文章编号:1005-0523(2025)02-0110-09

基于多尺度Transformer特征的道路场景语义分割网络

彭 洋,吴文欢,张浣坤

(湖北汽车工业学院智能网联汽车学院, 湖北 十堰 442002)

摘要:道路场景中图像通常内容复杂,不同物体之间的尺度和形态差异较大,并且光照阴影等情况会让场景变得难以识别。而现有语义分割方法通常不能有效提取并充分融合多尺度语义特征,泛化能力和鲁棒性较差。文章提出了一种融合多尺度Transformer特征的语义分割网络模型。首先,利用CSWin Transformer提取不同尺度的语义特征,并且引入特征细化模块(FRM)提升深层小尺度特征的语义辨析能力;其次,采用注意力聚合模块(AAM)对不同尺度特征分别进行聚合;最后,通过融合这些增强后的多尺度特征,进一步提升特征的语义表达能力,从而提高分割性能。实验结果表明:该网络模型在Cityscapes数据集上取得了82.3%的准确率,较SegNeXt和ConvNeXt分别提升了2.2个百分点和1.2个百分点;在目前最具挑战性的ADE20K数据集上取得了47.4%的准确率,较SegNeXt和ConvNeXt分别提升了3.2个百分点和1.8个百分点。所提出的融合多尺度Transformer特征模型不仅具有较高的语义分割精度,能准确预测道路场景图像的像素语义类别,而且具有较强的泛化性能和鲁棒性。

关键词:语义分割;Transformer特征;特征融合;空间期望最大化注意力;通道注意力

中图分类号: TP391.41;U491.1 **文献标志码:**A

本文引用格式:彭洋,吴文欢,张浣坤. 基于多尺度Transformer特征的道路场景语义分割网络[J]. 华东交通大学学报,2025,42(2):110-118.

Road Scene Semantic Segmentation Network Based on Multi-Scale Transformer Features

Peng Yang, Wu Wenhuan, Zhang Haokun

(School of Intelligent and Connected Vehicle, Hubei University of Automotive Technology, Shiyan 442002, China)

Abstract: Image contents in road scenes are usually complex, with significant differences in scale and shape between different objects, and lighting and shadows can make the scenes difficult to recognize. However, existing semantic segmentation methods often fail to effectively extract and fully integrate multi-scale semantic features, resulting in poor generalization ability and robustness. To address these issues, this study proposes a semantic segmentation network model that fuses multi-scale Transformer features. Firstly, the CSWin Transformer was employed to extract semantic features at various scales, accompanied by the introduction of a feature refinement module (FRM) to enhance the semantic discrimination capability of deep, fine-grained features. Secondly, an attention aggregation module (AAM) was adopted to separately aggregate features across scales. Finally, by integrating these enhanced multi-scale features, the semantic expression ability of the features was further enhanced,

收稿日期:2024-09-14

基金项目:湖北省自然科学基金联合基金项目(2025AFD239);湖北汽车工业学院博士科研启动基金项目(BK202347)

thereby improving segmentation performance. Experimental results demonstrate that this network model achieves an accuracy of 82.3% on the Cityscapes dataset, outperforming SegNeXt and ConvNeXt by 2.2 percentage points and 1.2 percentage points, respectively. Moreover, it attains an accuracy of 47.4% on the highly challenging ADE20K dataset, surpassing SegNeXt and ConvNeXt by 3.2 percentage points and 2.8 percentage points, respectively. The proposed multi-scale Transformer feature fusion model not only achieves high semantic segmentation accuracy, accurately predicting pixel semantic categories of road scene images, but also has strong generalization performance and robustness.

Key words: semantic segmentation; Transformer features; feature fusion; spatial expectation maximizes attention; channel attention

Citation format: PENG Y, WU W H, ZHANG H K. Road scene semantic segmentation network based on multi-scale transformer features[J]. Journal of East China Jiaotong University, 2025, 42(2): 110−118.

语义分割的目标是识别出图像中每个像素所属的物体类别标签。作为计算机视觉中的一个基础任务，语义分割在自动驾驶[1]、智能交通管理[2]和视频分析[3]等许多领域得到广泛应用。但是，由于实际道路场景中的物体类别繁多，形态和尺度不一，准确识别出每个像素的物体类别是一项相当困难的任务。因此，研究如何提升语义分割性能进而帮助智能汽车感知其驾驶环境具有重要的意义。

近年来，随着深度学习技术的发展，人们提出许多基于深度学习的语义分割方法。SegNet和DeepLabv3+采用编码器和解码器结构，通过融合浅层细节特征和深度语义特征提升分割性能[4-6]。上述研究都取得了较好效果，但卷积核感受野大小是有限的，只能捕获局部短距离上下文信息，对于长距离依赖关系的捕获能力较弱，这对于解决语义歧义问题是非常不利的。

考虑到注意力机制能够聚合整个图像空间上的上下文信息，NLNet和DNLNet采用空间非局部自注意力机制来挖掘像素之间关联关系[7-8]。EncNet则通过通道注意力重新校准每个通道的权重进而对通道特征进行优化[9]。SegNeXt提出一种全新的卷积注意力机制，通过深度卷积来聚合局部特征[10]。

