引文格式:韩汀,陈思宇,马津,等.可学习深度位置编码引导的车前图像道路可行驶区域检测[J].武汉大学学报(信息科学版),2024,49(4):582-594.DOI:10.13203/j.whugis20230252
Citation:HAN Ting,CHEN Siyu,MA Jin,et al.Road Image Free Space Detection via Learnable Deep Position Encoding[J].Geomatics and Information Science of Wuhan University,2024,49(4):582-594.DOI:10.13203/j.whugis20230252

可学习深度位置编码引导的车前图像道路可行驶区域检测

韩 汀[1] 陈思宇[2] 马 津[1] 蔡国榕[2] 张吴明[1] 陈一平[1]
1 中山大学测绘科学与技术学院,广东 珠海,519082
2 集美大学计算机工程学院,福建 厦门,361021

摘 要:道路可行驶区域检测是汽车辅助驾驶系统中场景感知的关键基础。基于卷积神经网络的方法因难以获取全局上下文信息而易产生道路空洞和中断等完整性问题,而基于Transformer的方法缺乏局部理解,容易造成边界的错位和越界问题。为了克服上述两类方法的缺陷,提出了一种可学习深度位置编码引导的金字塔Transformer网络架构,融合卷积神经网络与Transformer进行道路可行驶区域检测。该框架建立金字塔Transformer主干网从全局感受野提取道路特征,并结合局部窗口注意力弥补细节损失,以收缩自注意力提升特征计算效率。针对Transformer中传统位置编码忽略像素与实际场景空间关联性的问题,提出用深度图像卷积特征构建可学习位置编码的方法,解决现实关联性脱节引起的注意力偏移和语义不对齐问题。在KITTI道路、Cityscapes与自建厦门市道路数据集上对该方法进行了测试和评估,结果表明,该方法在保证较高效率的同时,具有较高的稳定性和精确性,其最大F值在KITTI和Cityscapes数据集上分别达到97.53%和98.54%,优于目前KITTI道路基准测试的所有方法。此方法可为汽车驾驶辅助系统的路径规划与轨迹预测等任务提供高精度的语义先验信息。

关键词:Transformer;位置编码;道路感知;可行驶区域检测;自动驾驶
中图分类号:P208 **文献标识码**:A **收稿日期**:2023-07-14
DOI:10.13203/j.whugis20230252 **文章编号**:1671-8860(2024)04-0582-13

Road Image Free Space Detection via Learnable Deep Position Encoding

HAN Ting[1] *CHEN Siyu*[2] *MA Jin*[1] *CAI Guorong*[2] *ZHANG Wuming*[1] *CHEN Yiping*[1]
1 School of Geospatial Engineering and Science, Sun Yat-Sen University, Zhuhai 519082, China
2 School of Computer Engineering, Jimei University, Xiamen 361021, China

Abstract: **Objectives**: The freespace detection is a crucial foundation for scene perception in advanced driver assistance systems. Convolutional neural network-based methods are unable to build global contextual infortmation that generate voids and interruptions in predicted results. At the same time, Transformer-based methods lack local understanding resulting in boundary misalignment and exceed. **Methods**: To this end, we propose a pyramid Transformer architecture with learnable deep position encoding for road freespace detection. First, the pyramid Transformer backbone is designed to extract road features from global perspectives. Second, local window attention is employed in dual-Transformer blocks to compensate for detail loss. Finally, to address the problem that traditional unlearnable position encoding ignores the spatial correlation between pixels and the real world, a learnable position encoding from deep convolutional features is constructed to solve the attention and semantic misalignment. **Results**: This model is tested and evaluated on KITTI road, Cityscapes, and Xiamen road datasets. The results show that our method achieves maximum F measure of 97.53% and 98.54% in KITTI and Cityscapes, respectively. **Conclusions**: Our method

基金项目:国家自然科学基金(42371343)。
第一作者:韩汀,博士生,主要从事图像和点云的语义分割理论与方法研究。ting.devin.han@gmail.com
通讯作者:陈一平,博士,副教授。chen79@mail.sysu.edu.cn

outperforms existing algorithms in the KITTI road benchmark by ensuring higher efficiency while providing higher stability and accuracy. Meanwhile, our method provides high-precision semantic prior information for tasks such as path planning and trajectory prediction in automotive driving assistance systems.

Key words: Transformer; position encoding; road perception; freespace detection; autonomous driving

近年来,随着自动驾驶和车辆辅助驾驶系统的普及,对复杂道路环境感知的需求愈加迫切[1-3]。现有方法一般基于可见光摄像头、深度相机、激光雷达等车载传感器获取道路场景数据,并在此基础上深入分析场景和数据特征,通过构建二元语义分割算法实现道路可行驶区域检测。可行驶区域检测结果可以提供道路场景的关键信息,改善自动驾驶中与其相互关联或依赖的任务,如路径规划[4]、轨迹预测[5]和道路跟踪[6]等。

早期的研究主要利用数字图像处理[7]、几何与拓扑学[8]、概率图模型[9-11]等机器学习方法[12]实施可行驶区域检测。深度学习被提出以来,基于卷积神经网络的方法(convolutional neural network,CNN)以强大的特征自学习能力为可行驶区域检测带来新的发展方向。例如文献[13-14]利用全卷积网络(fully convolutional network,FCN)、文献[15]利用UNet进行端到端的可行驶区域语义分割工作。然而,这些方法均被彩色图像输入本身的固有缺陷制约,当道路环境在光照、天气条件变化下引起明显的同质变异性问题(纹理变化、尺度变化、视角变化,透视变化等),或无视觉纹理线索的情况下会出现明显的道路空洞和边界错位等分割错误,无法适应多变的道路环境。

激光雷达三维点云具有强大的空间信息表征能力,可以提供对环境的精确三维测量[16],弥补了彩色图像受限于纹理线索的缺陷。文献[17-19]尝试在彩色图像中融合三维激光雷达点云的几何信息辅助道路分割。然而,由于激光雷达造价昂贵,不具备部署在车辆上的通用性,以及三维点云本身计算量大的问题,始终难以有效应用于需要高实时性和低成本普及性的汽车驾驶辅助系统。因此,文献[20]探索生成带有空间距离信息的视差图像辅助进行道路可行驶区域检测,利用可表征三维空间信息的图像数据代替点云,大幅提升了该任务的计算效率。文献[21-22]以三维点云生成高度差图像与彩色图像进行线性融合,文献[23-26]利用以表面法线信息辅助的深度图像[27]结合彩色图像进行可行驶区域检测。以彩色图像和表征三维的图像构造数据融合网络,结合二维视觉纹理信息与三维的空间结构信息,已经成为了可行驶区域检测的主流方法。

