计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(12):135−146 [doi: 10.15888/j.cnki.csa.008824]

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

改进 U-Net 的高分辨率遥感图像轻量化分割①

胡　伟, 文　武, 魏　敏

(成都信息工程大学 计算机学院, 成都 610225)
通信作者: 文　武, E-mail: wenwu@cuit.edu.cn

摘　要: 针对传统图像分割方法分割效率低下, 遥感图像特征复杂多样, 复杂场景下分割性能受到限制等问题, 在基于 U-Net 网络架构的基础上, 提出一种能够较好提取遥感图像特征并兼顾效率的改进 U-Net 模型. 首先, 以 EfficientNetV2 作为 U-Net 的编码网络, 增强特征提取能力, 提高训练和推理效率, 然后在解码部分使用卷积结构重参数化方法并结合通道注意力机制, 几乎不增加推理时间的前提下提升网络性能, 最后结合多尺度卷积融合模块, 提高网络对不同尺度目标的特征提取能力和更好地结合上下文信息. 实验表明, 改进的网络在遥感图像分割性能提升的同时分割效率也提高.

关键词: 遥感图像; 图像分割; U-Net; EfficientNetV2; 结构重参数化; 多尺度卷积; 注意力机制; 卷积神经网络

引用格式: 胡伟,文武,魏敏.改进 U-Net 的高分辨率遥感图像轻量化分割.计算机系统应用,2022,31(12):135–146. http://www.c-s-a.org.cn/1003-3254/8824.html

Lightweight Segmentation for High Resolution Remote Sensing Image Based on Improved U-Net

HU Wei, WEN Wu, WEI Min

(School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Considering the problems of low segmentation efficiency of traditional image segmentation methods, complex and diverse features of remote sensing images, and limited segmentation performance in complex scenes, an improved U-Net model is proposed on the basis of the U-Net network architecture, which can satisfactorily extract the features of remote sensing images while maintaining efficiency. First, EfficientNetV2 is used as the encoding network of U-Net to enhance the feature extraction ability and improve the training and inference efficiency. Then, the convolutional structural re-parameterization method is applied in the decoding network and is combined with the channel attention mechanism to improve the network performance without increasing the inference time. Finally, the multi-scale convolution fusion module is employed to improve the feature extraction ability of the network for objects with different scales and the utilization of context information. The experiments reveal that the improved network can not only improve the segmentation performance of remote sensing images but also promote segmentation efficiency.

Key words: remote sensing image; image segmentation; U-Net; EfficientNetV2; structural re-parameterization; multi-scale convolution; attention mechanism; convolutional neural network (CNN)

随着现代遥感技术的快速发展, 遥感图像已经成为一种不可或缺的重要资源, 被广泛应用于各个领域. 遥感图像语义分割是解释遥感图像信息的重要基础环节, 所以, 遥感图像分割技术的研究, 对充分利用遥感图像的价值起着重要的意义. 而随着遥感图像分辨率的逐步提高, 使得遥感图像所包含的地物类型也逐步复杂, 对于高分辨率的遥感图像分割算法的要求也在不断提高. 传统的分割算法例如阈值分割、边缘检测

① 基金项目: 四川省科技计划重点研发项目 (2020YFG0442, 2020YFG0453)
收稿时间: 2022-03-17; 修改时间: 2022-04-14; 采用时间: 2022-04-25; csa 在线出版时间: 2022-07-25

分割等, 但这些方法效率较低, 对于特征复杂多样的图像分割效果比较差.

卷积神经网络 (CNN), 具有局部感知和参数共享的特点, 可以有效降低网络训练的参数量和复杂度, 便于网络的训练和优化改进[1]. 2015 年, Long 等人[2] 提出了著名的 FCN 结构, 实现了像素级的分类和端到端的训练, 简化了分割过程, 解决了语义级图像分割问题. Ronneberger 等人[3] 在基于 FCN 的基础上, 对 FCN 进行了改进, 提出了一种新的网络结构 U-Net, 作者将该网络结构用于医学图像分割处理中, 在分割细胞壁的任务中取得了很好的结果.

U-Net 最早用于医学图像分割任务中, 以简洁的结构和优秀的性能著称, 因此根据不同的问题进行了不同的改进, 例如 UNet++[4]、Attention U-Net[5]、U2-Net[6] 等优秀模型. 为了在遥感图像分割任务中能更好、更精确地获取地物信息, 国内外学者将 U-Net 模型用于遥感图像分割任务中, 并进行了很多尝试改进. 苏健民等人[7] 将 U-Net 卷积过滤器改为统一深度, 尺寸为 3×3×64, 并将 ReLU 激活函数改为 ELU 激活函数, 提升了对噪声的鲁棒性, 使网络能够更好地收敛, 同时提高分割准确率. 范自柱等人[8] 提出 W-Net 网络结构, 在解码阶段使用 GCNet[9] 中的全局上下文模块来提升网络的分割准确率, 使用 FPN 结构来整合高层和低层的特征信息, 提升分割能力, 因整体的网络结构呈 W 型, 所以网络被称为 W-Net. 王曦等人[10] 提出 U-Net 和 FPN 相结合的网络模型, 通过在跳跃连接部分使用 FPN 结构, 更好地融合高层特征和低层特征, 通边引入界标签松弛损失函数 (BLR)[11], 提升模型对各目标类别边缘的分割能力, 使分割结果更加精细化.

