doi:10.11920/xnmdzk.2025.02.009

# 基于 Transformer 的道路场景语义分割综述

黄天云[1],向明建[2],邵世霖[2]
(1.西南民族大学数学学院,四川 成都 610225;2.西南民族大学计算机与人工智能学院,四川 成都 610225)

**摘　要**:在自动驾驶领域,通过对道路场景进行高质量的语义分割,可以为自动驾驶汽车的安全行驶提供重要保障.近年来,随着自动驾驶技术的不断进步,人们对语义分割模型在尺寸、计算成本和分割精度等方面的要求也日益提高,这促使研究者们探索更为先进的算法.首先介绍了语义分割技术在深度学习快速发展下取得的显著进展与不足,从而引出基于 Transformer 的道路场景语义分割方法.相较于传统的深度学习算法,Transformer 具备全面理解复杂场景中上下文关系的能力,尤其在处理多对象和复杂环境时表现出显著优势.接着,根据不同的特征处理策略和模型架构,将基于 Transformer 的道路场景语义分割方法分为四类:基于全局特征提取的方法、基于局部特征增强的方法、基于混合架构的方法以及基于自监督学习的方法.最后,分析和对比了每类方法的代表性算法,概括总结了各类方法的技术特点和优缺点.

**关键词**:语义分割;Transformer;全局特征提取;局部特征增强;混合架构;自监督学习
**中图分类号**:TP391.41　　**文献标志码**:A　　**文章编号**:2095-4271(2025)02-0193-13

# Review of semantic segmentation of road scene based on Transformer

HUANG Tianyun[1], XIANG Mingjian[2], SHAO Shilin[2]
(1.School of Mathematics, Southwest Minzu University, Chengdu 610225, China;
2.School of Computer and Artificial Intelligence, Southwest Minzu University, Chengdu 610225, China)

**Abstract**: In the field of autonomous driving, high-quality semantic segmentation of road scenes is crucial for ensuring the safe operation of self-driving cars. In recent years, with the continuous improvement of autonomous driving technology, the demand has been increasing for semantic segmentation models in terms of size, computational cost, and segmentation accuracy, prompting researchers to explore more advanced algorithms. This paper first introduced the significant progress and shortcomings of semantic segmentation technologies under the rapid development of deep learning, leading to the discussion of Transformer-based methods for road scene semantic segmentation. Compared to traditional deep learning algorithms, Transformers possessed the ability to comprehensively understand contextual relationships in complex scenes, demonstrating significant advantages, particularly in handling multi-objects and complex environments. Subsequently, based on different feature processing strategies and model architectures, the Transformer-based road scene semantic segmentation methods were categorized into four types: methods based on global feature extraction, methods based on local feature enhancement, methods based on hybrid architectures, and methods based on self-supervised learning. Finally, the paper analyzed and compared representative algorithms from each category, summarizing their technical characteristics, advantages and disadvantages.

**Keywords**: semantic segmentation; Transformer; global feature extraction; local feature enhancement; hybrid architecture; self-supervised learning

**收稿日期**:2025-01-07
**作者简介**:黄天云(1973-),男,教授,博士,研究方向:图形图像处理、计算机视觉.E-mail:huang.t.y@163.com
**通信作者**:向明建(1998-),男,湖北恩施人,研究方向:图像处理、计算机视觉.E-mail:xiang_m_j@163.com
**基金项目**:中央高校基本科研业务费专项资金研究生创新项目(YCYB2024001)

图像语义分割技术在实际应用中广泛存在,典型应用场景包括医学图像识别[1]和自动驾驶[2]等领域.其中,针对道路场景的语义分割技术是自动驾驶的核心之一,通过将采集到的道路场景图像中的每个像素划分到对应的类别,实现图像的像素级分类[3].在自动驾驶系统中,环境信息的处理至关重要,需要高水平的道路场景语义分割技术为智能车辆提供关键的路况信息,从而确保自动驾驶汽车的安全行驶.因此,随着技术的进步,语义分割的应用需求在复杂道路场景中不断增长,成为当前研究的热点.

随着深度学习的快速发展,语义分割技术取得了显著进展,尤其是在数据集和模型架构的丰富性方面.例如,卷积神经网络(CNN)[4]在早期的语义分割任务中取得了良好的效果,但传统CNN在捕捉长距离依赖和全局上下文信息方面存在不足,导致物体边界不清晰.在处理不同尺度物体时,特征融合困难,难以有效提取多尺度信息;并且在复杂场景中,对遮挡、光照变化和背景杂乱的适应能力较差.这些局限性促使研究者们探索更先进的模型和方法,如基于Transformer[5]的模型.这些模型利用自注意力机制去捕捉图像中的全局上下文信息以建模准确的长距离依赖关系,从而提升分割的准确性和鲁棒性.此外,Transformer的灵活性使其能够处理不同尺度的特征,适应复杂的道路场景,提升对动态对象的识别能力.

在自动驾驶领域中,基于Transformer的道路场景语义分割方法仍在不断发展中,尚未形成全面的综述性文献.因此,本文旨在关注基于Transformer的道路场景语义分割的研究进展,系统总结当前的技术方法、应用实例及其面临的挑战.同时,本文还将探讨未来的研究方向,以期为相关领域的研究者提供参考和启发.

# 1 道路场景语义分割的发展历史

## 1.1 从传统方法到深度学习

道路场景语义分割是计算机视觉领域的一个关键任务,旨在将图像中的每个像素分类为特定的语义类别.这项技术在自动驾驶、智能交通系统及城市规划领域具有广泛的应用前景.道路场景语义分割经历了从传统方法[6-8]到深度学习[9-12]主导的发展历程.传统方法主要依赖手工设计的特征分类器,这些方法多采用传统的图像处理技术,包括边缘检测、区域生长和阈值分割等.在这些方法中,研究者们通常需要提取图像的颜色、纹理和形状等特征,以便进行分类.这些手工设计的特征在处理简单场景时效果较好,然而在复杂环境中表现有限,且需要大量的人工参与,增加了时间和成本.

深度学习时期,基于神经网络的语义分割方法逐渐兴起.这些方法通过自动学习特征,显著减少了人工干预,能够处理更复杂的视觉信息.这一时期出现了多种基于神经网络的语义分割方法,发展出多尺度特征融合[13]、空间注意力机制[14]等创新技术,能够有效捕捉图像中的上下文信息,从而在复杂场景中实现更好的分类效果,提升分割精度和鲁棒性.

## 1.2 基于深度学习的道路场景语义分割

随着深度学习技术的迅猛发展,尤其是卷积神经网络(CNN)的应用,道路场景语义分割算法经历了显著的变革.这些算法能够有效处理复杂的视觉信息,提供高精度的像素级分类.全卷积网络(FCN)[15]是第一个将全卷积用于语义分割的模型,采用跳跃连接融合多尺度特征,提升了分割精度.U-Net[16]最初为医学图像分割设计,其对称的编码器-解码器结构同样适用于道路场景分割,特别适合处理小样本数据.MIF-Net[17]同样采用编码器-解码结构来完成语义分割,在编码部分利用分离策略和非对称卷积设计轻量型特征提取结构,在解码部分引入通道注意力机制恢复特征图尺寸和细节信息.DeepLab[18]系列通过引入空洞卷积和条件随机场,增强了感受野和边界精细化.陈晔等[19]对STDC网络进行改进,引入残差连接来更好地融合多尺度语义信息,提出一种嵌入位置注意力模块的空洞空间卷积池化金字塔(PA-ASPP)来增强网络对道路等特定区域的位置感知能力.Mask R-CNN[20]在Faster R-CNN[21]基础上增加了生成物体二进制掩码的分支,结合了实例分割和语义分割.PSP-Net[22]和HRNet[23]等先进模型,通过金字塔池化和高分辨率特征并行处理,进一步增强了场景理解能力.