基于CNN的方法因卷积核的大小固定而将感受野限制在有限的范围内,导致模型只能依靠局部信息进行图像理解,从而影响编码器所提取特征的可区分性。在局部关系下进行密集的像素分割,难以适应同质像素周围变化强度大的场景,因此研究人员开始探索利用注意力机制和视觉Transformer建模像素的长距离依赖,从全局理解图像优化无人驾驶任务[28]。文献[29]构建了一种双分支网络,在其中的一条分支内加入全局注意力模块,开始尝试全局感受野下的语义分割。但是该方法利用全局池化构造全局上下文的策略,无法获取空间维度下的关键信息。文献[30]在网络中嵌入通道注意力,建立通道强化依赖,以学习更丰富的小目标语义类别特征。文献[31]提出了反向注意力和边界注意力网络,逐层进行双分支计算,对非道路区域至道路边界的距离进行约束和细化,强化边界分割效果。ViT是将Transformer用于视觉任务的开山之作,其主要思想是将输入图像划分为多个图像块序列,投影成固定长度的向量后计算多头注意力,在不改变图像分辨率的情况下进行全局建模。文献[32]提出了首个基于金字塔结构的视觉Transformer架构PVT,兼具CNN和Transformer的优点进行像素级语义分割。文献[33]提出的SegFormer在层次化Transformer的基础上,仅用全连接构建解码器用于道路区域语义分割。文献[34]提出Swin-Transformer构建层次结构,并引入局部窗口思想,对无重合的窗口区域进行注意力计算。尽管视觉Transformer具有全局归纳建模能力,但是CNN具有归纳偏置的特性优势,如果仅从全局角度理解道路场景,难以有效建模道路边缘和障碍物轮廓等精细结构。同时,视觉Transformer普遍基于固定、不可学习的相对/绝对位置编码建模像素关系,却忽略了像素与其本身所在实际空间的相关性,因此造成了注意力的偏移和语义不对齐等问题。

总体而言,CNN忽略了像素的长距离依赖,无法从全局理解图像;Transformer缺乏局部的归纳偏置,造成细节丢失,同时不可学习的位置编码易引起注意力与语义的偏移和不对齐问题。

在数据融合架构下,构造可学习位置编码,辅助Transformer对齐像素与实际空间场景,同时精细化细节分割是本文考虑的关键问题。针对当前道路可行驶区域分割任务中分割结果的空洞、中断问题,道路边界的模糊和越界问题,以及Transformer不可学习位置编码造成的注意力偏移问题,本文设计了可学习深度位置编码引导的金字塔Transformer网络。首先,通过Transformer全局上下文建模能力,网络从更大的感受野对道路区域进行感知,有效消除道路分割结果中的空洞问题;其次,双分支Transformer模块利用全局-局部窗口融合的方式加强对细节信息的把握,对边界、轮廓进行细化,同时将传统Transformer中的多头自注意力修改为多头收缩注意力,减少高分辨率特征图所引入的计算量;最后,本文设计用深度图像的卷积特征构造可学习位置编码,减少注意力偏移和语义不对齐问题,优化可行驶区域的分割。

通过上述方式,本文构建了高精度的道路可行驶区域二元语义分割算法,实现对交通环境的精确感知,作为语义先验引导高级驾驶辅助系统中的路径规划和轨迹预测等任务,并进一步实现对无人驾驶车辆行驶的控制和调整。本文的主要贡献如下:

1) 构造了一种用于道路可行驶区域检测的双分支金字塔Transformer主干网络,以并行方式融合基于全局和窗口的收缩注意力,在全局感知的基础上顾及局部细节的精确分割;

2) 提出可学习深度位置编码辅助网络,基于深度图像固有的空间位置信息,以其卷积特征构造辅助网络,生成可学习的位置编码,建模彩色图像像素与实际场景的空间位置联系,消除注意力偏移;

3) 设计多尺度特征融合模块,利用级联融合的上采样渐进式恢复特征分辨率,有效融合多尺度特征丰富的语义信息,并以加权深监督约束中间层特征,实现高精度、稳健的可行驶区域检测。

1 方法原理

本文提出的可学习深度位置编码引导的道路可行驶区域检测网络模型主要包含3个模块:双分支金字塔Transformer主干网络,与主干网逐层对应的基于卷积的可学习深度位置编码辅助网络,带有加权深监督的多尺度特征融合模块,整体网络模型如图1所示。

1.1 双分支金字塔Transformer主干网络

金字塔Transformer主干网络:主干网接收彩色图像输入,将彩色图像划分为不重叠的一系列图像块和窗口。本文将图像块设置成大小为4×4个像素,将窗口设置成大小为4×4个图像块。主干网为包括4个Transformer层的金字塔结构,每层的结构统一设置为并行双分支形式,融合基于图像块的全局注意力和窗口注意力。其中,注意力计算方式由传统Transformer模块中的多头自注意力修改为多头收缩自注意力,以有效减少高分辨率特征图引入的计算量。主干网络共设置4个阶段,为了构建金字塔结构以提取多尺度特征,本文在每个阶段的Transformer计算层中添加图像块特征,利用卷积计算缩小图像特征分辨率并扩展特征通道数。因此,网络接收尺寸为$H\times W\times 3$的彩色图像作为输入,经过金字塔网络提取特征后,4个阶段可分别获得尺寸为$\frac{H}{2}\times\frac{W}{2}\times C$、$\frac{H}{4}\times\frac{W}{4}\times 2C$、$\frac{H}{8}\times\frac{W}{8}\times 4C$和$\frac{H}{16}\times\frac{W}{16}\times 8C$的多尺度金字塔特征,其中$H$和$W$为图像长和宽的像素数,3为原始图像的RGB通道,$C$为图像特征通道数。

双分支Transformer模块:为了有效聚合全局和局部窗口特征,本文设计了双分支Transformer模块,以并行双分支方式构建全局和窗口化的注意力表示。给定每一阶段的输入x,将其划分为不同的图像块和窗口序列,而后图像块序列以全分辨率感受野从全局理解图像上下文关系,同时窗口序列在窗口内部计算注意力,优化细节信息,模块示意图如图2所示。

每一层Transformer模块共享相同的结构和相同的注意力计算方式。每个Transformer模块与传统Transformer相比,设置收缩注意力以减少计算量,其他层均保持一致,其包含一个收缩注意力层、两个归一化层和一个前馈神经网络,并应用残差结构链接注意力模块和前馈神经网络。为了有效融合全局注意力和窗口注意力,提升特征表示的能力,本文引入了一个可学习的特征变换模块利用CNN估计尺度参数和偏置参数,对全局注意力特征和窗口注意力特征进行线性变换融合。该计算过程为:

