CN 51-1346/O4 ISSN 1003-501X (印刷版) ISSN 2094-4019 (网络版)

光电工程
Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

基于多尺度特征增强的高效Transformer语义分割网络

张艳，马春明，刘树东，孙叶美

引用本文:

张艳，马春明，刘树东，等. 基于多尺度特征增强的高效Transformer语义分割网络[J]. 光电工程，2024， **51**(12): 240237.

Zhang Y, Ma C M, Liu S D, et al. Multi-scale feature enhanced Transformer network for efficient semantic segmentation[J]. *Opto-Electron Eng*, 2024, **51**(12): 240237.

https://doi.org/10.12086/oee.2024.240237

收稿日期: 2024-10-10; 修改日期: 2024-11-16; 录用日期: 2024-11-19

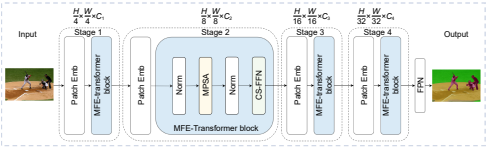
Opto-Electronic Engineering
光 电 工 程

Article

2024 年, 第 51 卷, 第 12 期

DOI: 10.12086/oee.2024.240237 CSTR: 32245.14.oee.2024.240237

基于多尺度特征增强的高效 Transformer 语义分割网络



张 艳，马春明，刘树东，孙叶美*

天津城建大学计算机与信息工程学院，天津 300380

摘要：针对现有基于 Transformer 的语义分割网络存在的多尺度语义信息利用不充分、处理图像时生成冗长序列导致的高计算成本等问题，本文提出了一种基于多尺度特征增强的高效语义分割主干网络 MFE-Former。该网络主要包括多尺度池化自注意力模块 (multi-scale pooling self-attention, MPSA) 和跨空间前馈网络模块 (cross-spatial feed-forward network, CS-FFN)。其中，MPSA 利用多尺度池化操作对特征图序列进行降采样，在减少计算成本的同时还高效地从特征图序列中提取多尺度的上下文信息，增强 Transformer 对多尺度信息的建模能力；CS-FFN 通过采用简化的深度卷积层替代传统的全连接层，减少前馈网络初始线性变换层的参数量，并在前馈网络中引入跨空间注意力 (cross-spatial attention, CSA)，使模型更有效地捕捉不同空间的交互信息，进一步增强模型的表达能力。MFE-Former 在数据集 ADE20K、Cityscapes 和 COCO-Stuff 上的平均交并比分别达到 44.1%、80.6% 和 38.0%，与主流分割算法相比，MFE-Former 能够以更低的计算成本获得具有竞争力的分割精度，有效改善了现有方法多尺度信息利用不足和计算成本高的问题。

关键词：语义分割；Transformer；深度学习；注意力机制

中图分类号：TP391.4 **文献标志码：**A

张艳，马春明，刘树东，等. 基于多尺度特征增强的高效 Transformer 语义分割网络 [J]. 光电工程，2024，**51**(12): 240237
Zhang Y, Ma C M, Liu S D, et al. Multi-scale feature enhanced Transformer network for efficient semantic segmentation[J]. *Opto-Electron Eng*, 2024, **51**(12): 240237

Multi-scale feature enhanced Transformer network for efficient semantic segmentation

Zhang Yan, Ma Chunming, Liu Shudong, Sun Yemei*

College of Computer and Information Engineering, Tianjin Chengjian University, Tianjin 300380, China

Abstract: To address the issues of insufficient utilization of multi-scale semantic information and high computational costs resulting from the generation of lengthy sequences in existing Transformer-based semantic segmentation networks, this paper proposes an efficient semantic segmentation backbone named MFE-Former, based on multi-scale feature enhancement. The network mainly includes the multi-scale pooling self-attention (MPSA) and the cross-spatial feed-forward network (CS-FFN). MPSA employs multi-scale pooling to downsample the feature map sequences, thereby reducing computational cost while efficiently extracting multi-scale contextual information, enhancing the Transformer's capacity for multi-scale information modeling. CS-FFN replaces the traditional fully connected layers with simplified depth-wise convolution layers to reduce the parameters in the initial

收稿日期：2024-10-10；修回日期：2024-11-16；录用日期：2024-11-19
基金项目：天津市哲学社会科学规划项目 (TJGL19XSX-045)
*通信作者：孙叶美，sunyemei1216@163.com。

linear transformation of the feed-forward network and introduces a cross-spatial attention (CSA) to better capture different spaces interaction information, further enhancing the expressive power of the model. On the ADE20K, Cityscapes, and COCO-Stuff datasets, MFE-Former achieves mean intersection-over-union (mIoU) scores of 44.1%, 80.6%, and 38.0%, respectively. Compared to mainstream segmentation algorithms, MFE-Former demonstrates competitive segmentation accuracy at lower computational costs, effectively improving the utilization of multi-scale information and reducing computational burden.

Keywords: semantic segmentation; transformer; deep learning; attention mechanism

1 引　言

近年来，随着人工智能领域的发展，生成式大模型开始逐渐受到重视，ChatGPT 和 SORA 就是其技术成功的实际应用案例。通过学习大量数据，生成训练数据相似的内容，在自然语言处理、图像生成等领域中展现出了巨大的潜力和应用价值。例如，生成对抗网络 (GAN)[1] 被用于生成更真实的训练数据，以增强模型的泛化能力；变分自编码器 (VAE)[2] 通过学习数据的潜在表示，被用于半监督任务中，以提高模型在有限标注数据下的性能。生成式模型擅长生成数据和内容，但在深度学习的有监督学习领域中，卷积神经网络 (CNN) 依然占据主导地位，并在处理图像分类[3-4]、语义分割[5-6]、实例分割[7-8] 等计算机视觉任务的性能优异。目前，基于 CNN 的语义分割算法[9-10] 的卷积操作存在固有的局部性，导致模型的感受野受限。然而，感受野对于语义分割任务至关重要，较大的感受野可以为网络提供更丰富的上下文信息，帮助模型做出更准确的判断。尽管通过堆叠多个卷积层可以扩大模型的感受野[11]，增强模型对深层特征的学习能力，但仍不能有效缓解 CNN 在建模长距离依赖关系方面的短板问题。

受自然语言处理 (natural language processing, NLP) 领域的启发[12]，基于 Transformer 的主干网络在计算机视觉任务中展现出巨大的潜力。Transformer 通过其独特的自注意力机制，能够有效地处理全局范围内的信息交互，相较于 CNN 所依赖的受限有效感受野而言，展现出了显著的优势。自 Dosovitskiy 等人[13] 将 Transformer 引入到图像视觉任务以来，诸多学者尝试利用 Transformer 模型来解决语义分割问题[14-16]，并取得了显著的成果。在视觉 Transformer 中，图像被划分为序列化的图像子块 (patch)，这些图像子块类似自然语言处理中的词序列被进行编码，但编码后得到的序列长度远长于自然语言处理中的序列，导致在多头自注意力机制层 (multi-head self-attention, MHSA) 中，需要进行大规模的矩阵乘法运算，极大地增加了计算负担。这也是将 Transformer 直接从 NLP 领域引入计算机视觉领域所面临的主要挑战。为了解决上述问题，Wang 等人[17] 提出了一种通过单一池化操作缩短序列长度来减少计算量的方案，但图像中不同元素和位置的相对重要性各异，单一池化操作无法充分捕捉不同感受野下的多尺度特征，导致原始序列中的部分信息丢失。此外，传统的前馈网络 (feed-forward network, FFN) 采用多层感知器 (multilayer perceptron, MLP) 来增强模型的表示能力，但其全连接架构导致每个 Transformer 块中存在大量参数，并且其不擅长学习空间关系。因此，如何在优化计算资源的同时更好地捕获空间关系，成为了当前亟待解决的关键问题。

为了解决上述问题，本文提出了一种基于多尺度特征增强的高效 Transformer 语义分割网络 MFE-Former。该网络由编码器和解码器 FPN[18] 构成，其中编码器由 4 个 Patch Embed 块和 MFE-Transformer 块组成。具体而言，MFE-Transformer 块主要包括多尺度池化自注意力模块和跨空间前馈网络模块。本文的主要贡献总结如下：

1) 提出 MFE-Former 网络，该网络融合了 Transformer 和 CNN 的优点，采用多尺度池化操作和跨空间特征融合方法，捕获了丰富的特征信息，并减少了计算资源消耗。该模型在三个标准语义分割的数据集上表现出色，同时保持了较低的计算成本和高性能的表现。