然而,这些模型也存在一些局限性.首先,复杂的网络结构导致较高的计算和内存需求,限制了它们在实时应用中的可行性;其次,尽管这些模型在特定数据集上表现出色,但其泛化能力在面对不同场景和条件时可能不足;此外,许多现有模型依赖于大量标注

数据,然而在实际应用中,获取高质量标注数据往往是一个挑战.这些局限性促使研究者们探索更先进的模型和方法,特别是基于 Transformer 的技术,以提升语义分割的性能和准确性.

### 1.3　基于 Transformer 的道路场景语义分割

Transformer 模型最初在自然语言处理(NLP)[24]领域推出,但近年来逐渐在计算机视觉中崭露头角,尤其是在语义分割任务中取得了显著进展.其引入的自注意力机制能够有效捕捉输入数据中不同部分之间的长距离依赖关系,使模型能够关注图像中的重要区域,而不仅限于局部特征.此外,Transformer 比传统深度学习网络更擅长处理全局信息,能够全面理解复杂场景中的上下文关系,特别是在多对象和高复杂度环境中.通过不同层次的特征图融合,Transformer 实现了对多尺度信息的有效处理,这对于道路场景中物体大小和形状的多样性至关重要.

基于 Transformer 的模型展现出优越的分割精度,能够有效应对遮挡和光照变化等挑战,通过全局上下文信息的建模提高物体识别和分类的准确性.这一转变不仅推动了语义分割技术的发展,也为解决传统 CNN 在处理复杂场景时的局限性提供了新的思路.Transformer 架构的端到端学习能力进一步简化了传统分割模型中的复杂预处理和后处理步骤,从而提升了训练和推理的效率.

## 2　基于 Transformer 的道路场景语义分割方法

基于 Transformer 的道路场景语义分割方法,借助其自注意力机制和全局上下文建模能力,为图像分割任务带来了新的可能性.全局上下文建模主要通过自注意力机制实现.

首先,模型接收一个输入序列(如图像特征),每个元素被表示为向量.对于序列中的每个元素,计算其与所有其他元素的相关性,这通过生成查询(Query)、键(Key)和值(Value)向量来完成.查询向量表示当前元素,键向量表示其他元素,值向量则包含实际信息.接着,计算查询与所有键的点积以获得注意力分数,衡量相关性,并使用 Softmax 函数将这些分数转换为权重,归一化为概率分布.然后,利用获得的权重对值向量进行加权求和,生成当前元素的上下文表示,从而实现全局上下文建模.

与传统的深度学习网络相比,Transformer 能够获取更多、更高级的语义信息,从而更精确地表达图像中的内容.Transformer 模型架构如图 1 所示[5].

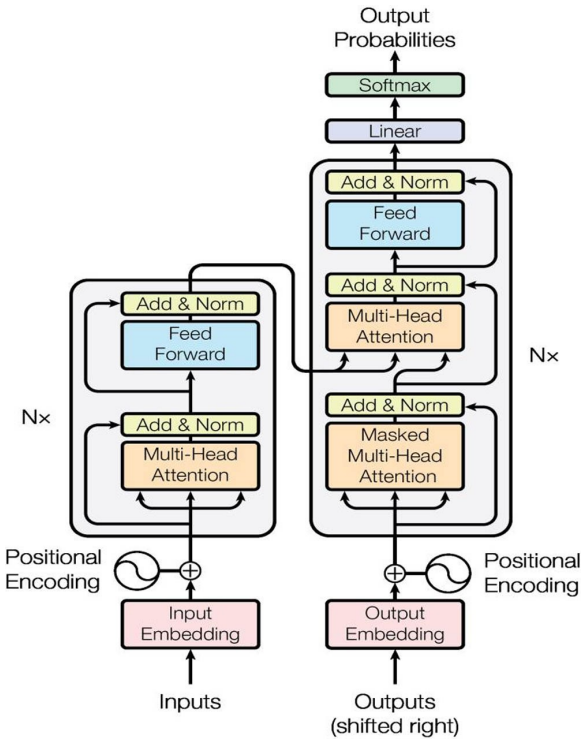


**图 1　Transformer 模型架构[5](源于文献[5])**

Fig.1　Transformermodel architecture[5]

图 1 中,Inputs 表示输入图像.Input Embedding 表示将输入的图像通过嵌入层映射为固定维度的向量(即序列)表示.Positional Encoding 表示将输入序列加上位置编码,以保留位置信息.Multi-Head Attention 表示多头注意力机制,用于捕捉输入序列中的全局依赖关系.Add & Norm 表示对多头注意力的输出添加残差连接,并进行归一化处理.Feed Forward 表示前馈神经网络,该网络由两个全连接层组成:第一个全连接层作为隐藏层,用于提取输入特征的高级表示;第二个全连接层作为输出层,将隐藏层的特征映射到目标输出.Shifted Right 表示将解码器的输入序列向右移动一个位置,以避免目标序列的像素块直接看到自身,从而实现自回归特性.Output Embedding 表示解码器生成的目标序列的输入(如训练时的标签)通过嵌入层转化为向量(即序列)表示.Masked Multi-Head Attention 用于目标序列的自注意力计算,确保模型在生成时只看到先前的输出.Linear 表示解码器的输出通

过一个线性层,将高维向量转换为所有像素块大小的分布.Softmax 表示将线性层的输出转换为概率分布,用于预测下一个像素块的概率.Output Probabilities 表示最终的输出概率.

本文根据不同的特征处理策略和模型架构,将基于 Transformer 的道路场景语义分割方法分为四类:基于全局特征提取的方法、基于局部特征增强的方法、基于混合架构的方法以及基于自监督学习的方法.

### 2.1 基于全局特征提取的方法

基于全局特征提取的方法利用 Transformer 的自注意力机制识别和理解不同区域之间的关系,例如在道路场景中理解行人、车辆和道路之间的空间关系.从全局角度捕捉图像中的信息,能够在整个图像范围内建立长距离的依赖关系,有效提高语义分割的精度.通过全局特征提取,模型同时考虑图像中的所有信息,理解整个场景的结构和语义,这对复杂的道路场景尤其重要.

Segmenter[25] 专门为语义分割设计,采用了标准的 Transformer 架构,利用自注意力机制在全局范围内建模图像特征,以充分捕捉图像中的长距离依赖关系,增强对复杂场景的理解.然而,该模型通常具有较高的计算复杂度,尤其是在处理高分辨率图像时.这导致训练和推理时间较长,增加了对硬件资源的需求.