遥感图像纹理信息复杂, 各类别尺度多变, 背景复杂等特殊原因, 导致遥感图像分割仍存在很多问题, 例如复杂结构的目标无法完整识别, 各目标之间存在错分和漏分现象, 并且深度神经网络往往计算量大, 为了能更加效率、更精确地实现地物信息的提取, 本文提出一种基于 U-Net 网络的图像分割方法来对遥感图像中的地物进行分割. 以 EfficientNetV2[12] 为编码网络替换原来 U-Net 网络的编码部分, 在解码部分使用卷积结构重参数化, 灵感来源于 Ding 等人[13] 提出的 RepVGG 网络, 并且在解码网络的下层分支结构中结合 SE 通道注意力机制[14], 命名为 RepVGG-SE 模块, 通过多尺度卷积获得不同感受野的特征图并通过下采样模块注入到网络深层, 更好地融合上下文信息. 实验证明, 本方法在所有对比网络中指标均为最高, 并且参数量和计算量大大减小, 提升了训练和推理效率.

1 网络结构

1.1 原始 U-Net 网络结构

U-Net 网络结构如图 1 所示, 由编码部分和解码部分组成. 在编码网络中, 在每次下采样之前使用 2 个卷积核为 3×3 的卷积层进行特征提取, 卷积之后使用 ReLU 激活函数, 使用大小为 2×2 的最大池化操作减少特征维度, 增大感受野. 每经过一次下采样, 图像尺寸缩小一半, 维度加倍, 通过这种重复的操作可以充分提取图像的高层特征和过滤掉不需要的信息. 解码部分使用反卷积进行上采样, 上采样之后同样使用 2 个卷积核为 3×3 的卷积层, 逐步恢复图像的细节信息, 并最终恢复特征图至输入图片的尺寸. 每经过一次上采样, 图像尺寸增大一倍, 维度减半. 编码部分和解码部分对应阶段之间使用跳跃连接结构, 复用低层次特征信息, 更好地还原图像细节信息.

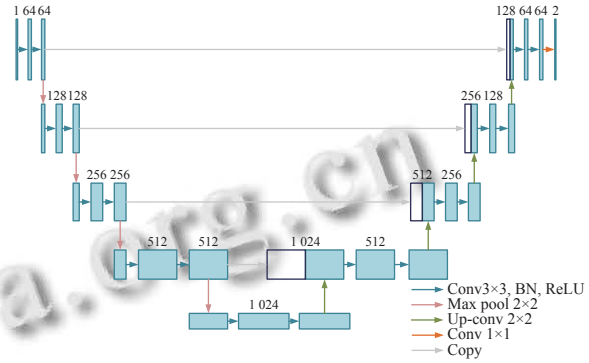


图 1 U-Net 网络结构

1.2 EfficientNetV2 网络结构

2021 年, Tan 等人[12] 提出 EfficientNetV2 网络模型, 是 EfficientNet 网络模型[15] 的改进版本, 不仅特征提取能力非常强, 参数量、计算量和推理效率都得到了很好的权衡. EfficientNet 网络, 主要存在以下问题, 当训练图像尺寸很大时, 训练速度非常慢; 在网络浅层中使用 depthwise convolutions (DWConv) 速度会很慢; 每个 stage 的深度和宽度同等放大是次优的. 对此, EfficientNetV2 网络做了以下改进.

在 EfficientNet 中使用了很多卷积核为 5×5 的卷积层, 而在 EfficientNetV2 中更偏向使用卷积核为

3×3 的卷积层, 由于 3×3 的卷积层感受野要小于 5×5 的卷积层, 因此需要堆叠更多的层结构来增加感受野. expansion ratio 由原来的 6 减小为 4, 目的是减少内存访问的开销. 由于每个 stage 对网络的训练速度和参数数量的贡献并不相同, 因此采用非均匀的缩放策略来缩放模型. 虽然 DWConv 结构比普通卷积拥有更少的参数和计算量, 但是通常无法充分利用一些加速器, 因此在 MBConv 模块的基础上, 使用 Fused-MBConv 模块来解决这一问题, 通过将 DWConv 替换成普通的 3×3 卷积, 能够明显提升训练速度. MBConv 模块结构和 Fused-MBConv 模块结构如图 2 所示. 在浅层网络部分, 堆叠 Fused-MBConv 模块, Fused-MBConv 模块包含普通 3×3 卷积进行升维, 1×1 卷积降维, Swish 激活函数, Dropout 层, 并将 SE 注意力机制嵌入到模块中. MBConv 由 DWConv、1×1 卷积升维和降维、Swish 激活函数、Dropout 层、SE 通道注意力机制组成.

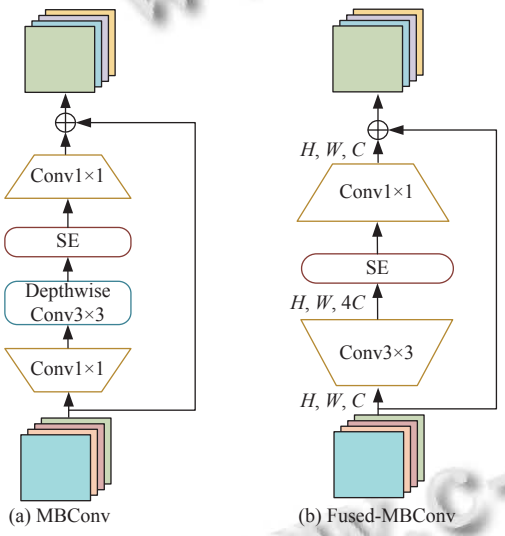


图 2 Fused-MBConv 和 MBConv 结构

本实验选用 EfficientNet 网络的 EfficientNetV2-S 结构作为改进 U-Net 模型的编码网络.