2) 设计多尺度池化自注意力模块，在 Transformer 的自注意力模块中采用并行池化操作，有效减少了图像特征的序列长度，提高了模型运算效率。此外，利用不同大小的池化因子同时对输入特征进行降采样得到不同尺度的特征图，捕捉多尺度的上下文信息，MPSA 对获得的多尺度特征图进行加权融合，增强模

型对于复杂背景中多尺度目标的识别能力，从而提高语义分割的整体性能。

3) 设计跨空间前馈网络模块，首先，利用简化的深度卷积层替代全连接层，有效减少了模型的参数量，同时保持了输入数据的空间结构。然后，引入跨空间注意力，通过局部和通道交互分支的协同作用，建立不同空间的依赖关系，从而改善前馈网络在捕捉空间关系方面的不足，为前馈网络提供更为精确的特征表示。最后，利用跨空间特征融合方法使得特征具有更丰富的细节信息和语义信息，提高网络对不同尺寸目标的分割精度。

2 相关工作

2.1 基于 CNN 的模型

近年来，卷积神经网络强大的特征提取能力使语义分割技术取得了显著的进步。FCN[19] 是语义分割领域中的开创性工作，为后续的研究奠定了基础。为解决传统卷积操作感受野受限的问题，DeepLab[6] 和 PSPNet[9] 引入了空洞卷积和空洞空间金字塔池化模块，有效地扩大了模型的感受野，并且能捕捉到不同尺度的上下文信息。此外，注意力机制因其在特征依赖关系建模方面的超强能力而受到广泛关注，如 SENet[20] 通过引入通道注意力机制，增强了网络对重要特征的关注。Wang 等人[21] 进一步提出了 ECANet，一种无需降维的局部跨通道交互方法，并自适应地选择一维卷积核大小，以增强特征的表达能力。为了更好地处理长程依赖关系，Non-local[22] 通过计算不同位置特征之间的相似性来加权特征，有效地建模了这种依赖性。DANet[23] 通过在空间和通道维度上并行引入注意力模块，成功捕捉了长程特征依赖性，从而提升了分割性能。

尽管这些方法显著提升了模型的性能，但模型计算成本相对较高，特别是在处理高分辨率图像时，会显著增加计算成本。为了解决该问题，轻量级网络应运而生。ICNet[24] 通过使用多尺度图像作为输入，并结合了低级空间细节和高级语义信息，提高了计算效率。EspNet[25] 通过将卷积分解为点卷积和扩张卷积，大幅减少了参数量和计算量。然而，由于传统的 CNN 通过局部区域的滑动窗口卷积操作来提取特征，这种操作本质上是局部的，模型在处理长程依赖关系方面仍存在局限性。

2.2 基于 Transformer 的模型

自从 Dosovitskiy 等人[13] 首次提出视觉 Transformer 模型并将图像视为 16×16 像素块的集合以来，该理论在大规模数据集上的图像分类任务中取得了突破性的成果。ViT 的自注意力机制能计算所有位置的像素关联性，可以有效地获取整个场景的全局上下文信息，从而促进了基于 Transformer 的骨干网络的快速发展。DPT[26] 引入了一种新颖的 Patch 嵌入方法，将输入图像划分为不重叠的图像块，并通过可学习的位置嵌入来计算每个图像块的空间位置，实现自适应地划分图像块。Liu 等人[27] 提出了一种新的内存高效设计，采用“三明治”式布局构建模块，并引入级联组注意力模块，通过为不同的注意力头提供输入特征的不同部分来减少计算冗余，有效提升了计算效率并增强了模型的表示能力。在语义分割领域中，基于 Transformer的模型也展示出了巨大的潜力。SETR[16] 使用 ViT 作为语义分割的骨干网络，在特征提取方面取得了显著效果。Mask2Former[28] 将交叉注意力限制在预测掩码的前景区域，从中获取局部特征，提高了注意力机制的效率。SegFormer[29] 引入了一种分层编码器结构获得多尺度特征，并在解码器中对这些特征进行融合，进一步提升了分割性能。SeaFormer[30] 结合了轴向压缩和细节增强的设计，提高了移动语义分割的效率。尽管这些基于 Transformer 的方法在捕捉全局上下文方面表现出色，但目前的视觉 Transformer 及其衍生模型仍具有大量的参数和较高的计算复杂度。此外，它们在增强多尺度特征表示的能力方面仍有改进的空间。

2.3 轻型视觉 Transformer 模型

Transformer 的计算复杂度会随着图像尺寸呈现平方式增长，导致计算负荷显著增加。早期的研究算法主要通过优化 Transformer 的自注意力机制来减少计算量，如 Swin Transformer[14] 通过局部窗口机制减少了图像块序列长度，有效降低了自注意力的计算开销，但该策略违背了 Transformer 的全局性特征，限制了模型全局感受野。Zhang 等人[31] 提出了一种高效的多头自注意力方法，使用简单的深度卷积来减少内存使用，进一步提高了效率。CvT[32] 模型利用深度卷积生成查询、键和值张量，并将其应用于多头自注意力的计算，以提高模型计算效率。MobileViT[33] 和 Mobile-Former[34] 将 Transformer 架构引入到 MobileNet 中，以增强全局语义理解能力。PVTv2[35] 和 MViT[36]

通过单一池化操作减少了 MHSA 层中图像块的数量，从而降低了计算复杂度。但这类单尺度池化操作可能会牺牲模型整体的鲁棒性，导致细节丢失。

3 本文模型及网络结构

3.1 网络结构设计

本文提出的基于多尺度特征增强的高效 Transformer 语义分割网络如图 1 所示。对于输入图像 $\boldsymbol{X}\in\mathbb{R}^{H\times W\times 3}$，首先经过四个层次的递进式特征提取，获得多尺度的输出特征 $\boldsymbol{X}_{\text{out}}^{i}\in\mathbb{R}^{H_i\times W_i\times C_i}$，其中 H_i、W_i 和 $C_i(i=\{1,2,3,4\})$ 分别代表高度、宽度和通道；最后，解码器采用 FPN 自上而下的融合多尺度特征，生成语义分割结果，MFE-Former 的各阶段详细设置如表 1 所示。

3.2 多尺度池化自注意力模块

多尺度池化自注意力 (Multi-scale pooling self-attention, MPSA) 模块由通道缩减操作、多尺度池化操作和多头自注意力模块组成，MPSA 模块的结构如图 1(a) 所示。首先，输入特征 $\boldsymbol{X}_{\text{in}}\in\mathbb{R}^{H\times W\times C}$，通过 1×1 卷积实现通道缩减操作。该操作通过减少输出通道的维度，将参数量降低至原来的 $1/r$，有效减少了模型的参数量，通道缩减率 r 的具体数值如表 1 所示，经过通道缩减操作后，得到低维特征 $\boldsymbol{X}_{\text{in}}'\in\mathbb{R}^{H\times W\times \frac{C}{r}}$：

