朱发汛, 孙伟, 汤勃, 等. 基于 Segformer 的非显著性目标语义分割算法研究[J]. 智能计算机与应用,2025,15(10):10-15.
DOI:10.20169/j.issn.2095-2163.251002

基于 Segformer 的非显著性目标语义分割算法研究

朱发汛, 孙 伟, 汤 勃, 赵晓柯
(武汉科技大学 机械自动化学院, 武汉 430081)

摘 要: 语义分割是计算机视觉领域的重要分支之一,然而在实际工作场景中,非显著性目标由于其尺寸较小和边界模糊的特性,容易被语义分割网络忽略和分割不完整。本文基于 Segformer 提出了一种针对非显著性目标的语义分割算法。首先,为获取非显著性目标更加细致的特征,本文添加了小尺度特征,保证非显著性目标能够被精准识别。然后,为进一步提高算法特征提取能力并缓解计算负担,设计了 XM-FFN 模块,该模块采用了高效的 XSepConv 并使用了计算更为简便的 ReLU 激活函数,进一步减少了算法参数量。最后,实验采用 VOC2012 数据集,以 Segformer 为参照,在 mIoU、mF1 和 MPA 指标上分别提升了 2.58%、2.35%和 2.49%,验证了该算法的有效性。

关键词: 语义分割; Segformer; 非显著性目标
中图分类号: TP242 **文献标志码**: A **文章编号**: 2095-2163(2025)10-0010-06

Research for semantic segmentation algorithm of non-salient targets based on Segformer

ZHU Faxun, SUN Wei, TANG Bo, ZHAO Xiaoke
(School of Machinery and Automation, Wuhan University of Science and Technology, Wuhan 430081, China)

Abstract: Semantic segmentation is one of the important branches in the field of computer vision. However, in practical working scenarios, non-salient targets are easy to be ignored and incompletely segmented by semantic segmentation networks due to their small size and fuzzy boundaries. In this paper, a semantic segmentation algorithm for non-salient targets is proposed based on Segformer. Firstly, in order to obtain more detailed features of non-salient targets, this paper adds small-scale features to ensure that insignificant objects can be accurately recognized. Then, in order to further improve the algorithm's feature extraction ability and alleviate the computational burden, the XM-FFN module is designed. The XM-FFN module adopts the highly efficient XSepConv and uses the ReLU activation function, which is easier to compute, to further reduce the number of algorithmic parameters. Finally, the experiment adopts the VOC2012 dataset with Segformer as the reference, and improves 2.58%, 2.35% and 2.49% in mIoU, mF1 and MPA indexes, respectively, which verifies the effectiveness of the algorithm.

Key words: semantic segmentation; Segformer; non-salient targets

0 引 言

语义分割是计算机视觉领域中的一项重要任务,其目的是给图像中的每个像素分配一种语义类别,使得计算机系统能够更加深入、更加全面地理解图像内容。语义分割在许多领域都得到了广泛关注与应用,如在自动驾驶领域中帮助车辆理解障碍物、行人、交通标识等不同元素,提高车辆在复杂车况下决策的鲁棒性;在医学图像领域中分割和识别不同的组织结构,提高医生对病情诊断的效率;在城市规划领域中识别建筑物、道路、公园和水域等不同区域的特征,协助相关人员制定城市后续的规划与发展。

主流的语义分割主要分为基于 CNN 的语义分割和基于 Transformer 的语义分割。其中,CNN 通过卷积操作实现了对局部信息的感知,这使得分割网络能够有效地捕捉图像中的局部特征。此外,深度 CNN 可以学习高层次的抽象特征表示,有助于提高语义分割模型对图像中复杂对象边界和结构的识别

基金项目: 国家自然科学基金(51874217)。
作者简介: 朱发汛(1998—),男,硕士研究生,主要研究方向:机器视觉。
通信作者: 孙 伟(1990—),男,博士,讲师,主要研究方向:机器人,机器视觉。Email:sw@wust.edu.cn。
收稿日期: 2024-01-03