$$\alpha=f_{\alpha}(F_{\text{window}};W_{\alpha}) \tag{1}$$

$$\beta=f_{\beta}(F_{\text{window}};W_{\beta}) \tag{2}$$

式中,f_{α}和f_{β}分别表示用以产生尺度参数α和偏置

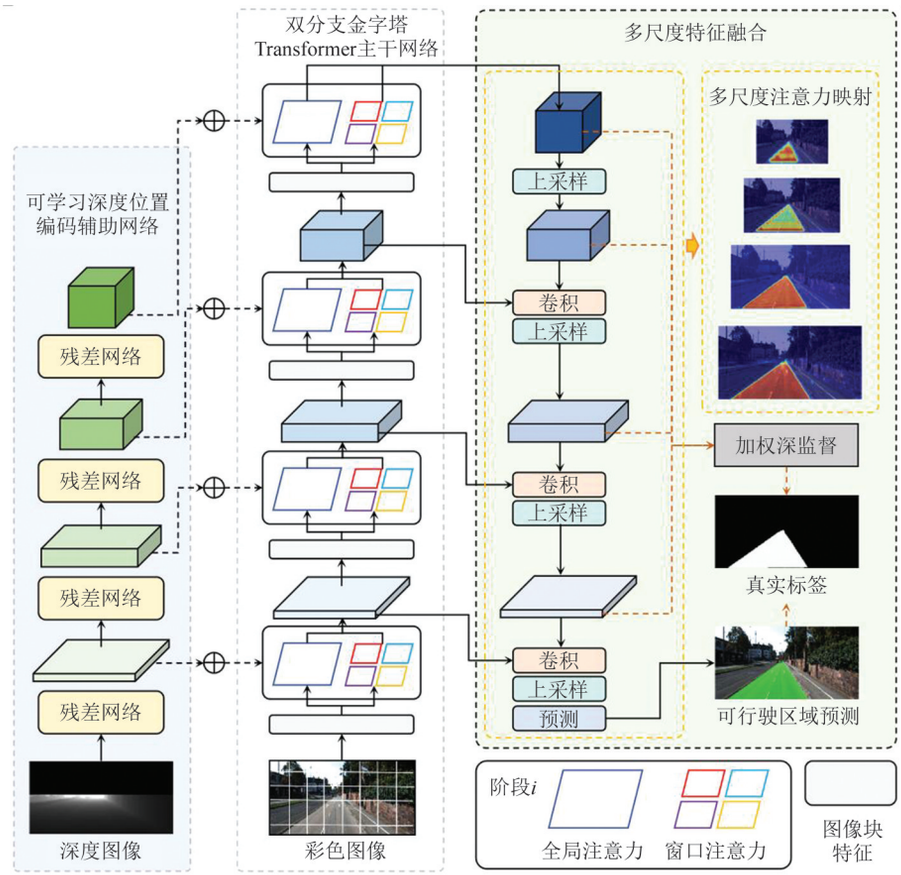


图 1　本文模型架构示意图

Fig. 1　Architecture Diagram of the Proposed Method

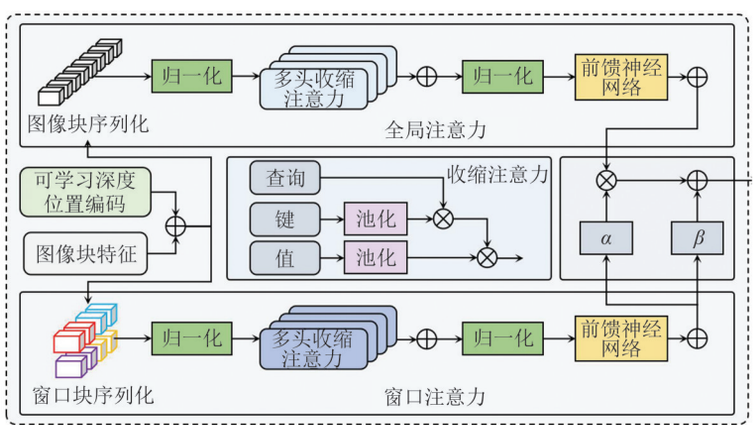


图 2　双分支 Transformer 模块示意图

Fig. 2　Diagram of Dual-Transformer Blocks

参数β的卷积函数;W_α和W_β为权重参数;F_{window}表示窗口注意力特征。

在注意力变换融合模块中,利用计算得到的尺度参数和偏置参数将窗口注意力特征和全局注意力特征进行融合。计算公式为:

$$F_{\text{fusion}}=\alpha F_{\text{global}}+\beta \tag{3}$$

式中,F_{fusion}和F_{global}分别表示融合后的输出特征以及全局注意力特征。

收缩注意力模块:为了减少注意力机制在全局计算中的平方次计算量,本文探索了收缩自注意力以适应稠密的像素级预测和高分辨率特征提取。对于每一个收缩自注意力计算层,遵循了传统 Transformer 注意力层中的多头策略,驱动模型关注不同特征空间的信息。多头收缩自注意力A_{MHS}的计算过程可以表示为:

$$\begin{aligned}A_{\text{MHS}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=&\\(\text{head}_0\oplus,\cdots,\oplus\text{head}_h)\boldsymbol{W}^O&\end{aligned} \tag{4}$$

式中,$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$是输入特征线性投影后产生的查

询、键和值矩阵；W^O表示注意力计算权重；$\oplus$表示特征拼接；head表示头空间；h设置为8，即设置8个头空间。将特征空间分为多个子空间以驱动模型感知不同方面的信息，该方式为将收缩自注意力过程计算h次，再将输出合并起来。本文使用池化收缩输入特征序列的空间维度，减少自注意力层所产生的参数数量和计算开销，则在每个注意力头中的相似性计算规则可以改写为：

$$\text{head}_i = \text{Attention}(\boldsymbol{Q}\cdot W_i^Q, \text{Pooling}(\boldsymbol{K})\cdot W_i^K, \text{Pooling}(\boldsymbol{V})\cdot W_i^V) \tag{5}$$

式中，W^Q、W^K、W^V分别表示$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$的权重；Attention表示点积注意力计算方式；Pooling表示池化操作，池化收缩因子为P。其中，全局注意力分支使用平均池化，窗口注意力分支使用最大池化。最大池化取池化区域中像素点的最大值，对纹理特征信息更加敏感；平均池化对池化区域内的像素取平均值，更关注特征整体分布。从理论角度出发，全局注意力模块使用平均池化获取全局上下文信息，窗口注意力模块中以最大池化为操作方式获取显著性特征。本文通过实证实验验证窗口注意力分支通过最大池化可以获取显著特征，并在网络更深层近似等效于提取全局上下文信息。$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别是来自于彩色图像特征R与深度图像特征D融合后的线性投影所产生的查询、键和值矩阵，具体表示为：

$$\boldsymbol{Q} = W^Q(\boldsymbol{x}_i^R + \boldsymbol{x}_i^D) \tag{6}$$

$$\boldsymbol{K} = W^K(\boldsymbol{x}_i^R + \boldsymbol{x}_i^D) \tag{7}$$

$$\boldsymbol{V} = W^V(\boldsymbol{x}_i^R + \boldsymbol{x}_i^D) \tag{8}$$

式中，i表示金字塔结构的第i个阶段。本文仅针对键矩阵和值矩阵进行收缩，因此，最终计算生成的注意力映射与原始输入的维度是保持一致的。

假设给定的输入$x\in\mathbf{R}^{H\times W\times C}$，经过大小为$P\times P$的池化核收缩后，原始多头注意力$A_{\text{MH}}$计算方式的时间复杂度计算公式：

$$O(A_{\text{MH}}) = H^2W^2C \tag{9}$$

变为：

$$O(A_{\text{MHS}}) = HWP^2C \tag{10}$$

式中，P为收缩尺度因子。由于P的维度远小于H和W，因此，收缩注意计算方式可以有效减少来自高分辨率特征和密集像素预测的计算量。

1.2 可学习深度位置编码辅助网络

目前主流的Transformer方法主要依靠不可学习的相对/绝对位置编码描述像素或图像块序列的先验位置关系。然而，这种不可学习的位置编码忽略了像素在现实环境中的空间特性，因此目标物体的注意力表示会出现偏差。

与彩色图像对齐的深度图像可以反映像素的空间信息，因此像素之间的关系可以依靠CNN提取的局部特征构建。深度特征是一种可以提供控制位置偏置的可学习位置编码。为了与金字塔Transformer结构对应，本文通过利用ResNet-34残差网络建立辅助网络提取了多尺度深度特征，并且逐层的深度特征维度与彩色图像特征保持一致。因此，计算注意力时深度特征作为位置编码偏置以逐像素方式与彩色图像特征相加。所提取的用于构造位置编码的深度图像特征可以在反向传播中进行优化。

1.3 多尺度特征融合模块

对于逐像素的密集语义预测任务而言，图像特征必须恢复至与原始输入一致的分辨率。一步直接完成的上采样会在插值过程中引入噪声，因此本文设计了多尺度级联融合的上采样模块。主干网络提取的多尺度特征逐层以双线性插值的形式上采样后与主干网提取的多尺度特征进行拼接，拼接后利用卷积操作将通道数压缩到与原始特征一致。通过逐层渐进的方式，每一个阶段对应的特征都可以计算得到原始分辨率下的双通道特征进行预测。本文设计以深监督方式对特征进行预测，因此每层恢复到原始分辨率的特征都会带有一个交叉熵分类器，最终深监督的损失函数为各层级交叉熵损失函数的加权求和，4层特征的权重分别为0.1、0.2、0.3和0.4。