随着Vision Transformer架构在各种视觉任务中取得了比卷积神经网络更具竞争性的性能，为解决语义歧义提供了新方法，该架构通过利用多头自注意力机制获得了很强的远程建模能力[11]。SETR用Transformer替代CNN的编码器部分来完成图像语义分割任务，取得了不错的分割效果[12]。Swin Transformer将每个token的关注区域限制为局部窗口，并且采用halo和shift操作来交换相邻窗口之间的信息，由此扩大关注区域[13]。SegFormer在SETR的基础上去掉了位置编码，并对解码器进行轻量化设计，进而降低网络模型复杂度[14]。ConvNeXt将Swin Transformer的思想融入到经典的ResNet中，通过结合这两种网络架构的优势，对卷积神经网络进行改进，取得与Swin Transformer相匹敌的性能[15-16]。GSS提出了生成语义分割模型，将语义分割当作图像条件掩码生成问题，通过最小化分割掩码的后验分布和输入训练图像的潜在先验分布之间的差异来实现图像语义分割[17]。

尽管Transformer具有较强的上下文建模能力，但在解决语义分割任务时，大都作为编码器构建特征金字塔，并没有对不同尺度的特征进行有效融合，使得网络通常难以区分物体与背景或者物体之间的边界，并可能会忽略小物体特征，进而影响语义分割性能。

考虑到CSWin Transformer[18]不仅在特征提取方面具有优势，而且计算效率非常高，本文利用CSWin Transformer作为编码器提取多尺度特征，并且引入特征细化模块（feature refinement module，FRM）对深层小尺度特征进行细化增强。其次，与其他方法通常将多尺度特征同时上采样后进行拼接融合不同，本文在解码器中采用将低分辨率的深层特征逐级与高分辨率的浅层特征进行融合，即将特征上采样后与编码器相等尺度特征进行拼接，随后采用注意力聚合（attention aggregation module，AAM）模块对拼接特征进行融合。在AAM模块中，利用空间期望最大化注意力构建像素之间全局语义关联关系，而且采用多头通道注意力进一步优化语义通道特征。通过对多尺度特征进行逐

级融合，能更好地挖掘不同分辨率特征的语义信息，使得语义分割的性能得到提升。

1 本文网络

1.1 整体网络架构

本文网络主要由编码器和解码器组成，如图1所示。编码器部分首先将原始图像输入到CSWin Transformer中进行多尺度特征提取，然后把提取出的特征图分别送入解码器部分。在解码器中，采用ASPP（atrous spatial pyramid pooling）模块扩展深层特征的感受野，捕获更多的上下文信息，然后用FRM模块强化特征的语义表征能力。将增强后的特征图进行逐级上采样，并与浅层特征进行融合。随后，将不同尺度的特征输入到AAM模块中，在空间维度上捕获像素的上下文信息，抑制不相关区域的干扰，在通道维度上建模通道之间的依赖性，增强重要通道特征。最后，经过聚合的不同尺度特征图通过上采样操作进行融合，并生成最终的分割结果。图1中，Unsampling表示上采样，Conv表示卷积运算，Fusion表示算子合并，C，H，W分别表示通道数，高度和宽度。

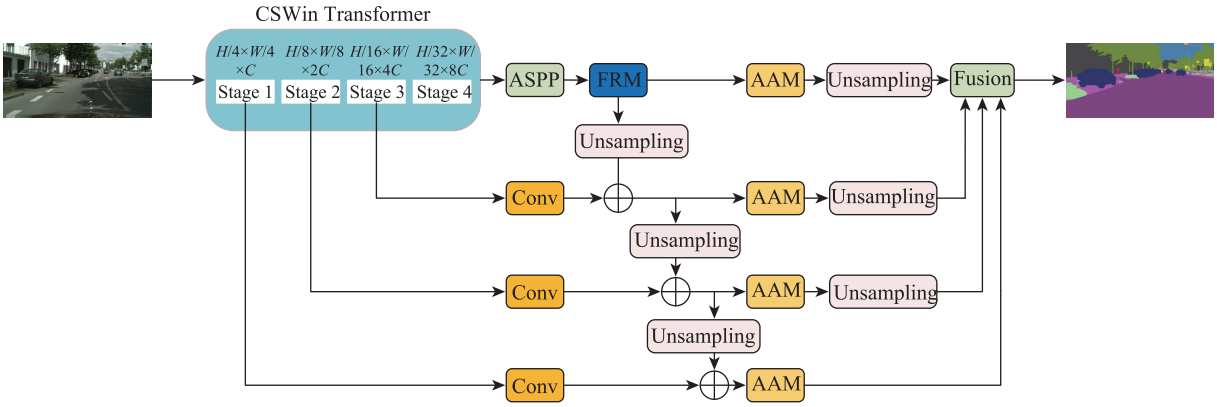

图1 网络整体架构

Fig. 1 Overall structure of the network model

1.2 特征细化模块

在语义分割任务中，深层特征的语义信息对提升网络性能至关重要，语义信息越清晰，融合后的分割效果就越好。为了增强深层特征的语义表达能力，本文在经过ASPP扩大感受野后，进一步使用了特征细化模块（FRM）对深层特征的语义信息进行细化处理。通过该模块，深层特征的语义辨析能力得以加强，使其在多级融合过程中能够提供更明确的语义指导，从而有效提升分割结果的精度。