DETR[26] 将目标检测的框架扩展至语义分割,通过引入可变形自注意力机制,能够有效提高目标检测模型的性能和灵活性.Deformable DETR 为未来的目标检测研究提供了新的方向,展示了 Transformer 在视觉任务中的广泛应用潜力.尽管该方法在捕捉长距离依赖关系和提供全局视野方面具有明显优势,但其计算复杂度较高,限制了模型的应用场景.

在语义分割任务中,PVT[27] 采用金字塔结构,通过自注意力机制在不同尺度上提取特征,增强了对各种尺度物体的检测能力.这种方法可以在图像的密集分区上训练,以实现高输出分辨率,这对于密集预测至关重要.然而,PVT 在处理高分辨率图像时的计算复杂度高,且需要大量的训练数据.

ViT Segmentation[28] 基于 Vision Transformer 架构,在不同尺度上进行特征提取,并通过特征融合策略整合这些信息,从而提高分割精度.该模型应用于语义分割任务时,利用全局自注意力机制实现高效特征提取,但超参数调整过程复杂,需要大量实验来优化性能.

SETR[29] 将输入图像视为一个序列,采用全局自注意力来处理图像特征,有效捕捉全局上下文信息和长距离依赖,并通过简单的特征转换层进行分割.这种方法能够更灵活地处理复杂场景,提高分割精度.然而,该方法在处理高分辨率图像时,模型计算量较大以至于模型达到收敛时所需的训练时间较长.5 个代表模型的比较如表 1 所示.

**表 1 基于全局特征提取的代表方法**

Table 1 Representative methods based on global feature extraction

| 方法 | 年份 | 特点 | 优点 | 缺点 |
|---|---|---|---|---|
| Segmenter[25] | 2021 | 通过图像块的输出嵌入获取类别标签,采用点对点线性解码器和掩码 Transformer 解码器. | 有效利用全局信息,在大模型和小图块尺寸配置下表现好. | 对计算资源需求较高,限制在某些特定场景中的应用. |
| DETR[26] | 2020 | 通过让注意力模块仅关注参考点周围的一小组关键采样点. | 在小物体检测上表现好,并且训练时收敛速度快. | 对关键采样点的选择和设置提出更高的要求,增加了模型的复杂性. |
| PVT[27] | 2021 | 通过在图像的密集分区上训练,实现高分辨率输出. | 能够直接替代卷积神经网络,分割精度高. | 依赖于精确的设计和实现,模型对计算资源的需求高. |
| ViT Segmentation[28] | 2022 | 通过通道连接融合不同特征图,并设计一个维度注意力模块聚合图像特征上下文信息. | 模型性能表现好,分割任务实现了优越的平均交并比. | 对计算资源的需求较高,增加了模型的复杂性. |
| SETR[29] | 2021 | 在每一层中建模全局上下文,并设计了一种简单的解码器. | 在分割测试中,性能表现上具有显著优势. | 对计算资源的需求较高,在实际应用中处理大规模图像时效率低. |

以 SETR[29] 为例,SETR 模型利用自注意力机制来有效捕捉图像中的全局上下文信息,从而提升语义分割的性能.实现过程中,图像被划分为多个像素块/补丁(Patch),并通过线性变换编码为特征向量.在每

个补丁中,模型计算其与其他补丁的相似度,生成注意力权重,从而确定各补丁在特征表示中的重要性.这些权重经过 Softmax 归一化后,用于对所有补丁的特征进行加权求和,形成新的特征表示.SETR 还采用多头注意力机制,通过在多个子空间中独立计算注意力,增强模型的表达能力.SETR 模型架构如图 2 所示.

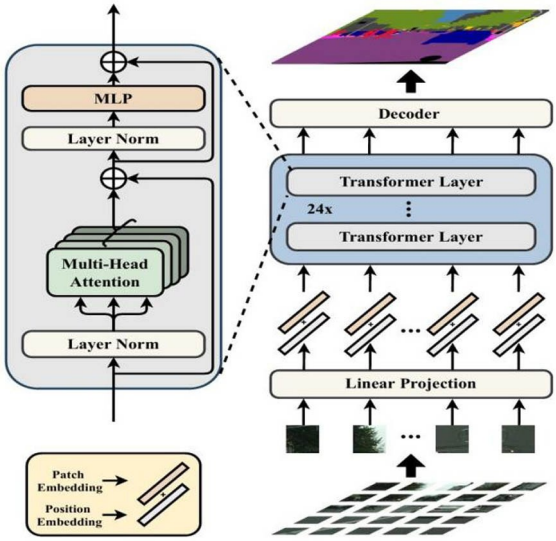


**图 2　SETR 模型架构[29](源于文献[29])**

Fig.2　SETR model architecture[29]

## 2.2　基于局部特征增强的方法

基于局部特征增强的方法利用 Transformer 的自注意力机制,结合局部特征与全局上下文信息,以提升道路场景语义分割的能力.这种方法特别适合处理复杂和细粒度的图像分割任务,能够有效捕捉图像中的细微局部信息,同时保持全局上下文的理解.局部特征增强方法强调多尺度特征的融合,通过从不同网络层提取特征并进行融合,以在保持局部细节的同时获取全局信息,尤其在包含多种元素的复杂道路场景中表现突出.此外,自注意力机制的应用使模型能够关注重要的局部区域,并根据上下文动态调整注意力权重,从而更好地理解局部区域的重要性及其与其他区域的关系.

Swin Transformer[30]引入了“移动窗口”自注意力机制,通过限制为不重叠的本地窗口,同时允许跨窗口连接,模型能够在不同层次上处理图像特征,从而有效捕捉局部和全局信息,提升了复杂道路场景中的分割性能.尽管该方法在细粒度分割和上下文理解方面具有显著优势,但也面临较高的计算开销和模型复杂性等挑战,其泛化能力在特定应用场景下仍需进一步验证.

PVTv2[31]继承了 PVT 的结构,该模型使用局部特征增强和金字塔结构,优化了多尺度特征提取过程.同时,使用自注意力机制可以有效捕捉全局上下文信息,保持较低的计算成本.模型适用于多种视觉任务,如目标检测和语义分割.尽管 PVTv2 在多个基准数据集上展现了优越的性能,但在处理高分辨率图像时的计算复杂度仍然较大.

UNetFormer[32]结合使用 U-Net 架构与 Transformer 架构,通过引入多尺度特征融合和自注意力机制,同时进行局部特征提取和全局上下文建模,有效捕捉图像的细节和上下文信息,该模型适用于医学图像分割和自动驾驶下的语义分割任务.然而,其在数据稀缺场景下的表现不佳,需要大量实验来优化性能.

TransBTS[33]结合了 Transformer 和 U-Net 的思想,通过引入与模态相关的交叉注意力机制,有效整合不同医学成像模态的信息,从而提高分割的准确性和可靠性.该方法通过局部特征提取和全局上下文建模进行医学图像分割,同时适用于道路场景语义分割任务.尽管该方法在多个图像数据集上表现出色,但模型对大量高质量标注数据的依赖限制了其在数据稀缺场景下的有效性.