2 可行驶区域检测实验结果与分析

2.1 测试数据

本文在城市道路数据集KITTI道路[35-36]和Cityscapes[37]两个大规模公开数据集上进行训练、验证和测试，同时为了进一步评估本文模型的有效性，还在自行采集的厦门市道路数据中进行了测试。

KITTI道路数据集为可行驶区域检测提供了开放访问的数据集和标准评估方法，通过高分辨率彩色摄像机采集卡尔斯鲁厄的城市公路数据。KITTI道路包含3种不同类别的道路场景，分为城市无标记道路(urban unmarked road，UU_Road)、城市标记道路(urban marked road，UM_Road)和城市多标记车道(urban multiple marked road，UMM_Road)，最终在城市道路中进行总体综合评估。

Cityscapes数据从50个城市采集了多个季节的道路图像,该数据集包含了30个类别的道路场景对象,因此为了更贴近本文研究,将该数据集的标签修改为两个类别,即道路可行驶区域为前景,其他像素点均为背景点。

厦门市道路数据由车载摄像机进行采集,包含了学校、城市公路等多个场景,本文将采集到的图像统一裁剪为与KITTI道路数据集图像大小一致的384×1 248像素。该自行采集的日间道路数据包括单车道、多车道、交叉路口等多种车道形态。

2.2　硬件条件与评价指标

本文网络模型部署在Ubuntu 20.04系统中,使用Python 3.7、CUDA 11.1以及PyTorch 1.11进行实验。训练过程在单张NVIDIA RTX 3090显卡上进行,批处理单元为4,数据加载线程设为6。模型初始训练次数设置为300迭代轮次,初始学习率设定为2×10^{-4},优化器使用AdamW,模型在第220轮次左右达到收敛。

本文所有的评价指标遵循了KITTI道路公开的评估标准:最大F值、基于PASCAL计算机视觉挑战中使用的平均精度、精确度、召回率、假阳性率、假阴性率对本文模型进行评估。

2.3　可行驶区域检测结果定性比较

为了有效评估本文方法在实际场景中进行可行驶区域检测的有效性,开展了一系列定性评估实验。随后的实验结果展示中,真阳性、假阳性以及假阴性像素点分别用绿色、蓝色和红色标记。本文在KITTI道路3个道路场景中进行实验,3个场景的代表性结果可视化如图3所示,代表性场景分别为UU_Road_82(图3(a))、UM_Road_95(图3(b))和UMM_Road_66(图3(c))。所选取的对比算法包括SNE-RoadSeg[24]、USNet[38]和SNE-RoadSeg+[26]。从图3中可以看出,本文方法在道路边界处的轮廓精细分割明显优于其他算法(如图中红框标识处),明确识别了道路的边缘以及车辆的边缘,以保证自动驾驶的安全性。同时,与其他方法相比,本文方法计算分割结果产生的假阳性和假阴性点更少,保证了自动驾驶的稳定性。

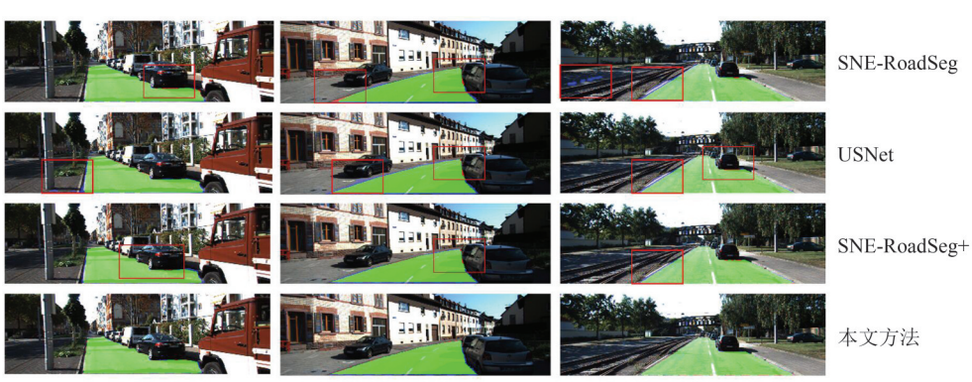


图3　KITTI道路数据集不同方法的可行驶区域检测结果比较

Fig. 3　Comparison of Different Methods in KITTI Road Dataset

为了探究本文方法在实际场景中应用的稳定性,以厦门市道路作为输入进行测试,结果可视化如图4所示。从图4中可以看出,本文方法充分利用视觉Transformer的全局感知能力,将道路的可行驶部分和不可行驶区域进行了明确的区分,并且不受道路形态变化与道路纹理变化的影响。在道路交叉区域和车道变窄区域都实现了具有鲁棒性的分割。

2.4　可行驶区域检测结果定量比较

针对KITTI道路数据集,本文在3种细分的道路场景中进行测试,所取得的精度见表1。本方法在城市多标记道路UMM_Road类别中各项计算指标达到最佳。该场景为多标记车道,即场景中目标实体较多,尤其是道路和非道路区域(人行道、铁路等)存在着更明显的分界。因此,本方法在全局感受野下保证了可行驶区域分割完整性的基础上,以局部窗口提取了实体与道路边缘处更丰富的可区分特征,形成了精细化的边缘,减少了错误分类的问题,故而产生了更好的验证精度。

同时,本文选取了KITTI道路基准测试中21个最先进算法在城市道路上的测试结果进行比较。对比算法包括LidCamNet[39]、LC-CRF[40]、TVFNet[19]、RGB36-Cotrain[41]、HID-LS[42]、Road-

Net3[43]、OFANet[44]、ALO-AVG-MM[45]、RBANet[31]、PLARD[21]、SNE-RoadSeg[24]、RoadNetRT[29]、NIM-RTFNet[23]、Hadamard-FCN[46]、BJN[15]、HA-DeepLabv3+[20]、CLCFNet[47]、DFM-RTFNet[25]、SNE-RoadSeg+[26]、USNet[38]以及HEAT[48]。

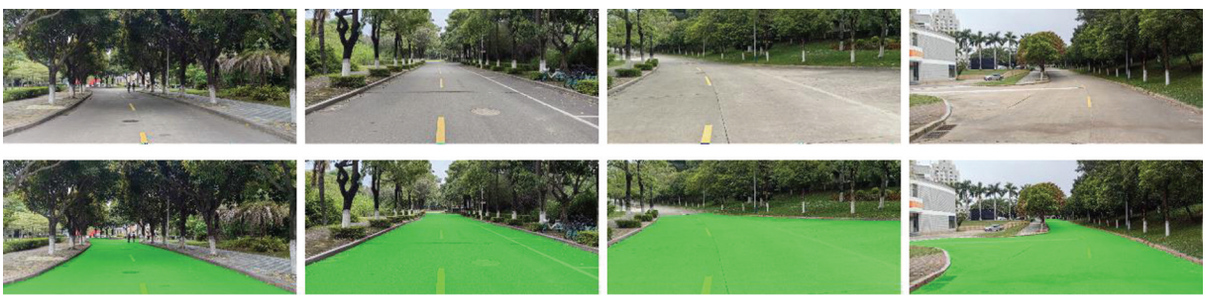


图4 厦门市道路数据集测试结果可视化

Fig. 4 Results Visulazition in Xiamen City Road Dataset

表1 KITTI 3种道路场景验证结果/%

Tab. 1 Test Results of 3 KITTI Road Scenes/%

道路类别	最大F值	平均精度	精确度	召回率
UM_Road	97.28	92.66	97.37	97.19
UMM_Road	98.09	94.74	97.74	98.45
UU_Road	96.85	91.42	96.73	96.96

