计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(5):94−102 [doi: 10.15888/j.cnki.csa.009513]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合注意力和多尺度特征的街景图像语义分割①

洪　军, 刘笑楠, 刘振宇

(沈阳工业大学 信息科学与工程学院, 沈阳 110870)
通信作者: 刘笑楠, E-mail: liuxiaonan@sut.edu.cn

摘　要: 为了解决在街道场景图像语义分割任务中传统 U-Net 网络在多尺度类别下目标分割的准确率较低和图像上下文特征的关联性较差等问题, 提出一种改进 U-Net 的语义分割网络 AS-UNet, 实现对街道场景图像的精确分割. 首先, 在 U-Net 网络中融入空间通道挤压激励 (spatial and channel squeeze & excitation block, scSE) 注意力机制模块, 在通道和空间两个维度来引导卷积神经网络关注与分割任务相关的语义类别, 以提取更多有效的语义信息; 其次, 为了获取图像的全局上下文信息, 聚合多尺度特征图来进行特征增强, 将空洞空间金字塔池化 (atrous spatial pyramid pooling, ASPP) 多尺度特征融合模块嵌入到 U-Net 网络中; 最后, 通过组合使用交叉熵损失函数和 Dice 损失函数来解决街道场景目标类别不平衡的问题, 进一步提升分割的准确性.实验结果表明, 在街道场景 Cityscapes 数据集和 CamVid 数据集上 AS-UNet 网络模型的平均交并比 (mean intersection over union, *MIoU*) 相较于传统 U-Net 网络分别提高了 3.9% 和 3.0%, 改进的网络模型显著提升了对街道场景图像的分割效果.

关键词: 图像语义分割; 街道场景; U-Net; 注意力机制; 多尺度特征融合

引用格式: 洪军,刘笑楠,刘振宇.融合注意力和多尺度特征的街景图像语义分割.计算机系统应用,2024,33(5):94–102. http://www.c-s-a.org.cn/1003-3254/9513.html

Semantic Segmentation of Street View Image Based on Attention and Multi-scale Features

HONG Jun, LIU Xiao-Nan, LIU Zhen-Yu

(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: This study aims to solve the problems faced by traditional U-Net network in the semantic segmentation task of street scene images, such as the low accuracy of object segmentation under multi-scale categories and the poor correlation of image context features. To this end, it proposes an improved U-Net semantic segmentation network AS-UNet to achieve accurate segmentation of street scene images. Firstly, the spatial and channel squeeze & excitation block (scSE) attention mechanism module is integrated into the U-Net network to guide the convolutional neural network to focus on semantic categories related to segmentation tasks in both channel and space dimensions, to extract more effective semantic information. Secondly, to obtain the global context information of the image, the multi-scale feature map is aggregated for feature enhancement, and the atrous spatial pyramid pooling (ASPP) multi-scale feature fusion module is embedded into the U-Net network. Finally, the cross-entropy loss function and Dice loss function are combined to solve the problem of unbalanced target categories in street scenes, and the accuracy of segmentation is further improved. The experimental results show that the mean intersection over union (*MIoU*) of the AS-UNet network model in the Cityscapes and CamVid datasets increases by 3.9% and 3.0%, respectively, compared with the traditional U-Net network. The improved network model significantly improves the segmentation effect of street scene images.

Key words: image semantic segmentation; street scene; U-Net; attention mechanism; multi-scale feature fusion

① 基金项目: 辽宁省应用基础研究计划 (2023JH2/101300225)
收稿时间: 2023-12-06; 修改时间: 2024-01-09; 采用时间: 2024-01-18; csa 在线出版时间: 2024-04-07
CNKI 网络首发时间: 2024-04-10

街道场景图像语义分割[1]的准确性和处理速度对于自动驾驶[2]具有重要意义. 传统图像分割技术如边缘检测分割法[3]、区域分割法[4]、阈值分割法[5]等只能从图像中获取低级的语义信息, 存在无法划分多语义类别且分割精度低等缺陷.

近年来, 深度学习方法在图像语义分割任务中的应用使得图像分割的准确性得到较大提升. 2015 年, Long 等[6]提出全卷积网络 (fully convolutional network, FCN), 该网络采用卷积层代替全连接层, 实现了端到端的预测, 在语义分割领域产生了重要影响. 为了解决 FCN 在分割结果精细度和边界连续性等方面存在的问题, Badrinarayanan 等[7]在 FCN 的基础上提出 SegNet, 该网络采用编解码结构, 其中编码网络部分用于提取图像的特征, 解码网络部分用于提取图像特征并恢复图像维度, 该网络可以在尽可能减少信息损失的前提下完成同尺度的输入输出. U-Net 网络也是一种经典的编解码结构, 该网络是由 Ronneberger 等[8]提出. 传统的编解码网络结构在下采样操作时容易失去图像的细节信息, 从而导致分割结果的边界模糊和细节丢失. 为了克服这个问题, U-Net 网络采用了对称的“U 型”结构, 并借助跳跃连接将不同层级的特征进行融合. U-Net 网络最初被广泛用于医学图像分割任务, 之后在 U-Net 网络基础上进行了一系列改进措施. 例如 U-Net++[9]引入了密集跳跃连接结构, 实现了更好的上下文感知能力, 但是没有关注关键特征信息, 并且对于街道场景语义分割来说模型的计算量过大. Attention U-Net[10]引入了注意力机制, 关注图像中关键区域的目标, 提高了分割结果的质量, 但是对浅层特征信息的提取较少.

