基于 DeepLabv3+ 改进的道路场景语义分割方法研究

袁世鹏[1] 李松松[1*] 毛晗宇[1] 郭天宇[1] 国津荣[1]
YUAN Shipeng LI Songsong MAO Hanyu GUO Tianyu GUO Jinrong

摘 要 针对 DeepLabv3+ 模型在语义分割中未充分利用主干网络的语义信息，以及在编码器 - 解码器结构中分割小物体目标边界性能差的问题，文章提出了一种改进的 DeepLabv3+ 语义分割方法。首先，在浅层特征融合分支结构中，利用 MobileNetv2 网络替换原先的 Xception 网络，并添加 VS-FRM 模块和 SA 模块，使网络最大程度识别到物体的边缘。其次，在中间层特征融合分支结构中，增加一条特征融合分支结构，并添加 CA 模块，使网络充分提取图像特征。最后，在深层特征融合分支结构中，用 C-ASPP 模块替换 ASPP 模块，并且将中间层特征融合分支的输出与 C-ASPP 的输出进行拼接，来弥补训练过程中丢失的语义特征。实验结果表明，在 CamVid 数据集上的 MIoU 为 66.72%，MPA 为 76.85%，Precision 为 77.77%，与 DeepLabv3+ 模型相比 MIoU 提升了 6.79%，MPA 提升了 8.62%，Precision 提升了 8.64%。有效提高了道路场景语义分割的准确性。

关键词 语义分割；DeepLabv3+；VS-FRM；C-ASP

doi：10.3969/j.issn.1672-9528.2025.05.002

0 引言

在计算机视觉领域当中，图像语义分割作为近几年的研究热点之一，目的是将图像分割成不同的颜色区域，得到相应的像素语义标注图像 [1]。传统的图像语义分割方法 [2] 主要是从图像的浅层信息出发，通过提取图像的颜色、纹理、几何等特征来实现的。其中，阈值分割算法 [3] 根据图像灰度值的级别，将不同的对象类别划分为不同的灰度级，然后识别出相应的区域。而边缘检测算法 [4] 是利用滤波器来获取图像的边缘灰度值检测边缘算子。这两种方法容易受到各种噪声的影响，得到的结果往往达不到目前语义分割的要求。

随着深度学习的不断发展，图像语义分割模型也取得了越来越优异的效果。其中，Long 等人 [5] 提出的全卷积神经网络（fully convolutional networks, FCN）是第一个像素到像素的分割模型，将卷积神经网络（convolutional neural networks, CNN）的最后一层替换为卷积层，让网络的输出与输入图像规格是相同的矩阵，不同于传统的分割算法，这种方式得到了优异的效果。但 FCN 在执行卷积操作时使用的滤波器步幅过大、特征图尺寸逐层减小，导致具有丰富空间信息的浅层信息并没有被充分利用，从而造成分割结果无法达到预期。为克服该问题，Ronneberger 等人 [6] 于 2015 年提出了 U 形对称的语义分割模型 U-Net，通过上采样代替池化层，设计压缩路径来捕捉上下文，并通过对称扩张路径来实现准定位。Badrinarayanan 等人 [7] 提出了一种基于深度卷积神经网络（SegNet）的算法，该算法在编码器中引入最大池化运算，以提高图像分辨率；与传统的 FCN 算法不同，该算法考虑了空间位置信息。在此基础上，提出了一种基于深度的图像分割方法。Zhao 等人 [8] 以 ResNet 为底层网络，通过构建“金字塔”模型，突破了传统分类任务的尺寸限制，实现了多尺度的上下文信息融合。Zhao 等人 [9] 于 2018 年提出 ICNet 进行实时语义分割，采用渐进的特征融合和串联指导网络进行合理的预测，并通过加入辅助损耗函数加速算法的收敛，实现对正、负样本的均衡。

Chen 等人 [10] 又提出了 DeepLab 系列语义分割模型，DeepLabv1 模型基于深度实验室序列的深度实验室语义划分方法，该方法利用尺度膨胀卷积与条件随机场（CRF）两种方法同时获得更多的空间信息，以增强感知野并获得更精确的影像轮廓。DeepLabv2[11] 利用一种新的基于凹腔的分层模型，通过融合具有多尺度特性的空腔模型，提高其对多个尺度物体的预报能力。DeepLabv3[12] 则利用一种新型的 ASPP 结构，通过一组连续的凹卷积上、下采样，并在骨干网络后加入一个非线性运算层，实现对多尺度数据的获取，从而有