图2详细介绍了FRM的整体结构。对于特征图 $\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$，$C$，$H$，$W$ 分别表示通道数，高度和宽度。特征图 $\boldsymbol{X}$ 在经过 $1\times W$ 和 $H\times 1$ 的条形池化层后，对每个通道在水平和垂直方向进行特征编码，以提取全局上下文信息。由此可得第 c 个通道上第 i 行的水平池化输出 $\boldsymbol{G}_c^{\mathrm{h}}(i)$ 为

$$\boldsymbol{G}_c^{\mathrm{h}}(i)=\frac{1}{W}\sum_{0\leqslant j<W}\boldsymbol{X}_c(i,j) \tag{1}$$

式中：$\boldsymbol{X}_c(i,j)$ 为特征图 $\boldsymbol{X}$ 在第 c 个通道、第 i 行、第

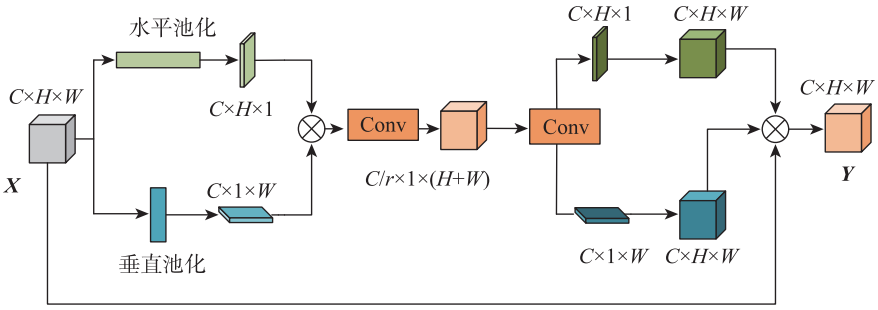

图2 特征细化模块

Fig. 2 Feature refinement module

j 列处的值。类似地，第 c 个通道上第 j 列的垂直池化输出 $\boldsymbol{G}_c^{\mathrm{w}}(j)$ 可以表示为

$$\boldsymbol{G}_c^{\mathrm{w}}(j)=\frac{1}{H}\sum_{0\leqslant i<H}\boldsymbol{X}_c(i,j) \tag{2}$$

上述两个变换分别沿着水平方向和垂直方向进行特征聚合，产生一对特征图。将这对特征图融合后输入到卷积变换函数 f_1 中，输出 $\boldsymbol{F}$ 可以表示为

$$\boldsymbol{F}=\alpha\left(f_1\left(\mathrm{concat}\left[\boldsymbol{G}^{\mathrm{h}},\boldsymbol{G}^{\mathrm{w}}\right]\right)\right) \tag{3}$$

式中：$\mathrm{concat}\left[\boldsymbol{G}^{\mathrm{h}},\boldsymbol{G}^{\mathrm{w}}\right]$ 为水平方向输出与垂直方向输出进行拼接融合；α 为非线性激活函数；$\boldsymbol{F}\in\mathbb{R}^{C/r\times 1\times(H+W)}$ 为编码后的中间特征图。使用缩放比 r 来控制通道大小，然后将特征图 $\boldsymbol{F}$ 按照空间维度拆成两个独立的张量 $\boldsymbol{F}^{\mathrm{h}}\in\mathbb{R}^{C/r\times H}$，$\boldsymbol{F}^{\mathrm{w}}\in\mathbb{R}^{C/r\times W}$。随后通过卷积变换 f_{h}，f_{w} 将张量的通道数变换为与输入 $\boldsymbol{X}$ 相同，从而得到新的张量 $\boldsymbol{Q}^{\mathrm{h}}$，$\boldsymbol{Q}^{\mathrm{w}}$ 为

$$\boldsymbol{Q}^{\mathrm{h}}=\beta\left(f_{\mathrm{h}}\left(\boldsymbol{F}^{\mathrm{h}}\right)\right) \tag{4}$$

$$\boldsymbol{Q}^{\mathrm{w}}=\beta\left(f_{\mathrm{w}}\left(\boldsymbol{F}^{\mathrm{w}}\right)\right) \tag{5}$$

式中：β 为激活函数。最后在第 c 个通道中，位置 (i,j) 上的结果 $\boldsymbol{Y}_c(i,j)$ 可以表示为

$$\boldsymbol{Y}_c(i,j)=\boldsymbol{X}_c(i,j)\times\boldsymbol{Q}_c^{\mathrm{h}}(i)\times\boldsymbol{Q}_c^{\mathrm{w}}(j) \tag{6}$$

则最终输出 $\boldsymbol{Y}$ 为所有通道的输出组成的集合

$$\boldsymbol{Y}=\left[\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_C\right] \tag{7}$$

1.3　注意力聚合模块

本文使用注意力聚合模块（AAM）来增强网络对不同特征的表达能力。AAM模块应用于逐级融合后的各尺度特征，使模型能够更加精准地聚焦于不同大小的目标，提升边缘分割能力和对小目标的识别效果。

AAM的结构如图3所示。在空间维度上，本文使用了空间期望最大化注意力（SEMA）挖掘整个图像上像素之间的关联关系。在通道维度上，本文设计了多头通道注意力（MCA），利用多个并行的注意力“头”来从多维视角评估和提炼通道信息。