表2给出了各个方法在KITTI道路基准测试中的综合评估结果。本文方法在KITTI道路基准测试中综合精度位列第一名,最大F值、平均精度、精确度、召回率分别达到了97.53%、92.97%、97.32%和97.74%,其中最大F值和召回率均为最佳性能。本文方法的平均精度比最佳方法低大约1.1%,但是其他各项指标均比其高1%左右。精确度比最佳方法仅低0.09%。所对比的方法,例如SNE-RoadSeg[24]系列、DFM-RTFNet[25]、USNet[38]与HEAT[48]等方法均为彩色图像与深度图的数据融合方法,本文方法均获得了比上述方法更好的表现。高最大F值表现出本文方法在可行驶区域检测中的高准确性,高召回率反映了本文方法在分割中的稳定性。上述指标表明本文方法在可行驶区域检测任务中具有高精度,可以保证自动驾驶任务的安全性。

本文同时对该模型的计算效率进行了验证。尽管提出了用收缩注意力的方式有效减少高分辨率特征以传统自注意力计算的复杂时间成本,但是序列-序列的方式不可避免地产生了比卷积神经网络更大的计算量。本文方法总体参数量为191.54 MB,浮点计算量为404 GB,计算速度可达12.5帧/s,即可以在0.08 s的时间内处理单幅图像,计算得到可行驶区域。城市道路整体评估如图5所示,本文方法以较高的计算效率达到了最佳性能。目前,本文方法提取深度卷积特征的网络为ResNet-34残差网络,如果选用ResNet-18残差网络,本文方法的计算效率将会进一步提升,但是会以损失精度为代价。相反,如果采用更深的特征提取网络如ResNet-50残差网络和ResNet-101残差网络等,计算效率将会进一步下降。为了有效平衡计算精度与效率的关系,本文方法选择使用ResNet-34残差网络作为可学习深度位置编码特征的编码器。

为了进一步验证本文方法提出的网络架构在多种不同交通环境下进行可行驶区域检测的有效性,本文验证了其在Cityscapes数据集上的性能。由于使用Cityscapes数据集的算法大多进行道路场景多类别分割,因此本文仅选取面向二分类的道路分割算法进行对比,对比算法包括FCN[49]、SegNet[50]、RBANet[31]以及USNet[38],对比结果见表3。由表3可知,本文方法在Cityscapes数据集上的最大F值为98.54%,精确度为98.35%,召回率为98.73%,均优于其他方法。其中,本文方法在最大F值上较RBANet[31]和USNet[38]分别有0.54%和0.27%的提升。

2.5 消融实验与分析

本文对各个模块对网络整体性能的影响进行了评估。定性评估结果如图6所示,其中①~④为特征提取层的注意力映射,分辨率依次降低。图6(a)表示KITTI道路验证集的UMM_Road_39,图6(b)表示KITTI道路验证集中的UU_Road_37,每个子图分为无局部窗口注意力和加入局部窗口注意力后的可视化结果。从下至上可以看出,随着网络的加深,红色反馈在道路区域中扩散并不断覆盖到整个道路区域,说明注意力逐渐汇聚在道路部分,注意力反馈在反向传播中不断增强。然而,图6中的无局部窗口注意力的可视化部分中明显观察到会出现部分边

界越界问题,这证明了全局特征变换对局部变换是不敏感的,尤其是在道路与人行道这种纹理相似的部分。从同层级的特征可以看出,在加入窗口注意力后特征响应为红色的部分占比更多,覆盖的道路区域更广,即注意力反馈变得更强,所有的注意力完全在道路区域中,消除了道路中的分割空洞问题。

从图6还可以看出边界部分在加入局部窗口注意力后要比无局部感知更精细,说明局部窗口的加入使得网络对细节信息的把握能力更强。在任何像素位置,局部窗口注意力增强后的响应都要比未加入更深,尤其是在未加入时响应表现为浅色的区域(例如图6中无局部窗口注意力所产生的空洞)在加入局部窗口注意力后都会被增强。综上所述,窗口注意力作为全局注意力的补充,一方面增强了全局特征响应反馈,另一方面增强了局部细节感知能力。

本文方法通过局部注意力对道路和障碍物形成了一个非常明确的边缘轮廓,边缘细节展示如图7所示。从图7中可以看出,本文方法可以准确识别车辆、行人的边缘,并且不受距离、遮挡以及形态等因素的制约。此外,本文方法在弯道处可以分割出平滑的曲线轮廓,消除锯齿状和模糊的道路边界,准确区分了人行道与车辆的可行驶道路。对于交通场景中的细小标志物(如路灯杆、立柱等),本文方法也可以准确识别,即使细长的道路标识在图像中将道路截断,本文方法依靠全局上下文信息也可以推理出道路区域,从而保证了道路识别的连续性。由此可以看出,双分支Transformer模块以全局和局部注意力的方式使得网络捕获了更丰富、更有区分度的特征,以保证可行驶区域检测的鲁棒性。

表2　KITTI道路中的综合评估结果

Tab. 2　Comprehensive Comparison of KITTI Road

方法	最大F值/%	平均精度/%	精确度/%	召回率/%	假阳性率/%	假阴性率/%	推理时间/s
LidCamNet[39]	96.03	93.93	96.23	95.83	2.07	4.17	0.15
LC-CRF[40]	95.68	88.34	93.62	97.33	3.67	2.67	0.18
TVFNet[19]	95.34	90.26	95.73	94.94	2.33	5.06	0.04
RGB36-Cotrain[41]	95.55	93.71	95.68	95.42	2.37	4.58	0.10
HID-LS[42]	93.11	87.33	92.52	93.71	4.18	6.29	0.25
RoadNet3[43]	94.44	93.45	94.69	94.18	2.91	5.82	0.30
OFANet[44]	93.74	85.37	90.36	97.38	5.72	2.62	0.04
ALO-AVG-MM[45]	92.03	85.64	90.65	93.45	5.31	6.55	0.03
RBANet[31]	96.30	89.72	95.14	97.50	2.75	2.50	0.16
PLARD[21]	97.03	94.03	97.19	96.88	1.54	3.12	0.16
SNE-RoadSeg[24]	96.75	94.07	96.90	96.61	1.70	3.39	0.18
RoadNetRT[29]	92.55	93.21	92.94	92.16	3.86	7.84	0.08
NIM-RTFNet[23]	96.02	94.01	96.43	95.62	1.95	4.38	0.05
Hadamard-FCN[46]	94.85	91.48	94.81	94.89	2.85	5.11	0.02
BJN[15]	94.89	90.63	96.14	93.67	2.07	6.33	0.02
HA-DeepLabv3+[20]	94.83	93.24	94.77	94.89	2.88	5.11	0.06
CLCFNet[47]	96.38	90.85	96.38	96.39	1.99	3.61	0.02
DFM-RTFNet[25]	94.78	94.05	96.62	96.93	1.87	3.07	0.08
SNE-RoadSeg+[26]	97.50	93.98	97.41	97.58	1.43	4.24	0.08
USNet[38]	96.89	93.25	96.51	97.27	1.94	2.73	0.02
HEAT[48]	97.00	93.09	96.53	97.48	1.93	2.51	0.08
本文方法	97.53	92.97	97.32	97.74	1.48	2.26	0.08