1. 大连海洋大学 辽宁大连 116000
[基金项目] 国家自然科学基金资助项目（51778104）；辽宁省教育厅科学研究项目（DL202005）

效地解决了条件随机场模型复杂度高的问题。DeepLabv3+[13] 基于 DeepLabv3 模式，引入了一个编码器 - 解码器的架构，实现了对视频的高质量重建，并通过双重分枝进行了浅层与深层的语义融合，提高了分割结果准确性，但无法高效地解决多类别对象的问题。

针对城市道路数据集 [14] 中存在小目标物体和长条形物体分割难度大等问题，本文提出一种改进的 DeepLabv3+ 语义分割方法。首先，选择 MobileNetv2 网络作为 DeepLabv3+ 语义分割模型的骨干网络，并且增加一条特征融合分支结构，提高了网络模型的分割准确度。其次，在浅层特征结构中添加 VS-FRM 模块和 SA 模块，增强模型识别物体边缘和局部细节的能力。之后，在中间层特征结构中添加 CA 模块，进一步增强对语义特征的提取能力。最后，在深层特征结构中，用 C-ASPP 替换 ASPP，提升网络在深层特征融合分支结构中获取精细的语义信息能力。

1 相关理论

DeepLabv3+ 网络是标准的编码器 - 解码器结构，如图 1 所示。在编码器端，输入图像送入骨干网络完成高级语义信息特征提取，之后经过特征融合模块 ASPP 后会获得 5 个具有不同信息的特征图，将得到的特征图在通道维度上进行拼接，获得新的特征图，最后使用固定的卷积将获得的新的特征图通道变为 256 并且送入解码端。在解码过程中，通过 4 次升取样，对已有的特征图进行分段还原，从而达到对目标区域的精确定位。

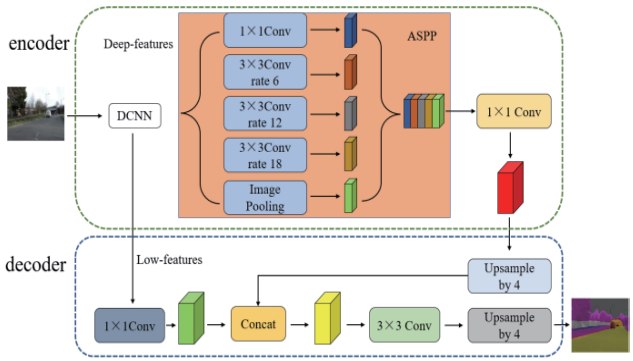

图 1 DeepLabv3+ 网络

2 改进的 DeepLabv3+ 网络框架

2.1 整体网络框架

本文提出的改进 DeepLabv3+ 网络如图 2 所示，首先，利用 MobileNetv2 网络替换原始的 Xception 网络，减少模型的计算复杂度，重新定义 MobileNetv2 网络层次并且增加一条特征融合分支结构 Mid-features，使网络充分提取图像特征。其次，在浅层特征融合分支结构中加入新的特征细化模块——VS-FRM，并且加入位置注意力机制模块 SA，来提升模型在编码阶段获取语义信息的能力；在中间层特征融合分支结构中加入空间注意力机制模块 CA，来弥补训练过程中丢失的深层语义信息；最后，在深层特征融合分支结构中用 C-ASPP 替换 ASPP，达到补充深层网络语义信息的目的，同时，将中间层处理之后的语义信息与深层处理之后的语义信息进行叠加，增加编码阶段的特征信息量。