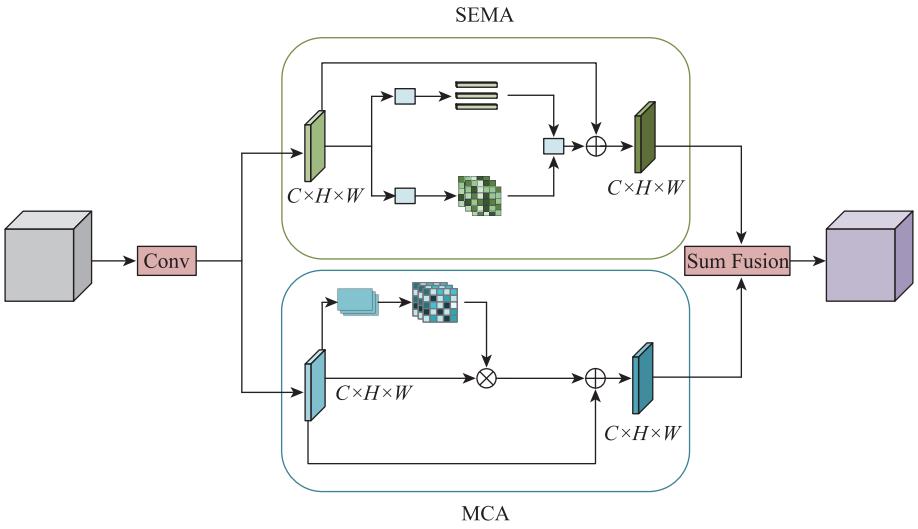

图3　注意力聚合模块的整体结构

Fig. 3　Overall structure of the attention aggregation module

1.3.1　空间期望最大化注意力

以往的空间注意力机制通常通过对所有位置的特征进行加权求和来计算每个位置的表示。虽然这种方式能够捕获远程的上下文关系，但是它增加空间复杂度并消耗大量计算资源。为了解决这个问题，本文采用空间期望最大化注意力（SEMA）。SEMA首先通过期望最大化算法生成一组紧凑的基，这组基能够有效代表原始特征的主要信息，然后在生成的基上实施注意力机制，避免了冗余信息的干扰，降低了计算复杂度。

SEMA的结构图如图4所示。首先，给定一个输入特征图 $\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$，$\boldsymbol{X}$ 被重塑为 $\boldsymbol{X}\in\mathbb{R}^{N\times C}$，

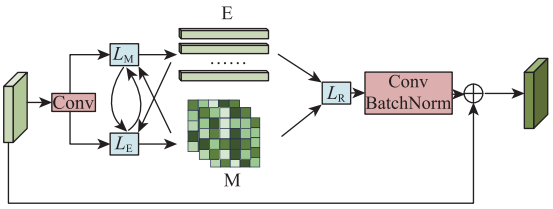

图4　空间期望最大化注意力的整体结构

Fig. 4　Overall structure of the attention aggregation module

$N=H\times W$。然后初始化一个基 $\boldsymbol{\mu}\in\mathbb{R}^{K\times C}$，$K$ 是基类的数量。SEMA方法包括以下3个步骤：权责估计 L_{E}、似然最大化 L_{M} 和数据重新估计 L_{R}，前两步分别对应EM算法的E步和M步。在 L_{E} 步中，目标是估计隐变量 $\boldsymbol{Z}\in\mathbb{R}^{N\times K}$，即每个基对像素的权责。第 k 个基对像素 $\boldsymbol{x}_n$ 的权责 z_{nk} 可表示为

$$z_{nk}=\frac{\varphi(\boldsymbol{x}_n,\boldsymbol{\mu}_k)}{\sum_{j=1}^{k}\varphi(\boldsymbol{x}_n,\boldsymbol{\mu}_j)} \tag{8}$$

式中：$\varphi(\boldsymbol{a},\boldsymbol{b})$ 为指数内积 $\exp(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b})$。在第 t 次迭代中，$\boldsymbol{Z}^{(t)}$ 可以表示为

$$\boldsymbol{Z}^{(t)}=\mathrm{softmax}\left(\gamma X\left(\boldsymbol{\mu}^{(t-1)}\right)^{\mathrm{T}}\right) \tag{9}$$

式中：γ 作为超参数来控制 $\boldsymbol{Z}$ 的分布，且每一个注意力图的大小为 $H\times W$。L_{M} 步的作用是更新基 $\boldsymbol{\mu}$。为了保证 $\boldsymbol{\mu}$ 和 $\boldsymbol{X}$ 处在同一表征空间内，L_{M} 步使用 $\boldsymbol{X}$ 的加权平均来更新 $\boldsymbol{\mu}$。在第 t 次迭代中，第 k 个基 $\boldsymbol{\mu}_k^{(t)}$ 更新为

$$\boldsymbol{\mu}_k^{(t)}=\frac{\sum_{n=1}^{N}z_{nk}^{(t)}\boldsymbol{x}_n}{\sum_{m=1}^{N}z_{mk}^{(t)}} \tag{10}$$