定量评估结果如表4所示,所对比的部分包括原始Transformer、金字塔Transformer、深度位置编码及双分支Transformer模块。从表4中可以看出,当建立多尺度金字塔Transformer后,相比于原始Transformer,最大F值提升大约3.31%(83.69 %vs. 87.00%)。为了建模像素与实际场景的位置关系,本文利用可学习的深度位置编码替换了传统的相对/绝对位置编码,将最大F值从87.00%直接提升至94.53%。更关键的是,传统的视觉Transformer只以全局感受野进行语义感知,限制了其提取和分割精细细节的能力。相比之下,本文所设计的具有局部窗口感知能力的双

分支Transformer模块以非全局而又非局部的方式感知细节,在全局感受野下优化边缘等结构,进一步增强分割的准确性。从数据可以看出,相比之前的模型,加入后最大F值提升到了97.53%,平均精度达到了93.97%,精确度和召回率分别提升至97.32%和97.74%。同时本文模型的假阳性率和假阴性率两项指标分别降低至1.48%和1.26%。

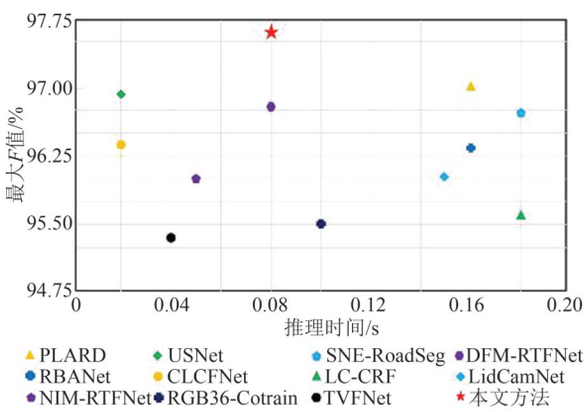


图5 KITTI道路数据集上运行时间与最大F值对比

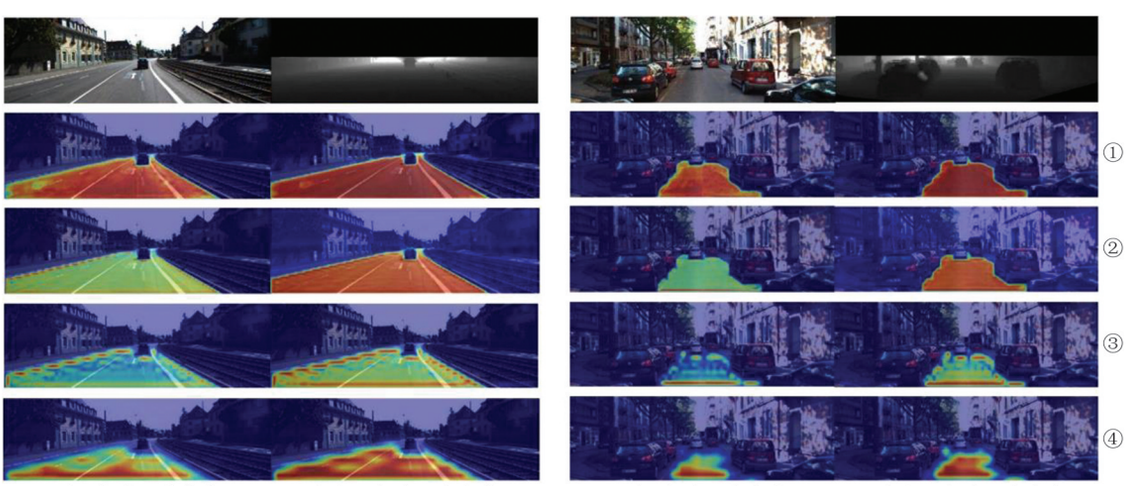
Fig. 5 Comparison of Runtime and Max F in KITTI Road Dataset

表3 Cityscapes数据集中的检测结果/%

Tab. 3 Test Results in Cityscapes Dataset/%

方法	最大F值	精确度	召回率
FCN[49]	94.68	93.69	95.70
SegNet[50]	95.81	94.55	97.11
RBANet[31]	98.00	97.87	98.12
USNet[38]	98.27	98.26	98.28
本文方法	98.54	98.35	98.73

为了进一步验证本文提出的Transformer框架更适合于可行驶区域检测任务,本文将所提方法与3个最先进的金字塔Transformer方法(PVT[32]、SegFormer[33]、CMX[51])进行了比较。图8是4种方法的分割结果定性比较,整体来看,由于Transformer方法具有全局特性,所以道路区域的完整性很高,但其问题在于全局上下文引入了纹理相似的道路区域的干扰,使得在人行道附近的边界存在错位问题。相比之下,本文方法可以更好地识别道路和障碍物轮廓,由此划分了更精确的可行驶区域,保证自动驾驶任务的安全性和稳定性。



(a) UMM_Road_39 (b) UU_Road_37

图6 定性评估的注意力映射图

Fig. 6 Attention Representation of Qualitative Evaluation

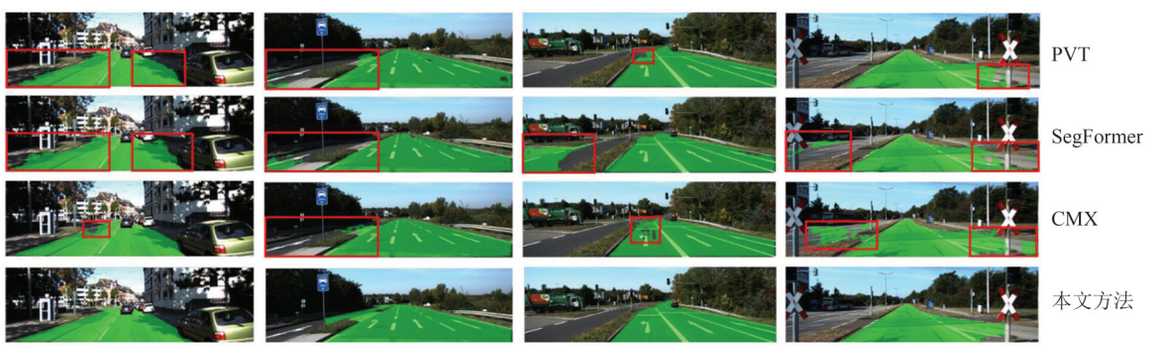
图7 道路边缘轮廓细节可视化结果

Fig. 7 Road Boundary Details Visualization Results

表 4　各模块对网络整体性能的影响/%

Tab. 4　Performance Impacts of Different Modules on Whole Network/%

原始 Transformer	金字塔 Transformer	深度位置编码	双分支 Transformer模块	最大F值	平均精度	精确度	召回率	假阳性率	假阴性率
√				83.69	87.59	81.34	86.18	10.89	13.82
	√			87.00	90.51	85.77	88.27	8.07	11.73
	√	√		94.53	93.68	94.62	94.45	2.96	5.55
	√	√	√	97.53	93.97	97.32	97.74	1.48	1.26



(a) UM_Road_72　(b) UMM_Road_13　(c) UMM_Road_16　(d) UU_Road_03

图 8　不同场景下与不同 Transformer方法的定性比较

Fig. 8　Qualitative Comparison of Different Transformer-Based Methods on Various Scenes

定量比较结果如表 5 所示。相比于目前最主流的 Transformer方法,本文方法在可行驶区域检测中展示了更好的优越性。例如 CMX[51]设计为一种用于数据融合的 Transformer网络,但本文方法可以在最大F值、精确度和召回率 3 项指标中明显超越 CMX[51]。其主要原因在于本文方法在全局感受野下加入了窗口化模式,加强了对图像细节信息的感知能力,同时以可学习的深度位置编码引入空间位置、结构信息,有效解决了传统不可学习的绝对/相对位置编码造成的注意力偏移问题,以及由此引发的语义不对齐问题。本文以上述方式强化了金字塔 Transformer的语义解译能力,并且在道路可行驶区域检测这项二元语义分割任务中达到了最佳的性能表现。

