ISSN 1003-501X CN 51-1346/O4

光电工程

Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

面向道路场景语义分割的移动窗口变换神经网络设计

杭昊，黄影平，张栩瑞，罗鑫

引用本文:

杭昊，黄影平，张栩瑞，等. 面向道路场景语义分割的移动窗口变换神经网络设计[J]. 光电工程，2024， **51**(1): 230304.

Hang H, Huang Y P, Zhang X R, et al. Design of Swin Transformer for semantic segmentation of road scenes[J]. *Opto-Electron Eng*, 2024, **51**(1): 230304.

https://doi.org/10.12086/oee.2024.230304

收稿日期: 2023-12-14; 修改日期: 2024-01-23; 录用日期: 2024-01-24

相关论文

角度差异强化的光场图像超分网络

吕天琪，武迎春，赵贤凌

光电工程 2023, **50**(2): 220185 doi: 10.12086/oee.2023.220185

基于自相似特征增强网络结构的图像超分辨率重建

汪荣贵，雷辉，杨娟，薛丽霞

光电工程 2022, **49**(5): 210382 doi: 10.12086/oee.2022.210382

群稀疏高斯洛伦兹混合先验超分辨率重建

马子杰，赵玺竣，任国强，雷涛，杨虎，刘盾

光电工程 2021, **48**(11): 210299 doi: 10.12086/oee.2021.210299

基于通道注意力与迁移学习的红外图像超分辨率重建算法

孙锐，章晗，程志康，张旭东

光电工程 2021, **48**(1): 200045 doi: 10.12086/oee.2021.200045

更多相关论文见光电期刊集群网站

OEE 光电工程 Opto-Electronic Engineering

http://cn.oejournal.org/oee

OE_Journal Website

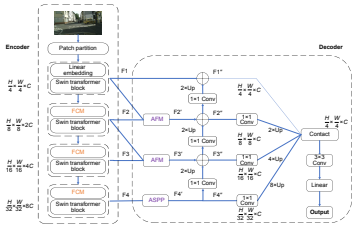
Opto-Electronic Engineering
光 电 工 程

Article

2024 年, 第 51 卷, 第 1 期

DOI: 10.12086/oee.2024.230304

面向道路场景语义分割的移动窗口变换神经网络设计



杭　昊，黄影平*，张栩瑞，罗　鑫

上海理工大学光电信息与计算机工程学院，上海 200093

摘要：道路场景语义分割是自动驾驶环境感知的一项重要任务。近年来，变换神经网络 (Transformer) 在计算机视觉领域开始应用并取得了很好的效果。针对复杂场景图像语义分割精度低、细小目标识别能力不足等问题，本文提出了一种基于移动窗口 Transformer 的多尺度特征融合的道路场景语义分割算法。该网络采用编码-解码结构，编码器使用改进后的移动窗口 Transformer 特征提取器对道路场景图像进行特征提取，解码器由注意力融合模块和特征金字塔网络构成，充分融合多尺度的语义特征。在 Cityscapes 城市道路场景数据集上进行验证测试，实验结果表明，与多种现有的语义分割算法进行对比，本文方法在分割精度方面有较大的提升。

关键词：语义分割；移动窗口变换神经网络；注意力机制；自动驾驶；深度学习

中图分类号：TP391.4　　**文献标志码：**A

杭昊，黄影平，张栩瑞，等. 面向道路场景语义分割的移动窗口变换神经网络设计 [J]. 光电工程，2024，**51**(1): 230304

Hang H, Huang Y P, Zhang X R, et al. Design of Swin Transformer for semantic segmentation of road scenes[J]. *Opto-Electron Eng*, 2024, **51**(1): 230304

Design of Swin Transformer for semantic segmentation of road scenes

Hang Hao, Huang Yingping*, Zhang Xurui, Luo Xin

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract: Road scene semantic segmentation is a crucial task in autonomous driving environment perception. In recent years, Transformer neural networks have been applied in the field of computer vision and have shown excellent performance. Addressing issues such as low semantic segmentation accuracy in complex scene images and insufficient recognition capabilities for small objects, this paper proposes a road scene semantic segmentation algorithm based on Swin Transformer with multiscale feature fusion. The network adopts an encoder-decoder structure, where the encoder utilizes an improved Swin Transformer feature extractor for road scene image feature extraction. The decoder consists of an attention fusion module and a feature pyramid network, effectively integrating semantic features at multiple scales. Validation tests on the Cityscapes urban road scene dataset show that, compared to various existing semantic segmentation algorithms, our approach demonstrates significant improvement in segmentation accuracy.

Keywords: semantic segmentation; Swin Transformer; attention mechanism; autonomous driving; deep learning

收稿日期：2023-12-14；修回日期：2024-01-23；录用日期：2024-01-24

基金项目：国家自然科学基金资助项目 (62276167)

*通信作者：黄影平，huangyingping@usst.edu.cn。

1 引 言

图像语义分割[1]在计算机视觉领域扮演着至关重要的角色，它被视为模式识别领域的核心研究议题[2]。这一任务的实质在于进行密集的预测，要求对图像中的每个像素进行准确的类别预测，以确保系统能够深刻理解对象的轮廓、位置和具体类别等关键信息。随着深度学习技术在计算机视觉领域的飞速发展，语义分割技术获得了广泛的应用，自动驾驶[3]、精准农业[4]、医学影像分析[5]等领域都在积极应用这一技术。在自动驾驶领域尤为显著，对道路场景进行的语义分割为汽车提供了关键的技术支持。通过对车辆前方场景进行精准的语义分割，系统能够准确地识别和定位道路、车辆和行人，从而提高自动驾驶汽车在环境感知方面的准确性。

现有的图像语义分割主要分为传统方法和基于深度学习的方法。传统的语义分割算法依赖于手工标注特征，实施过程繁琐、效率低，在一些简单的场景下效果不错，但难以应付较为复杂的语义分割任务。基于深度学习的方法利用大量的数据对模型进行训练，自动提取数据特征，逐渐成为了语义分割的主流方法。目前，应用于图像语义分割的深度学习算法主要分为两类，分别是基于卷积神经网络 (convolutional neural network, CNN)[6]和基于 Transformer[7]。

2015 年，Long 等人基于编码器-解码器结构提出了全卷积神经网络 (fully convolutional networks for semantic segmentation, FCN)[8]，首次实现了端到端的像素级语义分割。Ronneberger 等提出 UNet[9]，在解码时使用编码阶段不同尺度的特征进行融合，获取丰富的上下文信息和空间位置信息。Zhao 等提出 PSPNet[10](pyramid scene parsing network)，通过金字塔池化模块融合不同池化尺度和不同子区域之间的上下文信息，提升相似物体的检测精度。Chen 等设计了 DeepLab[11-14]算法，其中 DeepLabV3[13]设计了空洞空间金字塔池化模块，采用不同膨胀速率的空洞卷积在多个尺度捕获目标信息及其上下文信息，解决图片多尺度问题的同时也提升了分割效果。