ASPP 模块和 scSE 模块在聚合图像多尺度特征方面和关注图像关键区域特征信息方面具有很好的效果. 注重捕捉图像的上下文信息有助于获得高质量的分割结果, 许多方法采用了扩大感受野或融合不同层次的上下文信息来提升网络的准确性. Zhao 等[11]引入了 PSPNet, 该网络利用金字塔池化模块 (pyramid pooling module, PPM) 来提取目标图像的全局信息. Chen 等[12–15]通过引入空洞卷积[16], 提出 DeepLab 结构的几种变体, 通过使用 ASPP 模块, 可以对输入的特征图进行并行采样, 以获取不同尺度的图像上下文信息. 在卷积神经网络中引入注意力机制可以使网络学习到图像中需要关注的区域, 从而增强了卷积特征表达, 获取更多的全局上下文信息. Woo 等[17]提出了 CBAM 注意力模块, 该模块由通道和空间两个注意力子模块串行组合构成, 能够增强网络对有用特征的关注. SENet 是由 Hu 等[18]提出, 该网络利用通道注意力机制来学习每个通道的权重, 并将这些权重应用于对特征图的调整. Roy 等[19]提出了基于 SE 模块的 3 种变体: sSE 模块、cSE 模块以及将两者并行加和构成的 scSE 模块, 其中 scSE 模块通过构建空间注意力和通道注意力两个子模块, 综合信息以获得更全面可靠的注意力信息, 通过实验验证该模块能够对图像语义分割准确率带来较大提升. 并且与 CBAM 注意力机制相比, scSE 注意力机制的计算量较少, 能够加快网络的推理速度, 更适用于街道场景图像语义分割的应用环境.

传统 U-Net 网络适用于目标单一和目标尺寸较小的图像, 而对于目标种类较多且目标尺度变化较大的街道场景图像, 它不能提供足够的特征来支持精确的图像语义分割, 因此本文提出一种基于改进 U-Net 网络的 AS-UNet 网络模型. 针对街道场景图像目标种类较多而导致目标分割率低的问题, 通过将 scSE 注意力机制模块融入编解码器的卷积层, 模型可以同时利用通道注意力机制和空间注意力机制, 以增强对关键目标特征信息的获取能力; 针对街道场景图像目标尺度变化较大而导致上下文特征信息关联性较差的问题, 模型采用 ASPP 多尺度特征融合模块将不同尺度目标的信息进行融合从而增强语义信息; 针对街道场景图像目标类别不平衡的问题, 在训练阶段, 引入了交叉熵损失和 Dice 损失两者结合的组合损失函数, 进一步提升分割的效果.

1 网络结构

1.1 传统 U-Net 网络

U-Net 网络模型结构图如图 1 所示.

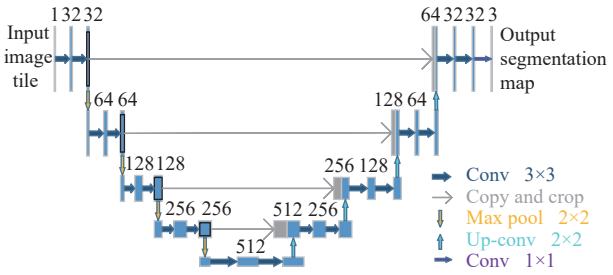


图 1 U-Net 网络模型结构图

U-Net 网络主要包含 2 个部分: 左侧下采样的过程, 可称为编码器结构, 是模型深化的过程; 右侧上采

样的过程可以称为解码器结构, 它是图像分辨率恢复的过程. 编码器中的每个下采样模块通过两次3×3卷积和一次2×2的最大池化来实现, 特征图每经过一次下采样操作, 通道数翻倍, 同时特征图尺寸减半; 解码器中的每个上采样模块通过一次2×2反卷积和两次3×3卷积来实现, 特征图每经过一次上采样操作, 特征图尺寸翻倍, 同时通道数减半, 然后在通道维度上将编解码器之间的特征图进行拼接, 最后通过1×1卷积输出分割图.

1.2 本文提出的 AS-UNet 网络

针对传统 U-Net 网络在处理街道场景图像分割任务时分割的准确率较低和图像上下文的关联性较差等问题, 本文在 U-Net 网络的基础上提出了 AS-UNet 网络模型. AS-UNet 网络模型相对于传统 U-Net 网络融入了 scSE 注意力机制模块和 ASPP 多尺度特征融合模块, 从而提升网络分割精度. scSE 注意力机制模块注重上下文特征的联系, 可捕获更多的特定信息, 增加分割精度; ASPP 多尺度特征融合模块注重扩大感受野, 通过不同程度的池化级别来解决分割目标大小不均的问题.

AS-UNet 网络模型结构如图 2 所示, 该网络模型采用了 5 层“U 型”的编解码结构. 与传统的 U-Net 网络不同, 首先, 该网络在编码器和解码器的每个下采样模块和上采样模块的3×3卷积层后嵌入 scSE 注意力机制模块, 经过卷积提取的特征图通过该模块可以在通道和空间两个维度引导卷积神经网络关注与分割任务相关的语义类别, 以提取更多有效的语义信息; 其次, 经过多次下采样操作后, 编码器提取的特征图中包含丰富的语义信息, 这些特征图经过 ASPP 模块处理, 利用不同扩张率的空洞卷积, 生成多个具有不同尺度感受野的特征图, 从而有效地捕获多尺度特征; 在解码器的每个上采样阶段, 进行了特征图的融合操作, 将其与编码器阶段的特征图进行融合, 将低层的语义信息再次应用于分割任务中, 最后通过1×1卷积输出分割图.