$$\boldsymbol{X}_{\text{in}}'=Conv(\boldsymbol{X}_{\text{in}})\,. \tag{1}$$

然后，将低维特征 $\boldsymbol{X}_{\text{in}}'$ 送入多尺度池化操作，具体操作如图 2 所示，生成多尺度分层表示特征

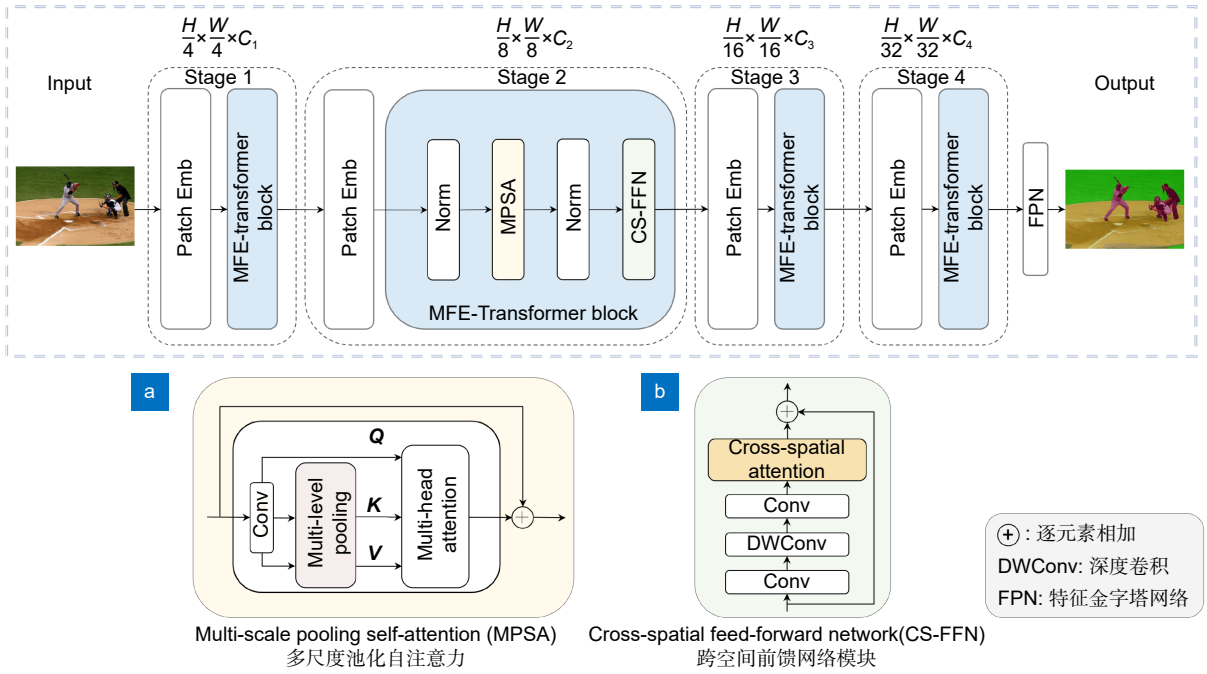


图 1 基于多尺度特征增强的高效 Transformer 语义分割网络。(a) 多尺度池化自注意力模块；(b) 跨空间前馈网络模块

Fig. 1 An efficient Transformer-based semantic segmentation network enhanced by multi-scale features. (a) Multi-scale pooling self-attention module; (b) Cross-spatial feed-forward network module

表 1 MFE-Former 的详细设置，输出大小代表每个阶段输出的图片分辨率，操作代表第 i 阶段所使用的操作。此外，参数设置包括每个 MFE-Transformer 块使用通道缩减率 r 和池化因子 p_l

Table 1 Detailed settings of the MFE-Former, output size refers to the resolution of the output at each stage, and operation represents the operations used in the i stage. Additionally, the parameter settings include the channel reduction ratio r and the pooling factor p_l used by each MFE-Transformer block

阶段	输出大小	操作	参数设置
1	128×128×48	Patch Embed, MFE-Transformer block×2	r=1, {p_1=8, p_2=16, p_3=24, p_4=32, p_5=40}
2	64×64×96	Patch Embed, MFE-Transformer block×2	r=2, {p_1=4, p_2=8, p_3=12, p_4=16, p_5=20}
3	32×32×260	Patch Embed, MFE-Transformer block×6	r=4, {p_1=2, p_2=4, p_3=6, p_4=8, p_5=10}
4	16×16×384	Patch Embed, MFE-Transformer block×3	r=8, {p_1=1, p_2=2, p_3=3, p_4=4, p_5=5}

$\boldsymbol{P}_j(j\in 1,2,...,n)$，其中 n 为池化层数。

$$\boldsymbol{P}_j = Pooling_j(\boldsymbol{X}_{\text{in}}'), \tag{2}$$

式中：$Pooling_j(j\in 1,2,...,n)$为第 j 层的池化操作。在多尺度池化操作中，采用了一组特定的池化因子 $p_l(l=\{1,2,3,4,5\})$。因此，这些不同大小的池化因子允许模型在不同尺度上捕捉图像特征。较小的池化因子更侧重于捕捉局部细节和边缘信息，而较大的池化因子则有助于捕捉更广泛的上下文和全局特征。网络中每个阶段的具体值如表 1 所示。经过池化操作后，$\boldsymbol{P}_j$的空间分辨率远小于$\boldsymbol{X}_{\text{in}}'$，因此多尺度池化操作有效降低了图像块序列长度，提高了计算效率；同时$\boldsymbol{P}_j$中的每个元素代表$\boldsymbol{X}_{\text{in}}'$中大小为 $\frac{H\times W}{p_l^2}$ 个像素的图像区域，从而可以提取丰富的上下文信息。随后，利用深度卷积对$\boldsymbol{P}_j$进行编码获得相对位置信息，为后续自注意力计算提供特征间的空间关系，更多地保留了特征的空间结构，增强特征表示能力。同时，添加残差连接缓解梯度消失问题，增加信息流动，得到相对位置编码的输出$\boldsymbol{P}_j^*$：

$$\boldsymbol{P}_j^* = DWConv(\boldsymbol{P}_j) + \boldsymbol{P}_j, \tag{3}$$

其中：DWConv代表 3×3 深度卷积。对输出结果$\boldsymbol{P}_j^*$进行扁平化和连接，得到输出多尺度特征图 $\boldsymbol{P}$：

$$\boldsymbol{P} = LayerNorm(Concat(\boldsymbol{P}_1^*, \boldsymbol{P}_2^*, \cdots, \boldsymbol{P}_j^*)), \tag{4}$$

式中：$LayerNorm$ 为归一化层，$Concat$ 为拼接操作。

将多尺度池化操作的输出 $\boldsymbol{P}$ 送入多头自注意力模块，计算键和值张量，具体计算方法为

$$(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = (\boldsymbol{X}'\boldsymbol{W}^q, \boldsymbol{P}\boldsymbol{W}^k, \boldsymbol{P}\boldsymbol{W}^v), \tag{5}$$

其中：$\boldsymbol{W}^q$、$\boldsymbol{W}^k$和$\boldsymbol{W}^v$为分别用于生成查询、键和值张量的权重矩阵。之后，注意力模块接收 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 计算注意力特征$Attention$：

$$Attention = Softmax\left(\frac{\boldsymbol{Q}\times\boldsymbol{K}^{\text{T}}}{\sqrt{d_K}}\right)\times\boldsymbol{V}, \tag{6}$$

其中：$\sqrt{d_K}$用于对点积进行缩放，d_K是键向量 $\boldsymbol{K}$ 的通道维度。最后，多尺度池化自注意力模块输出为

$$\boldsymbol{X}_{\text{att}} = LayerNorm(\boldsymbol{X}_{\text{in}} + Attention), \tag{7}$$

其中：$\boldsymbol{X}_{\text{att}}$为多尺度池化自注意力模块的输出。

3.3 跨空间前馈网络模块

Transformer 通常使用多层感知器 (multilayer perceptron，MLP) 作为前馈网络，但其对空间关系的学习能力有限，且 MLP 中的全连接层会产生较大的参数量。为了解决该问题，本文提出了跨空间前馈网络模块 (cross-spatial feed-forward network, CS-FFN)，如图 1(b) 所示，由卷积操作和跨空间注意力 (cross-spatial attention, CSA) 组成，可表述为

$$\boldsymbol{X}_{\text{out}} = CSA(Conv(DWConv(Conv(\boldsymbol{X}_{\text{in}})))) + \boldsymbol{X}_{\text{in}}. \tag{8}$$

首先，使用 1×1 卷积替代传统 MLP，与传统 MLP 相比，1×1 卷积通过独立学习每个输出通道的权重，而非为每个输入-输出连接定义权重，有效降低了参数量。此外，CS-FFN 通过深度卷积的参数共享机制，进一步减少参数量，同时保留空间结构，以高效提取局部特征。

然后，引入跨空间注意力改进前馈网络，捕捉空间和通道依赖性。如图 3 所示，跨空间注意力由局部

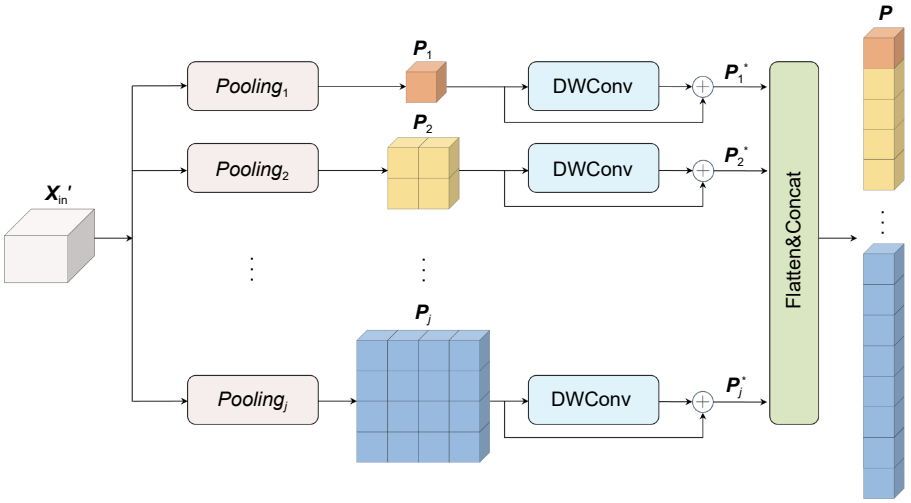


图 2　多尺度池化操作

Fig. 2　Multi-scale pooling operation

交互分支、通道交互分支和跨空间特征融合组成。对于输入特征图$\boldsymbol{X}_{\mathrm{FFN}} \in \mathbb{R}^{H\times W\times C}$，将其沿通道维度划分为$N$个子特征$\boldsymbol{X}_i=[\boldsymbol{X}_{\mathrm{FFN}_0},\boldsymbol{X}_{\mathrm{FFN}_1},...,\boldsymbol{X}_{\mathrm{FFN}_{N-1}}],\boldsymbol{X}_i \in \mathbb{R}^{\frac{C}{N}\times H\times W}$，其中每个子特征$\boldsymbol{X}_i$对应一个特定的组，使空间语义特征在每个特征组内均匀分布，在局部范围内增强跨通道交互，从而有效降低计算复杂度并提升特征表示能力。跨空间注意力的特征提取过程具体可以分成三个步骤完成：