1.3 R epVGG-SE 模块

对于一般通用模型来说, 例如 VGG[16] 网络, 相对于各种多分支架构 (ResNet[17], Inception[18], DenseNet[19]), 性能更差, 然而, 多分支结构的缺点是推理时需要消耗大量的计算资源, 因此, 通过卷积结构重参数化的方式, 将多分支模型等价转换为单路模型, 使得整个网络结构在推理上更加简单, 这样既利用了多分支模型训练时性能高的优势, 又利用了推理时单路模型速度快、节省内存的特点. 在 RepVGG 网络中, 借鉴 ResNet 网络的思想, 构建 RepVGG 模块, RepVGG 模块如图 3 所示. ResNet 的 ResBlock 构建了一个短连接模型信息流 $y = x + f(x)$, 当x和$f(x)$维度不匹配时, 上述信息流则转变成$y = g(x) + f(x)$, 在 RepVGG 模块中, 信息流为$y = b(x) + g(x) + f(x)$, 若x和$f(x)$维度不匹配, $y = g(x) + f(x)$, 其中$b(x)$、 $g(x)$和$f(x)$分别为x通过 batch normalization (BN)[20]、1×1 卷积和 3×3 卷积的同层分支连接, 在推理时, 网络分支架构等价于$y = h(x)$, 其中$h(x)$仅由一个 3×3 卷积层实现, 参数通过线性组合方式从已训练好的模型中转换得到. 相比于其他卷积核, 3×3 卷积计算密度更高, 更加有效, 单路架构并行度也更高, 更节省内存.

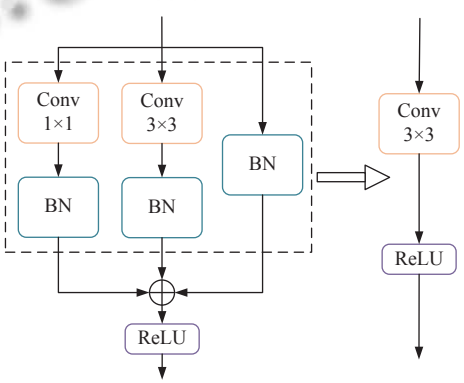


图 3 RepVGG 模块结构图

卷积结构重参数化训练后, 通过代数进行变换. RepVGG 模块中 1×1 卷积相当于一个退化的 3×3 卷积, 将 1×1 卷积核用 0 进行填充, 即可得到 3×3 卷积, identity 分支是一个特殊 1×1 卷积, 以单位矩阵为矩阵核, 整个分支在训练期间都通过了 BN 层. 此外, 还整合了卷积层和 BN 层之间的特征. 式 (1)、式 (2) 分别为卷积层和 BN 层公式, 将卷积层带入 BN 公式, 得到式 (3), 在训练时卷积层不设置*bias*参数, 即式 (1) 中没有b. 经过 BN 后的卷积由式 (3) 控制, 得到一个新的卷积, 只不过考虑了 BN 的参数, 训练阶段训练$W(x)$、γ和β参数, 最终融合的结构由式 (4) 表示. 通过变换, RepVGG 模块仅具有一个 3×3 卷积核, 两个 1×1 卷积核以及 3 个*bias*参数, 3 个*bias*参数可以直接通过 add 方式合并为一个*bias*, 新的卷积核可以将 1×1 卷积核参数加到 3×3 卷积核的中心点得到, 所有分支特征和最终偏置将分配给一个新的 3×3 的偏置卷积, 再进入 ReLU 激活函数.

$$Conv(x) = W(x) + b \tag{1}$$

$$BN(x)=\gamma\times\frac{(x-mean)}{\sqrt{var+\varepsilon}}+\beta \tag{2}$$

$$BN(Conv(x))=\frac{\gamma\times W(x)}{\sqrt{var+\varepsilon}}+\left(\beta-\frac{(\gamma\times mean)}{\sqrt{var+\varepsilon}}\right) \tag{3}$$

$$BN(Conv(x))=W_{\text{fused}}(x)+B_{\text{fused}} \tag{4}$$

在传统的卷积池化过程中, 默认特征图的每个通道都是同等重要的, 而实际问题中, 不同通道的重要性是有差异的. SE 通道注意力机制目的在于关注特征图不同通道之间的关系, 希望模型可以学习到不同通道特征的重要程度, 帮助网络模型更好地学习到有用特征. SE 模块结构如图 4 所示.