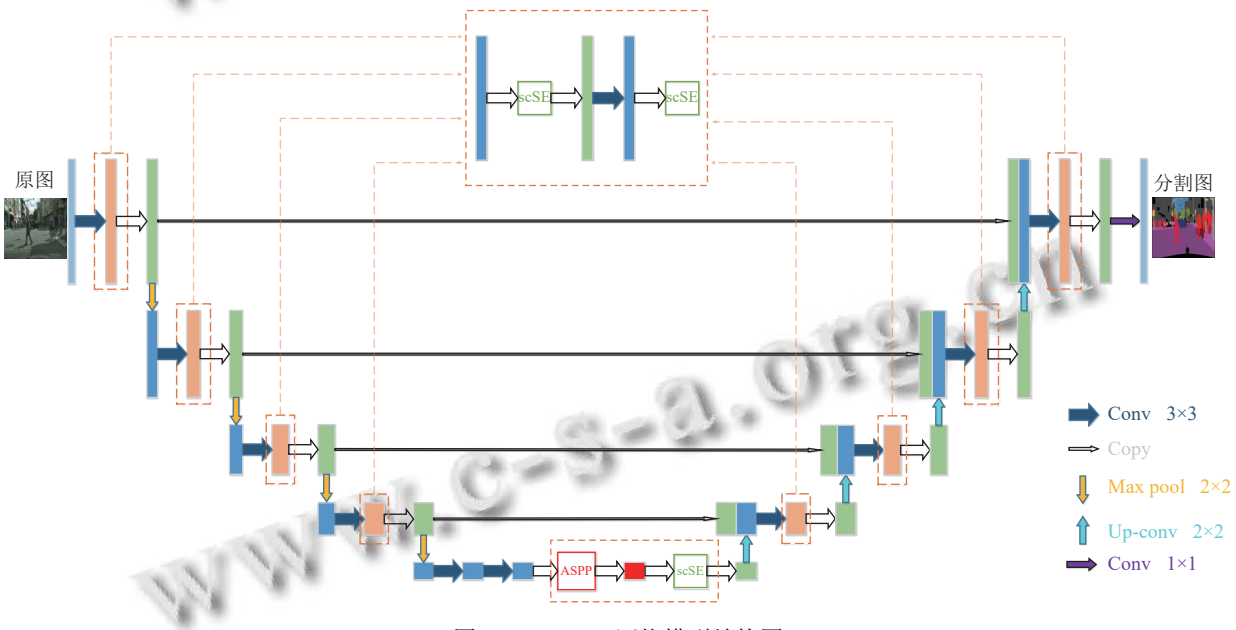


图 2 AS-UNet 网络模型结构图

1.3 AS-UNet 中的 scSE 模块

AS-UNet 中的 scSE 模块用于解决传统 U-Net 网络在提取特征的过程中不能关注图像关键区域的特征信息的问题. 为了适用于街道场景图像目标种类较多的应用环境, 本文使用 scSE 注意力机制模块对编解码器中经过卷积层输出的特征图进行特征的重标定, 以建立带有注意力机制的特征关联, 从而获取更丰富的全局上下文信息. scSE 注意力机制模块具备同时作用于空间和通道的双重特性, 由 sSE 模块和 cSE 模块以并行加和的形式组成, 其执行过程如图 3 所示.

在输入特征图U的基础上分别提取空间和通道间的关键性程度, 并对两个分支提取的特征进行加和, 如式 (1):

$$\hat{U}_{\text{scSE}} = \hat{U}_{\text{cSE}} + \hat{U}_{\text{sSE}} \tag{1}$$

在此基础上, 对有着重要通道和空间特征的子特征图赋予了较高的激活值, 从而使得网络能够学习到重要的特征信息.

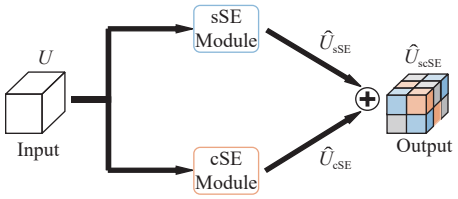


图 3 scSE 模块结构图

sSE 模块通过对特征图的通道特征进行压缩, 实现对各个尺度上的重要空间特征的学习, 并获得一组空间权重参数. 这些参数在空间位置上与输入特征相乘, 从而在空间层面上对特征信息进行重标定. sSE 模块的操作过程如图 4 所示, 输入特征图假定为 $U=[u_{1,1}, u_{1,2},\cdots,u_{i,j},\cdots,u_{H,W}]$, 其中 $u_{i,j}$ 表示在坐标为 (i,j) 处的全部通道特征. 首先, 对特征图 U 进行通道压缩后输出特征图 q:

$$q = V_{sq} * U \tag{2}$$

其中, *表示卷积运算, q 表示尺寸为 $H\times W\times 1$ 的输出特征图, V_{sq} 表示尺寸为 $1\times 1\times C$ 的卷积核.

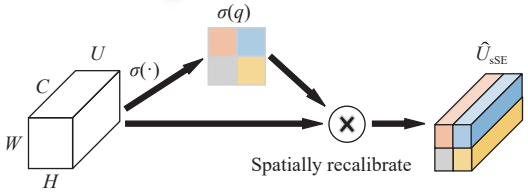


图 4 sSE 模块结构图

然后, 用 Sigmoid 归一化处理特征图 q, 把输出的结果值乘以输入特征中每个对应的空间位置, 最终得出了位置为 (i,j) 的空间信息重要性因子 $\sigma(q_{i,j})$, 用以提高任务相关空间位置的比重.

cSE 模块的操作过程如图 5 所示. 首先, 对特征图 U (尺寸大小为 $H\times W\times C$) 进行全局的平均池化运算, 将特征图 U 的空间全局特征嵌入至向量空间 z. 其中第 k 个通道的值为:

$$z_k = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_k(i,j) \tag{3}$$

其中, z_k 表示第 k 个通道经过全局平均池化后的值, i、j 分别为特征图 U 的第 k 个通道特征图的横、纵坐标参数, u_k 为特征图 U 的第 k 个通道的特征值.

