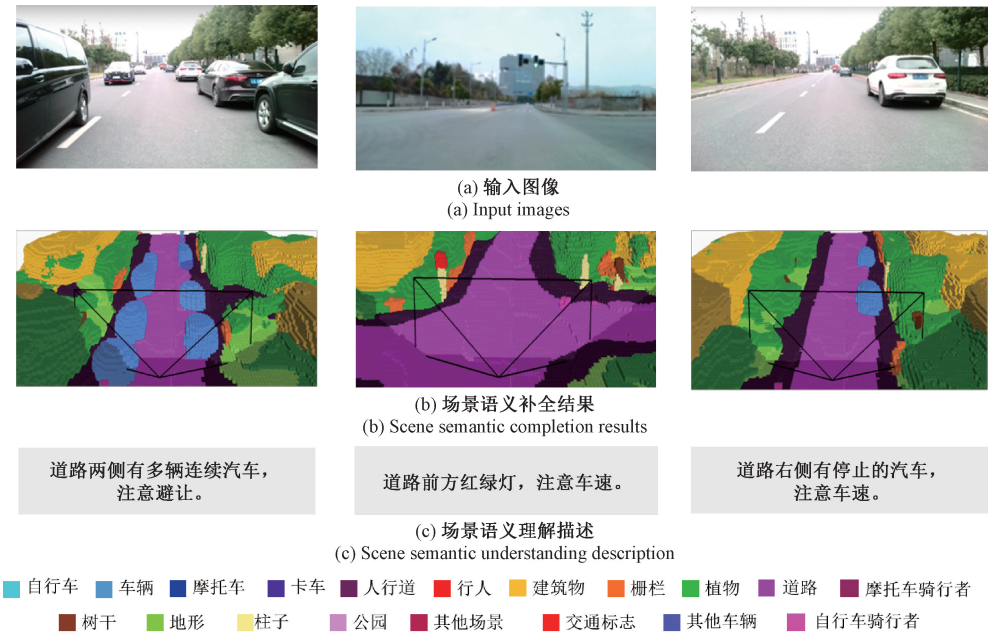


图 13　场景语义理解模型实车验证结果

Fig. 13　Real vehicle validation results of the scene semantic understanding model

实车实验结果验证了该研究在复杂场景语义理解上的有效性,突显了模型在补充交通场景缺失信息、感知丰富语义及提供驾驶建议方面的能力。

4　结　　论

1)本研究设计了基于特征视线投影模块 FLoSP 和改进 3D UNet 网络的场景补全语义理解模型及基于改进 VGG-16 的场景语义描述生成模型,实现了基于单帧视觉 RGB 图像的复杂交通场景语义补全和自然语言描述,为智能车辆驾驶辅助及无人驾驶决策规划提供了理论基础。

2)通过 KITTI 数据集与实车实验,对模型进行实验验证,结果表明,基于要素信息补全的自动驾驶复杂场景语义理解模型相较于 JS3C-Net 的 mIoU 相对提升了 11.27%。该模型能有效补全场景空间语义特征,对复杂交通场景理解具有良好的精确度和鲁棒性。

参考文献

[1] 王若萱,吴建平,徐辉. 自动驾驶汽车感知系统仿真的研究及应用综述[J]. 系统仿真学报,2022,34(12):2507-2521.
WANG R X, WU J P, XU H. Overview of research and application on autonomous vehicle oriented perception system simulation[J]. Journal of System Simulation, 2022,34(12):2507-2521.

[2] XIAO H H, XU H B, KANG W X, et al. Instance-aware monocular 3d semantic scene completion[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(7):6543-6554.

[3] LIU X ZH, XIE H ZH, ZHANG SH P, et al. 2D semantic-guided semantic scene completion[J]. International Journal of Computer Vision, 2024, 133(3):1306-1325.

[4] 徐晓龙,俞晓春,何晓佳, 等. 基于改进 U-Net 的街景图像语义分割方法[J]. 电子测量技术,2023,46(9):117-123.
XU X L, YU X CH, HE X J, et al. Semantic segmentation method of street view image based on improved U-Net[J]. Electronic Measurement Technology, 2023,46(9):117-123.

[5] 李利荣,丁江,梅冰, 等. 基于像素注意力特征融合的

城市街景语义分割算法研究[J]. 电子测量技术, 2023,46(20):184-190.
LI L R, DING J, MEI B, et al. Semantic segmentation method for urban street scenes based on pixel attention feature fusion[J]. Electronic Measurement Technology, 2023,46(20):184-190.

[6] ZHANG Y ZH, ZHANG X G, YU H. Triple-branch asymmetric network for real-time semantic segmentation of road scenes[J]. Instrumentation, 2024, 11(2):72-82.

[7] YU ZH, ZHANG R M, YING J CH, et al. Context and geometry aware voxel transformer for semantic scene completion[J]. ArXiv preprint arXiv: 2405.13675, 2024.

[8] LUO SH T, SUN ZH X, SUN Y H, et al. Resolution-switchable 3D semantic scene completion[J]. Computer Graphics Forum, 2022, 41(7):121-130.

[9] 樊博,高玮玮,单明陶, 等. 融合注意力机制与重影特征映射的无人机交通场景目标轻量级语义分割[J]. 电子测量与仪器学报,2023,37(3):21-28.
FAN B, GAO W W, SHAN M T, et al. Lightweight semantic segmentation of UAV traffic scene objects combining attention mechanism and ghost feature mapping[J]. Journal of Electronic Measurement and Instrumentation, 2023,37(3):21-28.