L_{E} 和 L_{M} 交替执行 T（本文中设置为3）步。在 L_{R} 步中，使用最终的 $\boldsymbol{Z}$ 和 $\boldsymbol{\mu}$ 来对 $\boldsymbol{X}$ 进行重新估计，得到 $\tilde{\boldsymbol{X}}$

$$\tilde{\boldsymbol{X}}=\boldsymbol{Z}\boldsymbol{\mu} \tag{11}$$

$\tilde{\boldsymbol{X}}$ 相比 $\boldsymbol{X}$，具有低秩的特性，能保留原有特征图的主要信息，不同类别之间差异也能进一步拉大。另外，将复杂度降低至 $\mathrm{O}(NKT)$，由于 T 为一个小常数可以被省去并且 $K<<N$，所以其复杂度得到有效降低。

1.3.2 多头通道注意力

在深度神经网络中，不同通道的特征图可以被视为对不同类别的响应，这些语义响应之间存在关联性。通过挖掘通道特征间的关联可以优化语义特征，使得重要的通道特征得到进一步增强。为此，本文提出了一个多头通道注意力（MCA），引入多头注意力策略来建立不同通道之间的关联关系。该方法通过从多个角度全面捕获通道之间的依赖性，挖掘出语义更显著的通道特征。

MCA模块结构如图5所示。首先给定一个输入特征图 $\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$，将其进行重塑为 $\boldsymbol{X}\in\mathbb{R}^{C\times N}$，其

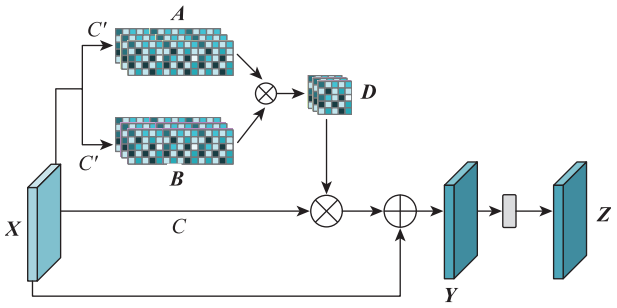

图5 多头通道注意力的整体结构

Fig. 5 Overall structure of multi-head channel attention

中 $N=H\times W$。根据预设的头的数量 h（设为4）来分组，并把通道设置为 C'，因此 $C=h\times C'$，可以得到 $\{\boldsymbol{A},\boldsymbol{B}\}\in\mathbb{R}^{h\times C'\times N}$。对于特征图 $\boldsymbol{A}$ 中第 k 个头的 u 个通道，可取得向量 $\boldsymbol{A}_{u,k}\in\mathbb{R}^{N}$。

同理，可得出特征图 $\boldsymbol{B}$ 中第 k 个头的 v 通道的向量 $\boldsymbol{B}_{v,k}\in\mathbb{R}^{N}$。由于每个头都有 C' 个通道，即可推出集合 $\boldsymbol{B}_k\in\mathbb{R}^{C'\times N}$。因此，在第 k 个头中通道 u 与通道 v 的关联性 $\boldsymbol{S}_{v,u,k}$ 可表示为

$$\boldsymbol{S}_{v,u,k}=\boldsymbol{A}_{u,k}\boldsymbol{B}_{v,k}^{\mathrm{T}} \tag{12}$$

由上述计算可推出在第 k 个头中，通道 u 与所有通道的关联性 $\boldsymbol{S}_{u,k}$ 为

$$\boldsymbol{S}_{u,k}=\boldsymbol{A}_{u,k}\boldsymbol{B}_{k}^{\mathrm{T}} \tag{13}$$

式中：$\boldsymbol{S}_{u,k}\in\mathbb{R}^{C'}$，那么在第 k 个头中不同通道之间的关系可表示为 $\boldsymbol{S}_k\in\mathbb{R}^{C'\times C'}$。进一步地，将所有头的通道关系进行整合，即可获得所有头的通道间关系为 $\boldsymbol{S}\in\mathbb{R}^{h\times C'\times C'}$。对 $\boldsymbol{S}$ 在通道维度上应用Softmax，就可得到关联矩阵 $\boldsymbol{D}\in\mathbb{R}^{h\times C'\times C'}$。

此外，将特征图 $\boldsymbol{X}$ 重构后按 h 分组可得到 $\boldsymbol{E}\in\mathbb{R}^{h\times C'\times N}$。在第 k 个头中，整合所有通道的关系后可得 $\boldsymbol{E}_k\in\mathbb{R}^{C'\times N}$，同理可得 $\boldsymbol{D}_k\in\mathbb{R}^{C'\times C'}$。将注意力应用于 k 头中的所有向量，其结果可表示为

$$\boldsymbol{H}_k=\boldsymbol{D}_k\boldsymbol{E}_k \tag{14}$$

式中：$\boldsymbol{H}_k\in\mathbb{R}^{C'\times N}$。在每个组重复上述计算过程后，将其输出结果整合。经过重构后，即可获得经过注意力整合的特征图 $\boldsymbol{Y}\in\mathbb{R}^{C\times H\times W}$，即

$$\boldsymbol{Y}=\mathrm{reshape}\left(\mathrm{concat}\left(\boldsymbol{H}_1,\cdots,\boldsymbol{H}_h\right)\right) \tag{15}$$