接下来, 将得到的特征 z 经过第 1 个全连接层 Q_1 后进行一次 ReLU 函数激活, 此时特征维度变为 $1\times 1\times C/2$, 然后再经过第 2 个全连接层 Q_2, 对输入特征图采用 Sigmoid 函数进行归一化处理, 将特征维度恢复至 $1\times 1\times C$, 得到第 k 个通道 u_k 的学习权重 $\sigma(\hat{z}_k)$, 其中 $\hat{z}$ 的值为:

$$\hat{z} = Q_2(\delta(Q_1(z))) \tag{4}$$

其中, $\delta(\cdot)$ 表示 ReLU 激活函数, $\sigma(\cdot)$ 表示 Sigmoid 函数.

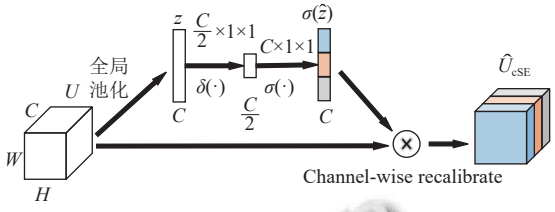


图 5 cSE 模块结构图

最后, 对初始输入的特征图重新分配通道权重, 通过将向量空间 $\sigma(\hat{z})$ 与特征图 U 相乘得到信息校准过的特征图, 实现对通道特征重要性的重新校准.

1.4 AS-UNet 中的 ASPP 模块

为了适用于街道场景图像目标尺度变化较大的应用环境, AS-UNet 中的 ASPP 模块用于解决传统 U-Net 网络对图像上下文特征关联性较差的问题, 因此本文在编码器末端嵌入 ASPP 多尺度特征融合模块, 特征图通过 ASPP 多尺度特征融合模块后生成多个含有不同尺度感受野的特征图, 从而实现对多尺度特征的捕获. ASPP 多尺度特征融合模块是在空洞卷积的理论基础上, 通过并行使用多个不同扩张率的空洞卷积, 对网络中单尺度提取的特征图信息进行重采样, 这些重采样后的特征图被融合在一起, 生成了最终的结果. 这种方式能够增加网络对不同尺度特征的感知能力, 提高特征提取的多样性.

空洞空间金字塔池化模块示意图如图 6 所示, 该模块是由 1 个 1×1 卷积层、3 个空洞卷积层和 1 个池化层所组成.

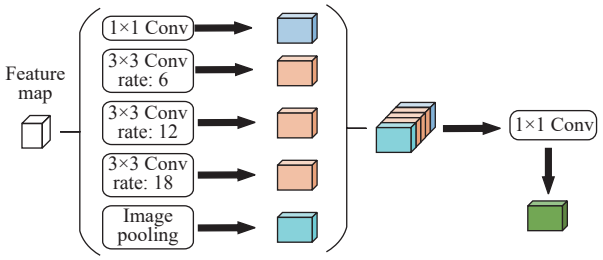


图 6 空洞空间金字塔池化模块结构图

第 1 个分支对从主干网络提取的特征图进行 1×1 卷积核处理, 以调整通道数量; 第 2–4 个分支采用了不同扩张率 (分别为 6、12、18) 的卷积核, 用于提取不同尺度的特征; 第 5 个分支使用全局平均池化进行下采样, 然后利用双线性插值方法将图像上采样, 使其达

到与其他分支相同的分辨率. 最后将以上 5 个分支所提取的特征叠加在一起, 以获得融合多尺度语义信息的特征层. 这样的设计能够有效地提升模型在多尺度任务上的性能.

1.5 组合损失函数

损失函数是训练过程中最重要的环节之一, 损失函数的好坏可以影响网络模型的整体性能. 语义分割常采用交叉熵损失函数, 但当目标类别不平衡时, 会导致网络模型的分割性能较差. 为了适用于街道场景图像目标类别不平衡的应用环境, 本文采用了交叉熵损失函数和 Dice 损失函数两者结合的组合损失函数.

交叉熵损失函数用来评价模型的预测结果与真实标签之间的差异, 计算公式如下:

$$L_{\mathrm{CE}} = -\sum_{i} y_i \log \hat{y}_i \tag{5}$$

其中, y_i表示真是标签, $\hat{y}_i$表示预测结果.

Dice 损失函数用来计算模型的预测值和真实值的相似度, 计算公式如下:

$$L_{\mathrm{Dice}} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{6}$$

其中, $|X+Y|$表示交集元素个数, $|X|$, $|Y|$分别表示其元素个数.

将交叉熵损失函数和 Dice 损失函数两者结合的损失函数作为本文新的损失函数, 计算公式如下:

$$L_{\mathrm{Total}} = L_{\mathrm{CE}} + L_{\mathrm{Dice}} \tag{7}$$

2 实验与分析

2.1 实验数据集与实验环境

2.1.1 实验数据集

在本文实验中, 使用 Cityscapes 数据集和 CamVid 数据集对网络模型进行结果分析和比较.