[10] 侯志强,程敏婕,马素刚, 等. 基于跨层次聚合网络的实时城市街景语义分割[J]. 光学 精密工程,2024, 32(8):1212-1226.
HOU ZH Q, CHENG M J, MA S G, et al. Real-time urban street view semantic segmentation based on cross-layer aggregation network[J]. Optics and Precision Engineering, 2024,32(8):1212-1226.

[11] 开志强,苗锡奎,马天磊, 等. 基于全局语义学习和显著目标感知的激光干扰图像修复[J]. 仪器仪表学报,2024,45(7):38-51.
KAI ZH Q, MIAO X K, MA T L, et al. Laser jamming image inpainting based on global semantic learning and salient target awareness[J]. Chinese Journal of Scientific Instrument, 2024,45(7):38-51.

[12] 靖永志,倪胜,贾兴科, 等. 基于列向语义分割的悬浮间隙视觉检测方法研究[J]. 仪器仪表学报,2024, 45(9):67-76.
JING Y ZH, NI SH, JIA X K, et al. Research on the visual detection method of levitation gap based on column-oriented semantic segmentation[J]. Chinese Journal of Scientific Instrument, 2024,45(9):67-76.

[13] KARANGWA J, LIU J, ZENG Z X. Vehicle detection for autonomous driving: A review of algorithms and datasets[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(11):11568-11594.

[14] YANG CH, ZHUANG K, CHEN M L, et al. Traffic sign interpretation via natural language description[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(11):18939-18953.

[15] 李国燕,田明达,董春华, 等. 面向遥感图像的结构化图像描述网络[J]. 电子测量与仪器学报,2024, 38(5):148-157.
LI G Y, TIAN M D, DONG CH H, et al. Structured image description network for remote sensing images[J]. Journal of Electronic Measurement and Instrumentation, 2024,38(5):148-157.

[16] KHURRAM I, FRAZ M M, SHAHZAD M, et al. Dense-CaptionNet: A sentence generation architecture for fine-grained description of image semantics[J]. Cognitive Computation, 2021, 13(3):595-611.

[17] PANG L, LI AI H. Design of an image content understanding and information extraction algorithm integrating natural language processing[J]. Traitement du Signal, 2024, 41(6):2839-2850.

[18] CAO A Q, DE CHARETTE R. Monoscene: Monocular 3D semantic scene completion[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022:3981-3991.

[19] 周勇,刘泓滨,侯亚东. 复杂城市交通场景下的自动驾驶语义分割方法[J]. 电子测量与仪器学报,2024, 38(4):241-247.
ZHOU Y, LIU H B, HOU Y D. Automatic driving semantic segmentation method for complex urban traffic scene[J]. Journal of Electronic Measurement and Instrumentation, 2024,38(4):241-247.

[20] WANG Y Q, QIN G H, ZOU M, et al. A lightweight intrusion detection system for internet of vehicles based on transfer learning and MobileNetV2 with hyper-parameter optimization[J]. Multimedia Tools and Applications, 2024, 83(8):22347-22369.

[21] YAN X, GAO J T, LI J, et al. Sparse single sweep LiDAR point cloud segmentation via learning contextual shape priors from scene completion[C]. 35th AAAI Conference on Artificial Intelligence, 2021, 35(4): 3101-3109.

[22] YAO J W, LI CH M, SUN K Q, et al. Ndc-scene:

Boost monocular 3D semantic scene completion in normalized device coordinates space[C]. 2023 IEEE/CVF International Conference on Computer Vision, 2023:9421-9431.

[23] WEI Y, ZHAO L Q, ZHENG W ZH, et al. Surroundocc: Multi-camera 3D occupancy prediction for autonomous driving[C]. 2023 IEEE/CVF International Conference on Computer Vision, 2023:21672-21683.

[24] SHI Y N, LI J S, JIANG K, et al. Panossc: Exploring monocular panoptic 3D scene reconstruction for autonomous driving[C]. 2024 International Conference on 3D Vision, 2024:1219-1228.

作者简介

赵树恩(通信作者),1997 年于长安大学获得学士学位,2005 年于重庆大学获得硕士学位,2010 年于重庆大学获得博士学位,现为重庆交通大学教授,主要研究方向为智能汽车与自动驾驶。

E-mail:zse0916@ 163. com

Zhao Shuen (Corresponding author) received his B. Sc. degree from Chang′an University in 1997, received his M. Sc. degree and Ph. D. degree both form Chongqing University in 2005 and 2010. He is currently a professor at Chongqing Jiaotong University. His main research interests include intelligent vehicles and autonomous driving.

袁亮,2023 年于长江师范学院获得学士学位,现为重庆交通大学硕士研究生,主要研究方向为图像处理与深度学习。

E-mail:2510157151@ qq. com

Yuan Liang received his B. Sc. degree from Yangtze Normal University in 2023. He is currently a master student at Chongqing Jiaotong University. His main research interests include image processing and deep learning.

赵东宇,2020 年于四川轻化工大学获得学士学位,2023 年于重庆交通大学获得硕士学位,现为四川大学博士研究生,主要研究方向为计算机视觉与图像处理。

E-mail:zdy19981636965@ 163. com

Zhao Dongyu received his B. Sc. degree from Sichuan University of Science and Engineering in 2020, and received his M. Sc. degree form Chongqing Jiaotong University in 2023. He is currently a Ph. D. candidate at Sichuan University. His main research interests include computer vision and image processing.