GCViT[34]在 Vision Transformer 中引入了局部和全局自注意力机制,能够更好地建模图像中不同区域

之间的关系,从而提升对复杂场景的理解能力,提高细节捕捉能力和上下文理解.然而,在处理高分辨率图像时,训练和推理时间会显著增加.5 个代表模型的比较如表 2 所示.

**表 2 基于局部特征增强的代表方法**

Table 2 Representative methods based on local feature enhancement

| 方法 | 年份 | 特点 | 优点 | 缺点 |
|---|---|---|---|---|
| Swin Transformer[30] | 2021 | 采用分层结构和移动窗口机制,并且将自注意力计算限制在不重叠的局部窗口内. | 在处理图像时具有线性计算复杂度,适用于多种视觉任务. | 对大规模数据集的训练需求较高,模型在某些应用场景中的实际部署较难. |
| PVTv2[31] | 2022 | 引入线性复杂度的注意力层、重叠补丁嵌入和卷积前馈网络. | 在分类、检测和分割等基础视觉任务上表现出色. | 模型训练的计算资源需求较高,尤其在大规模数据集上. |
| UNetFormer[32] | 2022 | 选择轻量级 ResNet18 作为编码器,设计高效的全局-局部注意力机制. | 在运行速度和准确性上表现良好. | 对特定硬件的依赖性,在更复杂的场景中可能面临性能瓶颈. |
| TransBTS[33] | 2023 | 将多模态图像按成像原理分为两组,采用双分支混合编码器和模态相关的交叉注意力机制提取特征. | 具有优越的分割性能. | 模型复杂性高,对计算资源的需求较高. |
| GCViT[34] | 2023 | 结合全局上下文自注意力与标准局部自注意力模块,使用改进的融合倒残差块来增强模型的归纳偏置. | 在图像分类、物体检测和语义分割任务上均取得了良好的结果. | 模型复杂性高,对计算资源需求高,影响在某些实时应用场景中的表现. |

以 GCViT[34] 为例,GCViT 模型的自注意力机制结合了全局上下文自注意力模块与标准局部自注意力模块,旨在有效建模长短距离空间交互.在实现过程中,查询生成器在每个阶段提取全局查询令牌,这些令牌通过与局部键值表示的交互,捕捉图像中的长距离信息.这种设计避免了传统方法中计算注意力掩码或移动局部窗口的昂贵操作,从而提高了计算效率.此外,GCViT 还引入了改进的融合倒残差块,增强了模型的归纳偏置,使得特征提取更加精细和全面.GCViT 模型架构如图 3 所示.

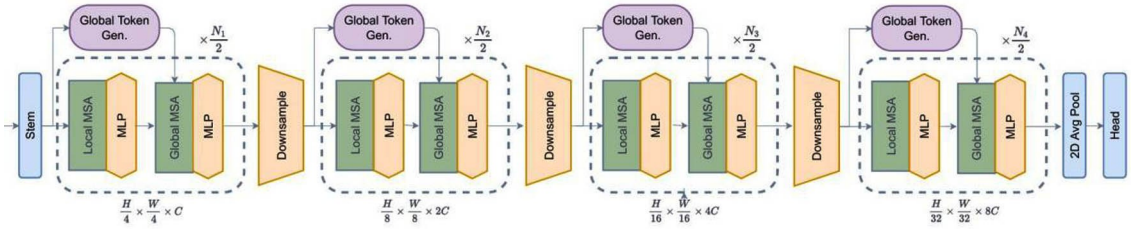


**图 3 GCViT 模型架构[34](源于文献[34])**

Fig.3 GCViT model architecture[34]

图 3 中,Local MSA 表示局部多头自注意力机制,Downsample 表示采用改进的 Fused-MBConv 块作为下采样操作,Global Token Gen 是用于提取全局查询令牌的一种机制.

## 2.3 基于混合架构的方法

基于混合架构的方法结合了卷积神经网络(CNN)和 Transformer 的优点,以提升道路场景语义分割的性能.混合架构通常采用 CNN 作为特征提取的基础,利用卷积层高效捕捉局部特征,随后将提取的特征传递给 Transformer 模块,以增强全局特征表示能力.这种结构使模型能够在处理复杂场景时保持细节信息,同时理解整体上下文.在特征融合策略上,通过在不同层次上结合卷积层和 Transformer 的输出,形成丰富的特征表示,提升了分割精度和模型鲁棒性.

SegFormer[35] 结合了 Transformer 与卷积层,采用层次化结构,结合多尺度特征提取和轻量级 Transformer 模块,能够有效捕捉图像的全局和局部信息,适用于高效的语义分割.但是,即使该模型是最轻量的模型,对于一些边缘设备而言,仍可能存在计算和存储资源的需求过高的问题;而且模型依赖于大量高质量标注数据,在数据稀缺的情况下,其效果会受到影响.

TransDeepLabV3+[36]结合了 DeepLabV3+的特征提取能力与 Transformer 的全局上下文建模,将 ASPP 模块替换为 Swin Transformer 块,以增强特征提取能力.该模型利用基于 CBAM 注意力机制的通道与空间融合注意力(CSFA)机制,提升了分割性能.

ViT-UNet[37]将 ViT 与 U-Net 架构结合,通过引入 Vision Transformer(ViT)来替换 U-Net 编码器中的卷积层,实现高效的特征提取和分割,能够有效提取分类所需的细节信息.但该模型的计算复杂度较高,特别是在处理高分辨率图像时.

RTS R-CNN[38]将 Transformer 集成到 Faster R-CNN 框架中,在特征提取阶段通过增加通道间交互来提高深度卷积网络的性能.其次,利用残差特征增强模块(RFA)和空洞空间金字塔池化(ASPP)来优化并行特征金字塔网络(PAFPN),并引入平衡特征金字塔模块(BFP)来处理不同分辨率下的不平衡特征信息,增强目标检测和语义分割性能,能够准确识别和定位复杂道路场景中的物体.但 R-CNN 网络的计算复杂度较高,整体计算量较大.

CCTNet[39]结合了卷积神经网络和 Transformer 的优势,通过紧凑的结构实现高效的特征提取.利用 CNN 的局部特征提取能力和 Transformer 的全局上下文建模能力,能够有效处理复杂的道路图像数据,提升分割精度.但该模型在处理高分辨率图像时,具有较高的计算复杂度,需要花费大量的计算资源和时间.5 个代表模型的比较如表 3 所示.