Transformer 最初应用于自然语言处理领域，是一种基于自注意力机制的神经网络模型。Vision in Transformer (ViT)[15]是第一个应用于图像分类的 Transformer 算法，不同于 CNN 的特征提取方式，Transformer 模型通过学习序列元素之间的相互依赖关系可以更有效地捕获全局交互信息。然而，ViT 存在输出单一尺度、低分辨率特征和高计算成本等缺点。为了应对这些挑战，研究者们提出了 Swin Transformer[16]算法。Swin Transformer 的核心思想是将图像分割成一系列的非重叠的图块，在图块内进行自注意力计算，从而减少计算量。然后在这些图块之间引入跨窗口的注意力机制，实现全局的信息交互。正是这些特点，让 Swin Transformer 在许多计算机视觉任务中表现出色。

道路场景的图像语义分割有 2 个难点：1) 道路场景的图像种类繁多，小目标物体不易识别。2) 道路场景的图像复杂多样，相似物体与重叠物体容易出现误判，图像中的边缘细节难以有效分割出。针对以上难点，本文借鉴 Swin Transformer 的思想提出了一个适用于道路场景语义分割的深度学习网络模型。该网络采用编码-解码结构，在编码部分中利用 Swin Transformer 网络优良的全局交互能力进行特征提取工作。由于层级式提取需要进行下采样，但下采样过程中会存在许多细节和结构信息的丢失，因此提出一种全新的特征压缩模块 (feature compression module，FCM) 来进行下编码阶段的采样操作，从而避免上述问题。在解码器部分，借鉴了特征金字塔网络 (feature pyramid network，FPN)[17]的思想，设计了一种多级特征融合的解码器结构。鉴于 FPN 并没有考虑不同尺度特征图之间的关联性，只是在上采样后将相同尺度的特征图进行简单的叠加，设计一种独特的注意力融合模块 (attention fusion module，AFM) 来考虑不同尺度特征之间的关联性，进而提高模型的全局与局部识别能力。

本文主要贡献如下：1) 借鉴了 Swin Transformer 的特征提取模块的思想，提出了一种编码-解码结构的适合于道路场景语义分割的深度学习网络模型。2) 提出了一种全新的特征压缩模块来进行编码阶段的特征提取操作，减少下采样过程中的特征信息损失，保留了尽可能多的边缘特征。3) 由于 Swin Transformer 没有固定的编码器，本文提出了一种多级特征融合解码器，使用注意力融合模块来充分利用编码器输出的多尺度特征信息，对不同尺度的信息进行选择性融合，再通过特征金字塔模块之后进行拼接，可以有效地恢复城市道路图像的细粒度细节。4) 在 Cityscapes 数据集上进行实验评估，与多种道路场景语义分割的经典算法对比，本文方法减少了特征细节的丢失，在分割精度方面具有一定优势。

2 相关工作

2.1 经典方法

传统的图像语义分割主要分为基于阈值、基于边缘和基于图论分割的方法。基于阈值的方法通过设定像素灰度值的阈值来将图像分割为不同的区域。灰度图的阈值差，图像的模糊程度，都会影响该方法的分割效果。基于边缘的方法则是对图像不同语义类别的边缘进行检测，根据这些边缘将图像分割成不同的区域[18]。其中，Canny 边缘检测和边缘链接算法是常用的技术。基于图论的图像分割方法利用图模型来表示图像中的像素，并通过对图模型分析来实现图像分割[19]。

2.2 基于卷积神经网络的语义分割

在语义分割的发展历程中，基于卷积神经网络的图像语义分割方法对该领域做出的贡献不容小觑[20]。2012 年 Krizhevsky 等人提出的 AlexNet[21] 架构掀起了卷积神经网络在各个领域的研究热潮。2015 年微软研究院提出的 ResNet[22] 模型，解决网络深层架构的问题。同年，Long 等人提出全卷积神经网络 FCN，标志着分割领域进入了全新的发展时期。与之前的所有图像语义分割算法最大的不同，FCN 使用卷积层代替了分类模型中全部的连接层，可以接受任意尺寸的图像输入，并实现了逐像素级的类别预测，极大地推动了分割方法的进展。2016 年，针对 FCN 网络中由于感受野有限而无法充分捕捉上下文信息的问题，PSPNet[10] 模型被提出了，该模型通过在网络中间添加金字塔池化操作，聚合不同尺度上的特征，进而提升网络的全局感知能力。Chen 等人提出的 DeepLab V2[12] 使用空洞卷积来增大感受野，同时保持图像分辨率不变，使用多个不同膨胀因子的空洞卷积构建空洞池化金字塔, 以获取多尺度的特征信息。Ronneberger 等人提出经典的对称编-解码结构 U-Net 网络，该网络一开始用于医学图像分割，采用编解码特征图拼接，通过跳过连接来学习相应编码级的空间相关性。

2.3 基于 Transformer 的语义分割

Transformer 最初应用在自然语言处理，后来被引入计算机视觉领域作为 VIT (vision Transformer)[15] 骨干网络。VIT 将图像划分成小块，并编码成令牌向量，然后进行全局自注意力计算。实验证明，VIT 在计算机视觉中有很大潜力，但需要大量训练样本和高算力支持。为了解决这些问题， Liu 等[16] 提出 Swin Transformer 这一种骨干网络，它借鉴了 CNN 的多层级思想，并采用层级式结构输出多尺度特征，能够较好地处理图像的多尺度问题。其给 patch 划分小窗口，在窗口内计算局部自注意力，并通过移位窗口操作实现窗口间的信息交互。此外，Swin Transformer 通过合并相邻的小块来缩小特征图的大小，适用于密集预测任务。这些改进使得 Swin Transformer 在计算机视觉任务中表现出色，并具有更高的计算效率。

3 本文方法

3.1 网络结构

本文提出的道路场景语义分割网络模型如图 1 所示，它由两部分构成：基于 Swin Transformer 特征提取编码器和基于 FPN 的多尺度特征融合的解码器。编码器利用 Swin Transformer Block 对输入的图像进行序列建模以获取多尺度特征，下采样过程中通过特征压缩模块 FCM 代替原来的图像块合并 (patch merging) 模块，减少下采样过程中的信息损失，保留更多细节特征。解码器使用空洞空间金字塔池化 (atrous spatial pyramid pooling, ASPP) 模块消除产生的噪声，进一步提取上下文信息。通过注意力融合模块 AFM 对特征通道之间的依赖关系进行建模的同时缩小高低层语义差距，并利用 FPN 融合低层细节特征和高层语义特征，将新融合生成的四层特征图分别进行上采样，进行拼接融合，得到分割特征映射图。

3.2 编码器

编码器的设计借鉴了 Swin Transformer 架构，其原始构架如图 2 所示。它是一种层级式的 Transformer 结构，通过重复堆叠的 Swin Transformer 模块和下采样模块，会生成四层不同尺度大小的特征图。