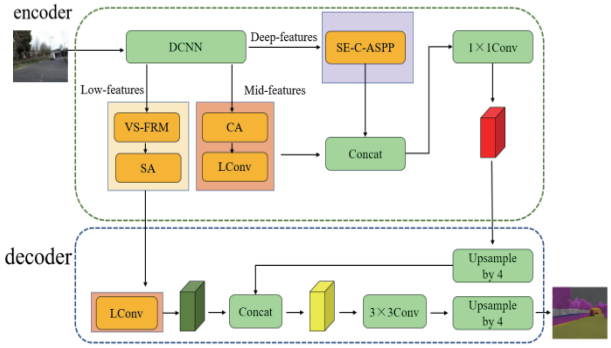

图 2 改进的 DeepLabv3+ 网络

2.2 MobileNetv2

MobileNetv2[15] 骨干网络可以达到降低参数量并且提高分割效率，核心部分是用深度可分离卷积（depthwise separable convolution）进行语义信息提取并且将各通道之间的语义信息进行融合。残差支路将信息的输入与信息的输出相联系，在全网结构中，为实现低维空间语义信息向高维空间语义信息的映射，并在骨干网中引入数据扩充层与数据压缩层，通过引入线性激活函数 ReLu6，有效地解决由于激活函数所带来的语义信息丢失问题。

2.3 浅层语义特征融合分支

2.3.1 特征细化模块（VS-FRM）

特征细化模块由两部分组成，结构如图 3 所示。

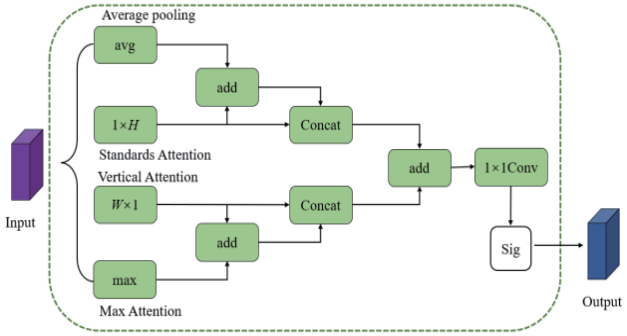

图 3 设计的特征细化模块

第一部分是平均池化模块和水平方向的注意力机制进行语义信息的累加，将累加得到的特征与水平方向注意力机制模块得到的特征进行拼接，一方面将每个区域得到的信息进行像素的复合叠加，虽然描述图像特征的信息量增多，但描

述图像的维度本身并未增加，对最后的图像分类是有益的。另一方面为有效地调整通道数，各个特征下的信息量并未增加，仅将横向或纵向空间上语义信息进行叠加。因为平均池化模块关注点为图像的整体信息，可以帮助网络更好地进行各个标签信息的分类。同时，水平方向的注意力机制模块可以帮助网络捕捉对象的边缘和连续性。

第二部分是将最大池化模块和垂直方向的注意力机制模块进行语义信息的累加，并将累加得到的结果再次与垂直方向的注意力机制模块得到的信息进行通道数的拼接，所实现的原理和前一部分保持一致。因为最大池化模块可以帮助网络捕捉局部图像的结构和纹理，对纹理特征信息更加敏感，同时，垂直方向的注意力机制模块也可以捕捉图像中的高度和纹理。最终，将第一部分和第二部分所测得到的结果进行语义信息的累加并经过固定的卷积来调整通道数，接着用 Sigmoid 激活函数来增加网络的非线性表达能力。

2.3.2 位置注意力机制（SA）

利用位置注意力机制（spatial attention, SA），调整各区域的重要程度，通过对各区域的加权因子进行调整，从而实现对关键区域的关注，SA 结构如图 4 所示。

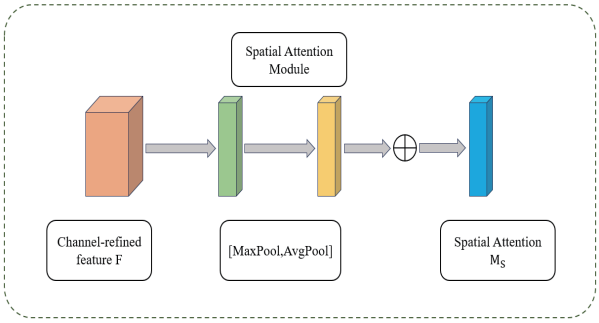

图 4 位置注意力机制

首先，利用 SA 的注意力机制，将各输入的特性映射通过一组含有卷积网络和 ReLu6 激励功能的模块，得到最大池化和平均池化两个特征图，之后进行通道维度上的拼接，最后经过卷积降维和 Sigmoid 激活函数得到最终的输出结果，计算过程用公式表示为：

$$M_s(F)=\sigma\left(f^{7\times7}([\mathrm{Avg}(F);\mathrm{Max}(F)])\right)=\sigma\left(f^{7\times7}\left([F^s_{\mathrm{avg}};F^s_{\mathrm{max}}]\right)\right) \quad (1)$$