**表 3　基于混合架构的代表方法**

Table 3　Representative methods based on hybrid architectures

| 方法 | 年份 | 特点 | 优点 | 缺点 |
|---|---|---|---|---|
| SegFormer[35] | 2021 | 采用新颖的分层结构 Transformer 编码器,输出多尺度特征,解码器设计简单. | 轻量化设计提升了分割效率. | 在处理复杂场景时的表现不如一些更复杂的模型. |
| TransDeepLabV3+[36] | 2024 | 用 Swin-Transformer 块替换了 ASPP 模块,采用基于 CBAM 注意力机制的通道和空间融合注意力机制. | 显著提升了小目标的检测性能. | 计算成本较高. |
| ViT-UNet[37] | 2024 | 将 Transformer 的输出与内部注意特征链接到相应的解码层,通过双线性聚合池化融合提取的特征信息. | 分割精度高、效果好. | 复杂结构导致计算成本增加. |
| RTS R-CNN[38] | 2023 | 在特征提取阶段采用 CSPDarkNet53_ECA,在特征融合部分引入 GR-PAFPN. | 提高了小目标检测的准确性和整体性能 | 模型复杂性增加了计算资源需求. |
| CCTNet[39] | 2024 | 采用编码器-解码器结构,基于高效的交叉自注意力构建解码器. | 能够实现高效的特征提取和轻量化设计. | 计算复杂度高、资源消耗较大. |

以 SegFormer[35]为例,SegFormer 模型在编码器部分设计了一种分层注意力机制,以捕捉图像的多尺度特征.接着,SegFormer 通过轻量级的多层感知机(MLP)解码器来聚合来自编码器不同层的信息,有效结合局部注意力和全局注意力,以生成强大的特征表示.这种简单而高效的设计使得 Transformer 在语义分割任务中表现出色,同时保持了较低的参数量.SegFormer 模型架构如图 4 所示.

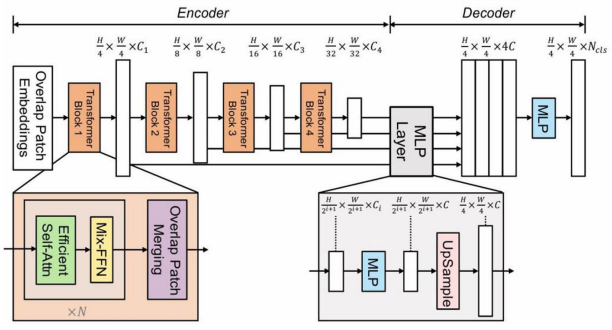


**图 4　SegFormer 模型架构[35](源于文献[35])**

Fig. 4　SegFormer model architecture[35]

图 4 中,Efficient Self-Attention 表示优化的自注意力机制,Mix-FFN 表示混合前馈网络,Overlapped Patch Merging 表示特征融合策略.

### 2.4 基于自监督学习的方法

基于自监督学习的方法在道路场景语义分割中,通过减少对标注数据的依赖,利用未标注数据进行训练,从而提升模型性能.这种方法显著降低了对大量像素级标注的需求,同时能够更有效地利用可用数据.自监督学习是一种无监督学习形式,通过自动生成标签,使模型从未标注的数据中学习特征,通常依赖于数据本身的结构和特性.模型可以通过设计辅助任务,如图像重建和对比学习,来提取有用的特征.

DINO[40](一种无标签的自蒸馏方法)展示了自监督学习在特征提取和语义分割任务中的有效性.该方法通过利用类感知的相似性,增强模型对不同类别的理解,从而在缺少详细标注的情况下实现更准确的分割.引入语义引导机制,使模型能够更好地利用未标记数据和弱标注信息,从而提升分割精度.尽管该方法在减少标注需求和利用未标注数据方面具有明显优势,但任务设计的复杂性和训练稳定性仍然是其面临的挑战.

BEIT[41]通过自监督学习任务生成图像的视觉表征,将原始图像“标记化”为视觉标记.然后随机屏蔽一些图像补丁并将它们输入到主干 Transformer 中.在预训练 BEIT 编码器之后,通过在预训练编码器上附加任务层来直接微调下游任务的模型参数,适用于多种视觉任务,包括语义分割.然而,模型的超参数调整过程较为复杂,这在实际应用中可能会增加工作量.

L-MAE[42]通过对输入图像进行遮蔽并重建未遮蔽部分来学习图像特征,利用标签中的现有信息生成完整的标签.采用将标签和相应图像堆叠的融合策略,提出了 Image Patch Supplement 算法以补充掩模重建过程中的缺失信息,适用于后续的分割任务.然而,该方法对数据质量的依赖较高,若原始数据集存在噪声或标注不准确,可能会影响增强效果和最终的分割性能.

G-SimCLR[43]结合了伪标签技术,通过引导投影优化对比学习过程,提升模型在无标签数据上的特征学习能力.与 Transformer 结合应用于自监督学习任务,增强特征提取能力.G-SimCLR 在自监督学习中取得了积极成果,但模型对伪标签的质量要求很高.

Contextual Transformer[44]通过自监督学习的方法,利用上下文信息增强特征表示.通过协同学习策略,利用多个视角的信息进行模型训练,从而增强对复杂道路图像数据的特征学习能力,适用于图像分割和其他视觉任务.然而,模型的协同训练对数据质量要求较高,若输入数据存在噪声或标注不准确,可能影响最终结果的准确性.5 个代表模型的比较如表 4 所示.

**表 4 基于自监督学习的代表方法**

Table 4 Representative methods based on self-supervised learning

| 方法 | 年份 | 特点 | 优点 | 缺点 |
|---|---|---|---|---|
| DINO[40] | 2021 | 设计动量编码器,进行多裁剪和小块处理. | 具有强大的特征表达能力. | 需要合理的训练策略和架构设计才能发挥其优势. |
| BEIT[41] | 2021 | 将原始图像“标记化”为视觉标记,并随机掩蔽一些图像块,通过主干变换器进行训练. | 具有良好的特征学习能力和灵活性. | 对掩蔽策略和训练数据的选择有一定依赖,影响其在特定任务上的表现. |
| L-MAE[42] | 2022 | 采用融合策略,将标签与对应图像叠加为融合图,并提出了图像块补充算法. | 在数据集补全方面表现出色,能有效提升模型性能. | 依赖于标签质量和掩蔽策略,且在特定应用场景中仍需验证其通用性. |
| G-SimCLR[43] | 2020 | 使用去噪自编码器生成的潜在空间表示进行聚类,获得伪标签. | 在不依赖大量标注数据的情况下,能够提升模型的表示能力. | 效果受到聚类质量和去噪自编码器性能的影响. |
| Contextual Transformer[44] | 2024 | 引入灵活的共训练程序,能够利用多个标注的通用数据集进行训练. | 模型的适应性和泛化能力方面表现出色. | 效果依赖于数据集的质量和任务间的知识共享效率. |

以 BEIT[41]为例,BEIT 模型采用自注意力机制,通过掩蔽图像建模任务来预训练视觉变换器.具体而言,模型首先将原始图像“标记化”为视觉标记,并将其划分为多个图像块(例如 16×16 像素).在预训练过

程中,随机掩蔽一些图像块后,将其输入到主干变换器中.自注意力机制在此过程中发挥关键作用,它能够动态地关注输入特征的不同部分,计算块之间的相似性,从而在重建过程中恢复被掩蔽的视觉标记.通过这种方式,BEIT 利用自注意力机制有效地捕捉图像的全局上下文信息,提升了特征表示的质量.预训练完成后,模型可以通过附加任务层直接在下游任务上进行微调,实现图像分类和语义分割等任务时的竞争性表现.BEIT 模型架构如图 5 所示.