能力。FCN[1]是语义分割领域中的开创性工作，首次将深度学习引入该任务，并以端到端的预测为特点，为后续语义分割研究奠定了基础。PSPNet[2]提出了池化金字塔结构，充分地整合了上下文信息，再通过对各尺度特征进行级联融合，在语义级别实现了像素预测分类。Deeplabv3+[3]通过添加一个简化的解码器模块解决了修复物体边缘信息的问题。此外，引入了深度可分离卷积到空洞空间金字塔池化（ASPP）和解码器模块中，从而使网络在性能上更为高效。杨大伟等学者[4]提出一种基于边界辅助的弱监督语义分割网络，通过结合边界信息和语义信息，为种子区域的生长提供了指导，使得这些种子区域能够自然生长至目标边界，有效地解决了在目标被遮挡或重叠时无法对目标类别正确区分的问题。DECANet[5]引入改进的空洞空间金字塔池化结构，通过对多尺度特征进行融合，从而避免了图像细节信息的丢失，并在加快模型收敛速度方面取得了显著的效果。SegNet[6]提出 Pooling Indices，能够在编码器特征图中捕获和存储边界信息。LAC-UNet[7]在 U-Net[8]语义分割模型中引入了胶囊结构，利用胶囊向量获得更为精细的空间结构，进一步采用了局部路由算法、并引入了空间与通道权重，以增强对局部上下文线索的捕捉能力。

Transformer 结构通过自注意力机制（Self-Attention）可以有效地捕获输入序列中任意 2 个位置之间的关系。在语义分割任务中，这意味着模型可以更好地理解图像中不同区域之间的全局上下文关系，而不仅仅是局部特征。越来越多的学者尝试将 Transformer 应用于语义分割。Segmenter[9]以 ViT[10]（Vision Transformer）为基础扩展到语义分割模型，并提出了一种基于 Transformer 的解码器生成类掩码的方法，以有效地捕捉上下文信息。孙万春等学者[11]通过引入空间注意力交换层，扩大了类激活图的覆盖范围，并设计了一个注意力自适应模块来引导模型增强弱区域的类响应，从而有效提高了弱监督图像的语义分割性能。DCaT[12]利用深度可分离卷积提取图像的局部语义，并通过基于坐标感知和动态稀疏混合注意力的轻量级 Transformer 获取全局语义，最后融合局部和全局语义信息构建了一个轻量级的语义分割模型。AFFormer[13]提出了一种专门用于语义分割的无头轻量级架构，从频率角度学习聚类原型的局部描述表示，而不是直接学习所有的像素嵌入特征。兰建平等学者[14]采用短期密集级联（STDC）网络提取图像特征构建了一种轻量的实时语义分割网络，在计算和实时性方面取得平衡。Swin-UNet[15]构建了一种跳跃连接的对称编解码器结构，实现从局部到全局的自关注，具备良好的性能和泛化能力。TransMANet[16]提出了一种双分支解码器的网络结构，能够增强浅层网络的语义信息并融合局部和全局上下文。TransUNet[17]采用了 CNN-Transformer 混合结构，通过对 Transformer 编码的自注意特征进行上采样并与不同分辨率的 CNN 特征结合，巧妙地利用低层 CNN 特征实现了精确定位。李俊杰等学者[18]提出了多窗口注意力聚合的方式，结合 Swin Transformer[19]能够实现更加精确的注意力计算。

然而，当前已有的语义分割模型在处理非显著性目标时，仍然存在着诸多挑战与不足，究其原因就在于：

（1）非显著性目标通常尺寸比较小。语义分割模型对小尺寸目标进行处理时，目标可能被忽视，没有得到有效的分割。

（2）非显著性目标通常具有模糊的边界，尤其是目标与周围环境相似或者存在光照变化时。在这些区域语义分割模型可能会产生不准确的分割结果，边界定位可能不够清晰。

针对以上问题，本文设计了一种基于 Segformer[20]的语义分割网络模型，在原本的模型基础上添加了更小的尺度信息，保证在分割非显著性目标时，能捕捉到不同目标的细节特征。更加丰富的尺度信息有助于捕捉目标边界的细微特征，提高分割的精度，减少边界模糊现象造成的影响。通过该方法可以有效地提高网络对图像内容的全局理解，从而改善分割性能。为降低算法的计算参数，本文采用 XM-FFN 模块来平衡算法效率和性能。