本文的特征提取模块也即编码器如图 1 左侧所示。图片通过图像块分割层 (patch partition) 被转换成 4×4 固定大小的、互不重叠的图像块。然后在通过线性嵌入层 (linear embedding) 将图像块序列化，将处理后的图像块送入到 Swin Transformer Block 中。 Swin Transformer Block 如图 3 所示。可以看出， Swin Transformer Block 包括了窗口多头注意力机制模块 (window multi-head self attention，W-MSA)，滑动窗口多头注意力机制模块 (shifted-window multi-head self-attention，SW-MSA)。Swin Transformer Block 中的自注意力计算过程如式 (1) ~ (4) 所示：

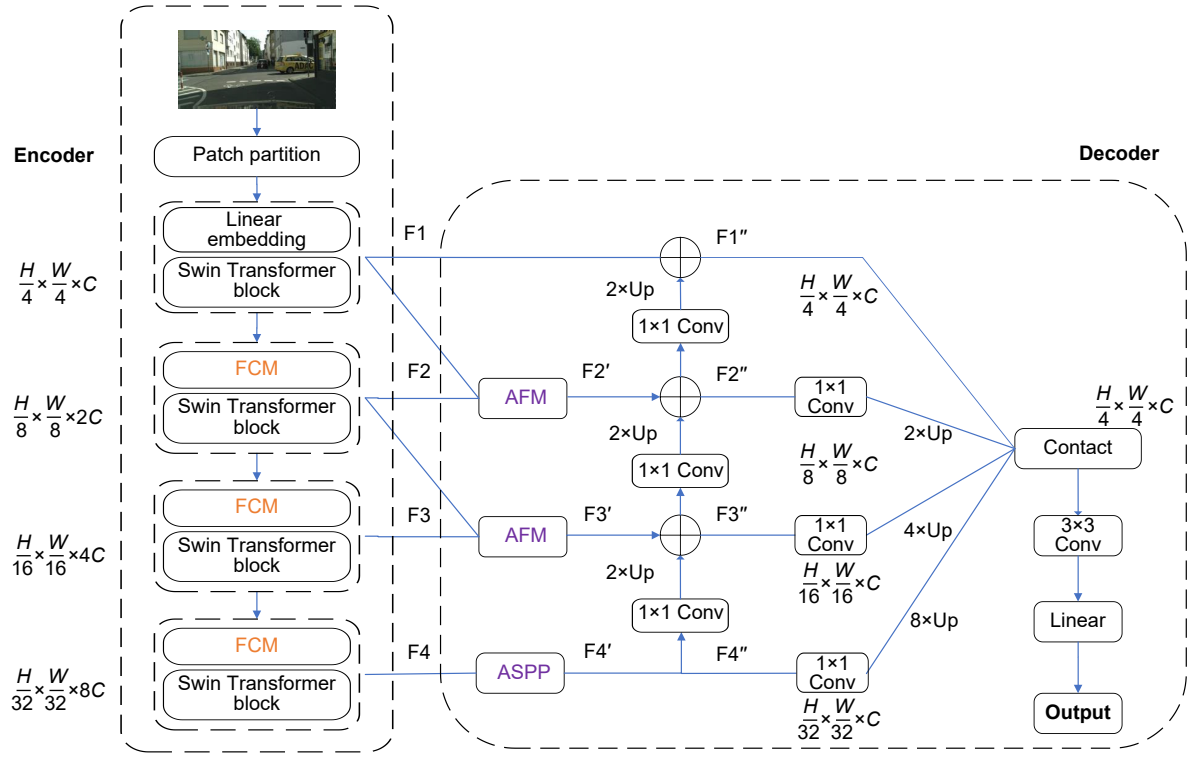


图 1 网络架构

Fig. 1 Network architecture

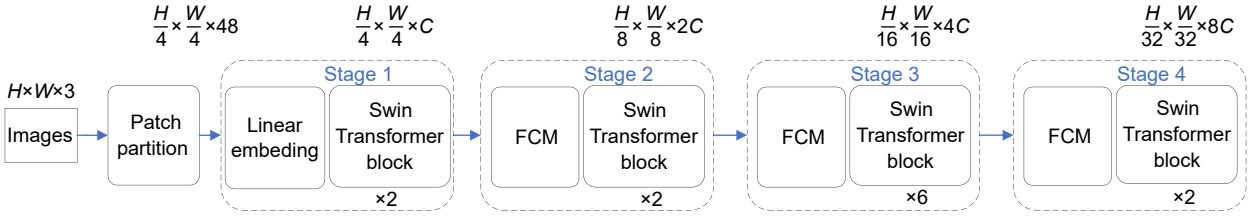


图 2 Swin Transformer 架构

Fig. 2 Swin Transformer architecture

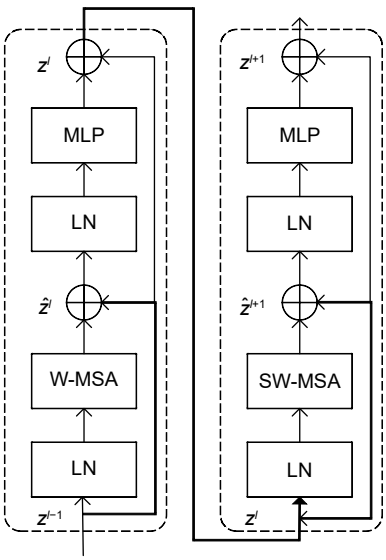


图 3 Swin Transformer 模块

Fig. 3 Swin Transformer block

$$\hat{z}^{l} = \text{W-MSA}\left(LN\left(z^{l-1}\right)\right) + z^{l-1}, \tag{1}$$

$$z^{l} = \text{MLP}\left(LN\left(\hat{z}^{l}\right)\right) + \hat{z}^{l}, \tag{2}$$

$$\hat{z}^{l+1} = \text{SW-MSA}\left(LN\left(z^{l}\right)\right) + z^{l}, \tag{3}$$

$$z^{l+1} = \text{MLP}\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1}, \tag{4}$$

其中：z^{l} 表示 W-MSA 模块的输出特征，z^{l+1}表示 SW-MSA 模块的输出特征；LN 为归一化层；MLP 为多层感知机，目的是将序列映射到多维通道，经过卷积后再映射回序列值。

原始的 Swin Transformer 采用 Patch Merging 模块来进行窗口特征映射，具体操作如图 4 所示。然而，这种方式进行下采样容易导致一些细节和上下文结构信息的丢失，不利于小尺度对象的语义分割。对此，我们提出了一种新的特征压缩模块 FCM，如图 5 所示。用特征压缩模块 FCM 替换 Patch Merging 模块，