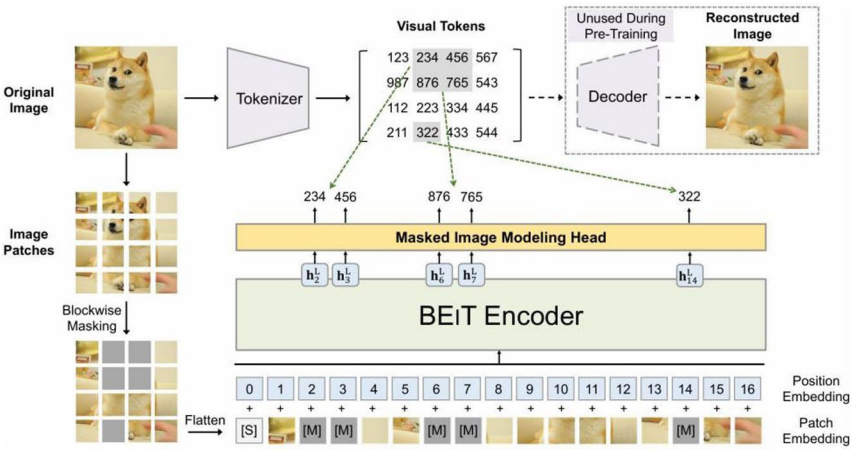


**图 5　BEIT 模型架构[41](源于文献[41])**

Fig.5　BEIT model architecture[41]

图 5 中,Tokenizer 表示将输入图像转换为模型可以处理的格式,Visual Tokens 表示将输入图像分割成小块后生成的离散表示,Blockwise Masking 表示 BEIT 预训练过程中采用的一种策略,Masked Image Modeling Head 表示用于处理掩蔽图像建模任务的特定网络组件.

## 3　道路场景数据集及性能评价

### 3.1　道路场景数据集

研究道路场景的大量人员首先通过实地拍摄和合成数据的方式收集图像和视频,以确保数据的多样性.接下来,对于小规模数据集,采用手动标注方式以确保高准确性;而对于大规模数据集,则利用众包平台将任务分配给大量用户进行标注.接着,使用去噪声和标准化过程对数据进行清理与处理.随后,将数据集划分为训练集、验证集和测试集三部分.通过这些步骤,研究人员构建出高质量的道路场景数据集.高质量道路场景数据集如表 5 所示.

**表 5　道路场景数据集**

Table 5　Road scene datasets

| 数据集 | 类别数 | 样本数 | 地区 | 环境 |
|---|---|---|---|---|
| KITTI[45] | 10 | 40 000 | 德国、美国 | 白天 |
| Cityscapes[46] | 34 | 20 000 | 法国、德国和瑞士 | 春、夏、秋 |
| Camvid[47] | 32 | 700 | 欧洲 | 白天 |
| Apollo Scape[48] | 28 | 142 906 | 中国 | 复杂天气 |
| BDD100K[49] | 10 | 10 000 | 世界上多个城市 | 多种场景 |
| NuScenes[50] | 23 | 1 400 000 | 波士顿和新加坡 | 白天 |
| Mapillary Vistas[51] | 66 | 25 000 | 欧洲、美洲、亚洲、非洲和大洋洲 | 复杂天气 |
| ADE20K[52] | 150 | 20 210 | 美国和加拿大 | 白天 |
| COCO 2017[53] | 80 | 118 287 | 美国 | 多种场景 |

### 3.2 性能评价指标

语义分割性能的评价指标主要用于衡量模型在图像中对不同物体类别的分割效果.常用准确度和运算速度两个指标评估分割网络的性能.

准确度:在道路场景语义分割任务中,一些研究者使用像素准确率(PA)及其平均值(mPA)来评估像素级语义分割性能的好坏.此外,另一些研究者则使用平均交并比(mIoU)作为评价模型效果的重要参考,来评估语义分割模型的性能.mIoU 通过先计算所有类别的交并比(IoU),然后取平均得到.其中,IoU 通过计算[模型正确预测的样本数量]/[真实样本数量+预测样本数量-正确预测的样本数量]得到.这一方法有效地综合了模型的分割效果,帮助评估其在实际应用中的表现.

运算速度:在道路场景语义分割任务中,运行时间是一个非常重要的性能标准.运行时间反映语义分割任务的实时性.通常使用单位时间内(1 秒)能够处理的图像帧数(FPS)来间接评估计算速度.FPS 通过训练一轮时所处理的图像数据除以所用时间获得.

### 3.3 算法性能对比

本文对基于 Transformer 的道路场景语义分割网络进行分析,主要从平均交并比 mIoU 和平均准确率 mPA 两个方面评估网络性能,对道路场景语义分割的结果进行深入探讨.这些分割方法的对比如表 6 所示.

**表 6 基于 Transformer 的道路场景语义分割方法对比**

Table 6 Comparisons of road scene semantic segmentation methods based on Transformer

| 方法 | 分类 | 主干网络 | Cityscapes (mIoU(%)) | ADE20K (mIoU(%)) | COCO 2017 (mPA(%)) |
|---|---|---|---|---|---|
| Segmenter[25] | 全局特征提取 | ViT-L/16 | 81.3 | 53.63 | |
| DETR[26] | | ResNet-101+DCN | | | 54.4 |
| PVT[27] | | PVT-Large | | 44.8 | 61.9 |
| SETR[29] | | T-Large | 82.15 | 50.28 | |
| Swin Transformer[30] | 局部特征增强 | Swin-L | | 53.5 | 51.1 |
| PVTv2[31] | | PVTv2-B5 | | 48.7 | 65.7 |
| UNetFormer[32] | | ResNet-50+ViT | 84.1 | | |
| GCViT[34] | | GCViT-B | | 49.2 | 69.2 |
| SegFormer[35] | 混合架构 | MiT-B5 | 84.0 | 51.8 | |
| TransDeepLabV3+[36] | | ResNet-101 | 84.9 | 55.1 | |
| ViT-UNet[37] | | ViT | 87.3 | | |
| RTS R-CNN[38] | | ResNet-101 | 86.56 | | |
| CCTNet[39] | | ResNet-18 | 83.39 | | |
| DINO[40] | 自监督学习 | ViT-B/8 | 80.1 | 44.1 | |
| BEIT[41] | | BEIT-B | | 47.7 | |
| L-MAE[42] | | ViT | 85.6 | | |
| Contextual Transformer[44] | | Swin-L | 81.6 | 46.3 | |

注:表 6 中空白处表示模型没有在该数据集进行测试.

从表 6 可以看出,针对道路场景语义分割,基于 Cityscapes 数据集,Segmenter、SETR、UNetFormer、SegFormer、TransDeepLabV3 +、ViT-UNet、RTS R-CNN、CCTNet、DINO、L-MAE 和 Contextual Transformer 网络的平均交并比 mIoU 值均达到了 80%以上,表明这些模型在识别和分割不同类别的对象方面表现良好,能够有效地处理复杂的道路场景,满足实际应用的需求.其中,基于混合架构的方法:SegFormer、TransDeepLabV3+、ViT-UNet、RTS R-CNN、CCTNet 的 mIoU 值相对较好,特别是 ViT-UNet 的 mIoU 值达到了 87.3,这表明基于混合架构的方法在捕捉细节和全局上下文信息方面具有显著优势,展现出很好的发展前景,为未来自动驾驶和智能交通系统的研究提供了强大的解决方案.