1　Segformer 模型

1.1　Segformer 结构

Segformer 是 2021 年提出的基于 Transformer 框架的语义分割网络。Segformer 提供了多个模型选择、从 SegFormer-B0 到 SegFormer-B5，适用于不同规模和要求的任务，在多个数据集上都取得了显著的效果。本文采用的是 SegFormer-B0 模型，其结构如图 1 所示。Segformer 模型结构分为编码器和解码器两部分，其中编码器由分层 Transformer 模块组成，用于生成多尺度特征；解码器则使用的是轻量级的 MLP（Multi-Layer Perceptron），融合多层特征并上采样，最终解决分割任务。

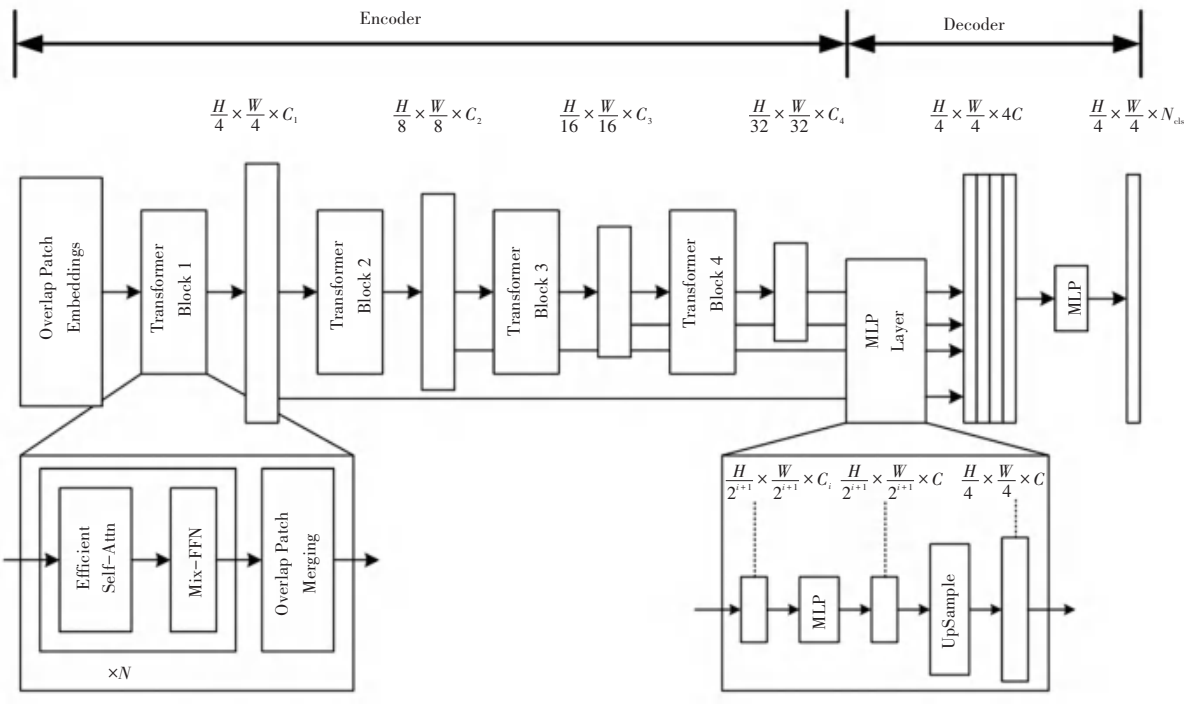

图 1 Segformer 模型结构图

Fig. 1 Segformer model structure diagram

1.2 Efficient Self-Attention 模块

Transformer 模型的计算量主要集中在自注意力机制层上，Segformer 为降低模型整体的计算复杂度，添加了缩放因子 R 来降低每一个自注意力机制模块的计算复杂度。传统的自注意力机制公式如下：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\text{head}}}}\right)\boldsymbol{V} \quad (1)$$