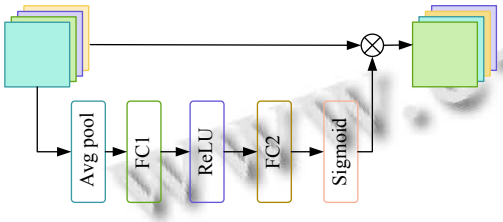


图 4 SE 模块

一个 SEBlock 分为 squeeze 和 excitation 两个过程, 首先 squeeze 过程是对输入的$C\times W\times H$图进行全局平均池化操作, 在全局感受野上对全局信息进行编码, 得到$C\times 1\times 1$压缩特征向量, 如式 (5) 所示:

$$F_{sq}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j) \tag{5}$$

其中, i和j表示特征图上像素点的位置, x_c表示具体位置的像素值. Excitation 过程为进入两个全连接层, 第 1 个全连接层进行降维处理, 第 2 个全连接层进行升维处理为原通道数, 目的是拟合通道间复杂的相关性, 建立起各通道之间的权重关系, 最后接入一个 Sigmoid 层, 生成通道之间 0–1 的注意力权重信息, 此过程如式 (6) 所示:

$$W_c=F_{ex}=\sigma(W_1(\delta(W_0(F_{sq})))) \tag{6}$$

其中, W_0和W_1表示两个全连接层, δ表示 ReLU 激活函数, σ表示 Sigmoid 激活函数, W_c为最终得到的通道注意力权重向量. 生成的权重向量与原始特征图做点乘操作, 增强信息量大的特征, 抑制无用的特征.

$$x=x\odot W_c \tag{7}$$

其中, $\odot$表示点乘操作, x为$C\times H\times W$大小的输入特征图.

Goyal 等人[21] 于 2021 年提出 ParNet 模型, 该模型深度仅为 12 层. ParNet 中有一个关键组件, 作者称为 RepVGG-SSE 模块, 这是一个经过修改的 RepVGG 模块, 在 identity 分支结构中带有一个专门构建的 skip-squeeze-excitation (SSE) 模块, 该模块基于上文提及的 SEBlock, 但 RepVGG-SSE 模块仅适用于层数较少的神经网络, 因此, 本文直接将 SSE 替换为 SEBlock, 称为 RepVGG-SE 模块, 将该模块用于解码网络. RepVGG-SE 模块如图 5 所示.

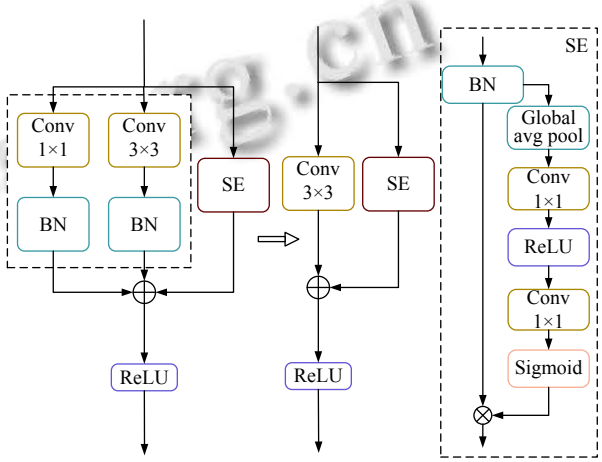


图 5 RepVGG-SE 模块

1.4 多尺度卷积融合模块

EfficientNetV2 网络开始阶段以卷积核为 3×3、步长为 2 的卷积进行特征提取和下采样, 在图像分割任务中, 存在的问题是感受野不够大, 无法获取足够多的特征信息和细节信息, 而遥感图像中目标类别尺度多变, 为了能够得到不同尺度更加丰富的特征信息, 采用多尺度卷积模块, 如图 6 所示.

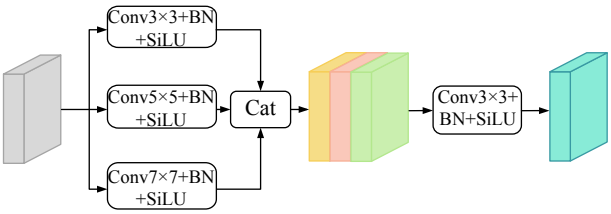


图 6 多尺度卷积模块

多尺度卷积模块中, 进行多尺度卷积操作, 卷积核大小分别为 3×3、5×5 和 7×7, 不同卷积核大小的卷积拥有不同大小的感受野, 然后将各个多尺度卷积操作后的特征图进行拼接, 以一个卷积核大小为 3×3 的卷积将多尺度特征进行融合, 形成一个包含不同尺度更多特征信息的特征图. 为了和 EfficientNetV2 网络更好

地融合, 采用 SiLU 函数进行非线性激活.

为了将多尺度特征信息更好地和网络进行融合, 同时 EfficientNetV2 网络使用了较多的 Fused-MBConv 模块和 MBConv 模块, 网络层数较深, 损失了较多的细节信息, 因此采用如图 7 所示的下采样模块, 将浅层特征信息注入到深层网络, 同时弥补深层网络的细节特征信息损失.

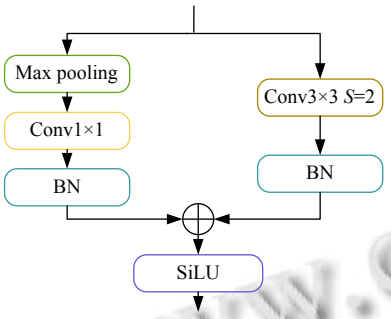


图 7 下采样模块

原始 U-Net 网络中采用最大池化的方式进行下采样, 但丢失了特征图较多的信息, 为了保留更多的上下文信息, 设计了下采样模块, 具体为包含两个分支, 一个分支是典型的最大池化下采样方式, 然后通过 1×1 卷积进行特征图维度的改变, 一个分支采用 3×3 卷积、步长为 2 的方式进行下采样, 最后两个分支通过对应元素求和的方式进行融合, 以 SiLU 函数进行非线性激活, 以保留更多的特征信息.