Cityscapes 数据集是一个大型城市街道场景数据集, 其中包括 5 000 张精细标注的图像, 图像分辨率为 1024×2048, 包含训练图 2 975 张、验证图 500 张和测试图 1 525 张, 本文选用其中 19 个目标类别进行训练和评估.

CamVid 数据集是一个以驾驶汽车角度来看的街道场景数据集, 其中包括 701 张图像, 图像分辨率为 720×960, 包含训练图 367 张、验证图 101 张和测试图 233 张, 本文选用其中 11 个目标类别进行训练和评估.

2.1.2 实验环境

本文实验在 Ubuntu 18.04 的操作系统下使用当前主流的深度学习框架 PyTorch 1.7.0, 利用 Python 3.8 编程语言构建网络框架模型. CPU 选用的型号为 Intel(R) Xeon(R) Platinum 8350C CPU @ 2.60 GHz, 硬件配置 GPU 显卡版本为 Nvidia RTX 3090 和显存为 24 GB 的设备上进行, 同时使用了计算框架 CUDA 11.0 和 GPU 加速库 cuDNN 进行高性能的并行计算.

2.2 训练过程及分析

2.2.1 数据集预处理

由于 Cityscapes 数据集和 CamVid 数据集存在数据量有限和类别不平衡等问题, 为了提高训练数据的多样性以及网络模型的鲁棒性, 在训练过程中对其进行了预处理. 这些预处理方法包括了随机裁剪、随机缩放以及镜像翻转等数据增强技术的使用. 首先, 对原始尺寸为1024×2048像素的 Cityscapes 数据集进行了随机裁剪操作, 使其尺寸调整为512×1024像素. 同样地, 对于原始尺寸为720×960像素的 CamVid 数据集, 也进行随机裁剪操作, 将其尺寸调整为352×480像素; 然后对两个数据集随机使用 (0.75, 1.00, 1.25, 1.50, 1.75, 2.00) 中的数值, 将图像转换至不同的尺寸输入到网络中训练; 最后再将两个数据集进行水平翻转来进行数据增强. 模型训练时采用的训练数据是 Cityscapes 和 CamVid 的 train+val 数据集, 测试时 Cityscapes 采用 val 数据集, CamVid 采用 test 数据集.

2.2.2 训练优化策略

当使用 Cityscapes 数据集和 CamVid 数据集进行实验时, 训练过程首先使用 warmup 预热策略, 然后再使用 Poly 策略来对学习率大小进行动态化调整. warmup 预热策略在前 10 个 epoch 训练时, 模型学习率会由初始值逐渐变大, 训练 10 个 epoch 后, 学习率达到预先设置的值之后, 再使用 Poly 策略使学习率随着 epoch 不断迭代而逐渐衰减. Poly 策略公式为:

$$lr = base\_lr \times \left(1 - \frac{iter}{max\_iter}\right)^{power} \tag{8}$$

其中, lr表示当前学习率, $base_lr$表示初始学习率, $iter$表示当前迭代次数, max_iter表示总迭代次数, 动量 $power = 0.9$.

通过采用 warmup 预热策略, 可以使模型在训练初

期使用较小的学习率, 经过一定数量的迭代次数后, 模型逐渐达到稳定状态, 然后会切换至预定的学习率进行进一步的训练. 这种方法有效地实现了预热学习率的效果, 避免了模型震荡, 同时也让网络的收敛速度更快, 从而提升了网络模型的性能.

2.2.3 训练过程参数设置及分析

当使用 CamVid 数据集进行实验时, 采用 Adam 优化器调整网络参数, 损失函数采用交叉熵损失与 Dice 损失两者结合的组合损失函数.

当使用 Cityscapes 数据集进行实验时, 采用 SGD 优化器调整网络参数, 损失函数采用交叉熵损失与 Dice 损失两者结合的组合损失函数. 实验其他部分的参数设置如表 1 所示.

表 1 实验参数设置

参数	Cityscapes	CamVid
输入尺寸	512×1024	352×480
优化器	SGD	Adam
损失函数	CE+Dice loss	CE+Dice loss
Batch size	8	8
学习率	5×10^{-2}	5×10^{-4}
Epoch	500	500

2.3 实验评价指标

本文采用的评价指标为交并比 (intersection over union, *IoU*)、平均交并比 (mean intersection over union, *MIoU*)、每秒处理帧数 (frames per second, *FPS*) 以及网络模型总参数量 (parameters).

IoU 和 *MIoU* 用于评估模型分割性能, 两者的值越大代表分割的结果越好. *IoU* 表示标签真实值和模型预测值交集和并集的比值; *MIoU* 表示所有类别标签真实值和模型预测值交集和并集的比值的平均值. *IoU* 和 *MIoU* 的计算公式如下:

$$IoU=\frac{p_{ij}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}(p_{ji}-p_{ii})} \tag{9}$$

$$MIoU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ij}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}(p_{ji}-p_{ii})} \tag{10}$$

其中, k 表示类别数量, p_{ii} 表示预测正确的像素个数, p_{ij} 表示预测错误的像素个数.

FPS 用于评价模型分割速度, 计算公式如下:

$$FPS=\frac{N}{\sum_{i=1}^{N}T_j} \tag{11}$$

其中, N 表示图像数量, T_j 表示模型处理第 j 幅图像的时间.

Parameters 是计算模型总的可训练参数量, 主要用来评估模型的空间复杂度.