由式(1)可知，传统的自注意力复杂度为 $O(N^2)$，导致无法计算高分辨率的图像，而 Segformer 为了减少计算量，对输入维度为 $N \times C$ 的 $\boldsymbol{K}$ 矩阵进行 Reshape，变成 $N/R \times (C \cdot R)$ 的大小，然后进行线性变换，将 $(C \cdot R)$ 的维度变成 C。这样输出的 $\boldsymbol{K}$ 大小就变成了 $N/R \times C$，$\boldsymbol{V}$ 矩阵也进行同样的操作。Efficient Self-Attention 公式如下：

$$\boldsymbol{K} = \text{Linear}(C \cdot R, C)(\hat{\boldsymbol{K}}) \quad (2)$$

$$\hat{\boldsymbol{K}} = \text{Reshape}\left(\frac{N}{R}, C \cdot R\right)(\boldsymbol{K}) \quad (3)$$

1.3 Mix-FFN 模块

ViT 使用位置编码来引入位置信息，Segformer 舍弃了位置编码，采用 Mix-FFN 替代。因为位置编码从开始训练后，其分辨率就已经固定，而当测试的时候给出一个不同分辨率的图片，位置编码就会做插值处理，导致精度下降。但 Mix-FFN 考虑了零填充对泄漏位置信息的影响，其做法是采用 3×3 卷积来动态表达 patch 间的位置关系，然后经由 GeLU 激活函数和一个 MLP 来处理。其计算方式如下：

$$x_{\text{out}} = \text{MLP}(\text{GeLU}(\text{Conv}_{3\times3}(\text{MLP}(x_{\text{in}})))) + x_{\text{in}} \quad (4)$$

其中，x_{in} 表示来自自注意力模块的特征；x_{out} 表示 Mix-FFN 模块的输出；$\text{Conv}_{3\times3}$ 表示 Mix-FFN 在每个 FFN 中将 1 个 3 × 3 卷积和 1 个 MLP 混合在一起。

2 改进 Segformer 的分割方法

在实际工作场景中，非显著性目标的小尺寸和边界模糊问题对语义分割任务都是巨大的挑战。为避免分割不精细导致在边缘出现覆盖不全的情况，本文网络添加了尺度更小特征。在 Segformer 的 Encoder 能提取到尺度为 1/4、1/8、1/16 和 1/32 的特征，本文网络模型的 Encoder 增添了 1/64 这个尺度的特征，然后将 5 级特征传递给 MLP 解码器，即可在图像上预测分割掩码。为进一步降低计算成本和大核的参数大小，本文改进了 Segformer 中的 Vix-FFN 模块，使用效率更高且性能更加优异的 XM-FFN 模块，在准确性和效率之间实现更好的平衡。本文的算法模型结构如图 2 所示。