1.5 网络整体结构

本文以 U-Net 网络为基础, 对 U-Net 网络模型进行改进, 网络结构如图 8 所示. 网络整体分为编码和解码, 编码部分, 利用 EfficientNetV2-S 网络替换原始 U-Net 网络的编码部分, 编码器有 5 种输出, 即 1/2、1/4、1/8、1/16 和 1/32 的 5 种特征图, 1/32 的特征图作为解码端的输入, 其余 4 种特征图作为跳跃连接的编码特征与解码端进行拼接融合. 解码部分, 通过反卷积的操作将特征图的尺寸扩大为原来的 2 倍并逐步将特征图恢复至原图片分辨率大小, 并与编码部分的 4 种编码特征进行拼接操作, 以融合获得更多的特征信息和细节信息, 最后通过输出层将特征图映射成特定数量的类别进行像素类别预测, 获得分割结果. 编码网络中, 通过引入 RepVGG-SE 模块进行加深处理, RepVGG-SE 模块是一个多分支结构, 训练时保持多分支结构, 和单路架构相比, 分支结构进一步融合了细节信息和语义信息, 提取的特征表达能力更强, 提高了训练性能, 推理时进行卷积结构重参数化转换为一个 3×3 卷积和 SEBlock, 提升推理速度, 节省内存, 在 identity 分支结构中的 SE 通道注意力机制, 能够建立起通道维度上的依赖关系, 增强有用信息的表达, 抑制无效特征. 经过实验, 在解码网络下层使用 RepVGG-SE 模块, 上层使用 RepVGG 模块效果最好. 当输入维度和输出维度不匹配时, RepVGG 模块没有 identity 分支, 只有 1×1 和 3×3 卷积分支. 多尺度卷积模块, 以不同卷积核大小的卷积进行特征提取, 然后将多尺度卷积获得的特征图进行拼接融合, 通过下采样模块, 将浅层信息注入到深层网路中, 使多尺度特征信息更好地与网络各层进行融合, 同时弥补了深层网络中丢失的细节信息.

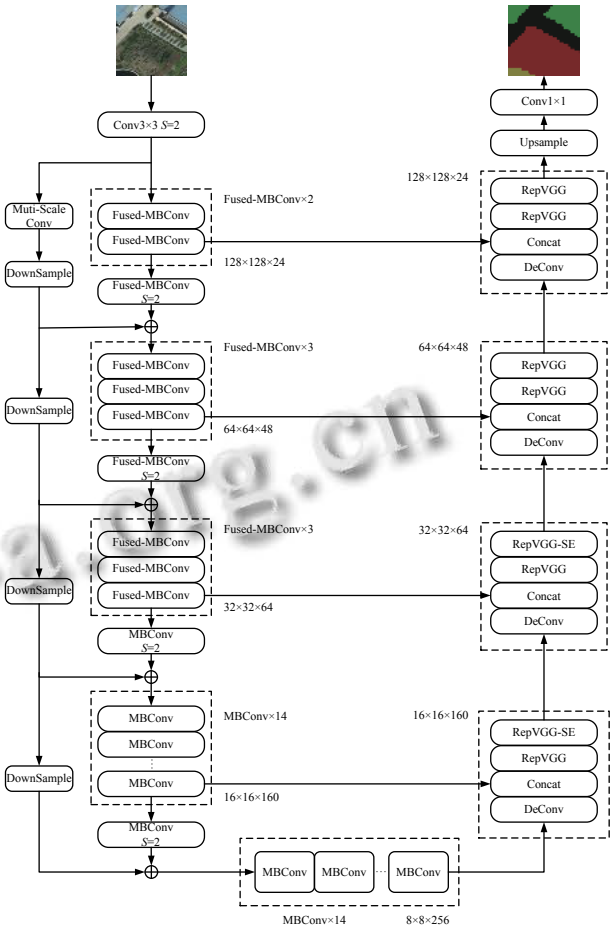


图 8 改进网络结构图

2 实验

2.1 实验环境

本实验基于 64 位 Ubuntu 20.04 系统, 使用 PyTorch

深度学习框架搭建模型, 实验代码基于 Python 3.6 实现, CPU 为 AMD Ryzen9 3900x, GPU 为一块 NVIDIA GeForce 2070S (8 GB).

2.2 数据集及预处理

本实验数据为“CCF 卫星影像的 AI 分类与识别竞赛”数据集. 数据集共 5 个类别, 为植被、建筑、水体、道路和其他. 数据集中包含人工标注的遥感图像共 5 张, 因此随机选取其中 4 张作为训练集, 1 张作为测试集. 由于图像较少, 并且每张图像尺寸各异, 单幅图像尺寸过大不能直接送入网络, 因此将遥感图像和标签图切割成 256×256 像素大小的图片, 为了缓解类别不均衡问题, 对数据中包含道路和水体的图片进行过采样, 测试集做同样处理. 为了增强网络的泛化性和鲁棒性, 采用数据增强的方式增加训练集的多样性, 具体为, 将图片翻转 90°、180°和 270°, 同时使用 PyTorch 深度学习框架的在线增强功能, 使同一批次的图像进行随机水平翻转, 到达丰富训练集数据的目的.