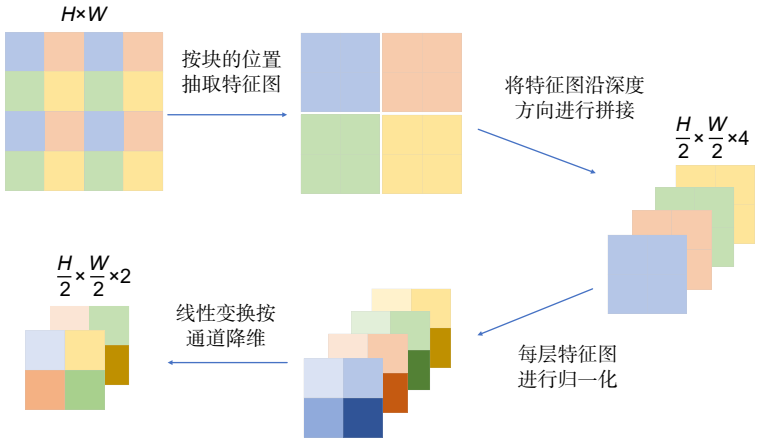


图 4 图像块合并模块

Fig. 4 Patch Merging module

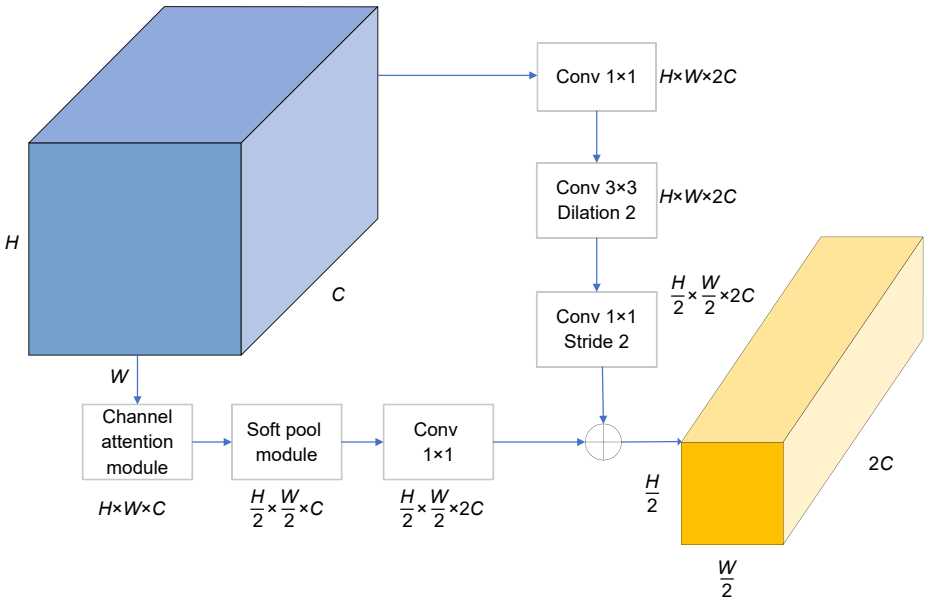


图 5 特征压缩模块

Fig. 5 FCM module

使网络在下采样过程中保留了尽可能多的详细特征。

本文设计的 FCM 模块的输入为 Swin Transformer Block 提取的特征，该模块分为上下两个支路。上支路利用膨胀卷积的感受野来广泛收集小尺度物体的特征和结构信息。在这个支路中，首先通过一个 1×1 的卷积层来调整通道数为 $2C$，接着通过一个 3×3 的膨胀卷积层，最后再通过一个 1×1 的卷积层来减小特征尺度。使用膨胀卷积能够扩大卷积操作的感受野，在减少参数量和计算成本的同时，保持对大范围信息的有效感知。这个分支主要用来获取大范围上下文的信息，以便深层网络继续特征提取，这个分支的输出是 $(H/2)\times(W/2)\times 2C$。

下支路首先通过通道注意力机制模块获取通道之间的联系，再通过软池化[23]的操作来获取更精细的下采样，最后通过 1×1 的卷积层来调整通道数为 $2C$。软池化在这个分支中起到了至关重要的作用，其核心功能是通过降低数据的空间维度，保留了更多的细节特征，使得输出更为紧凑。这个分支主要用来保持通道之间的联系并且在一定程度上减少细节特征的丢失，这个分支的输出也是 $(H/2)\times(W/2)\times 2C$。最后将这两个等维度的分支对应位置逐元素相加得到输出特征。

3.3 解码器

本文设计的解码器如图 1 右侧所示，解码器将经

过编码器处理所得到的四层特征图进一步优化和融合，得到分割结果图。本文的解码器借鉴了 FPN 思想，并加入了 AFM 模块和 ASPP 模块。传统的金字塔架构虽然在处理多尺度信息和实现特征融合方面取得了一定的成功，但也存在一些问题：不同层次的特征图具有不同的感受野，这可能导致融合后的特征在全局和局部信息的平衡上存在困难，浅层网络中无用信息带来的冗余等问题[24]。

针对上述问题，本文提出了 AFM 模块来聚合不同尺度特征，如图 6 所示，该模块将深层和浅层网络的信息进行结合，缩小不同尺度特征之间的语义差距，进而达到提高网络精度的目的。具体来说，首先使用 1×1 卷积来改变浅层网络输出的通道，接着在保持通道数目不变的情况下，采用双线性上采样改变特征尺度，使得通道和尺度大小都与通过横向连接提取的下级输出相匹配。然后我们将浅层特征和深层特征分别计算通道注意力机制再加权，这样做，可以调整原始特征图中每一个特征通道的重要程度，达到增强目标的特征、抑郁背景噪声的目的。最后将添加通道注意力机制的浅层网络和深层网络按照通道拼接的方式进行融合，获取含有丰富上下文语义信息的特征。

解码器的输入为编码器输出 4 个阶段的特征图 F={F1，F2，F3，F4}，其分辨率分别为原输入图像的 (1/4，1/8，1/16，1/32}。首先让最深层特征图 F4 通过 ASPP 模块，进一步提取上下文信息，得到本层特征图 F4'；让特征图 F3 通过 AFM 模块融合特征图 F2 得到本层特征图 F3'； 让特征图 F2 通过 AFM 模块融合特征图 F1 得到本层特征图 F2'。其次，构建自下而上的特征融合通道，让高层次的特征经过 2 倍上采样并调整通道数再与低层级特征进行逐像素相加，得到 FPN 四层特征图 F"={F1"，F2"，F3"，F4"}，最后将 FPN 各级输出分别调整通道数并按上采样率为 {1，2，4，8}进行上采样，并按照通道维度进行拼接，再通过线性层调整通道数为分类类别数，最后得到分割预测图。

4 实验结果与分析

4.1 数据集

实验采用广泛使用的语义分割数据集 Cityscapes [25] 数据集。Cityscapes 数据集包含了来自 50 个德国不同城市的街道景观图像。这些图像经过了像素级别的高质量标注，总共有 5000 张。其中，训练集包含了 2795 张图片，验证集包含了 500 张图片，测试集包含了 1525 张图片。数据集包含了 8 个类别，涵盖了 19 个子类别。

4.2 实验环境

实验环境如表 1 所示，实验所用的操作系统为 Ubuntu18.04，CPU 型号为 AMD5600Xd，GPU 型号为 NVIDIA RTX3070。网络是基于 MMSegmentation (Pytorch 1.10.0、Python 3.7) 开发框架下实现的，编译环境采用 Python3.7 编译环境。模型使用 AdamW