将 $\boldsymbol{Y}$ 经过线性投影 $\boldsymbol{G}\in\mathbb{R}^{C\times C}$ 后，与可学习的标量 α（初始值为0）相乘，最后与 $\boldsymbol{X}$ 相加得到最终的输出为

$$\boldsymbol{Z}=\alpha\boldsymbol{Y}\boldsymbol{G}+\boldsymbol{X} \tag{16}$$

经过以上计算，实现了多头通道注意力机制。其可帮助网络更好的捕获特征之间的关联性，提升特征表示的强度和灵活性。

2　实验与分析

2.1　数据集

Cityscapes数据集[19]是一个广泛使用的大规模街景图像数据集，特别是在语义分割领域。它包含大约5 000张高分辨率图像，每张图的分辨率为2 048*1 024像素。这些图像捕获了多样的城市环境，包括不同的天气条件、时间段和季节变化。每张图像都附有精细的人工标注，包含像素级别的语义标签，如道路、建筑物、行人、车辆等。数据集被划分为3个主要部分，训练集包含约2 975张图像，验证集约500张图像，测试集约1 525张图像。

ADE20K[20]是一个专为语义分割任务设计的大规模图像数据集，包含大22 462张高分辨率图像，并给每个像素赋予了对应的语义标签，涵盖了超过150种不同的类别。数据集由训练集、测试集、验证集组成，分别包含1 3151,1 817,3 376张图像。

2.2　评价指标及实验设置

与其他语义分割方法一样，使用mIoU（平均交并比）、FPS（实时处理速度）、Param（参数量）作为评价指标来评估模型性能。mIoU用于衡量预测分割结果与真实标签之间的重叠程度，可表示为

$$mIoU=\frac{1}{N+1}\sum_{i=1}^{N}\frac{X_{ii}}{\sum_{j=0}^{N}X_{ij}+\sum_{j=0}^{N}X_{ji}-X_{ii}} \tag{17}$$

式中：N 为图像类别数；i，j 分别为不同类别；X_{ii} 为正确预测的像素数目；X_{ij} 为错误的将 i 预测为 j 的像素数目；X_{ji} 为错误的将 j 预测为 i 的像素数目。

本文实验是在Ubuntu22.04环境下基于Pytorch框架实现的。实验设备采用Intel Core i9-13900k，Nvidia GeForce RTX 4090。

在训练设置方面，损失使用交叉熵损失代价函数来计算，每次训练的批处理大小在Cityscapes和ADE20K中均设置为8。使用多项式衰减策略，在每次迭代后更新全局学习率

$$lr=lr_{\text{base}}\times\left(1-\frac{iter}{iter_{\max}}\right)^{power} \tag{18}$$

式中：lr_{base} 为初始学习率，设置为0.000 06，lr 为当前的学习率；$iter$ 为当前迭代次数；$iter_{\max}$ 为最大迭代次数，设置为160 000次；$power$ 为衰减控制参数，设置为0.9，优化器使用Adam，一阶动量参数与二阶动量系数分别为0.9和0.999，权重衰减系数为0.01。在数据增广方面，在两个数据集上使用随机裁剪和随机左右翻转。Cityscapes数据集图片尺寸被统一为769*769像素，ADE20K数据集图片尺寸被统一为512*512像素。

2.3　消融实验

为了验证本文所提的FRM和AAM模块的有效性，本文进行了对不同组件有效性的消融研究。在实验中，数据集使用Cityscapes数据集，主干网络使用CSWin Transformer，将使用空洞卷积的FCN作为Baseline。依次单独将FRM和AAM添加到网络中进行实验，以及同时将其添加到网络中进行实验，其余训练设置保持一致。表1给出了FRM和AAM模块的消融结果。

表1　在Cityscapes验证集上的消融实验结果

Tab.1　Results of ablation experiments on the Cityscapes

方法	Backbone	FRM	AAM	mIoU/%
Baseline	CSWin-T	–	–	78.6
本文	CSWin-T	√	–	81.5
本文	CSWin-T	–	√	81.8
本文	CSWin-T	√	√	**82.3**

根据实验结果显示，当仅使用FRM或AAM时，相较于Baseline，语义分割准确率分别提升了2.9个百分点和3.2个百分点。而在同时使用FRM和AAM时，相较于基线网络提升了3.7个百分点，与单独使用FRM或AAM相比分别提升了0.8个百分点与0.5个百分点。实验结果验证了FRM与AAM模块能有效提升语义分割性能，且两者联合作用能进一步提高网络精度。

图6展示了消融实验的可视化图像。第1列的图6(a)和图6(e)为原始图像；第2列的图6(b)和图6(f)为真值；第3列的图6(c)为去除FRM的结果，图6(g)为去除AAM的结果；第4列的图6(d)为加载FRM的结果，图6(h)为加载AAM的结果。通过比较图6(c)和图6(d)红框标注的部分，我们可以观察到使用本文提出的FRM后，网络减少了错误识别道路边缘区域的情况。“墙壁”与“道路白线外边缘”，“草坪”与“道路”，“道路”与“人行道”这些不同部分的边缘处分割更加清晰。这表明FRM强化了特征的深层语义，有效提升了网络的边缘分割能力。