2.4 深层特征融合分支

ASPP 模块由 4 个不同空洞率的深度可分离卷积和 1 个最大池化模块组成。改进的 ASPP 将第一个卷积得到的特征信息与第二个空洞率为 6 的卷积得到的特征信息进行通道数的拼接融合，之后将上一步得到的特征信息与第三个空洞率为 12 的卷积得到的特征信息进行通道数的拼接融合，接着将上一步得到的特征信息与空洞率为 18 的卷积得到的特征信息进行拼接融合，最后将上一步得到的特征信息与最大池化模块得到的特征信息进行通道数的拼接融合。由于在图像的深层语义信息中，通道数的大小会影响网络所提取到的图像像素信息的多少，用拼接来控制通道数在一个合适的范围，会给整个网络带来很好的表征能力。将每一层的语义信息和上一层语义信息逐层叠加，并且之后使用 3×3 的卷积来进行通道数的调整。结构如图 5 所示。

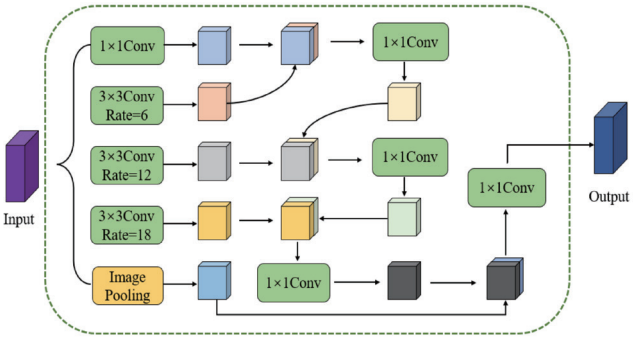

图 5 设计的特征金字塔融合模块

2.5 中间层特征融合分支

空间注意力机制（channel attention, CA），结构如图 6 所示。

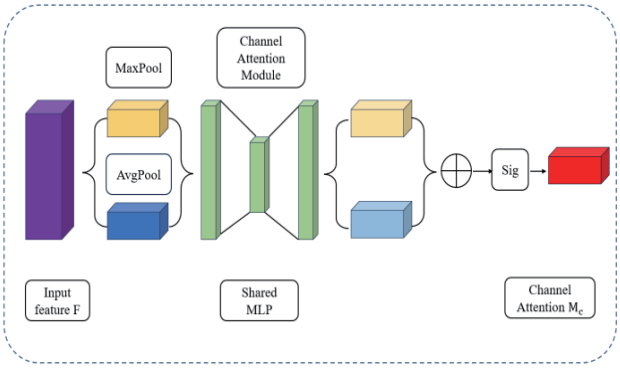

图 6 通道注意力机制模块

首先，将输入的特征图经过一维的全局平均池化之后，可以得出所有通道维度上的平均值，之后再使用全连接层来进行处理，最后得出一个向量，每一元素都表示了一条固定路径的权重函数。ReLu6 作为网络的激活函数，目的是将所有的权重系数转化为正数。之后对全部通道进行调整，将得到的权重系数通过映射的方式扩展到所有特征图的所有通道上，接着，通过一个 Sigmoid 激活函数进行批量正则化，以便得到每个通道的权重，最终得到调整之后的网络输出特征图，整个计算过程公式为：

$$M_c(F)=\sigma\left(\mathrm{MLP}(\mathrm{Avg}(F))+\mathrm{MLP}(\mathrm{Max}(F))\right)=\sigma(W_1\left(W_0(F^c_{\mathrm{avg}})\right)+W_1(W_0(F^c_{\mathrm{max}}))) \quad (2)$$

3 实验及结果分析

3.1 实验数据集及参数配置

本文采用 CamVid 公开数据集，实验是在 Windows11 操作系统上使用深度学习开源框架 PyTorch 实现的，处理器为 AMD Ryzen5 5600X 6-Core Processor，显存为 16 GB，GPU 为 NVIDIAGeforce RTX 3070 Ti；软件环境为 PyCharm2021，PyTorch2.0，Python3.8，CUDA 11.8 等。