基于 ADE20K 数据集,Segmenter、PVT、SETR、Swin Transformer、PVTv2、GCViT、SegFormer、TransDee-

pLabV3+、DINO、BEIT 和 Contextual Transformer 网络的 mIoU 值在 44 到 56 之间,这是由于 ADE20K 数据集包含多种场景,涵盖城市、农村、室内等多种环境,并且包含较多的语义类别;虽然类别丰富,但在道路场景的细节上标注不够精细.相较之下,Cityscapes 数据集专注于城市街道场景,聚焦于与城市交通相关的类别,更加关注道路场景细节,提供高质量的像素级标注,特别适合城市道路场景的研究.

基于 COCO 2017 数据集,DETR、PVT、Swin Transformer、PVTv2 和 GCViT 网络中,GCViT 的平均准确率 mPA 值达到了 69.2,这表明 GCViT 网络在该数据集上的表现优于其他模型,显示出其在道路场景分割任务中的相对优越性.

针对运算速度,大部分文献并未专门做对比分析,因此本文仅列出包含有该指标的几种代表性算法,如表 7 所示.鉴于各算法在模型训练和测试时的硬件配置、网络参数设置等各不相同,因此表格中的 FPS 指标横向对比价值不大,仅供参考.

**表 7　算法速度分析**

Table 7　Algorithm speed analysis

| 方法 | 主干网络 | Cityscapes | | ADE20K | | COCO 2017 | |
|---|---|---|---|---|---|---|---|
| | | FPS | mIoU(%) | FPS | mIoU(%) | FPS | mPA(%) |
| DETR[26] | ResNet-101+DCN | | | | | 19 | 54.4 |
| SETR[29] | T-Large | 0.5 | 82.15 | 5.4 | 50.28 | | |
| Swin Transformer[30] | Swin-L | | | 6.2 | 53.5 | 9.9 | 51.1 |
| SegFormer[35] | MiT-B5 | 2.5 | 84.0 | 9.8 | 51.8 | | |

注:表 7 中准确度数据和表 6 一致;表 7 中空白处表示模型没有在该数据集进行测试.

## 4　总结与展望

本文介绍了基于 Transformer 的道路场景语义分割方法,根据不同的特征处理策略和模型架构,将基于 Transformer 的道路场景语义分割方法分为四类:基于全局特征提取的方法、基于局部特征增强的方法、基于混合架构的方法以及基于自监督学习的方法,概括总结了各类方法的技术特点和优缺点,分析对比了这些主流分割方法在不同数据集上的性能指标及分割效果.

与基于深度学习的传统道路场景语义分割算法相比,基于 Transformer 的道路场景语义分割方法更擅长处理全局信息,能够全面理解复杂场景中的上下文关系,具有获取多对象和高复杂度环境中更多、更高级的语义信息的优势,能够更精确地表达图像中的内容,展现出良好的发展前景.

### 4.1　存在的问题

首先,在处理高分辨率图像时,计算量较大且计算复杂度较高,尤其是在处理大规模数据集时,模型训练和推理的时间成本显著增加.同时,模型的超参数调整过程可能较为烦琐,需要进行多次实验以优化性能,这会极大影响实际应用中的部署;其次,部分模型对局部特征的捕捉能力不足,可能导致在复杂场景下的分割精度下降;再者,模型对数据标注的依赖较大,训练数据的多样性和标注质量对模型性能影响显著,尤其是在不同道路条件下,数据集的不足和标注错误可能导致模型泛化能力降低.此外,模型的可解释性仍然是一个亟待解决的问题,尤其是在自动驾驶等安全要求较高的应用场景中,理解模型的决策过程至关重要.

### 4.2　未来的发展方向

基于 Transformer 的道路场景语义分割方法通常需要大量的计算资源,这导致分割效率较低.因此,如何在保持模型性能的同时提高分割效率,成为当前的主要挑战.

未来,基于 Transformer 的道路场景语义分割方法可以朝以下几个方向发展:首先,探索更高效的模型架构,以降低计算复杂度,例如引入新型的轻量化网络结构或优化算法;其次,增强模型对局部特征的捕捉能力,并结合多尺度特征融合技术,以提高在复杂场景中的分割精度;此外,利用自监督学习和生成对抗网络等技术,提升数据标注的效率和质量,进而增强模型的泛化能力;最后,进一步研究模型的可解释性,通过可视化技术或解释模型的方法,帮助理解模型决策背后的逻辑,为安全要求较高的应用提供更好的支持.

## 参考文献


[1]AL-SHAYEA Q K.Artificial neural networks in medical diagnosis[J].International Journal of Computer Science Issues,2011,8(2):150-154.

[2]GEIGER A,LENZ P,URTASUN R.Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012. Providence, RI: IEEE,2012:3354-3361.

[3]王龙飞,严春满.道路场景语义分割综述[J].激光与光电子学进展,2021,58(12):44-66.

[4]DENG L Y,YANG M,QIAN Y Q,et al.CNN based semantic segmentation for urban traffic scenes using fisheye camera[C]//2017 IEEE Intelligent Vehicles Symposium (IV). Los Angeles, CA, USA: IEEE, 2017:231-236.

[5]HAN K,WANG Y,CHEN H,et al.A survey on visual transformer[J].arXiv preprint:2012. 12556,2020.

[6]CREMERS D,ROUSSON M,DERICHE R.A review of statistical approaches to level set segmentation:Integrating color,texture,motion and shape[J].International Journal of Computer Vision,2007,72(2):195-215.

[7]李怡静,胡翔云,张剑清,等.影像与 LiDAR 数据信息融合复杂场景下的道路自动提取[J].测绘学报,2012,41(6):870-876.

[8]YANG W,ZHANG X,CHEN L J,et al.Semantic segmentation of polarimetric SAR imagery using conditional random fields[C]//2010 IEEE International Geoscience and Remote Sensing Symposium. Honolulu, HI,USA:IEEE,2010:1593-1596.

[9]张祥甫,刘健,石章松,等.基于深度学习的语义分割问题研究综述[J].激光与光电子学进展,2019,56(15):20-34.

[10]LIU D P,ZHANG D,WANG L,et al.Semantic segmentation of autonomous driving scenes based on multi-scale adaptive attention mechanism[J].Frontiers in Neuroscience,2023,17:1291674.

[11]LUO J B,WANG Q H,ZOU R R,et al.A heart image segmentation method based on position attention mechanism and inverted pyramid[J].Sensors,2023,23(23):9366.

[12]谷湘煜,刘晓熠,周仁彬.多特征融合的道路场景语义分割算法[J].科学技术与工程,2021,21(33):14251-14257.

[13]FAN J H,BOCUS M J,HOSKING B,et al.Multi-scale feature fusion:Learning better semantic segmentation for road pothole detection[C]//2021 IEEE International Conference on Autonomous Systems (ICAS). Montreal,QC,Canada:IEEE,2021:1-5.

[14]CHENG Z M,QU A P,HE X F.Contour-aware semantic segmentation network with spatial attention mechanism for medical image[J].The Visual Computer,2022,38(3):749-762.

[15]LONG J,SHELHAMER E,DARRELL T.Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.2015:3431-3440.

[16]SUGIRTHA T,SRIDEVI M.Semantic segmentation using modified U-Net for autonomous driving[C]//2022 IEEE International IOT,Electronics and Mechatronics Conference (IEMTRONICS). Toronto, ON, Canada:IEEE,2022:1-7.