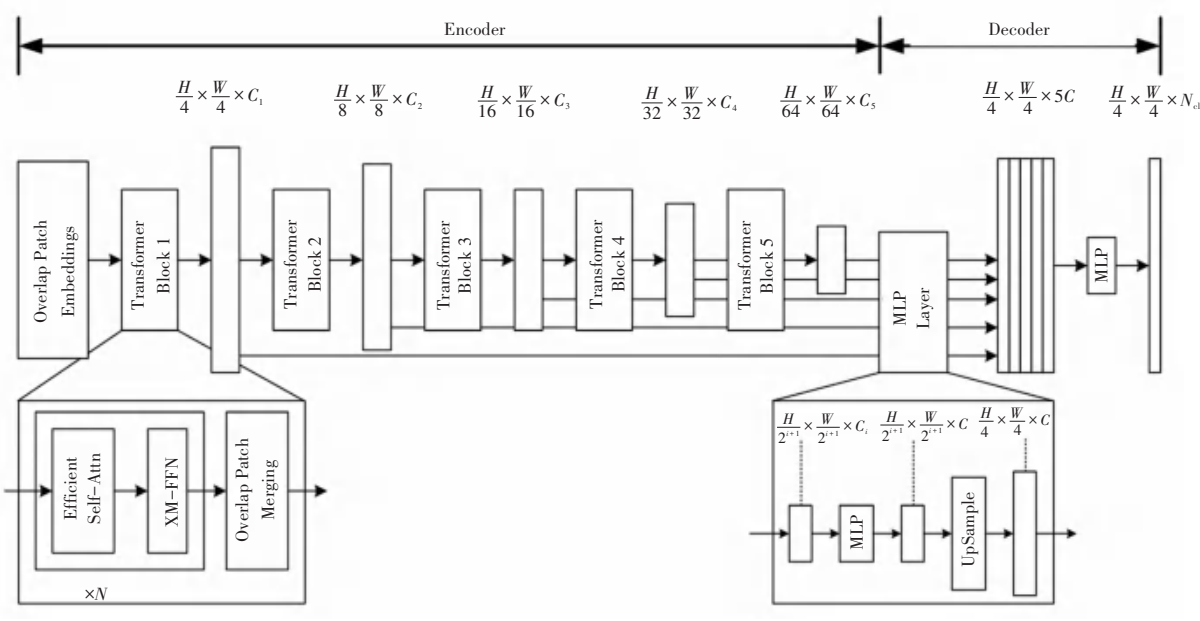

图 2　算法模型结构图

Fig. 2　Structure of the algorithm model

2.1　XM-FFN 模块

为减缓网络的计算负担，设计了一种名为 XM-FFN 的结构，以替代 SegFormer 中的 Mix-FFN。XM-FFN 和 Mix-FFN 的设计结构对比如图 3 所示。XM-FFN 采用了一种名为 XSepConv[21] 的高效卷积操作，为了提高计算效率，选择将 GeLU 激活函数替换为 ReLU 激活函数。ReLU 是一种简单高效的非线性激活函数，在正数部分返回输入值，而在负数部分返回零。通过采用 XSepConv 和 ReLU 激活函数，XM-FFN 能够在保持性能的同时减少计算复杂度，这种改进有助于提高模型的训练和推理效率，特别是在资源受限的环境中。

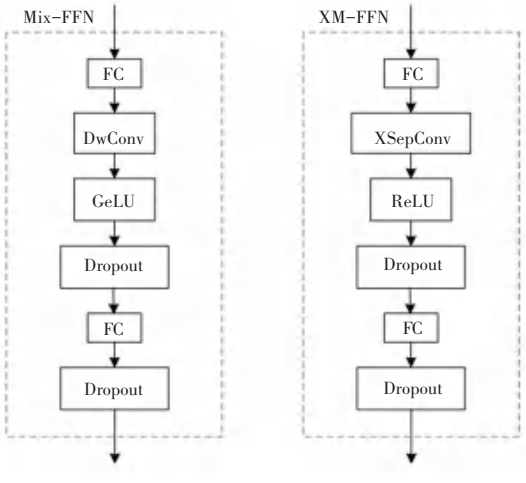

图 3　XM-FFN 和 Mix-FFN 结构图

Fig. 3　XM-FFN and Mix-FFN structures

2.2　XSepConv 模块

XSepConv 是一种新型的卷积神经网络模块，将空间可分离卷积融合为深度卷积，以进一步降低大内核的计算成本和参数大小，结构如图 4 所示。XSepConv 由 3 个部分组成：2×2 深度卷积核、1 个 k 深度卷积核和 $k\times 1$ 个深度卷积核。$1\times k$ 和 $k\times 1$ 深度卷积共同形成了空间分离深度卷积，其中 2×2 深度卷积用于捕获由于固有的结构缺陷而可能被空间分离深度卷积丢失的信息。

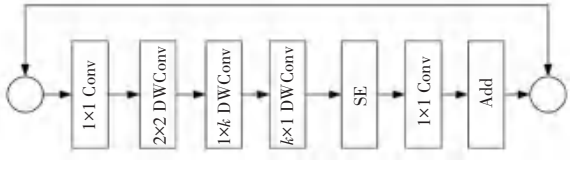

图 4　XSepConv 结构图

Fig. 4　XSepConv structure diagram

3　实验与分析

3.1　实验环境与数据集

本文的实验在 Ubuntu 18.04 上进行，主要硬件为 i5-12400f 16 G、NVIDA GeForce RTX2060 12 G，开发语言为 Python。数据集采用的是 VOC2012，该数据集中训练集、验证集和测试集分别有 1 464、1 449 和 1 456 张图片，一共有 20 类物体。