3.2 实验评价指标

图像语义分割中，常用平均交并比（MIoU）、平均像素精度（MPA）以及精确率（Precision）来评价模型的整体性能。

平均交并比（MIoU）：表示全部种类的交并比（IoU）取平均值得到的数值，计算公式为：

$$\mathrm{MIoU}=\frac{1}{k+1}\sum_{i=0}^{k}\mathrm{IoU}_i \tag{3}$$

平均像素精度（MPA）是指在所有类别中，正确预测的像素占总像素比例的平均值，该指标的计算公式为：

$$\mathrm{MPA}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{j=0}^{k}P_{ij}} \tag{4}$$

式中：P_{ij} 表示真实类别为 i 被预测为 j 的样本个数；k+1 表示全部的类别数；P_{ii} 表示真实类别为 i 且被正确预测的样本数。

精确率（Precision）是指预测为某个类别的像素中，实际上也属于该类别的像素所占的比例，计算公式为：

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

式中：TP 表示正确预测的正样本数；FP 表示错误预测的负样本数。

3.3 实验结果与分析

3.3.1 不同骨干网络对比

使用 MobileNetv2 和 Xception 的骨干网络模型进行对比，表 1 列出了两种骨干网络模型训练得到的实验数据。

表 1 不同骨干网络定性实验

Backbone	MIoU/%	MPA/%	Precision/%
MobileNetv2	63.41	74.29	75.93
Xception	56.62	65.67	67.28

从表 1 中可以看出，用 MobileNetv2 替换原有的 Xception 网络后，整个模型的 MIoU、MPA、Precision 有较大提升，分别提升了 6.79%、8.62%、8.64%，在小物体的边缘连续性方面效果显著。

3.3.2 消融实验

为验证所设计模块的有效性，本文依次增加 VS-FRM、SA、C-ASPP 和 CA 网络模块的消融实验，实验组编号和结果如表 2 所示。

表 2 模块消融定性实验

Model	VS-FRM	SA	C-ASPP	CA	MIoU /%	MPA /%	Precision /%
原模型					63.41	74.29	75.93
1	√				63.79	74.59	76.18
2	√	√			64.24	75.15	76.84
3			√		63.47	74.34	75.96
4			√	√	64.46	74.89	76.01
5	√	√	√	√	66.72	76.85	77.77

从表 2 中可以看出，第 1 组实验与原模型相比，MIoU、MPA、Precision 分别优于原始模型 0.56%、0.30%、0.25%；第 2 组实验与原模型相比，MIoU、MPA、Precision 分别优于原始模型 0.83%、0.86%、0.91%；第 3 组实验与原模型相比 MIoU、MPA、Precision 分别优于原始模型 0.07%、0.05%、0.03%；第 4 组实验与原模型相比，MIoU、MPA、Precision 分别优于原始模型 1.05%、0.6%、0.08%；第 5 组实验与原模型相比，MIoU、MPA、Precision 分别优于原始模型 3.31%、2.56%、1.84%。

3.3.3 不同模型对比

为更好地验证本文改进的网络模型的优越性，采用定性分析与定量分析两个不同的方面进行评价，并且与多组经典的分割网络模型进行对比实验，如 PSPNet、U-Net、HRNetv2、DeepLabv3+。表 3 为不同网络模型在定性实验分析的实验结果，可以看到本文模型在 MIoU、MPA、Precision 都优于现有的主流模型。与 PsPNet 相比，MIoU、MPA、Precision 分别领先 11.83%、11.67%、7.77%；与 U-Net 相比，MIoU、MPA、Precision 分别领先 2.68%、2.95%、0.26%；与 HRNet 相比，MIoU、MPA、Precision 分别领先 0.81%、1.15%、1.15%；与 DeepLabv3+ 相比，MIoU、MPA、Precision 分别领先 3.31%、2.56%、1.84%。

表 3 经典分割网络模型实验

Model	MIoU/%	MPA/%	Precision/%
PSPNet	54.89	65.18	70.00
U-Net	64.04	73.90	77.51
HRNet	65.91	75.70	78.79
DeepLabv3+	63.41	74.29	75.93
Ours	66.72	76.85	77.77