1) 局部交互分支

局部交互分支聚焦于局部区域内通道间的信息交互，通过卷积操作促进局部特征的融合，并利用Softmax激活函数和全局平均池化操作，得到局部特征增强和压缩后的通道特征$\boldsymbol{F}_{\mathrm{L}_1} \in \mathbb{R}^{\frac{C}{N}\times H\times W}$和$\boldsymbol{F}_{\mathrm{L}_2} \in \mathbb{R}^{\frac{C}{N}\times 1\times 1}$，提高模型对局部特征变化的敏感度，为跨空间特征融合提供了丰富的局部特征表示，局部交互分支的输出可描述为

$$(\boldsymbol{F}_{\mathrm{L}_1},\boldsymbol{F}_{\mathrm{L}_2})=\boldsymbol{F}_{\mathrm{Local}}(\boldsymbol{X}_i), \tag{9}$$

其中：$\boldsymbol{F}_{\mathrm{Local}}$表示局部交互分支处理过程的函数，接收输入特征图$\boldsymbol{X}_i$并输出特征图$\boldsymbol{F}_{\mathrm{L}_1}$和$\boldsymbol{F}_{\mathrm{L}_2}$。

2) 通道交互分支

首先，通过全局平均池化分别沿水平和垂直方向对特征图的每个通道进行编码，以获取每个通道在这两个方向上的平均值，从而捕获全局空间信息。随后，将特征向量在通道维度上进行拼接，形成了一个融合多尺度空间信息的特征。接下来，通过1×1卷积操作进一步混合通道信息，生成通道注意力图。通过逐元素相乘的方式，将每个组内的通道注意力图进行交互，实现跨通道的特征融合。最终，采用组归一化操作来稳定特征分布，并通过全局平均池化操作和Softmax激活函数进一步优化注意力权重图，通道交互分支的输出可描述为

$$(\boldsymbol{F}_{\mathrm{C}_1},\boldsymbol{F}_{\mathrm{C}_2})=\boldsymbol{F}_{\mathrm{Channel}}(\boldsymbol{X}_i), \tag{10}$$

其中：$\boldsymbol{F}_{\mathrm{Channel}}$表示通道交互分支处理过程的函数，接收输入特征图$\boldsymbol{X}_i$并输出特征图$\boldsymbol{F}_{\mathrm{C}_1} \in \mathbb{R}^{\frac{C}{N}\times 1\times 1}$和$\boldsymbol{F}_{\mathrm{C}_2} \in \mathbb{R}^{\frac{C}{N}\times H\times W}$。

3) 跨空间特征融合

跨空间特征融合主要作用是融合局部交互分支和通道交互分支的特征输出，通过计算局部交互分支的特征输出$\boldsymbol{F}_{\mathrm{L}_1}$和$\boldsymbol{F}_{\mathrm{L}_2}$以及通道交互分支的特征输出$\boldsymbol{F}_{\mathrm{C}_1}$和$\boldsymbol{F}_{\mathrm{C}_2}$，生成包含多尺度空间信息$\boldsymbol{A}_1$和丰富通道信息的注意力图$\boldsymbol{A}_2$。然后再将注意力图与原始特征图相结合，并通过Sigmoid函数进行加权，得到最终的输出特征图$\boldsymbol{X}_{\mathrm{out}}$。跨空间特征融合保留了特征的局部信息，而且通过注意力机制强化了特征表示，增强了模型对空间和通道依赖关系的捕捉能力。跨空间特征融合可描述为

$$\boldsymbol{X}_{\mathrm{out}}=\boldsymbol{X}_i\times Sigmoid(\boldsymbol{A}_1+\boldsymbol{A}_2). \tag{11}$$

跨空间注意力所需的额外计算成本很低，通过局部和通道交互分支的协同作用，实现了对特征图中空间和通道信息的融合，增强 Transformer 模型的特征表示能力，为其提供了更为丰富和精确的特征信息。

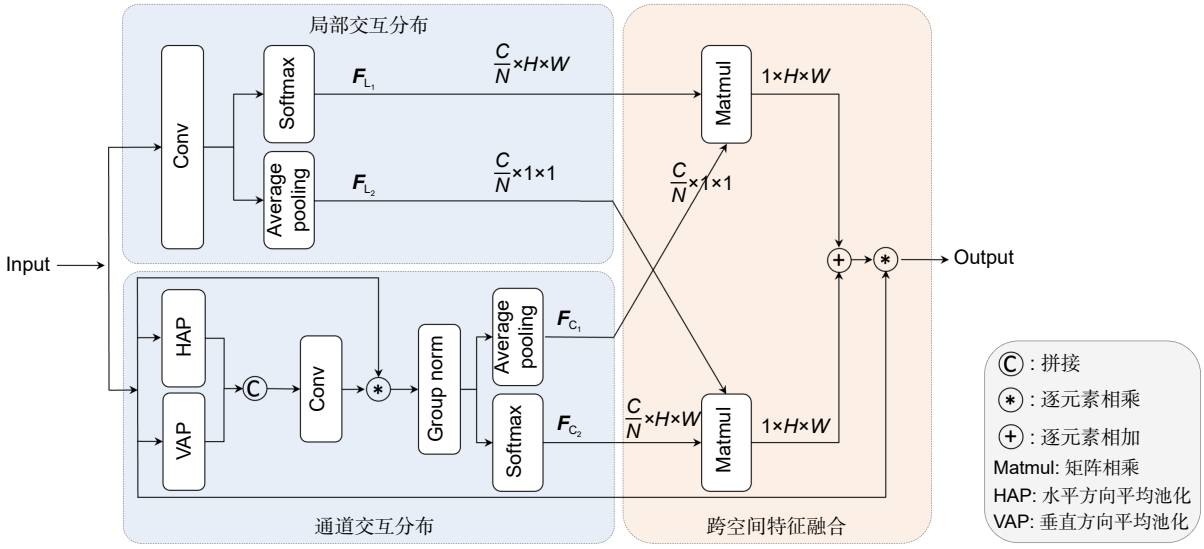


图 3 跨空间注意力

Fig. 3 Cross-spatial attention

4 实验与结果分析

4.1 数据集

本文使用三个被广泛使用的标准数据集进行了测试，主要包括 ADE20K[37]、Cityscapes[38] 和 COCO-Stuff[39]。ADE20K 的训练集中有 20210 张图片，验证集中有 2000 张图片，包含 150 个语义类别，该数据集场景解析难度较大。Cityscapes 包含 5000 张高清照片，分为 19 个类别，主要包含复杂的城市场景。在训练、验证和测试中，分别有 2975/500/1525 张图片。COCO-Stuff 有 1000 张测试图像和 9000 张训练图像，数据集中的 182 个类别并没有全部出现在测试分割中。为了进行实验，根据 MMsegmentation 的实现[40]，使用了 171 个类别。上述数据集场景多样、类别丰富、图像大小不一，对语义分割网络的泛化提出了很高的要求。

4.2 实验设置

实验所用的硬件平台为搭载 Intel Core i9-10900K 处理器并配置 NVIDIA GeForce RTX 3080 Ti 显卡的服务器。软件平台环境为 Ubuntu18.04 操作系统、Python 3.6 编译器环境、PyTorch 1.10 深度学习框架。采用商汤科技开源深度学习工具 MMsegmentation 框架来训练和测试网络。在训练过程中，连续使用了随机水平翻转、随机调整大小 (比例在 0.5 到 2.0 之间) 和随机裁剪作为数据增强方法。对于所有数据集，使用 AdamW[41] 优化器更新模型参数，初始学习率为 2×10^{-5}，权重衰减为 10^{-3}。