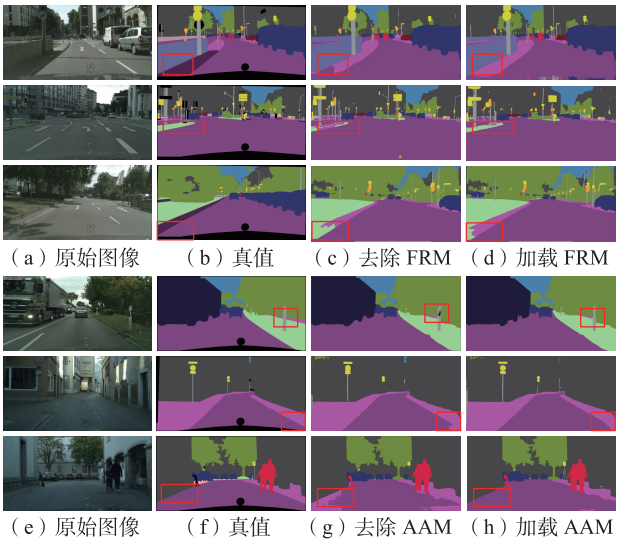

图6 消融实验的可视化

Fig. 6 Visualization of ablation experiments

同样地，观察图6(g)和图6(h)的红框标注部分并进行对比，我们可以发现当网络加载了AAM后，在“墙壁”与“道路”，“路标柱”与“草坪”这些物体的重叠处分割效果有明显提升。这说明对于多个对象重叠的复杂场景，网络对于相似类别的区分能力得到了提升。这表明AAM通过融合语义表征能力更强的通道特征与空间特征，使得网络能够更加有效地处理复杂场景下的类别模糊问题。

2.4 对比实验

2.4.1 在Cityscapes数据集上的评测结果

本文方法在Cityscapes数据集上与其他语义分割方法进行了对比。表2展示了这些方法在同一实验环境下的结果，包括分割精度(mIoU)、运行速度(FPS)以及参数量(Params)。

表2 不同方法在Cityscapes验证集上的实验结果

Tab.2 Results of different methods with the same experimental setup on the Cityscapes validation set

方法	Backbone	mIoU/%	FPS/(帧/s)	Params/(×10⁶个)
EncNet	ResNet50	74.2	1.04	35.89
NLNet	ResNet50	77.0	0.82	50.02
SETR-MLA	VIT-L	77.3	0.41	310.68
SegFormer	MIT-B1	78.6	4.30	13.68
Swin	Swin-T	79.2	7.40	59.94
GSS	Swin-L	80.0	2.78	65.76
SegNeXt	MSCAN-S	80.1	7.63	27.60
ConvNeXt	ConvNeXt-S	81.1	1.02	81.88
本文	CSWin-T	**82.3**	3.40	92.27

根据表中的数据，可以看出，本文的方法在分割精度上优于其他方法。与EncNet、NLNet、SETR-MLA、SegFormer、Swin、SegNeXt、ConvNeXt和GSS等方法相比，在mIoU上本文方法分别提高了8.1、5.3、5.0、3.7、3.1、2.2、1.2个百分点和2.3个百分点。这说明本文方法具有较高的分割精度。

在运行速度方面，本文方法的FPS为3.40帧/s，低于Swin-T的7.40帧/s，但快于其他多数方法。由此可见，本文方法在精度和运行速度方面能够保持较好的平衡。

图7展示了各方法的分割结果。通过观察红色框标记的区域，可以看出网络在处理边缘和保持物体完整性方面表现更为出色。特别是在“道路”与“草坪”的重叠处以及“道路分叉口”等复杂场景中，网络能够分割出更清晰的物体边界，避免了传统方法中常见的边缘分割不准问题。此外，在“墙壁”等建筑物的识别上，也能保持更好的完整性，错误率较少。此外，对于“草坪”与“树木”等相似类别物体，分割出来的轮廓清晰分明，避免了常见的类别模糊问题。同时，对于“路桩”等小物体的识别也非常精确，分割结果与人工标注几乎一致。这些视觉对比结果说明，本文网络不仅能够较好保持物体边缘，而且能有效提高小目标物体的识别准确率。

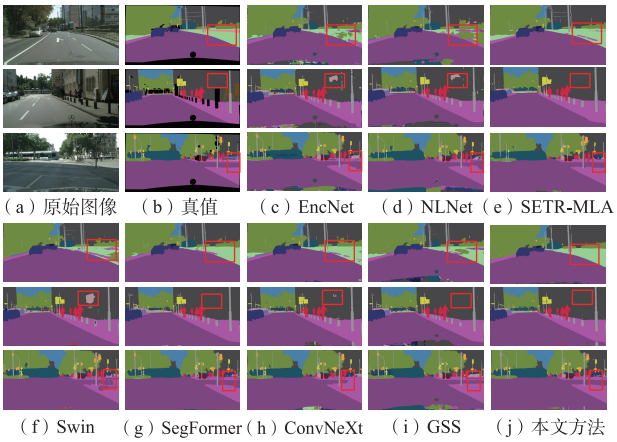

图7 语义分割结果

Fig. 7 Results of semantic segmentation

2.4.2 在ADE20K数据集上的评测结果

为了进一步验证本文网络的适应性和有效性，本文在具有高度挑战性的ADE20K数据集上与其他先进方法进行了比较。实验结果如表3所示，与EncNet、Segformer、NLNet、SegNeXt、Swin、