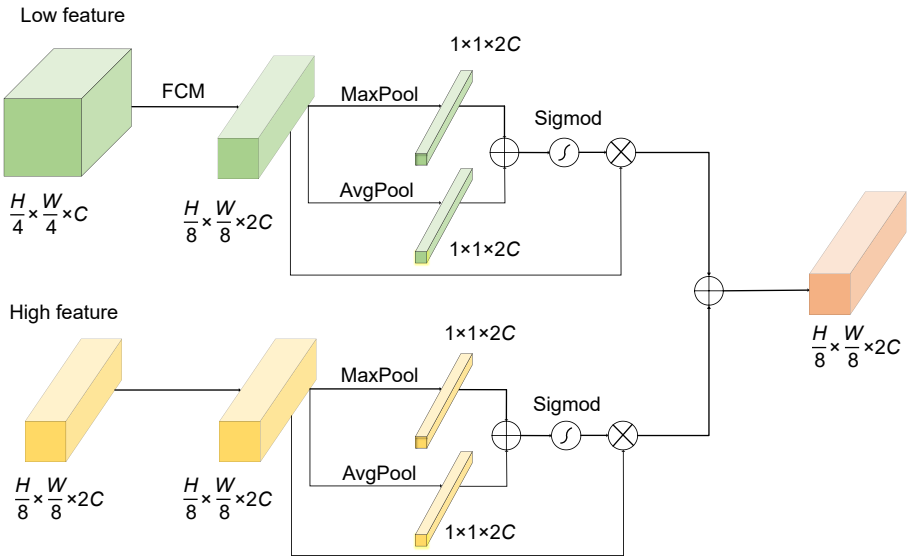


图 6 注意力融合模块

Fig. 6 AFM module

表 1 实验环境

Table 1 Experimental environment

实验环境	配置	实验环境	配置
CPU	AMD5600Xd	CPU核心数	6
GPU	NVIDIA RTX3070	主频	3.7 GHz
内存	32 G	显存	11 G
操作系统	Ubuntu18.04	编程语言	Python 3.7
深度学习框架	Pytorch 1.10.0	CUDA	10.2

优化器，其中初始学习率设置为 0.00006，权重衰减率为 0.01，损失函数使用交叉熵损失，进行 16 万次迭代。

4.3 评价标准

为了评价所提出的语义分割算法的性能，本文选取交并比 (intersection over union， IoU)、平均交并比 (mean intersection over union，MIoU)、像素准确率 (pixel accuracy，PA)、平均像素准确率 (mean pixel accuracy，MPA) 评估网络的分割性能[26]，其计算方式分别如式 (5) ~ (8)：

$$IoU=\frac{\sum_{i=0}^{k}P_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}P_{ij}+\sum_{i=0}^{k}\sum_{j=0}^{k}P_{ji}-\sum_{i=0}^{k}P_{ii}}, \tag{5}$$

$$MIoU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}+\sum_{j=0}^{k}P_{ji}-P_{ii}}, \tag{6}$$

$$PA=\frac{\sum_{i=0}^{k}P_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}P_{ij}}, \tag{7}$$

$$MPA=\frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}}. \tag{8}$$

假设数据一共包括 $k+1$ 个类别，对于一个像素点，假设该像素点的真实标签值为类别 i，模型对该像素点的预测结果为类别 j。P_{ij} 表示将 i 类像素点预测成 j 类，P_{ii} 表示将 i 类像素正确预测，P_{ji} 表示将 j 类像素点预测成 i 类。IoU 通过计算标签真实值集合和模型预测值集合之间的交并集来评估预测结果与真实标签的重合程度；MIoU 表示对所有类的 IoU 按类计算后取平均值；PA 通过计算模型正确分类的像素点总数在所有像素点中所占的比例来评估模型的分类精度；MPA 表示图像中所有物体类别像素准确率的平均值。

除了以上几个评价指标之外，本文还选取参数量 (params)、浮点数运算次数 (floating-point operations，FLOPs)、画面每秒传输帧数 (frames per second，FPS) 来评估网络的计算效率。

4.4 性能评估及与其他方法的比较

将现有的几种语义分割模型和本文模型在 Cityscapes 数据集上进行对比，具体的比较结果见表 2、表 3 和表 4 所示。表 2 展示的是各个网络在 Cityscapes 数据集上不同类别的交并比和平均交并比，表 3 展示的是各个网络在 Cityscapes 数据集上不同类别的像素准确率和平均像素准确度，表 4 综合展示了各类语义分割网络的性能。

从表 2 和表 3 参数结果来看，本文网络比 FCN 网络、PSPNet 网络、UNet 网络、DeepLabv3 网络，Swin Transformer 网络在 MIoU 方面分别提高了 10.2%、5.9%、4.7%、1.4%、2.0%，在 MPA 方面分别提升了 10.2%、4.9%、4.8%、2.5%、3.2%。从具体类别上来看，本文网络相比其他网络，在人行道、栏杆、柱子、植被、摩托车这几个类尺度较小且形状较不规则的目标类别上更胜一筹，在建筑物、汽车、火车这几类大型且形状比较规则的类别上不是很突出。分析其原因可能如下：首先，传统的池化操作会导致目标的特征在网络的不断加深中逐渐消失，而本文方法采用特征压缩模块尽可能多地保留特征细节和上下文信息；其次，本文在解码器器中引入 ASPP 模块，使得模型更关注于道路图像中目标所在区域，有效抑制复杂背景信息的干扰；最后，其他方法只利用网络最深层的特征捕获目标的多尺度特性，而本文方法基于 FPN 设计多尺度特征融合解码器融合网络所有输

表 2 各类模型在 Cityscapes 数据集上的 IoU 和 MIoU

Table 2 IoU and MIoU of various models on the Cityscapes dataset

Classes	FCN	PSPNet	UNet	DeepLabv3	SwinT	Ours
Road	97.1	98.0	98.0	98.1	98.0	98.1
Sidewalk	79.9	81.8	84.2	84.5	84.7	86.2
Building	89.3	91.1	91.1	91.7	91.4	91.6
Wall	44.2	48.2	48.7	51.2	54.4	55.5
Fence	48.3	50.3	51.5	53.6	57.3	59.9
Pole	30.6	45.7	48.2	50.3	55.5	57.2
Traffic Light	44.7	50.0	51.7	53.7	61.9	63.2
Traffic Sign	56.8	62.3	65.8	68.2	73.5	74.4
Vegetation	87.1	89.2	90.1	90.1	90.2	92.4
Terrain	60.4	62.8	65.3	64.2	61.3	63.2
Sky	90.8	94.2	93.8	95.3	94.2	95.1
Person	64.1	71.2	72.6	74.5	75.5	76.9
Rider	38.2	45.6	46.1	49.5	55.7	55.9
Car	90.4	92.0	92.2	92.6	93.8	93.5
Truck	51.3	68.5	63.4	74.4	73.6	72.5
Bus	72.0	80.3	77.6	83.2	79.4	79.9
Train	74.4	77.4	78.5	81.5	77.7	78.1
Motocycle	52.5	50.1	55.5	53.5	56.5	59.2
Bicycle	59.1	60.1	63.4	64.2	71.2	73.2
MIoU/%	64.92	69.28	70.45	73.71	73.17	75.18