2.4 实验结果与分析

2.4.1 不同损失函数对比

为了验证交叉熵损失和 Dice 损失两者结合的组合损失函数对 AS-UNet 改进网络模型性能的贡献, 在 Cityscapes 和 CamVid 数据集上对比了常用的交叉熵损失 (CE) 与本文组合损失函数 (CE+Dice loss), 实验结果如表 2 所示. 可以看出, 采用组合损失函数后 AS-UNet 网络模型的平均交并比在 Cityscapes 和 CamVid 数据集上分别提升了 0.3% 和 0.2%, 进而证明了组合损失函数对本文 AS-UNet 改进网络模型的有效性.

表 2 不同损失函数 *MIoU* 比较 (%)

损失函数	Cityscapes	CamVid
CE	64.2	59.6
CE+Dice loss	**64.5**	**59.8**

2.4.2 不同模型对比

为了验证 AS-UNet 网络模型对于街道场景图像分割的有效性, 将本文方法在 Cityscapes 数据集和 CamVid 数据集上进行测试, 并与近年流行的语义分割模型进行对比, 实验结果如表 3 和表 4 所示.

表 3 Cityscapes 验证集上不同模型比较

模型	输入尺寸	*MIoU* (%)	Parameters (M)	*FPS*
U-Net	512×1024	60.6	7.85	67.4
FCN	512×1024	63.1	134.50	2.0
SegNet	360×640	57.0	29.50	14.6
ESPNet	512×1024	60.3	**0.36**	**112.0**
ENet	512×1024	58.3	**0.36**	76.9
AS-UNet	512×1024	**64.5**	17.38	38.2

表 4 CamVid 测试集上不同模型比较

模型	输入尺寸	*MIoU* (%)	Parameters (M)	*FPS*
U-Net	352×480	56.8	7.85	92.8
FCN	360×480	57.0	134.50	12.0
SegNet	360×480	55.6	29.50	16.7
ESPNet	360×480	55.6	**0.36**	**132.0**
ENet	360×480	51.3	**0.36**	105.7
AS-UNet	352×480	**59.8**	17.38	61.3

表 3 是在 Cityscapes 验证集上将本文模型 AS-UNet 与 U-Net、FCN、SegNet、ESPNet[20]、ENet[21]模型在

平均交并比、模型参数量和分割速率方面进行对比. 实验结果表明, 本文提出的 AS-UNet 相较于传统 U-Net 平均交并比提升了 3.9%, 分割速率依然能满足实时性要求. 相较于 FCN 和 SegNet 这两种经典模型, 改进模型在平均交并比方面相较于 FCN 模型提升了 1.4%, 且相较于 SegNet 模型提升了 7.5%, 另外改进模型分割速率更高并且参数量更少. 相较于 ESPNet 和 ENet 这两种轻量级模型, 模型的分割速率虽然有一定下降并且参数量增加, 但是改进模型在平均交并比方面相较于 ESPNet 模型提升了 4.2%, 且相较于 ENet 模型提升了 6.2%.

表 4 是在 CamVid 测试集上同样的将改进模型 AS-UNet 与 U-Net、FCN、SegNet、ESPNet、ENet 模型在平均交并比、模型参数量和分割速率方面进行对比. 实验结果表明, 本文提出的 AS-UNet 相较于传统 U-Net 平均交并比提升了 3.0%, 分割速率依然能够满足实时性要求. 相较于 FCN 和 SegNet 这两种经典模型, 改进模型不仅在平均交并比方面分别提升了 2.8% 和 4.2%, 而且模型的分割速率更高且参数量更少. 相较于 ESPNet 和 ENet 这两种轻量级模型, 模型的分割速率虽然有一定下降且参数量增加, 但是平均交并比分别提升了 4.2% 和 8.5%.

综合以上分析, 本文改进的模型在平衡分割精度和分割速度方面表现较好, 实现了高效而准确的分割结果, 验证了本文改进的模型更能够适用于目标种类较多和目标尺度变化较大条件下的街道场景图像语义分割.

2.4.3 消融实验

本文还进行了消融实验, 证明 scSE 注意力机制模块和 ASPP 多尺度特征融合模块对街道场景图像分割的有效性, 实验结果如表 5 所示.

当在传统 U-Net 中仅加入 scSE 注意力机制模块来同时增强通道和空间特征之间的表达时, 在 Cityscapes 和 CamVid 数据集上模型的 *MIoU* 较原始 U-Net 模型分别提升了 1.5% 和 1.3%; 当仅加入 ASPP 多尺度特征融合模块来获取图像的全局上下文信息时, 在 Cityscapes 和 CamVid 数据集上模型的 *MIoU* 较原始 U-Net 模型分别提升了 2.8% 和 2.5%; 当同时加入 scSE 注意力机制模块和 ASPP 多尺度特征融合模块时, 在 Cityscapes 和 CamVid 数据集上模型的 *MIoU* 分别提升到 64.5% 和 59.8%. 进而证明了在传统 U-Net 网络中加入 scSE 注意力机制模块和 ASPP 多尺度特征融合模块能够关注图像关键区域特征信息和聚合图像多尺度特征.

表 5 消融实验 (%)

模型	Cityscapes	CamVid
U-Net	60.6	56.8
U-Net+scSE	62.1	58.1
U-Net+ASPP	63.4	59.3
AS-UNet	**64.5**	**59.8**

2.4.4 单类别实验结果及分析

为了进一步证明 AS-UNet 网络模型的有效性, 本文还在 Cityscapes 数据集和 CamVid 数据集上进行各类别分割结果比较实验, 如表 6 和表 7 所示. 从表 6 和表 7 中数据可以发现, AS-UNet 网络模型分别有 13 个类别和 5 个类别目标的分割结果超过其他对比模型, 尤其是像电线杆、信号标、行人等这种在图片中所占像素比较低的小目标. 而与传统 U-Net 模型相比, AS-UNet 网络模型所有类别的分割结果均超过原始 U-Net 网络. 根据表 6 和表 7 中数据, 证明本文改进模型对于街道场景图像中远景的小目标具有良好的分割效果.