3.2　评价指标

为准确评估本文算法对非显著性目标的实际分割效果，采用 3 个针对语义分割性能的评价指标：mIoU、mF1 和 MPA。其中，mIoU 指标衡量了像素级

别的分割精度,mF1 指标则关注了精细分割和边缘检测的准确性,而 MPA 指标综合考虑了多目标场景下的整体性能。

(1)mIoU 为平均交并比,数学公式如下:

$$\mathrm{mIoU}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{TP}+\mathrm{FP}} \tag{5}$$

其中,TP 表示真阳性;FP 表示假阳性;FN 表示假阴性;k 表示分类的种类数。

(2)mF1 是各类别基于其准确率和召回率的数学计算,公式如下:

$$\mathrm{mF1}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{2\times\mathrm{Precision}_k\times\mathrm{Recall}_k}{\mathrm{Precision}_k+\mathrm{Recall}_k} \tag{6}$$

其中,Precision 和 Recall 分别表示准确率和召回率,计算公式具体如下:

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{7}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{8}$$

(3)MPA 为类别平均像素准确率,公式如下:

$$\mathrm{MPA}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{9}$$

3.3 实验结果与分析

本文算法是在 Segformer 语义分割网络的基础上进行改进的。为了提高对非显著性目标分割的准确率,引入了小尺度特征信息,并采用了 XM-FFN 模块来进一步提升模型性能,同时有效地减少了计算参数量。这些改进的关键点旨在优化模型的感知力和分辨率,关注非显著性目标的识别和分割任务,使得算法在复杂环境下能够取得显著的性能提升。Segformer 改进前后的网络模型的分割效果见表 1。

表 1 VOC2012 数据集分割效果对比

Table 1 Comparison of segmentation effect of VOC2012 dataset %

算法	mIoU	mF1	MPA
Segformer	77.56	85.92	87.36
改进的 Segformer	80.14	88.27	89.85

由表 1 可知,相较于原始 Segformer,在 mIoU、mF1 和 MPA 指标上,本文算法分别实现了 2.58%、2.35%、2.49%的提升,证明其在 VOC2012 数据集上取得了显著的成效。

在实际场景下,Segformer 改进前后分割效果对比如图 5 所示。图 5(a)~(c)中,左侧图片为原图像,中间为 Segformer 算法;右侧为本文算法分割效果。

从图 5 可以看出,图 5(a)黑色背景下的黑猫与环境之间的边界模糊不清,Segformer 进行分割时对背景产生了错误的判断,本文算法准确地区分了猫与背景。图 5(b)中人的主体部分被狗遮挡,Segformer 将人与狗划分为同一类,本文算法有效地将两者进行了区分。图 5(c)中狗的运动变化较大造成图像模糊,本文算法更加细致地分割出目标。

(a) 黑猫

(b) 行人

(c) 狗

图 5 实验效果图

Fig. 5 Experimental effect diagram

图 6 为训练损失和验证损失的变化图。图 6 中 Train loss 衡量模型在训练集上的拟合能力,Val loss 衡量在未见过数据上的拟合能力,Smooth train loss 和 Smooth val loss 是平滑处理后的结果。从图 6 可以看出,本文算法对训练集和未训练过的数据均取得了良好的效果。