2.3 网络配置

2.3.1 标签平滑

标签平滑[22] 是一种损失函数的修正, 神经网络训练过程中往往对预测变得“过于确信”, 造成过拟合现象, 同时, 数据集可能存在标注错误的情况, 需要神经网络一定程度上去减少对错误答案的建模. 标签平滑技术可以提高神经网络对新数据的预测能力, 增强泛化能力, 是降低模型过拟合程度的一种正则化方法. 本实验将标签平滑用于交叉熵损失函数.

2.3.2 损失函数

图像分割一般采用交叉熵损失函数, 多分类交叉熵损失函数如式 (8) 所示:

$$loss = -\frac{1}{N}\sum_{i}\sum_{c}^{M} y_{ic}\log(p_{ic}) \tag{8}$$

其中, N表示样本总数, M表示类别数量, y_{ic}表示样本i的真实类别, 如果样本的真实类别为c则取 1, 否则取 0, p_{ic}表示样本i属于类别c的预测结果.

Dice 系数用于计算两个图像之间的相似度, 如式 (9) 所示. $|X\cap Y|$表示集合X和Y的交集, X和Y表示其元素的个数, 对于分割任务, X表示 ground truth 分割图像, Y表示预测的分割图像. Dice 损失函数如式 (10) 所示:

$$s = \frac{2|X\cap Y|}{|X|+|Y|} \tag{9}$$

$$L_{\text{Dice}} = 1 - \frac{2|X\cap Y|}{|X|+|Y|} \tag{10}$$

当使用交叉熵损失函数时, 标签分布越不平衡, 训练就越困难, 会使模型偏向占比更高的类别, Dice 损失函数能够缓解类别不均衡的情况, 但一般情况下, Dice 损失函数会对反向传播造成不利的影响, 容易使训练变得不稳定, 因此, 本实验使用带标签平滑的交叉熵损失函数和 Dice 损失函数组合的方式, 两个损失函数权重占比分别为 50%, 计算公式如下:

$$L_{\text{all}} = 0.5\times L_{\text{SCE}} + 0.5\times L_{\text{Dice}} \tag{11}$$

2.3.3 训练策略

神经网络在刚开始训练时是非常不稳定的, 模型权重会迅速改变, 当 batch size 较小时, 样本之间的方差较大, 可能会出现提前过拟合现象, 采用 warmup 训练策略, 在 batch size 较少情况下保持数据分布的稳定性和模型深层的稳定性.

训练过程中, 算法可能会陷入局部最优点, 无法得到更好的训练结果, 通过使用余弦退火策略, 使模型能够“跳出”当前的局部最优点.

训练策略为前 5 个 epoch 使用 warmup 策略, 以保持模型的稳定性, 后面使用余弦退火策略, 整个训练过程学习率如图 9 所示.

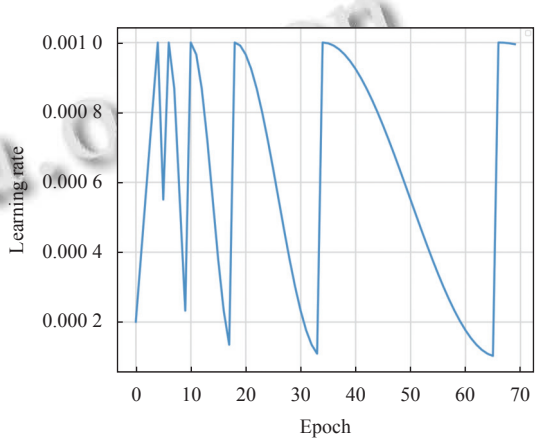


图 9 学习率变化曲线图

2.3.4 参数设置

本实验使用 AdamW 优化器[23], AdamW 优化器是 Adam 优化器[24] 的改进, 指数衰减率β_1=0.9, β_2=0.999, 权重衰减为 1E–3, 设置批次输入大小为 12, 训练 70 个轮次, 余弦退火学习率最高为 1E–3, 最低为 1E–4, 变化区间为 0.1 倍, 改进模型的编码网络中加入了预训练模

型, 这部分网络的学习率最高为 5E–4, 最低为 5E–5.

2.4 评价指标

假设数据集有$k+1$个类别, 则p_{ii}表示实际为i类且预测为i类的像素数量 (TP), p_{jj}表示实际为j类且预测为j类的像素数量 (TN), p_{ij}表示实际为i类而预测为j类的像素数量 (FP), p_{ji}表示实际为j类而预测为i类的像素数量 (FN). 通过计算像素准确率 (*PA*)、*Kappa* 系数、平均交并比 (*MIoU*)、*Macro-F*1 对各网络模型进行评估和比较. 各项指标计算公式如下:

$$PA=\frac{\sum_{i=0}^{k}p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}} \tag{12}$$

$$p_o=\frac{\sum_{i=0}^{k}p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}} \tag{13}$$

$$p_e=\frac{\sum_{i=0}^{k}\left(\sum_{j=0}^{k}p_{ij}\times\sum_{j=0}^{k}p_{ji}\right)}{\left(\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}\right)^2} \tag{14}$$