表 6 Cityscapes 验证集上各类别的 *IoU* 值比较 (%)

类别	U-Net	FCN	SegNet	ESPNet	ENet	Ours
道路	96.7	97.2	96.4	97.0	96.3	97.0
人行道	75.6	78.2	73.2	77.5	74.2	**78.8**
建筑物	85.0	88.0	84.0	76.2	75.0	**88.6**
墙体	29.5	34.7	28.4	35.0	32.2	**35.5**
围栏	45.9	42.1	29.0	36.1	33.2	**46.4**
电线杆	47.0	47.2	35.7	45.0	43.4	**48.8**
交通灯	46.6	45.2	39.8	35.6	34.1	**47.2**
交通标	56.2	50.4	45.1	46.3	44.0	**56.7**
植被	89.5	91.2	87.0	90.8	88.6	91.0
地面	57.2	61.7	63.8	63.2	61.4	61.5
天空	86.4	92.4	91.8	92.6	90.6	91.8
行人	72.2	74.9	62.8	67.0	65.5	**73.6**
骑行者	48.0	51.2	42.8	40.9	38.4	49.5
汽车	90.8	92.4	89.3	92.3	90.6	91.4
卡车	42.9	39.4	38.1	38.1	36.9	**46.4**
公交车	52.9	53.4	43.1	52.5	50.5	**55.9**
火车	29.8	51.3	44.1	50.1	48.1	**53.6**
摩托车	35.7	47.4	35.8	41.8	38.8	**47.9**
自行车	62.6	60.6	51.9	57.2	55.4	**63.7**

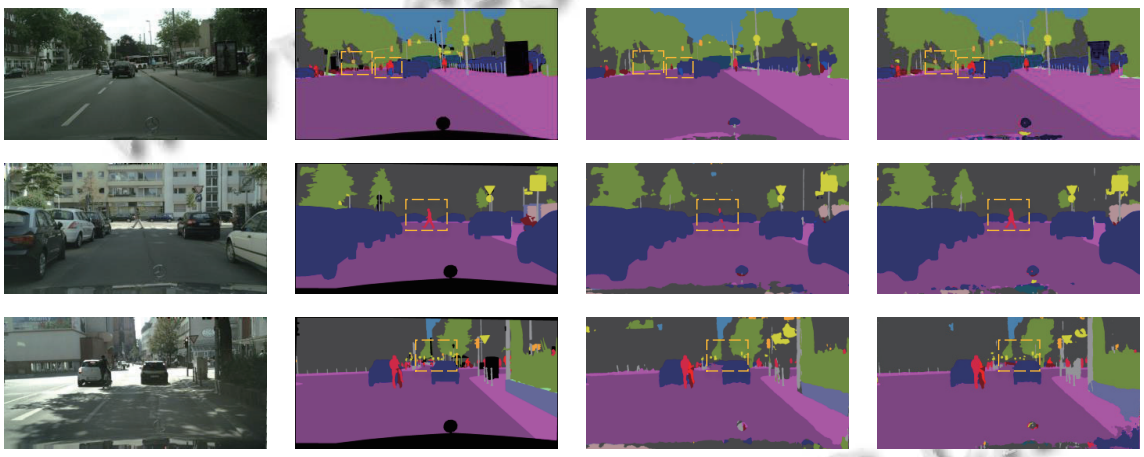
2.4.5 定性结果分析

为了验证 AS-UNet 网络模型的分割性能, 在

Cityscapes 数据集和 CamVid 数据集中选取 3 幅图像与原始 U-Net 网络进行定性结果分析, 结果如图 7 和图 8 所示. 从图 7 和图 8 的定性对比结果可以看出, 原始 U-Net 网络和改进的 AS-UNet 网络模型对于尺寸较大的物体基本能够正确分割, 但是对于 Cityscapes 数据集来说原始模型对于因遮挡或尺寸较小的交通标、行人、摩托车和自行车等目标没能获得准确分割, 而本文模型提高了分割的准确率; 对于 CamVid 数据集来说原始 U-Net 模型对于电线杆、信号标、行人和道路等目标未能很好地分割, 而通过使用改进的 AS-UNet 网络模型可以将这些目标完整地分割出来. 综合上述分析, AS-UNet 网络模型较原始 U-Net 网络的分割结果有显著提升, 进一步证明通过引入注意力机制模块和多尺度特征融合模块能够增强街道场景图像分割效果.