4.3 评价指标

为了对本文方法进行正确的评估，本文采用图像语义分割领域常用的度量指标：整体准确率 (overall accuracy, *aAcc*) 用于反映模型在所有预测中的整体准确性，其数学表达式为

$$aAcc=\frac{\sum_{i=1}^{n}TP_i}{\sum_{i=1}^{n}(TP_i+FN_i)}, \tag{12}$$

其中：TP_i表示第 i 类别正确预测结果的正类别数，FN_i表示第 i 类别错误预测结果的负类别数，n 是类别的总数。平均准确率 (mean accuracy, *mAcc*) 用于反映模型在每个类别上的准确性，其数学表达式为

$$mAcc=\frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i+FN_i}. \tag{13}$$

平均 Dice 系数 (Mean dice coefficient, *mDice*) 能反映类别级别的性能评估和衡量预测与真实区域的重叠程度，其数学表达式为

$$mDice=\frac{1}{n}\sum_{i=1}^{n}\frac{2\times|P_i\cap G_i|}{|P_i|+|G_i|}, \tag{14}$$

其中：P_i是预测结果中第 i 类的像素集合，G_i是真实标签中第 i 类的像素集合，$|P_i\cap G_i|$表示预测结果与真实标签交集的像素数，$|P_i|$和$|G_i|$分别表示预测结果和真实标签中第 i 类的像素总数，n 是类别的总数。平均交并比 (mean intersection over union, *mIoU*) 用于计算某个具体类别真实标签与预测值两个集合的交集和并集之比的平均值，其数学表达式为

$$mIoU=\frac{1}{n+1}\sum_{i=0}^{n}\frac{TP}{FN+FP+TP}, \tag{15}$$

其中：*TP* 代表正确预测结果的正类别数，*FP* 代表错误预测结果的正类别数，*FN* 代表错误预测结果的负类别数，*mIoU* 值越高，模型的分割性能越好。此外，采用每秒浮点运算次数 (floating-point operations per second, FLOPs) 和模型参数数量 (params) 来描述模型的复杂度，它们的值越大，模型越复杂，所需的计算成本也越高。

4.4 文中方法与经典方法对比分析

本文在三个标准数据集上进行了语义分割实验，并对 MFE-Former 和先前在上述数据集上建立的最先进的方法进行比较分析。此外，还进行了一系列消融研究和超参数设置实验，以找出该方法的不同部分对性能的潜在影响。

4.4.1 ADE20K 数据集实验结果分析

表 2 为图像大小为 512×512 分辨率下的性能结果。相较于如 FCN、ResNet、PSPnet 和 DeeplabV3+等基于 CNN 的语义分割网络，MFE-Former 的语义分割精度有显著提升，得益于 MFE-Transformer 的自注意力机制，使模型能够捕捉到图像中的长程依赖关系。与基于 Transformer 的 PVT-Tiny 相比，MFE-Former 参数量和每秒浮点运算次数分别降低了 1.1 M 和 2.1 G，mIoU 却提高了 8.4%。与轻量高效的 EdgeViT-XS(27.7 G) 相比，MFE-Former(31.1 G) 以更低的每秒浮点运算次数实现了更高的 mIoU(提高了 2.7%)。MFE-Former 的参数量仅为 PoolFormer-M48 的 20.6%，但实现的 mIoU 仍比 PoolFormer-M48 高 2.4%。实验结果表明，与其他取得最优分割效果的模型相比，MFE-Former 有一定的竞争力。

表 2 不同分割模型在 ADE20K 数据集上的模型评估结果

Table 2 Model evaluation results of different segmentation models on ADE20K dataset

Method	#Param/M	FLOPs/G	mIoU/%
FCN[19]	9.8	39.6	19.7
ResNet18[11]	15.5	32.2	32.9
PSPNet[9]	13.7	52.2	29.6
DeepLabV3+[6]	15.4	25.9	38.1
ViT[13]	10.2	24.6	37.4
PVT-Tiny[17]	17.0	33.2	35.7
PoolFormer-S12[42]	15.7	—	37.2
Conv-PVT-Tiny[43]	16.4	—	37.2
EdgeViT-XS[44]	10.3	27.7	41.4
Swin-tiny [14]	31.9	46.0	41.5
PVT-Large[17]	65.1	79.6	42.1
Segformer-B1[29]	13.7	15.9	42.2
PoolFormer-M48[42]	77.1	—	42.7
Twins-SVT-S[45]	28.3	37.0	43.2
Xcit-T12/16[46]	33.7	—	43.5
TBFormer-T[47]	20.5	24.3	42.8
SCTNet-B[48]	17.4	—	43.0
MFE-Former (Ours)	**15.9**	**31.1**	**44.1**

4.4.2 Cityscapes 数据集实验结果分析

表 3 为该模型与先进的方法在 Cityscapes 数据集上的结果。与 DeepLabV3+ (MV2) 和 EncNet 相比，MFE-Former 的每秒浮点运算次数分别降低了 316.4 G 和 1509.4 G，同时在准确性方面表现出色，mIoU 分别提高了 5.4% 和 3.7%，表明 MFE- Former 模型在保持高准确度的同时，还具有出色的计算效率。与 RTFormer-Base 模型 (16.8 M) 相比，MFE-Former 能以获得更高的 mIoU 精度，进一步证实了 MFE-Former 在语义分割任务中的具有良好的鲁棒性。

4.4.3 COCO-Stuff 数据集实验结果分析

COCO-Stuff 数据集中包含大量从 COCO 数据集中收集的困难分割示例，实验分割结果如表 4 所示。由于采用了多尺度融合和跨空间特征融合方法，MFE-Former 相比于其他卷积神经网络和 Transformer 模型取得了更好的结果。与参数量接近的 DeepLabV3+相比，MFE-Former 在 mIoU 上提高了 8.1%。MFE-Former 的参数量和每秒浮点运算次数分别为 MaskFormer 的 38.7% 和 58.6%，在大幅减少参数和计算量的同时提高了 mIoU(0.9%)，实现了 38.0% mIoU。实验结果表明，MFE-Former 模型在保持较低计算开销的同时，性能表现优异。

4.4.4 模型综合分析

本节在 ADE20K、Cityscapes 和 COCO-stuff 三个数据集上提供更全面的分析，选取平均交并比、整体准确率、平均准确率和平均 Dice 系数作为评价指标。针对具有相似参数量和每秒浮点运算次数的 Segformer-B1 模型和 SCTNet-B 模型进行对比分析，实验结果如表 5 所示。在 ADE20K 数据集上，MFE-Former 的 mIoU 较 Segformer-B1 提高了 2.0%，较 SCTNet-B 提高了 1.1%。同时，MFE-Former 在 *aAcc*、*mAcc* 和 *mDice* 上也分别超越了对比模型。尽管在 Cityscapes 数据集上 SCTNet-B 的 *mDice* 略高，但 MFE-Former 在 *mIoU*、*aAcc* 和 *mAcc* 上均表现更佳，尤其是在 *mAcc* 上比 Segformer-B1 高出 4.3%。在 COCO-Stuff 数据集上，MFE-Former 模型在多个关键性能指标上均优于 Segformer-B1。具体来说，MFE-Former 的 *mIoU*、*aAcc*、*mAcc*、*mDice* 比 Segformer-B1 分别高出 2.1%、2%、0.6%、0.2%。实验结果表明，MFE-Former 模型在参数量和每秒浮点运算次数接近的情况下，整体和平均准确率以及分割区域的重叠度方面表现良好。

4.5 可视化

本文将 MFE-Former 模型与实验中表现较好并且

参数量接近的 Segformer-B1 模型和 PoolFormer-S12 模型在 ADE20K 数据集上的实验结果进行了可视化展示，如图 4 所示。第一列中 MFE-Former 能够连续且完整地分割画框的轮廓与背景像素，而 Segformer-B1 和 PoolFormer-S12 则出现了将背景墙壁像素错分为画框一部分的情况。此外，MFE-Former 在处理人物边缘时也表现出了更高的精度，边缘更加精细，而其他两种方法则显得相对粗糙。第二列中 MFE-Former 能够准确地将人物、草地和比赛场地进行区分和分割，而 Segformer-B1 和 PoolFormer-S12 在这些区域的分割中存在误检和不完整的问题。这表明 MFE-Former 在处理复杂场景和多目标分割时具有更强的准确性和完整性。在第三列中 Segformer-B1 和 PoolFormer-S12 橱柜上的物体分割存在困难，且对地毯的分割也不够完整，而 MFE-Former 能够正确地分割橱柜上的物体，实现地毯边缘更清晰的分割。与对比方法相比，MFE-Former 实验结果整体的分割精度更高，对小目标的分割效果更好，不同类别的边界也更加清晰。这些优势源于 MFE-Former 在多尺度特征提取和跨空间融合的设计，其能够在处理高分辨率和