$$Kappa=\frac{p_o-p_e}{1-p_e} \tag{15}$$

$$IoU_i=\sum_{j=0,j\neq i}^{k}\frac{p_{ii}}{p_{ii}+p_{ij}+p_{ji}} \tag{16}$$

$$MIoU=\frac{1}{k+1}\sum_{i=0}^{k}IoU_i \tag{17}$$

$$precision_i=\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}} \tag{18}$$

$$recall_i=\frac{p_{ii}}{\sum_{j=0}^{k}p_{ji}} \tag{19}$$

$$F1_i=2\times\frac{precision_i\times recall_i}{precision_i+recall_i} \tag{20}$$

$$Macro\text{-}F1=\frac{1}{k+1}\sum_{i=0}^{k}F1_i \tag{21}$$

3 实验结果与分析

3.1 网络性能对比

本实验选取的对比网络如下, 3 个主流的图像分割网络, 分别为 FCN-8s、SegNet[25]、DeepLabV3+[26], 本实验的基线网络 U-Net, 以及基于 U-Net 网络的一些改进网络, 分别为 UNet++、U2-Net、Attention U-Net, 其中 DeepLabV3+训练加入了预训练模型, 由于显存大小限制, U2Net 批次输入大小为 10, 对比实验的数据集划分一致, 使用相同的训练策略, 每个网络训练 70 轮, 使用 *PA*、*Kappa*、*MIoU* 和 *Macro-F*1 作为评价指标, 对比结果如表 1 所示.

表 1 不同方法在测试集上的各项指标对比

方法	*PA* (%)	*Kappa*	*MIoU* (%)	*Macro-F*1 (%)
FCN-8s	83.59	0.7589	75.63	85.94
SegNet	82.41	0.7429	74.74	85.30
DeeplabV3+	85.81	0.7924	79.16	88.21
U-Net	86.99	0.8097	80.67	89.16
UNet++	85.31	0.7854	78.96	88.02
U2-Net	86.38	0.8004	80.50	89.02
Att U-Net	87.17	0.8123	80.88	89.29
本文	**87.56**	**0.8180**	**81.72**	**89.80**

如表 1 所示, 本文方法的 *PA* 为 87.56%, *Kappa* 系数为 0.818、*MIoU* 为 81.72%, *Macro-F*1 为 89.8%, 与主流分割网络 FCN-8s、SegNet 和 DeepLabV3+相比, 本文方法在各项指标上均高于这些网络, 在这些网络中, DeepLabV3+网络的分割性能最好. U-Net 是基于编码解码结构的网络, 并加入了大量的跳跃连接, 也是本次实验改进的基线网络, U-Net 网络分割结果 *PA* 为 86.99%, *Kappa* 系数为 0.8097, *MIoU* 为 80.67%, *Macro-F*1 为 89.16%, 同时, 实验表明 U-Net 网络分割结果优于对比的主流分割网络, 表明了 U-Net 网络相比这些分割网络更加适合遥感图像分割任务. 本文方法和 U-Net 网络相比, *PA* 高了 0.57%, *Kappa* 系数高了 0.0083, *MIoU* 高了 1.05%, *Macro-F*1 高了 0.64%, 表明了本文方法相比于原始 U-Net 网络学习能力更强, 分割性能更好. 在与 U-Net 改进的网络对比中, 本文方法在各项对比指标中也均取得最高值. 图 10 给出了本文改进的网络和进行对比的各个网络的实验分割对比结果.