表3　不同方法在ADE20K验证集上的实验结果
Tab.3　Results of different methods with the same experimental setup on the ADE20K validation set

方法	Backbone	mIoU/%	FPS/(帧/s)	Params/(×10⁶个)
EncNet	ResNet50	40.1	14.87	35.89
SegFormer	MIT-B1	40.2	47.66	13.68
NLNet	ResNet50	42.0	13.97	50.02
SegNeXt	MSCAN-S	44.2	56.2	27.60
Swin	Swin-T	44.3	21.06	59.94
ConvNeXt	ConvNeXt-S	45.6	7.44	81.88
GSS	GSS-FF	46.3	11.58	65.76
SETR-MLA	VIT-L	46.4	4.68	310.68
本文	CSWin-T	**47.4**	17.26	92.27

ConvNeXt、SETR-MLA以及GSS网络方法相比，本文网络的分割精度分别提高了7.3、7.2、5.4、3.2、3.1、1.8、1.0、1.1个百分点。而且，本文网络方法具有17.26帧/s的实时处理速度。这说明本文网络方法在ADE20K数据集上同样保持较好的性能，具有较强的泛化性能，能够实时准确完成复杂场景的语义分割任务。

3　结论

1）采用CSWin Transformer作为编码器提取多尺度特征，可增强所提取的特征的远程建模能力，提高计算效率。

2）通过将FRM集成到解码器中对深层特征进行细化增强，同时将低分辨率的深层特征逐级与高分辨率的浅层特征进行融合，充分利用特征的语义信息与细节信息，增强了网络的语义辨析能力。

3）采用AAM模块从空间与通道两个维度上对特征进行融合。通过空间期望最大化注意力高效地捕获每个像素的全局上下文信息，以及通过多头通道注意力挖掘语义更显著的通道特征。

参考文献：

[1] MUHAMMAD K, HUSSAIN T, ULLAH H, et al. Vision-based semantic segmentation in scene understanding for autonomous driving: recent achievements, challenges, and outlooks[J]. IEEE Transactions On Intelligent Transportation Systems, 2022, 23(12): 22694-22715.

[2] DEWANGAN D K, SAHU S P, SAIRAM B, et al. VLDNet: vision-based lane region detection network for intelligent vehicle system using semantic segmentation[J]. Computing, 2021, 103(12): 2867-2892.

[3] SUN G L, LIU Y, DING H H, et al. Learning local and global temporal contexts for video semantic segmentation[J]. IEEE Transactions on Pattern Analysis And Machine Intelligence, 2024, 46(10):6919-6934.

[4] 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.

[5] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]// Munich: Proceedings of the European Conference on Computer Vision, 2018: 801-818.

[6] 王晓明, 温锐, 姚道金, 等. 基于改进DeepLabv3+的接触网开口销缺陷检测[J]. 华东交通大学学报, 2023, 40(5): 120-126.
WANG X M, WEN R, YAO D J, et al. Defect detection of the split pins in catenary based on improved DeepLabv3 + [J]. Journal of East China Jiaotong University, 2023, 40(5):120-126.

[7] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Salt Lake: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7794-7803.

[8] YIN M H, YAO Z L, CAO Y, et al. Disentangled non-local neural networks[C]// Glasgow: Proceedings of the 16th European Conference on Computer Vision, 2020: 191-207.

[9] ZHANG H, DANA K, SHI J P, et al. Context encoding for semantic segmentation[C]// Salt Lake: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7151-7160.

[10] GUO M H, LU C Z, HOU Q, et al. SegNeXt: rethinking convolutional attention design for semantic segmentation [J]. Advances in Neural Information Processing Systems, 2022, 35: 1140-1156.

[11] 周丽娟, 毛嘉宁. 视觉Transformer识别任务研究综述[J]. 中国图象图形学报, 2023, 28(10): 2969-3003.
ZHOU L J, MAO J N. Vision Transformer-based recognition tasks:a critical review[J]. Journal of Image and Graphics, 2023, 28(10): 2969-3003.

[12] ZHENG S, LU J, ZHAO H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective

with transformers[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2021: 6881-6890.

[13] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C] // Montreal: Proceedings of the IEEE International Conference on Computer Vision, 2021: 10012-10022.

[14] 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.

[15] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s [C]//New Orleans: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2022: 11976-11986.

[16] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Las Vegas: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016: 770-778.

[17] CHEN J Q，LU J C，ZHU X T，et al. Generative semantic segmentation[C]//Vancouver: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023: 7111-7120.

[18] DONG X Y, BAO J M, CHEN D, et al. CSwin Transformer: a general vision transformer backbone with cross-shaped windows[C]//New Orleans: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 12124-12134.

[19] KIRILLOV A, HE K M, GIRSHICK R, et al. Panoptic segmentation[C]//Long Beach：Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 9404-9413.

[20] ZHOU B L, ZHAO H, PUIG X, et al. Scene parsing through ADE20K dataset[C]//Honolulu：Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 633-641.

第一作者：彭洋(2000—)，男，硕士研究生，研究方向为计算机视觉和语义分割。E-mail：1172390843@qq.com。

通信作者：吴文欢(1985—)，男，副教授，博士，硕士生导师，研究方向为计算机视觉和图像处理等。E-mail：wuwenhuan5@163.com。

（责任编辑:吴海燕）