表 5　不同 Transformer方法的定量比较/%

Tab. 5　Quantitative Comparison of Different Transformer-Based Methods/%

方法	最大F值	平均精度	精确度	召回率
PVT[32]	87.00	90.51	85.77	88.27
SegFormer[33]	91.67	92.47	89.85	93.56
CMX[51]	94.55	93.41	94.44	94.66
本文方法	97.53	93.97	97.32	97.74

3　结　语

针对现有基于 CNN 的方法缺失全局上下文先验信息而造成可行驶区域检测的空洞和中断问题,本文提出了多尺度金字塔 Transformer网络,从全局理解交通场景图像。同时,本文构建了提取可学习深度位置编码的辅助网络,以解决传统不可学习的位置编码造成的注意力偏移和语义不对齐问题,优化分割结果。通过融合全局和窗口化的双分支注意力计算模块,补充局部细节丢失,以感知精细的道路和障碍物边缘轮廓。综合实证研究表明,本文方法在 KITTI 道路和 Cityscapes 数据集上表现出了优于现有前沿方法的性能。本文方法在兼顾计算效率的同时,在 KITTI 道路数据集上以 97.53% 的精度排名第一,在 Cityscapes 数据集上验证为 98.54%,并在厦门市道路场景中进行了测试。总体而言,本文提出的框架形成了高精度的可行驶区域检测技术路线,可以更好地应用于智能驾驶辅助系统,以辅助自动驾驶任务。在未来,将会以本文模型为基础,探索面向道路场景的多类别目标语义分割与实例分割问题,以及研究以多帧图像和视频序列为数据的基于时序的道路可行驶区域检测研究,以应对复杂多变的道路交通环境,进一步服务于智能驾驶辅助系统以及自动驾驶任务。

参　考　文　献

[1]　Cui Mingyang, Huang Heye, Xu Qing, et al. Sur-

vey of Intelligent and Connected Vehicle Technologies: Architectures, Functions and Applications[J]. *Journal of Tsinghua University* (*Science and Technology*), 2022, 62(3): 493-508. (崔明阳, 黄荷叶, 许庆, 等. 智能网联汽车架构、功能与应用关键技术[J]. 清华大学学报(自然科学版), 2022, 62(3): 493-508.)

[2] Zhang Yanjie, Huang Wei, Liu Xintao, et al. An Approach for High Definition (HD) Maps Information Interaction for Autonomous Driving[J]. *Geomatics and Information Science of Wuhan University*, 2023, DOI: 10.13203/j.whugis20230166. (张焱杰, 黄炜, 刘信陶, 等. 自动驾驶高精地图信息交互方法[J]. 武汉大学学报(信息科学版), 2023, DOI: 10.13203/j.whugis20230166.)

[3] Ying Shen, Jiang Yuewen, Gu Jiangyan, et al. High Definition Map Model for Autonomous Driving and Key Technologies[J]. *Geomatics and Information Science of Wuhan University*, 2023, DOI: 10.13203/j.whugis20230227. (应申, 蒋跃文, 顾江岩, 等. 面向自动驾驶的高精地图模型及关键技术[J]. 武汉大学学报(信息科学版), 2023, DOI: 10.13203/j.whugis20230227.)

[4] Daoud M A, Mehrez M W, Rayside D, et al. Simultaneous Feasible Local Planning and Path-Following Control for Autonomous Driving[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(9): 16358-16370.

[5] Pan J C, Sun H Y, Xu K C, et al. Lane-Attention: Predicting Vehicles' Moving Trajectories by Learning Their Attention over Lanes[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, USA, 2020.

[6] Weber M, Xie J, Collins M D, et al. STEP: Segmenting and Tracking Every Pixel[C]//The 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track, New Orleans, USA, 2021.

[7] Shinzato P Y, Wolf D F. A Road Following Approach Using Artificial Neural Networks Combinations[J]. *Journal of Intelligent & Robotic Systems*, 2011, 62(3): 527-546.

[8] Alvarez J M, Gevers T, LeCun Y, et al. Road Scene Segmentation from a Single Image[C]//The 12th European Conference on Computer Vision: Volume Part VII, Florence, Italy, 2012.

[9] Passani M, Yebes J J, Bergasa L M. CRF-Based Semantic Labeling in Miniaturized Road Scenes [C]//The 17th International IEEE Conference on Intelligent Transportation Systems, Qingdao, China, 2014.

[10] Passani M, Yebes J J, Bergasa L M. Fast Pixelwise Road Inference Based on Uniformly Reweighted Belief Propagation [C]//IEEE Intelligent Vehicles Symposium, Seoul, 2015.

[11] Vitor G B, Victorino A, Ferreira J V. A Probabilistic Distribution Approach for the Classification of Urban Roads in Complex Environments [C]//IEEE Workshop on International Conference on Robotics and Automation, Hong Kong, China, 2014.

[12] Munoz D, Bagnell J A, Hebert M. Stacked Hierarchical Labeling[C]//The 11th European Conference on Computer Vision: Part VI, Heraklion, Crete, Greece, 2010.

[13] Mendes C C T, Frémont V, Wolf D F. Exploiting Fully Convolutional Neural Networks for Fast Road Detection [C]//IEEE International Conference on Robotics and Automation, Stockholm, Sweden, 2016.

[14] Muñoz-Bulnes J, Fernandez C, Parra I, et al. Deep Fully Convolutional Networks with Random Data Augmentation for Enhanced Generalization in Road Detection [C]//The 20th International Conference on Intelligent Transportation Systems, Yokohama, Japan, 2017.

[15] Che Manqiang, Li Shubin, Li Ming. Road Surface Semantic Segmentation Method Based on HarDNet Fully Convolutional Network[J]. *Journal of Computer Applications*, 2021, 41(S2): 76-80. (车满强, 李树斌, 李铭. 基于HarDNet全卷积网络的道路路面语义分割方法[J]. 计算机应用, 2021, 41(S2): 76-80.)

[16] Jiang Tengping, Yang Bisheng, Zhou Yuzhou, et al. Bilevel Convolutional Neural Networks for 3D Semantic Segmentation Using Large-Scale LiDAR Point Clouds in Complex Environments[J]. *Geomatics and Information Science of Wuhan University*, 2020, 45(12): 1942-1948. (蒋腾平, 杨必胜, 周雨舟, 等. 道路点云场景双层卷积语义分割[J]. 武汉大学学报(信息科学版), 2020, 45(12): 1942-1948.)

[17] Yu B, Lee D, Lee J S, et al. Free Space Detection Using Camera-LiDAR Fusion in a Bird's Eye View Plane[J]. *Sensors*, 2021, 21(22): 7623.

[18] Chen L, Yang J, Kong H. LiDAR-Histogram for Fast Road and Obstacle Detection[C]//IEEE International Conference on Robotics and Automation, Singapore, 2017.

[19] Gu S, Zhang Y G, Yang J, et al. Two-View Fusion Based Convolutional Neural Network for Urban

Road Detection[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems, Macau, China, 2019.

[20] Fan R, Wang H L, Cai P D, et al. Learning Collision-Free Space Detection from Stereo Images: Homography Matrix Brings Better Data Augmentation [J]. *IEEE/ASME Transactions on Mechatronics*, 2022, 27(1): 225-233.