为更加全面地评价各模型的分割效果，在客观评价的基础之上，加入可视化图片，各模型的分割效果图如图 7 所示。

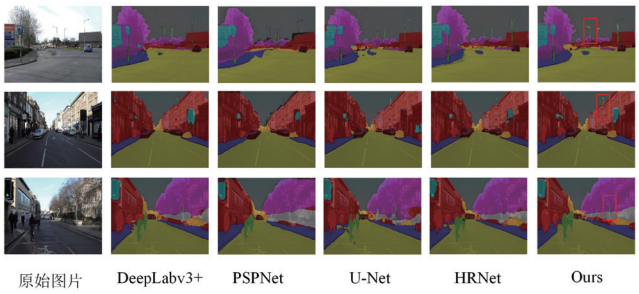

图 7 各语义分割模型分割效果图

4 结论

本文提出的语义分割模型能够有效提取图像的边缘特征，在道路场景数据集中取得了优异的图像分割效果。将主干网络替换为 MobileNetv2 一方面是为了降低参数量，另一方面是为提升模型的分割准确度。添加特征细化模块和 SA 模块是为增加模型在浅层结构中提取语义信息的能力。采用 C-ASPP 模块是为了增强深层结构语义特征的信息量，有助于模型识别物体边缘。实验结果表明，与 DeepLabv3+ 相比较，本文方法在 MIoU、MPA、Precision 分别提高了 6.79%、8.62%、8.64%。

参考文献：

[1] 青晨，禹晶，肖创柏，等 . 深度卷积神经网络图像语义分割研究进展 [J]. 中国图象图形学报 ,2020,25(6):1069-1090.

[2] IVANOVS M, OZOLS K, DOBRAJS A, et al. Improving semantic segmentation of urban scenes for self-driving cars with synthetic images[J]. Sensors, 2022, 22(6): 2252.

[3] KONTSCHIEDER P, BULO S R, BISCHOF H, et al. Structured class-labels in random forests for semantic image labelling[C]//2011 International Conference on Computer Vision. Piscataway: IEEE,2011: 2190-2197.

[4] HEUVEL M V D, MANDL R, POL H H. Nor-malized cut group clustering of resting-state FMRI da-ta[J]. Public library of science one, 2008, 3(4): 2001.

[5] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[J].IEEE transactions on pattern analysis and machine intelligence, 2017, 39(4): 640-651.

[6] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutionalnetworks for biomedical image segmentation[C]// Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Berlin: Springer, 2015:234-241.

[7] BADRINARAYANAN V, KENDALL A, CIPOLLA R.SegNet:a deepconvolutional encoder-decoder architecture for imagesegmentation[J].IEEE transactions on pattern analysisand machine intelligence, 2017, 39(12): 2481-2495.

[8] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsingnetwork[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 6230-6239.

[9] ZHAO H S, QI X J, SHEN X Y, et al. ICNet for real-time semantic segmentation on high-resolution images[C]//Computer Vision-ECCV 2018. Berlin:Springer,2018: 418-434.

[10] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs [EB/OL]. (2016-06-07) [2024-02-10].https://doi.org/10.48550/arXiv.1412.7062.

[11] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convo- lutional nets, atrous convolution, and fully connected CRFs[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 40(4): 834-848.

[12] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmenta-tion[EB/OL].(2017-12-05)[2024-02-25].https://doi.org/10.48550/arXiv.1706.05587.

[13] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encod-er-decoder with atrous separable convolution for semantic image segmentation[C]//Computer Vision-ECCV 2018. Berlin: Springer, 2018: 833-851.

[14] 李利荣，丁江，梅冰，等 . 基于像素注意力特征融合的城市街景语义分割算法研究 [J]. 电子测量技术 , 2023, 46(20): 184-190.

[15] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetv2: invertedresiduals and linear bottlenecks[EB/OL].(2019-03-21)[2024-09-15].https://doi.org/10.48550/arXiv.1801.04381.

【作者简介】

袁世鹏（1998—），男，青海海东人，硕士研究生，研究方向：深度学习语义分割。

李松松（1973—），通信作者（email:lisongsong@dlou.edu.cn），女，辽宁丹东人，博士，教授，研究方向：超声无损检测技术及信号处理技术。

（收稿日期：2025-01-10 修回日期：2025-05-16）