表 3　在 Cityscapes 数据集上的模型评估结果 (FLOPs 的测试在 1024×2048 分辨率下进行)

Table 3　The model evaluation results on the Cityscapes dataset (the FLOPs test was performed at a resolution of 1024×2048)

Method	#Param/M	FLOPs/G	mIoU/%
FCN[19]	9.8	317	61.5
PSPNet[9]	13.7	423	70.2
DeepLabV3+[6]	15.4	555	75.2
SwiftNetRN[49]	11.8	104	75.5
EncNet[50]	55.1	1748	76.9
PVT-Tiny[17]	17.0	—	71.7
MLT[51]	20.1	—	77.4
RTFormer-Bas[52]	16.8	—	79.3
SFNet(ResNet-18)[53]	12.87	247.0	78.9
DDRNet-39[54]	32.3	281.2	80.4
PIDNet-L[55]	36.9	275.8	80.6
MFE-Former (Ours)	**15.9**	**238.6**	**80.6**

表 4　与 COCO-stuff 数据集上先进的模型进行比较 (采用 512×512 输入分辨率的进行测试)

Table 4　Compare with the state-of-the-art models on the COCO stuff dataset (testing was conducted using an input resolution of 512×512)

Method	#Param/M	FLOPs/G	mIoU/%
PSPNet[9]	13.7	52.9	30.1
DeepLabV3+[6]	15.4	25.9	29.9
LR-ASPP[56]	—	2.37	25.2
MaskFormer[57]	41	53	37.1
TBFormer-T[47]	20.5	37.5	37.9
SCTNet-B[48]	17.4	—	35.9
MFE-Former (Ours)	**15.9**	**31.1**	**38.0**

表 5　三种模型在 ADE20K、Cityscapes 和 COCO-stuff 数据集上的分割性能指标

Table 5　The segmentation performance metrics of 3 models on the ADE20K, Cityscapes, and COCO-stuff datasets

Method	ADE20K				Cityscapes				COCO-stuff			
	mIoU/%	aAcc/%	mAcc/%	mDice/%	mIoU/%	aAcc/%	mAcc/%	mDice/%	mIoU/%	aAcc/%	mAcc/%	mDice/%
Segformer-B1	42.1	79.6	52.7	56.3	78.5	95.8	83.4	85.4	—	—	—	—
SCTNet-B	43.0	80.2	56.1	57.2	80.1	96.4	86.6	88.3	35.9	67.1	48.5	47.9
MFE-Former(Ours)	44.1	81.0	56.1	57.4	80.6	96.5	87.7	87.5	38.0	69.1	49.1	48.1

复杂场景图像时，展现出更优的分割性能。

4.6 消融实验

4.6.1 不同池化方法的影响

在研究 MPSA 性能的实验中，池化策略对模型的表现有着显著的影响。为了评估不同池化方法对 MPSA 性能的具体影响，本文设计了一系列消融实验，涵盖了四种不同的模型配置：采用最大池化的模型、采用多尺度最大池化的模型、采用平均池化的模型以及采用多尺度平均池化的模型。不同模型包含不同类型池化的统计结果如表 6 所示。

从表 6 可以看出，采用多尺度最大池化和多尺度平均池化的实验结果明显优于不采用多尺度池化的方法，其中池化方法没有可训练的参数，采用 FLOPs 和 mIoU 作为评估指标，评估在 ADE20K 数据集进行。虽然只增加了 0.4 G 每秒浮点运算次数，但精度分别提高了 2.3% 和 1.6%，充分说明多尺度池化在降低计算成本的同时，也带来了丰富的语义信息。多尺度平均池化的方法在四种策略中表现最为优异，该结果表明，多尺度平均池化能更有效地捕捉和利用图像内在的多尺度信息，并且能在不显著增加计算负担的情况下，有效地提高模型的识别性能和泛化力。

4.6.2 不同池化因子的影响

本节实验对比了不同池化因子以及并行多尺度池化操作在第一阶段的性能，实验结果如表 7 所示，其中 R 表示不同尺度池化前后特征图序列的大小比。由表 7 可以看出，如果使用较大的比率 (如 16、24、32、40)，虽然可以减少每秒浮点运算次数，但会对模型识别局部特征的能力产生负面影响，导致准确率显著下降。当使用较小的比率 (如 4) 时，可以观察到性能基本达到饱和。当比率设置为 8 时，单个池化操作可

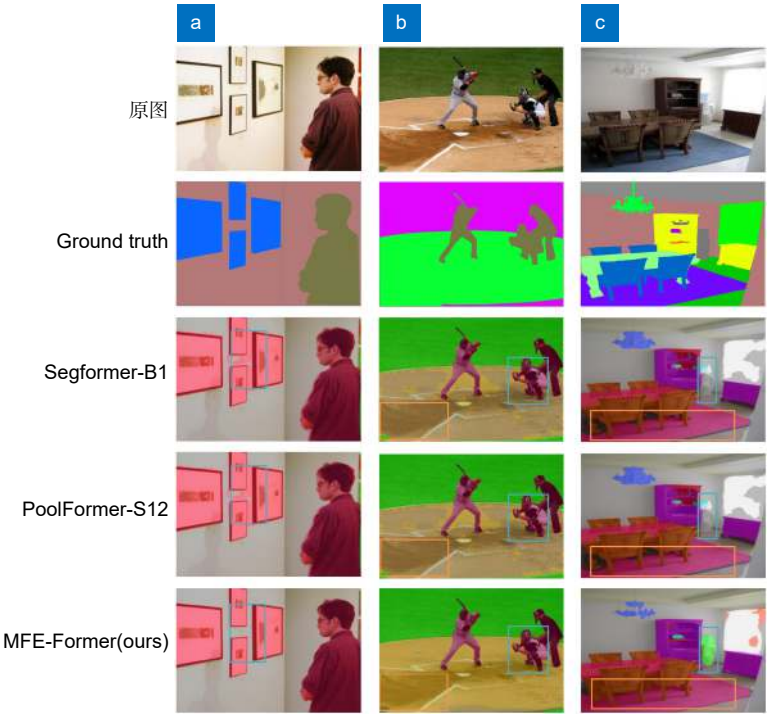


图 4 不同算法在 ADE20K 数据集上分割结果可视化

Fig. 4 Visualization of segmentation results of different algorithms on the ADE20K dataset

表 6 不同池化方法的消融实验

Table 6 Experiment results on ablation using different pooling methods

池化方法	FLOPs/G	mIoU/%
最大池化	30.7	40.8
多尺度最大池化	31.1	43.1
平均池化	30.7	42.5
多尺度平均池化	**31.1**	**44.1**

达到理想性能，而在基于 8 的比率实施并行多尺度池化后，有效地将不同尺度的信息整合到了模型中，这增强了模型对不同分辨率输入的适应能力，优化了模型识别局部特征的能力；而且并行应用三个池化操作的模型可以更显著提高模型的准确性，但当并行池化操作扩展到五个时，模型性能达到最优，且此时的每秒浮点运算次数依然低于采用比率为 4 的设置。该实验结果表面，适度增加并行池化操作的数量，可以在不显著增加计算负担的情况下，提高模型的分割性能。

4.6.3 跨空间注意力不同组成部分的影响

跨空间特征融合对于融合局部分支和通道分支特征之间的依赖关系至关重要。为了验证跨空间特征融合在局部交互分支和通道交互分支中采用跨空间注意力结构的有效性，本实验设置了三种对比：不添加跨空间注意力、仅添加通道交互分支和添加跨空间注意力，实验结果如表 8 所示。由表 8 可以看出，添加跨空间注意力的模型比不添加跨空间注意力模型的 mIoU 提升了 1%。仅添加通道交互分支的模型，由于没有采用跨空间注意力的特征融合，缺乏局部图像语义信息，精确度没有明显提升，而引入跨空间注意力的模型显著增强了分割细节和上下文信息的提取能力。实验结果表明，利用跨空间特征融合将通道和局部的优势结合起来，增强了模型对图像的理解能力。