[17]易清明,张文婷,石敏,等.多尺度特征融合的道路场景语义分割[J].激光与光电子学进展,2023,60(12):92-100.

[18]WANG W,HE H,MA C S.An improved Deeplabv3+ model for semantic segmentation of urban environments targeting autonomous driving[J].International Journal of Computers Communications & Control,2023,18(6):1-17.

[19]陈晔,杨长春,杨森,等.融合位置注意力机制与轻量化 STDC 网络的非结构化场景语义分割[J].计算机系统应用,2024,33(04):254-262.

[20]FANG S Q,ZHANG B,HU J Y.Improved mask R-CNN multi-target detection and segmentation for autonomous driving in complex scenes[J].Sensors,2023,23(8):3853.

[21]LIU T R,STATHAKI T.Faster R-CNN for robust pedestrian detection using semantic segmentation network[J].Frontiers in Neurorobotics,2018,12:64.

[22]KIM J H,LEE S H , HAN H H.Modified pyramid scene parsing network with deep learning based multi scale attention[J].Journal of the Korea Convergence Society,2021. 12(11):45-51.

[23]WU H S,LIANG C X,LIU M S,et al.Optimized HRNet for image semantic segmentation[J].Expert Systems with Applications,2021,174:114532.

[24]KANG Y,CAI Z,TAN C W,et al.Natural language processing (NLP) in management research:A literature review[J].Journal of Management Analytics,2020,7(2):139-172.

[25]STRUDEL R,GARCIA R,LAPTEV I,et al.Segmenter:Transformer for semantic segmentation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV).Montreal,QC,Canada:IEEE,2021:7242-7252.

[26]ZHU X,SU W,LU L,et al.Deformable detr:Deformable transformers for end-to-end object detection[J].arXiv preprint:2010. 04159,2020.

[27]WANG W H,XIE E Z,LI X,et al.Pyramid vision transformer:A versatile backbone for dense prediction without convolutions[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal,QC,Canada:IEEE,2021:548-558.

[28]WANG W,TANG C,WANG X,et al.A ViT-based multiscale feature fusion approach for remote sensing image segmentation[J].IEEE Geoscience and Remote Sensing Letters,2022,19:1-5.

[29]ZHENG S X,LU J C,ZHAO H S,et al.Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Nashville,TN,USA:IEEE,2021:6877-6886.

[30]LIU Z,LIN Y T,CAO Y,et al.Swin transformer:Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV).Montreal,QC,Canada:IEEE,

2021:9992-10002.

[31]WANG W H,XIE E Z,LI X,et al.PVT v2:Improved baselines with pyramid vision transformer[J].Computational Visual Media,2022,8(3):415-424.

[32]WANG L B,LI R,ZHANG C,et al.UNetFormer:A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery[J].ISPRS Journal of Photogrammetry and Remote Sensing,2022,190:196-214.

[33]LIN J W,LIN J T,LU C,et al.CKD-TransBTS:Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation[J].IEEE Transactions on Medical Imaging,2023,42(8):2451-2461.

[34]HATAMIZADEH A,YIN H X,HEINRICH G,et al.Global context vision transformers[C]//Proceedings of the 40th International Conference on Machine Learning.Honolulu,Hawaii,USA:ACM,2023:12633-12646.

[35]XIE E Z,WANG W H,YU Z D,et al.SegFormer:Simple and efficient design for semantic segmentation with transformers[J].Advances in Neural Information Processing Systems,2021,34:12077-12090.

[36]AILING D,JIANFENG W,SHANGZHEN S,et al.Semantic Segmentation of Road Traffic Sign based on Improved Deeplabv3+[C]//2024 9th International Conference on Computer and Communication Systems(ICCCS).IEEE,2024:149-154.

[37]ZHOU N,XU M M,SHEN B Q,et al.ViT-UNet:A vision transformer based UNet model for coastal wetland classification based on high spatial resolution imagery[J].IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2024,17:19575-19587.

[38]ZHANG G R,PENG Y M,WANG H.Road traffic sign detection method based on RTS R-CNN instance segmentation network[J].Sensors,2023,23(14):6543.

[39]WU H L,ZENG Z B,HUANG P,et al.CCTNet:CNN and cross-shaped transformer hybrid network for remote sensing image semantic segmentation[J].IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2024,17:19986-19997.

[40]CARON M,TOUVRON H,MISRA I,et al.Emerging properties in self-supervised vision transformers[C]//2021 IEEE/CVF International Conference on Computer Vision(ICCV).Montreal,QC,Canada:IEEE,2021:9630-9640.

[41]BAO H,DONG L,PIAO S,et al.Beit:Bert pre-training of image transformers[J].arXiv preprint:2106.08254,2021.

[42]JIA J R,LIU M Y,XIE J K ,et al.L-MAE:Masked Autoencoders are Semantic Segmentation Datasets Augmenter[J].arXiv preprint:2211.11242,2022.

[43]CHAKRABORTY S,GOSTHIPATY A R,PAUL S.G-SimCLR:Self-supervised contrastive learning with guided projection *via* pseudo labelling[C]//2020 International Conference on Data Mining Workshops(ICDMW).Sorrento,Italy:IEEE,2020:912-916.

[44]LI Q Y,CHEN Y S,HE X,et al.Co-training transformer for remote sensing image classification,segmentation,and detection[J].IEEE Transactions on Geoscience and Remote Sensing,2024,62:1-18.

[45]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.

[46]CORDTS M,OMRAN M,RAMOS S,et al.The cityscapes dataset for semantic urban scene understanding[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR).Las Vegas,NV,USA:IEEE,2016:3213-3223.

[47]BROSTOW G J,FAUQUEUR J,CIPOLLA R.Semantic object classes in video:A high-definition ground truth database[J].Pattern Recognition Letters,2009,30(2):88-97.

[48]HUANG X Y,CHENG X J,GENG Q C,et al.The ApolloScape dataset for autonomous driving[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPRW).Salt Lake City,UT:IEEE,2018:1067-10676.

[49]YU F,CHEN H F,WANG X,et al.BDD100K:A diverse driving dataset for heterogeneous multitask learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR).Seattle,WA,USA:IEEE,2020:2633-2642.

[50]CAESAR H,BANKITI V,LANG A H,et al.nuScenes:A multimodal dataset for autonomous driving[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR).Seattle,WA,USA:IEEE,2020:11618-11628.

[51]NEUHOLD G,OLLMANN T,BULO S R,et al.The mapillary vistas dataset for semantic understanding of street scenes[C]//2017 IEEE International Conference on Computer Vision(ICCV).Venice:IEEE,2017:5000-5009.

[52]ZHOU B L,ZHAO H,PUIG X,et al.Scene parsing through ADE20K dataset[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR).Honolulu,HI:IEEE,2017:5122-5130.

[53]LIN T Y,MAIRE M,BELONGIE S,et al.Microsoft COCO:Common objects in context[C]// Computer Vision - ECCV 2014.Cham:Springer International Publishing,2014:740-755.

(责任编辑:张阳,殷锋,付强,和力新,肖丽;英文编辑:周序林,郑玉才)