表 3 各类模型在 Cityscapes 数据集上的 PA 和 MPA

Table 3 PA and MPA of various models on the Cityscapes dataset

Classes	FCN	PSPNet	UNet	DeepLabv3	SwinT	Ours
Road	98.1	98.5	98.8	99.1	99.1	99.1
Sidewalk	89.9	89.3	90.2	92.0	91.2	92.7
Building	96.3	94.7	96.1	96.2	96.5	96.8
Wall	52.2	72.1	60.7	73.1	71.4	72.3
Fence	60.3	69.3	68.5	72.5	71.4	74.6
Pole	36.6	74.7	59.2	74.3	74.1	77.7
Traffic Light	56.7	72.0	62.7	69.2	70.4	72.1
Traffic Sign	68.8	79.3	75.8	76.5	76.7	79.3
Vegetation	94.1	93.2	95.1	93.6	95.3	97.7
Terrain	74.4	79.8	78.3	78.1	79.2	80.3
Sky	95.8	97.2	97.8	97.5	97.5	97.9
Person	77.1	82.2	84.6	84.2	86.3	87.9
Rider	58.2	68.6	55.1	71.2	72.4	73.7
Car	96.4	96.0	96.2	96.3	97.6	97.6
Truck	62.3	79.5	76.4	75.5	73.5	76.2
Bus	85.0	87.3	89.6	91.7	85.6	87.7
Train	78.4	83.4	92.5	88.4	79.3	82.9
Motocycle	66.5	73.5	67.5	77.5	77.3	79.2
Bicycle	77.1	73.1	80.4	76.2	80.3	84.2
MPA/%	74.64	79.97	80.06	82.31	81.59	84.83

表 4 各类语义分割算法性能比较

Table 4 Performance comparison of various semantic segmentation algorithms

方法	MIoU/%	MPA/%	Param/M	FLOPs/G	FPS
FCN	64.92	74.64	34.90	66.38	58.61
PSPNet	69.28	79.97	51.86	152.97	81.25
UNet	70.45	80.06	49.10	166.92	54.52
DeepLabv3	73.71	82.31	68.37	235.37	36.59
SwinT	73.17	81.59	121.25	297.57	12.22
Ours	75.18	84.83	123.77	305.46	14.83

出层中目标的特征，可以有效补充目标的细节信息。因此本文方法可以更好地提取和利用目标的特征，模型分割性能更优。

尽管我们的模型分割效果较为优秀，但表 4 也表明了我们的模型参数量较大，推理时间也更久，实时性一般，也需要消耗更多的内存。在实际应用中，仍需进行轻量化处理，降低参数量。

4.5 典型场景下可视化结果分析

为了更清楚地看到我们提出的方法与其他网络的优势，我们可视化了不同算法在 Cityscapes 场景中的结果。如图 7 所示，从第一行到第七行分别是场景图原图、FCN 网络分割结果图、PSPNet 网络分割结果图、UNet 网络分割结果图、DeepLabv3 网络分割结果图、Swin Transformer 网络分割结果图以及本文改进的基于 Swin Transformer 的模型分割结果图。

从 (a) 组图可以看出，本文方法在电线杆、路灯杆子、警示杆这种柱类图形上有更优秀的分割效果。从 (b) 组图可以看出，本文方法相比于其他的网络在人行道和车辆轮廓细节上效果更好，可以将较为完整的车道线的部分分割出来。在左下角人物和车辆重叠的部分，也可以将人与车辆的边缘分割清晰。但对于图片右下角处汽车阴影遮挡住了人行道，仍然存在漏检的现象。从 (c) 组图可以看出，对比关注右下角的花坛处，本文方法可以较为清晰地将地势高出的部分

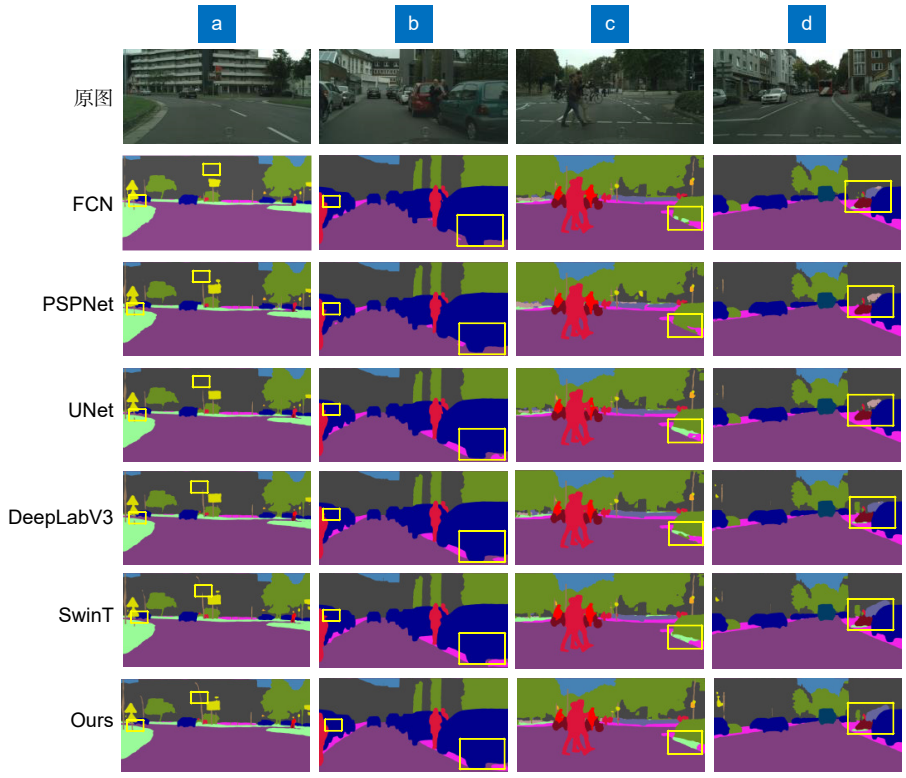


图 7 Cityscapes 场景中多种方法分割效果对比图

Fig. 7 Comparison of segmentation effects of multiple methods in Cityscapes scenes

展示出来，在边界处的分割效果更优。从 (d) 组图可以看出，本文方法可以将花坛护边分割出来，对于边缘的分割也更加平滑、整齐。其他网络并未将道路两侧隐约的路人分割出来，本文方法可以没有遗漏地把道路两侧的行人分割出来。