4.6.4 跨空间注意力的参数设置

超参数 N 控制着 CSA 模块中通道的大小，为了研究 CSA 中超参数 N 对模块的影响，通过改变超参数 N 以评估其对模型整体性能的影响，结果如表 9 所示。当 N 设为 16 时，减少了参数和每秒浮点运算次数，但压缩较低的维度限制了模型学习更多信息的能力，导致模型的语义分割性能降低。N 设为 4 与 N 设为 8 相比，参数增加了 0.4 M，浮点运算次数增加了 0.7 G，但性能并没有明显提升，反而下降，说明 N 的减小不仅不会单调地提高模型性能，还会增加参数量。当 N 设定为 8 时，不仅确保了模型的准确性，也兼顾了模型的计算效率，使模型在准确性和计算复杂性之间达到了最佳平衡，且在多个评估指标上都优于 N 设为 4 和 16 的情况。因此，本文最佳模型将 N 设为 8 作为 CSA 模块的默认配置。

5 结 论

本文针对现有语义分割网络在多尺度信息利用和计算效率方面的不足问题，提出了一种基于多尺度特征增强的高效语义分割主干网络 MFE-Former。设计的多尺度池化自注意力模块和跨空间前馈网络模块，

表 7 池化因子参数设置的消融实验和 ADE20K 数据集的参数设置实验结果

Table 7 Experiment results on ablation with multi-scale pooling ratio parameter settings and experimental results of parameter setting for ADE20K dataset

池化因子p_l					R	FLOPs/G	mIoU/%
4	—	—	—	—	16	31.5	42.3
8	—	—	—	—	64	30.9	42.5
16	—	—	—	—	256	30.6	41.9
24	—	—	—	—	576	30.5	41.3
32	—	—	—	—	1024	30.5	40.9
40	—	—	—	—	1600	30.5	40.4
8	16	24	—	—	47	31.1	43.7
8	**16**	**24**	**32**	**40**	**43**	**31.1**	**44.1**

表 8 在 ADE20K 数据集上对跨空间注意力的不同组成部分进行的消融实验

Table 8 An ablation experiment was conducted on different components of cross spatial attention on the ADE20K dataset

Setting	#Param/M	FLOPs/G	mIoU/%
不添加跨空间注意力	15.8	30.3	43.1
仅添加通道交互分支	15.8	30.3	43.3
添加跨空间注意力	**15.9**	**31.1**	**44.1**

表 9 在 ADE20K 数据集上对 CSA 模块中的超参数 N 的参数设置进行的消融实验

Table 9 Ablation study on the hyperparameter N in the CSA module on the ADE20K dataset

N	#Param/M	FLOPs/G	mIoU/%
4	16.3	31.8	43.9
8	**15.9**	**31.1**	**44.1**
16	15.8	30.8	43.3

不仅有效地提取了多尺度上下文信息，减少了计算资源消耗，还增强了模型对不同空间信息的捕捉能力。在 ADE20K、Cityscapes 和 COCO-Stuff 数据集上的实验结果显示，MFE-Former 在平均交并比上分别达到了 44.1%、80.6% 和 38.0%，均优于同量级经典模型算法。此外，虽然 MFE-Former 在提升语义分割的效率和性能方面取得了显著进展，但在生成式模型方面探索不足。为了进一步增强模型的泛化能力，未来的研究将探索整合生成式语义分割模型的优势。

利益冲突：所有作者声明无利益冲突

参考文献

[1] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. *Commun ACM*, 2020, **63**(11): 139−144.

[2] Kingma D P. Auto-encoding variational Bayes[Z]. arXiv: 1312.6114, 2013. https://doi.org/abs/1312.6114.

[3] Jiang W T, Dong R, Zhang S C. Global pooling residual classification network guided by local attention[J]. *Opto-Electron Eng*, 2024, **51**(7): 240126.
姜文涛, 董睿, 张晟翀. 局部注意力引导下的全局池化残差分类网络[J]. 光电工程, 2024, **51**(7): 240126.

[4] He F T, Wu Q Q, Yang Y, et al. Research on laser speckle image recognition technology based on transfer learning[J]. *Laser Technol*, 2024, **48**(3): 443−448.
贺锋涛, 吴倩倩, 杨祎, 等. 基于深度学习的激光散斑图像识别技术研究[J]. 激光技术, 2024, **48**(3): 443−448.

[5] Zhang C, Huang Y P, Guo Z Y, et al. Real-time lane detection method based on semantic segmentation[J]. *Opto-Electron Eng*, 2022, **49**(5): 210378.
张冲, 黄影平, 郭志阳, 等. 基于语义分割的实时车道线检测方法[J]. 光电工程, 2022, **49**(5): 210378.

[6] 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.

[7] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: common objects in context[C]//*Proceedings of the 13th European Conference on Computer Vision -- ECCV 2014*, 2014: 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

[8] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, 2017: 2980–2988. https://doi.org/10.1109/ICCV.2017.322.

[9] 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.

[10] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[Z]. arXiv: 1412.7062, 2014. https://doi.org/abs/1412.7062.

[11] 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.

[12] 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.

[13] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]//*Proceedings of ICLR 2021*, 2021.

[14] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//*Proceedings of ICCV 2021*, 2021.

[15] Strudel R, Garcia R, Laptev I, et al. Segmenter: transformer for semantic segmentation[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 7242–7252. https://doi.org/10.1109/ICCV48922.2021.00717.

[16] Zheng S X, Lu J C, Zhao H S, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 6877–6886. https://doi.org/10.1109/CVPR46437.2021.00681.

[17] Wang W H, Xie E Z, Li X, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 548–558. https://doi.org/10.1109/ICCV48922.2021.00061.

[18] 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.

[19] 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.

[20] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.

[21] Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155.

[22] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7794–7803. https://doi.org/10.1109/CVPR.2018.00813.

[23] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 3141–3149. https://doi.org/10.1109/CVPR.2019.00326.

[24] Zhao H S, Qi X J, Shen X Y, et al. ICNet for real-time semantic segmentation on high-resolution images[C]//*Proceedings of the 15th European Conference on Computer Vision* (*ECCV*), 2018: 418–434. https://doi.org/10.1007/978-3-030-01219-9_25.

[25] Mehta S, Rastegari M, Caspi A, et al. ESPNet: efficient spatial pyramid of dilated convolutions for semantic segmentation[C]//*Proceedings of the 15th European Conference on Computer Vision* (*ECCV*), 2018: 561–580. https://doi.org/10.1007/978-3-030-01249-6_34.

[26] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 12159–12168. https://doi.org/10.1109/ICCV48922.2021.01196.

[27] Liu X Y, Peng H W, Zheng N X, et al. EfficientViT: memory efficient vision transformer with cascaded group attention[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 14420–14430. https://doi.org/10.1109/CVPR52729.2023.01386.

[28] Cheng B W, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 1280–1289. https://doi.org/10.1109/CVPR52688.2022.00135.

[29] Xie E Z, Wang W H, Yu Z D, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[C]//*Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021: 924.

[30] Wan Q, Huang Z L, Lu J C, et al. SeaFormer: squeeze-enhanced axial transformer for mobile semantic segmentation[C]//*Proceedings of ICLR 2023*, 2023.

[31] Zhang Q L, Yang Y B. ResT: an efficient transformer for visual recognition[C]//*Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021: 1185.

[32] Wu H P, Xiao B, Codella N, et al. CvT: introducing convolutions to vision transformers[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision* (*ICCV*), 2021: 22–31. https://doi.org/10.1109/ICCV48922.2021.00009.

[33] Mehta S, Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer[Z]. arXiv: 2110.02178, 2021. https://doi.org/abs/2110.02178.

[34] Chen Y P, Dai X Y, Chen D D, et al. Mobile-Former: bridging MobileNet and transformer[C]//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 5260–5269. https://doi.org/10.1109/CVPR52688.2022.00520.

[35] Wang W H, Xie E Z, Li X, et al. Pvt v2: improved baselines with pyramid vision transformer[J]. *Comp Visual Media*, 2022, **8**(3): 415−424.

[36] Yuan L, Chen Y P, Wang T, et al. Tokens-to-token ViT: training vision transformers from scratch on ImageNet[C]//*Proceedings of CVPR 2021*, 2021: 558–567.

[37] Zhou B L, Zhao H, Puig X, et al. Scene parsing through ADE20K dataset[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 5122–5130. https://doi.org/10.1109/CVPR.2017.544.

[38] 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.