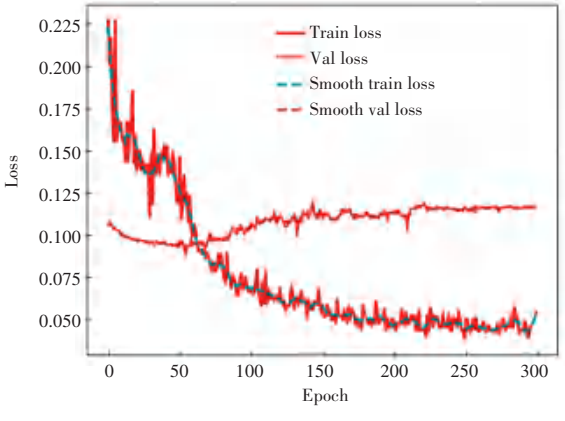

图 6 损失变化图

Fig. 6 Loss variation graph

4 结束语

本文基于 Segformer 语义分割模型针对非显著

性目标进行了改进。为增强算法对非显著性目标细节特征的分割能力并避免边界模糊问题，添加了更小的尺度信息，以保证非显著性目标不会被网络忽视。为更好地提升算法性能并减少参数，采用 XM-FFN 模块替代 Segformer 中的 Mix-FFN 模块，该模块采用了高效的 XSepConv、并使用了更为简便的 ReLU 激活函数，在一定程度上又降低了模型计算量。在 VOC2012 数据集上进行训练和测试，分别在 MIoU、mF1 和 MPA 指标上提升了 2.58%、2.35%和 2.49%，最后在实际场景中进一步证实了本文的策略对算法具有良好的效果。

参考文献

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

[2] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2881-2890.

[3] CHEN L C, ZHU Yukun, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[J]. arXiv preprint arXiv, 1802.02611, 2018.

[4] 杨大伟, 迟津生, 毛琳. 基于边界辅助的弱监督语义分割网络[J]. 计算机应用研究, 2024, 41(2): 623-628.

[5] 唐璐, 万良, 王婷婷, 等. DECANet: 基于改进 DeepLabv3+的图像语义分割方法[J]. 激光与光电子学进展, 2023, 60(4) 92-100.

[6] 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: 2481-2495.

[7] 仲诚诚, 周恒, 张梓童, 等. LAC-UNet: 基于胶囊表达局部整体特征关系的语义分割模型[J]. 山东大学学报(理学版), 2023, 58(11): 116-126.

[8] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]// Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015). Cham: Springer, 2015: 234-241.

[9] STRUDEL R, GARCIA R, LAPTEV I, et al. Segmenter: Transformer for semantic segmentation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 7262-7272.

[10] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv, 2010.11929, 2021.

[11] 孙万春, 冯欣, 马慧, 等. 细化 Transformer 网络的弱监督图像语义分割[J]. 计算机应用研究, 2023, 40(11): 3515-3520.

[12] 黄科迪, 黄鹤鸣, 李伟, 等. DCaT: 面向高分辨率场景的轻量级语义分割模型[J]. 计算机工程与应用, 2025, 61(1): 252-262.

[13] DONG Bo, WANG Pichao, WANG Fan. Afformer: Head-free lightweight semantic segmentation with linear transformer[C]// Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23). New York: ACM, 2023: 57.

[14] 兰建平, 董冯雷, 杨亚会, 等. 改进 STDC-Seg 的实时图像语义分割网络算法[J]. 传感器与微系统, 2023, 42(11): 110-113.

[15] CAO Hu, WANG Yueyue, CHEN Jieneng, et al. Swin-UNet: UNet-like pure transformer for medical image segmentation[C]// Proceedings of the European Conference on Computer Vision. Cham: Springer, 2022: 205-218.

[16] 宋熙睿, 葛洪伟. 基于 TransMANet 的遥感图像语义分割算法[J]. 激光与光电子学进展, 2024, 61(10): 309-320.

[17] CHEN Jieneng, LU Yongyi, YU Qihang, et al. TransUNet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv, 2102.04306, 2021.

[18] 李俊杰, 易诗, 何润华, 等. 基于窗口注意力聚合 Swin Transformer 的无人机影像语义分割方法[J]. 计算机工程与应用, 2024, 60(15): 198-210.

[19] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 10012-10022.

[20] XIE Enze, WANG Wenhai, Yu Zhiding, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[J]. arXiv preprint arXiv, 2105.15203, 2021.

[21] CHEN Jiarong, LU Zongqing, XUE Jinghao, et al. XSepConv: Extremely separated convolution[J]. arXiv preprint arXiv, 2002.12046, 2020.