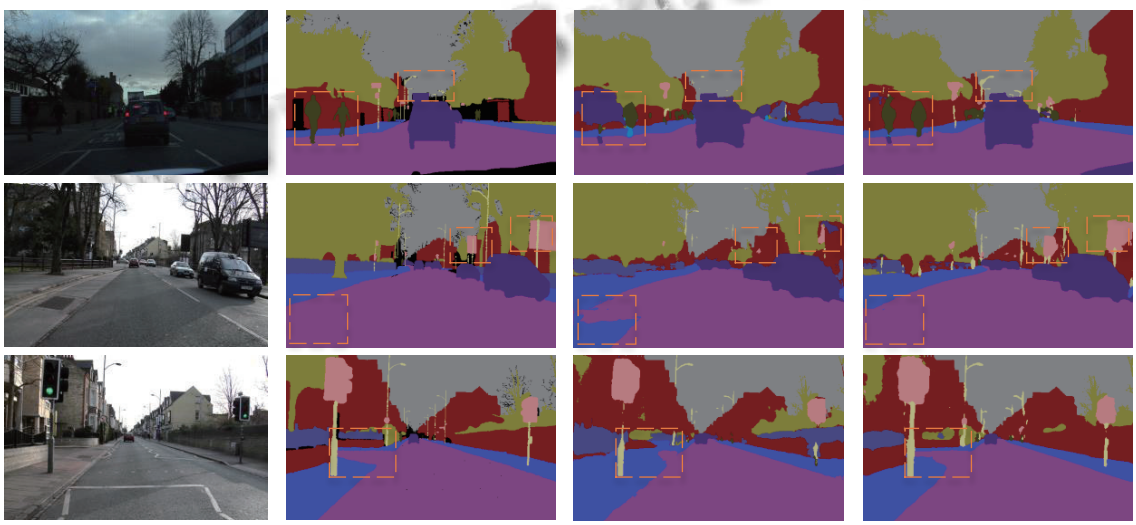
表 7 CamVid 测试集上各类别的 *IoU* 值比较 (%)

类别	U-Net	FCN	SegNet	ESPNet	ENet	Ours
天空	89.2	88.7	87.6	90.9	90.4	89.5
建筑物	72.6	77.8	78.8	73.9	66.4	73.9
电线杆	18.7	19.9	20.5	25.9	19.4	**27.7**
道路	89.3	91.2	87.5	92.5	91.6	**97.4**
人行道	71.2	72.7	70.0	75.5	71.9	73.2
树木	66.8	71.0	62.3	67.6	59.6	67.7
信号标	33.5	21.0	33.4	22.3	18.8	**34.5**
围栏	20.9	31.4	19.3	20.0	20.6	23.8
汽车	73.1	78.8	68.3	73.8	67.8	75.8
行人	47.6	50.5	43.3	38.9	26.8	**52.3**
骑手	41.9	31.0	40.6	41.7	32.1	**42.1**



(a) 原图 (b) 标签 (c) U-Net (d) AS-UNet

图 7 Cityscapes 数据集上的可视化对比结果



(a) 原图 (b) 标签 (c) U-Net (d) AS-UNet

图 8 CamVid 数据集上的可视化对比结果

3 结论

针对原始 U-Net 网络在处理目标种类较多且尺度变化较大的街道场景图像时存在分割精度不高、分割效果不好的问题, 本文提出 AS-UNet 网络模型来改进原始 U-Net 网络. 通过采用 ASPP 多尺度特征融合模块, 实现将街道场景图像不同尺度目标的语义信息进行有效融合, 提高特征图信息丰富度; 引入 scSE 注意力机制模块, 引导网络关注街道场景图像的目标区域, 提高分割的准确性; 采用组合损失函数解决街道场景图像目标类别不平衡的问题, 进一步提升分割性能. 在 Cityscapes 数据集和 CamVid 数据集上的实验表明, 与原始 U-Net 网络相比, AS-UNet 网络模型能够实现高效且准确的街道场景图像分割. 并且 AS-UNet 网络模型的参数量较低, 推理速度较快, 在分割精度、参数量以及分割速度之间取得了较优的平衡, 适合部署在嵌入式设备上. 下一步研究将专注于进一步提升网络的分割精度和分割速度, 使网络的性能更加高效.

参考文献

1 王龙飞, 严春满. 道路场景语义分割综述. 激光与光电子学进展, 2021, 58(12): 1200002.

2 Grigorescu S, Trasnea B, Cocias T, *et al*. A survey of deep learning techniques for autonomous driving. Journal of Field Robotics, 2020, 37(3): 362–386. [doi: 10.1002/rob.21918]

3 钮圣虓, 王盛, 杨晶晶, 等. 完全基于边缘信息的快速图像分割算法. 计算机辅助设计与图形学学报, 2012, 24(11): 1410–1419. [doi: 10.3969/j.issn.1003-9775.2012.11.005]

4 郑美珠, 赵景秀. 基于区域一致性测度的彩色图像边缘检测. 计算机应用, 2011, 31(9): 2485–2488, 2492.

5 付云凤. 基于阈值的图像分割研究 [硕士学位论文]. 重庆: 重庆大学, 2013.

6 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3431–3440.

7 Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481–2495. [doi: 10.1109/TPAMI.2016.2644615]

8 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.

9 Zhou ZW, Siddiquee MMR, Tajbakhsh N, *et al*. U-Net++: A nested U-Net architecture for medical image segmentation. Proceedings of the 4th International Workshop on Deep Learning in Medical Image Analysis. Granada: Springer, 2018. 3–11.

10 Oktay O, Schlemper J, Le Folgoc L, *et al*. Attention U-Net: Learning where to look for the pancreas. arXiv:1804.03999, 2018.

11 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6230–6239.

12 Chen LC, Zhu YK, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 833–851.

13 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

14 Chen LC, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFS. Computer Science, 2014, (4): 357–361.

15 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587, 2017.

16 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. Proceedings of the 4th International Conference on Learning Representations. San Juan: OpenReview.net, 2016.

17 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.

18 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

19 Roy AG, Navab N, Wachinger C. Concurrent spatial and channel ‘squeeze & excitation’ in fully convolutional networks. Proceedings of the 21st International Conference on Medical Image Computing and Computer-assisted Intervention. Granada: Springer, 2018. 421–429.

20 Mehta S, Rastegari M, Caspi A, *et al*. ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 561–580. [doi: 10.1007/978-3-030-01249-6_34]

21 Paszke A, Chaurasia A, Kim S, *et al*. ENet: A deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147, 2016.

(校对责编: 张重毅)