[21] Chen Z, Zhang J, Tao D C. Progressive LiDAR Adaptation for Road Detection [J]. *IEEE/CAA Journal of Automatica Sinica*, 2019, 6(3): 693-702.

[22] Khan A A, Shao J, Rao Y B, et al. LRDNet: Lightweight LiDAR Aided Cascaded Feature Pools for Free Road Space Detection[J]. *IEEE Transactions on Multimedia*, 2022, 99: 1-13.

[23] Wang H L, Fan R, Sun Y X, et al. Applying Surface Normal Information in Drivable Area and Road Anomaly Detection for Ground Mobile Robots[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, USA, 2020.

[24] Fan R, Wang H L, Cai P D, et al. SNE-RoadSeg: Incorporating Surface Normal Information into Semantic Segmentation for Accurate Freespace Detection [C]//The 16th European Conference, Glasgow, UK, 2020.

[25] Wang H L, Fan R, Sun Y X, et al. Dynamic Fusion Module Evolves Drivable Area and Road Anomaly Detection: A Benchmark and Algorithms [J]. *IEEE Transactions on Cybernetics*, 2022, 52(10): 10750-10760.

[26] Wang H L, Fan R, Cai P D, et al. SNE-RoadSeg: Rethinking Depth-Normal Translation and Deep Supervision for Freespace Detection [C]//IEEE/RSJ International Conference on Intelligent Robots and Systems, Prague, 2021.

[27] Song Shuang, Chen Chi, Yang Bisheng, et al. Large Field of View Array System Using Low Cost RGB-D Camerasin[J]. *Geomatics and Information Science of Wuhan University*, 2018, 43(9): 1391-1398. (宋爽, 陈驰, 杨必胜, 等. 低成本大视场深度相机阵列系统[J]. 武汉大学学报(信息科学版), 2018, 43(9): 1391-1398.)

[28] Meng Yiyue, Guo Chi, Liu Jingnan. Deep Reinforcement Learning Visual Target Navigation Method Based on Attention Mechanism and Reward Shaping [J]. *Geomatics and Information Science of Wuhan University*, 2023, DOI: 10.13203/j.whugis20230193. (孟怡悦, 郭迟, 刘经南. 基于注意力机制和奖励塑造的深度强化学习视觉目标导航方法[J]. 武汉大学学报(信息科学版), 2023, DOI: 10.13203/j.whugis20230193.)

[29] Bai L, Lyu Y C, Huang X M. RoadNet-RT: High Throughput CNN Architecture and SoC Design for Real-Time Road Segmentation[J]. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2021, 68(2): 704-714.

[30] Ai Qinglin, Zhang Junrui, Wu Feiqing. AF-ICNet Semantic Segmentation Method for Unstructured Scenes Based on Small Target Category Attention Mechanism and Feature Fusion[J]. *Acta Photonica Sinica*, 2023, 52(1): 0110001. (艾青林, 张俊瑞, 吴飞青. 基于小目标类别注意力机制与特征融合的AF-ICNet非结构化场景语义分割方法[J]. 光子学报, 2023, 52(1): 0110001.)

[31] Sun J Y, Kim S W, Lee S W, et al. Reverse and Boundary Attention Network for Road Segmentation [C]//IEEE/CVF International Conference on Computer Vision Workshop , Seoul, 2019.

[32] Wang W H, Xie E Z, Li X, et al. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions[C]//IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021.

[33] Xie E, Wang W, Yu Z, et al. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 12077-12090.

[34] Liu Z, Lin Y T, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows[C]//IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021.

[35] Fritsch J, Kühnl T, Geiger A. A New Performance Measure and Evaluation Benchmark for Road Detection Algorithms[C]//The 16th International Conference on Intelligent Transportation Systems, The Hague, Netherlands, 2013.

[36] Geiger A, Lenz P, Stiller C, et al. Vision Meets Robotics: The KITTI Dataset[J]. *International Journal of Robotics Research*, 2013, 32(11): 1231-1237.

[37] Cordts M, Omran M, Ramos S, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding [C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016.

[38] Chang Y C, Xue F, Sheng F, et al. Fast Road Segmentation via Uncertainty-Aware Symmetric Network[C]//International Conference on Robotics and Automation, Philadelphia, USA, 2022.

[39] Caltagirone L, Bellone M, Svensson L, et al. LiDAR-Camera Fusion for Road Detection Using Fully Convolutional Neural Networks[J]. *Robotics and*

Autonomous Systems, 2019, 111: 125-131.

[40] Gu S, Zhang Y, Tang J, et al. Road Detection Through CRF Based LiDAR-Camera Fusion[C]// 2019 International Conference on Robotics and Automation, Montreal, Canada, 2019.

[41] Han Z, Zhang C, Fu H, et al. Trusted Multi-view Classification [C]//International Conference on Learning Representations, New York, USA, 2020.

[42] Gu S, Zhang Y G, Yuan X, et al. Histograms of the Normalized Inverse Depth and Line Scanning for Urban Road Detection[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2019, 20(8): 3070-3080.

[43] Lyu Y C, Bai L, Huang X M. Road Segmentation Using CNN and Distributed LSTM[C]//IEEE International Symposium on Circuits and Systems , Sapporo, Japan, 2019.

[44] Zhang S C, Zhang Z, Sun L B, et al. One for All: A Mutual Enhancement Method for Object Detection and Semantic Segmentation [J]. *Applied Sciences*, 2019, 10(1): 13.

[45] Reis F A L, Almeida R, Kijak E, et al. Combining Convolutional Side-Outputs for Road Image Segmentation [C]//International Joint Conference on Neural Networks, Budapest, Hungary, 2019.

[46] Oeljeklaus M. An Integrated Approach for Traffic Scene Understanding from Monocular Cameras [M]. Düsseldorf: VDI Verlag, 2021.

[47] Gu S, Yang J, Kong H. A Cascaded LiDAR-Camera Fusion Network for Road Detection[C]//IEEE International Conference on Robotics and Automation, Xi'an, China, 2021.

[48] Han T, Li C M, Chen S Y, et al. HEAT: Incorporating Hierarchical Enhanced Attention Transformation into Urban Road Detection[J]. *IET Intelligent Transport Systems*, 2023(1): 1-20.

[49] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation [C]// IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015.

[50] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(12): 2481-2495.

[51] Zhang J M, Liu H Y, Yang K L, et al. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2023, 24(12): 14679-14694.

(上接第581页)

ny, 2020.

[16] Zhang Y F, Sun P Z, Jiang Y, et al. ByteTrack: Multi-object Tracking by Associating Every Detection Box [C]//The17th European Conference on Computer Vision, Tel-Aviv, Israel, 2022.

[17] Sun P Z, Cao J K, Jiang Y, et al. DanceTrack: Multi-object Tracking in Uniform Appearance and Diverse Motion [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022.

[18] Kalman R E. A New Approach to Linear Filtering and Prediction Problems[J]. *Journal of Basic Engineering*, 1960, 82(1): 35-45.

[19] Luiten J, Osep A, Dendorfer P, et al. HOTA: A Higher Order Metric for Evaluating Multi-object Tracking[J]. *International Journal of Computer Vision*, 2021, 129(2): 548-578.

[20] Bernardin K, Stiefelhagen R. Evaluating Multiple Object Tracking Performance: The CLEARMOT-Metrics[J]. *Journal on Image and Video Processing*, 2008(1): 1.

[21] Ristani E, Solera F, Zou R, et al. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking[C]//The 14th European Conference on Computer Vision, Amsterdam, Netherlands, 2016.