[39] Caesar H, Uijlings J, Ferrari V. COCO-stuff: thing and stuff classes in context[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 1209–1218. https://doi.org/10.1109/CVPR.2018.00132.

[40] Contributors M M S. MMSegmentation: openmmlab semantic segmentation toolbox and benchmark[EB/OL]. (2020). https://github.com/open-mmlab/mmsegmentation.

[41] Loshchilov I, Hutter F. Decoupled weight decay regularization[C]//*Proceedings of ICLR 2019*, 2019.

[42] Yu W H, Luo M, Zhou P, et al. MetaFormer is actually what you need for vision[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 10809–10819. https://doi.org/10.1109/CVPR52688.2022.01055.

[43] Zhang X, Zhang Y. Conv-PVT: a fusion architecture of convolution and pyramid vision transformer[J]. *Int J Mach Learn Cyber*, 2023, **14**(6): 2127−2136.

[44] Pan J T, Bulat A, Tan F W, et al. EdgeViTs: competing light-weight CNNs on mobile devices with vision transformers[C]//*Proceedings of the 17th European Conference on Computer Vision*, 2022: 294–311. https://doi.org/10.1007/978-3-031-20083-0_18.

[45] Chu X X, Tian Z, Wang Y Q, et al. Twins: revisiting the design of spatial attention in vision transformers[C]//*Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021: 716.

[46] El-Nouby A, Touvron H, Caron M, et al. XCiT: cross-covariance image transformers[C]//*Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021: 1531.

[47] Wei C, Wei Y. TBFormer: three-branch efficient transformer for semantic segmentation[J]. *Signal, Image Video Process*, 2024, **18**(4): 3661−3672.

[48] Xu Z Z, Wu D Y, Yu C Q, et al. SCTNet: single-branch CNN with transformer semantic information for real-time segmentation[C]//*Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 2024: 6378–6386.

https://doi.org/10.1609/aaai.v38i6.28457.

[49] Oršic M, Krešo I, Bevandic P, et al. In defense of pre-trained ImageNet architectures for real-time semantic segmentation of road-driving images[C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 12599–12608. https://doi.org/10.1109/CVPR.2019.01289.

[50] Zhang H, Dana K, Shi J P, et al. Context encoding for semantic segmentation[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7151–7160. https://doi.org/10.1109/CVPR.2018.00747.

[51] Zhou Q, Sun Z H, Wang L J, et al. Mixture lightweight transformer for scene understanding[J]. *Computers and Electrical Engineering*, 2023, **108**: 108698.

[52] Wang J, Gou C H, Wu Q M, et al. RTFormer: efficient design for real-time semantic segmentation with transformer[C]//*Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022: 539.

[53] Li X T, You A S, Zhu Z, et al. Semantic flow for fast and accurate scene parsing[C]//*Proceedings of the 16th* European Conference on Computer Vision, 2020: 775–793. https://doi.org/10.1007/978-3-030-58452-8_45.

[54] Pan H H, Hong Y D, Sun W C, et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes[J]. *IEEE Trans Intell Transp Syst*, 2023, **24**(3): 3448−3460.

[55] Xu J C, Xiong Z X, Bhattacharyya S P. PIDNet: a real-time semantic segmentation network inspired by PID controllers[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 19529–19539. https://doi.org/10.1109/CVPR52729.2023.01871.

[56] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3[C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 1314–1324. https://doi.org/10.1109/ICCV.2019.00140.

[57] Cheng B W, Schwing A G, Kirillov A. Per-pixel classification is not all you need for semantic segmentation[C]//*Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021: 1367.

作者简介

张艳 (1982-)，女，教授，主要研究方向为：图像处理、计算机视觉和深度学习。

E-mail：yanzhang0910@163.com

刘树东 (1965-)，男，教授，主要研究方向为：网络通信与物联网技术、嵌入式技术及其应用。

E-mail：liushudong@tuc.edu.cn

马春明 (2000-)，男，硕士研究生，主要研究方向为语义分割。

E-mail：machunming0201@163.com

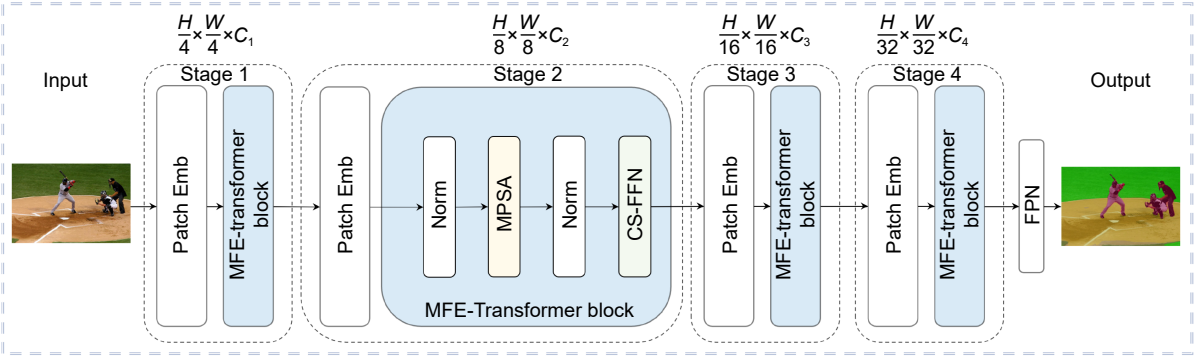
【通信作者】孙叶美 (1990-)，女，硕士，实验师，2019 年至今从事研究基于深度学习的小目标检测、语义分割、图像超分辨率重建等以及人工智能其他延伸应用方向。

E-mail：sunyemei1216@163.com

扫描二维码，获取PDF全文

Multi-scale feature enhanced Transformer network for efficient semantic segmentation

Zhang Yan, Ma Chunming, Liu Shudong, Sun Yemei*



An efficient Transformer-based semantic segmentation network enhanced by multi-scale features

Overview: In recent years, advancements in deep learning have propelled the field of semantic segmentation forward, resulting in the development of numerous innovative algorithms. The approach of employing extensive datasets to train deep learning models that automatically extract features has become the predominant method in semantic segmentation. Since Dosovitskiy introduced the Transformer to image vision tasks, many scholars have attempted to use Transformer models to address semantic segmentation issues, achieving notable results. In visual Transformers, the sequence length obtained after image encoding is much longer than the sequences in natural language processing. This leads to the need for large-scale matrix multiplication operations in the multi-head self-attention mechanism layers, significantly increasing the computational burden. This is also the main challenge faced when directly introducing Transformers from the NLP field to the computer vision field. PVT proposed a solution to reduce the computation by shortening the sequence length through a single pooling operation. However, the relative importance of different elements and positions in the image varies, and a single pooling operation cannot fully capture the multi-scale features under different receptive fields, leading to the loss of some information in the original sequence. Moreover, the traditional feed-forward network uses multi-layer perceptrons to enhance the model's representational power, but its fully connected architecture results in a large number of parameters in each Transformer block, and it is not adept at learning spatial relationships. In response to the aforementioned issues, this paper introduces an efficient semantic segmentation backbone network based on multi-scale feature enhancement, named MFE-Former. The network mainly includes the multi-scale pooling self-attention (MPSA) module and the cross-spatial feed-forward network (CS-FFN) module. The MPSA utilizes multi-scale pooling operations to downsample the feature map sequence, achieving a reduction in computational costs while efficiently extracting multi-scale contextual information from the feature map sequence, enhancing the Transformer's ability to model multi-scale information. The CS-FFN replaces the traditional fully connected layers with simplified depth convolutional layers, reducing the parameter count of the initial linear transformation layer in the feed-forward network, and introduces the cross-spatial attention module, enabling the model to more effectively capture interactions between different spatial regions and further enhancing the model's expressive power. The MFE-Former achieves mIoU of 44.1%, 80.6%, and 38.0% on the datasets ADE20K, Cityscapes, and COCO-Stuff, respectively. Compared to mainstream segmentation algorithms, MFE-Former can achieve competitive segmentation accuracy at a lower computational cost, effectively improving the issues of insufficient utilization of multi-scale information and high computational costs in existing methods.

Zhang Y, Ma C M, Liu S D, et al. Multi-scale feature enhanced Transformer network for efficient semantic segmentation[J]. *Opto-Electron Eng*, 2024, **51**(12): 240237; DOI: 10.12086/oee.2024.240237

Foundation item: Project supported by Tianjin Philosophy and Social Sciences Planning Project (TJGL19XSX-045)

College of Computer and Information Engineering, Tianjin Chengjian University, Tianjin 300380, China
* E-mail: sunyemei1216@163.com