4.6 消融实验

消融实验如表 5 所示，在实验 1 中，编码器中没有使用 AFM 模块而是直接使用原论文中的 Patch Merging 模块进行降采样和调整通道数，解码器中也不使用 FCM 模块和 ASPP 模块，让编码器得到的四层特征输入图直接传入到特征金字塔模块中。在实验 2 中，编码器中使用 AFM 模块代替原来的 Patch Merging 模块。在实验 3 中，在使用 AFM 模块的同时，在特征融合的时候加入 FCM 模块。在实验 4 中，在实验 3 的基础上再加上 ASPP 模块。

在图 8 消融实验对比图中可以看出，在图 8(a) 分割结果图中，使用 AFM 模块的实验 2 结果图中的人行道分割边界比实验 1 更好，加入 FCM 模块的实验 3，更是进一步优化了整体的分割精度，边缘的分割效果也进一步提高了。实验 4 相比实验 3，虽然提升的精度不大，但是对分割的鲁棒性有良好的增强效果。在图 8(b) 分割结果图中，随着实验模块的增加，右下角的人行道边界区域越发清晰，中间的马路的边界区域的清晰程度也是越来越好。在图 8(c) 实验分割结果图中，相比实验 1，实验 2 中可以分割出更多的细节物体，例如远处的指示牌。结合表 3 中的数据，可以看出相比原来的模块，AFM 模块巧妙地结合了细节特征和语义特征，有助于保留网络中小目标的细节，显著优化对于小目标的分割效果，MIoU 增长了 0.7%。实验 3 相比实验 2 在物体轮廓方面效果更好一些，

表 5 消融实验

Table 5 Ablation experiment

实验序号	AFM	FCM	ASPP	MIoU/%	MPA/%
①	×	×	×	73.1	81.6
②	√	×	×	73.8	80.5
③	√	√	×	74.9	83.3
④	√	√	√	75.2	84.8

注：“√”表示网络中包含该结构，“×”表示在网络中去掉该结构。

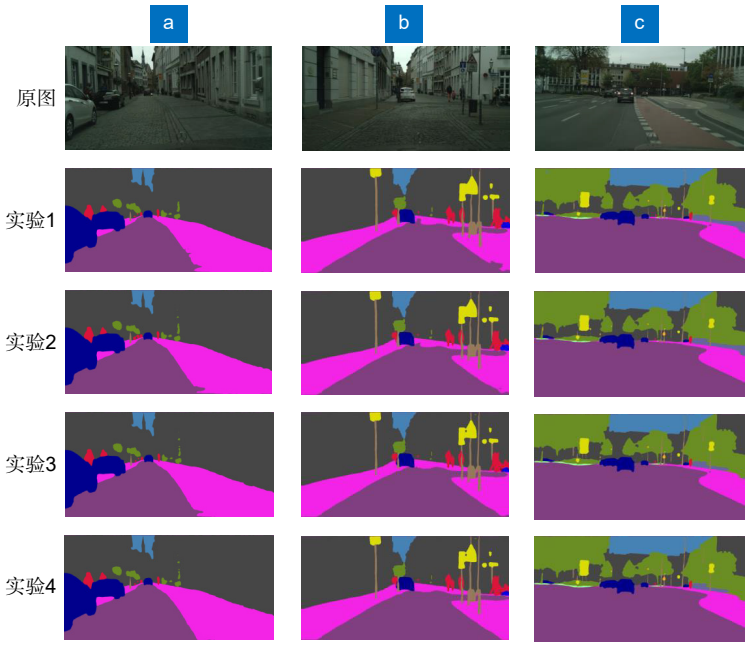


图 8 消融实验效果对比图

Fig. 8 Comparison of ablation experiment effects

FCM 模块相比不加此模块 MIoU 增长了 1.1%，优化了边缘分割效果。实验 4 相比实验 3，加入了 ASPP 模块进行捕获不同感受野尺度的信息，获得丰富的上下文信息，可以在图中看出在柱子和其他物品重叠的情况下，柱子的分割的连续性也是较好的，由此可以看出 ASPP 模块的作用。

5 总 结

本文提出一种多尺度特征融合的道路场景语义分割模型。模型采用编码-解码结构。编码器部分设计了一种全新的特征压缩模块 FCM 来优化编码阶段的下采样操作。实验表明该模块可以减少下采样过程中的特征信息损失，保留尽可能多的边缘特征信息。在解码器部分，本文设计出一种注意力融合模块 AFM 来充分利用不同尺度的特征信息，缩小高低层语义差距，充分联系上下文特征信息，再使用特征金字塔网络自下而上对特征融合的结果进一步叠加融合，这种解码方式有助于优化整体的边缘分割效果和分类精度。在 Cityscapes 城市道路场景数据集上进行定量和定性实验，实验结果表明，与多种语义分割算法相比，本文方法在分割精度方面有明显的优势。

参考文献

[1] Mo Y J, Wu Y, Yang X N, et al. Review the state-of-the-art technologies of semantic segmentation based on deep learning[J]. *Neurocomputing*, 2022, **493**: 626−646.

[2] Liu X L, Deng Z D, Yang Y H. Recent progress in semantic image segmentation[J]. *Artif Intell Rev*, 2019, **52**(2): 1089−1106.

[3] Zhang Y, Huang Y P, Guo Z Y, et al. Point cloud-image data fusion for road segmentation[J]. *Opto-Electron Eng*, 2021, **48**(12): 210340.
张莹, 黄影平, 郭志阳, 等. 基于点云与图像交叉融合的道路分割方法[J]. 光电工程, 2021, **48**(12): 210340.

[4] Chiu M T, Xu X Q, Wei Y C, et al. Agriculture-vision: a large aerial image database for agricultural pattern analysis[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 2825–2835. https://doi.org/10.1109/CVPR42600.2020.00290.

[5] Qureshi I, Yan J H, Abbas Q, et al. Medical image segmentation using deep semantic-based methods: a review of techniques, applications and emerging trends[J]. *Inf Fusion*, 2023, **90**: 316−352.

[6] Chua L O, Roska T. The CNN paradigm[J]. *IEEE Trans Circuits Syst I:Fundam Theory Appl*, 1993, **40**(3): 147−156.

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 6000–6010.

[8] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965.

[9] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C]//*18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015: 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

[10] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 6230–6239. https://doi.org/10.1109/CVPR.2017.660.

[11] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[C]//*3rd International Conference on Learning Representations*, 2015.

[12] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. *IEEE Trans Pattern Anal Mach Intell*, 2018, **40**(4): 834−848.

[13] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[Z]. arXiv: 1706.05587, 2017. https://doi.org/10.48550/arXiv.1706.05587.

[14] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 833–851. https://doi.org/10.1007/978-3-030-01234-2_49.

[15] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]//*9th International Conference on Learning Representations*, 2021.

[16] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.

[17] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 936–944. https://doi.org/10.1109/CVPR.2017.106.

[18] Xie S N, Tu Z W. Holistically-nested edge detection[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, 2015: 1395–1403. https://doi.org/10.1109/ICCV.2015.164.

[19] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation[J]. *Int J Comput Vis*, 2004, **59**(2): 167−181.

[20] Sehar U, Naseem M L. How deep learning is empowering semantic segmentation: traditional and deep learning techniques for semantic segmentation: a comparison[J]. *Multimed Tools Appl*, 2022, **81**(21): 30519−30544.

[21] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//*Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2012: 1106–1114.

[22] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778. https://doi.org/10.1109/CVPR.2016.90.

[23] Stergiou A, Poppe R, Kalliatakis G. Refining activation downsampling with SoftPool[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 10337–10346. https://doi.org/10.1109/ICCV48922.2021.01019.

[24] Ma L, Gou Y T, Lei T, et al. Small object detection based on multi-scale feature fusion using remote sensing images[J]. *Opto-Electron Eng*, 2022, **49**(4): 210363.
马梁, 苟于涛, 雷涛, 等. 基于多尺度特征融合的遥感图像小目标检测[J]. 光电工程, 2022, **49**(4): 210363.

[25] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 3213–3223. https://doi.org/10.1109/CVPR.2016.350.

[26] Ulku I, Akagündüz E. A survey on deep learning-based architectures for semantic segmentation on 2D images[J]. *Appl Artif Intell*, 2022, **36**(1): 2032924.

作者简介

杭昊 (1997-)，男，硕士研究生，研究方向为计算机视觉。

E-mail：212230402@st.usst.edu.cn

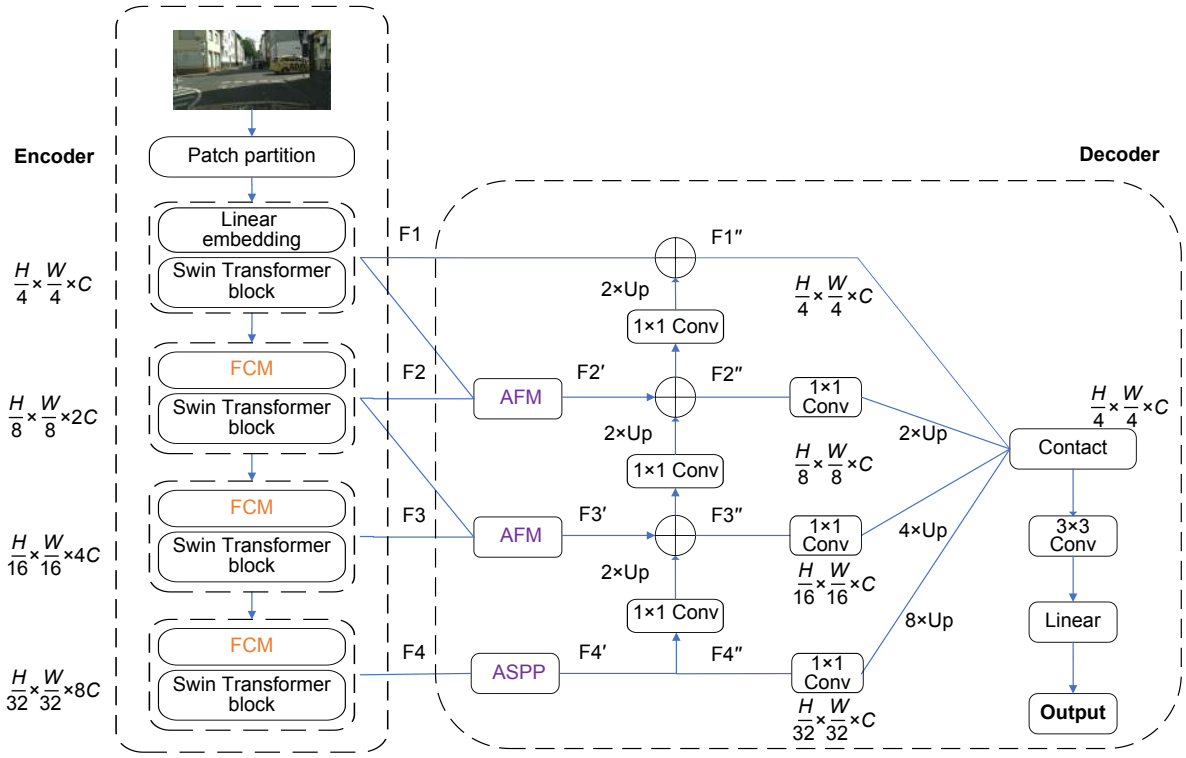
【通信作者】黄影平 (1966-)，男，教授，研究方向为汽车电子、计算机视觉。

E-mail: huangyingping@usst.edu.cn

扫描二维码，获取PDF全文

Design of Swin Transformer for semantic segmentation of road scenes

Hang Hao, Huang Yingping*, Zhang Xurui, Luo Xin



Overview: Semantic segmentation of road scenes is a crucial task for the perception of autonomous driving environments. In recent years, deep learning technologies have elevated research in semantic segmentation, leading to the emergence of numerous new algorithms. Methods based on deep learning train models with extensive data automatically extract data features and become the mainstream approach for semantic segmentation. Currently, deep learning algorithms applied to image semantic segmentation primarily fall into two categories: those based on CNN and those based on Transformer. CNN-based image semantic segmentation algorithms such as FCN, PSPNet, U-Net, and DeepLab have made significant contributions to the field. Transformer is a novel architecture based on self-attention, initially applied in the NLP domain. With powerful feature extraction capabilities, Transformer can capture long-range dependencies between feature vectors, acquiring richer contextual information. Researchers have gradually adapted Transformers to the computer vision domain, forming various Visual Transformers. Subsequently, the Swin Transformer stands out, employing a hierarchical structure to output multi-scale features, calculating local self-attention within a window, achieving information interaction between windows through shift-window operations, and demonstrating excellent performance in various visual tasks. Despite extensive research on semantic segmentation algorithms for road scenes, existing methods still face challenges in practical applications. Addressing issues such as low segmentation accuracy in complex scene images and inadequate recognition of small targets, this paper proposes a road scene semantic segmentation algorithm based on the SwinTransformer with multi-scale feature fusion. The network adopts an encoder-decoder structure, where the encoder employs an improved SwinTransformer feature extractor for feature extraction in road scene images, reducing information loss during downsampling and retaining as many edge features as possible. The decoder consists of an attention fusion module and a feature pyramid network, effectively integrating multi-scale semantic features and efficiently restoring fine-grained details in urban road images. We conduct quantitative and qualitative experiments on the Cityscapes urban road scene dataset. The results show that, compared to

Foundation item: Projected supported by National Natrual Science Foundation of China (62276167)

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

* E-mail: huangyingping@usst.edu.cn

various existing semantic segmentation algorithms, our method exhibits significant improvements in segmentation accuracy. However, our network structure is relatively complex, with a large number of computations and parameters. In practical applications, further refinement, optimization of the network structure, and lightweight processing to reduce parameters and computations are still required.

Hang H, Huang Y P, Zhang X R, et al. Design of Swin Transformer for semantic segmentation of road scenes[J]. *Opto-Electron Eng*, 2024, **51**(1): 230304; DOI: 10.12086/oee.2024.230304