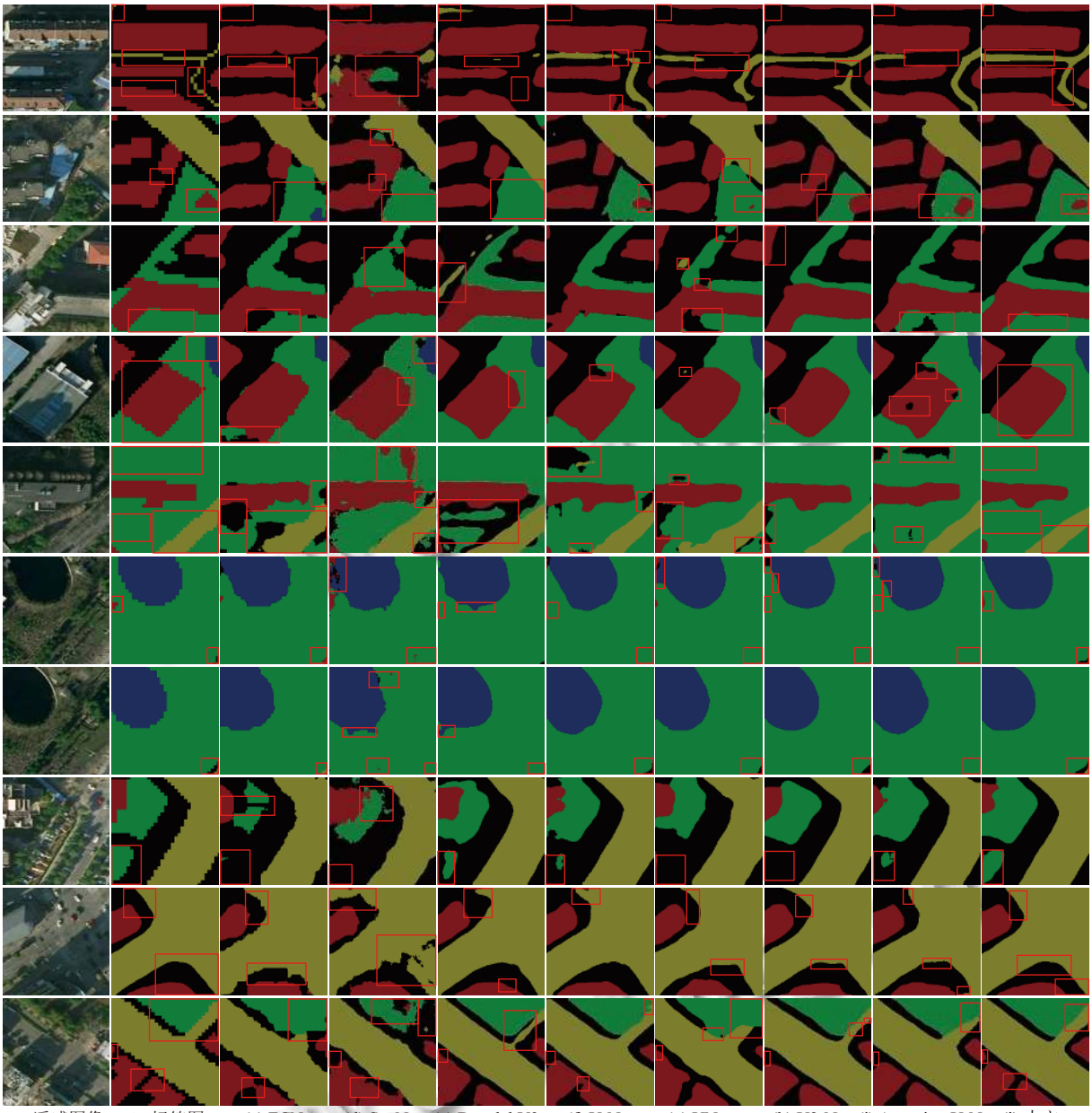


图 10 各网络模型的分割结果

图 10 中已将标签图片转换为 RGB3 通道图片, 其中植被为绿色 (0, 128, 0), 建筑为红色 (128, 0, 0), 水体为蓝色 (0, 0, 128), 道路为黄色 (128, 128, 0), 其他为黑色 (0, 0, 0). 从对比图片中可以看出, 本文方法更好地改善了各目标错分和漏分的现象, 同时对于不同尺度的目标能更好地识别并完整分割, 更加接近标签图片, 分割细节也比其他网络更加精细.

本文算法利用更好的编码网络、多尺度卷积融合模块、通道注意力机制和卷积结构重参数化的方法, 增强了网络特征提取的能力, 使网络对遥感图像中不同尺度目标的感知能力更强, 加强了有用通道信息的关注度, 因此, 本文算法在每项指标上都取得了最高的结果, 表明了本文网络模型的有效性.

3.2 网络轻量化

本实验通过 PyTorch 深度学习框架的轻量级网络分析工具 torchstat 对网络模型的参数量 (parameters, Params) 和计算量 (floating point operations, FLOPs) 进行统计, 统计结果如表 2 所示.

表 2 不同方法的参数量和计算量对比

方法	Params (M)	FLOPs (G)
FCN-8s	14.72	20.09
SegNet	29.44	40.25
DeepLabV3+	54.70	20.76
U-Net	31.04	46.14
UNet++	9.16	34.65
U2-Net	44.07	37.90
Att U-Net	31.39	47.23
本文	21.40	5.67

从表 2 可知, UNet++网络的参数量最低, 为 9.16 M, 主要原因为 UNet++网络从 32 通道开始至 512 通道, 相比于原始 U-Net 网络从 64 通道到 1 024 通道, 参数量大大减小, 参数量最高的网络为 U2-Net 网络, 为 44.07 M. 原始 U-Net 网络参数量为 31.04 M, 相较于原始 U-Net, 本实验改进的网络参数量为 21.4 M, 比原始 U-Net 网络的参数量减少了 9.64 M, 在所有对比网络中, 只比 UNet++网络和 FCN8s 网络参数量高. 计算量方面, 对比的网络中计算量普遍较高, 最高的网络为 Attention U-Net, 计算量为 47.23 G, 原始 U-Net 网络计算量为 46.14 G, 在所有网络中计算量从大到小排第二. 本实验方法的计算量最低, 为 5.67 G, 相较于原始 U-Net 网络 46.14 G 的计算量, 本文方法计算量减小了 40.47 G, 减小了接近 88% 的计算量, 同时相比于其他对比的网络, 计算量也大大减小.

本实验记录了训练过程所消耗的时间, 并对测试集进行测试, 设置批次输入大小为 12, 从加载数据后进行测试时开始计时, 所有图片全部预测后停止计时, 记为测试时间, 实验结果如表 3 所示.

表 3 不同方法训练时间和测试时间对比

方法	训练时间	测试时间 (s)
FCN-8s	5 h 55 min	6.08
SegNet	12 h 51 min	10.98
DeepLabV3+	10 h 46 min	11.56
U-Net	15 h 3 min	13.71
UNet++	17 h 8 min	15.63
U2-Net	32 h 47 min	17.83
Att U-Net	15 h 37 min	15.11
本文	6 h 49 min	6.52

从表 3 可知, FCN-8s 网络的训练时间和测试时间都为最短, 除 FCN-8s 和本文方法外, 其余网络训练时间都普遍较长, 训练时间最长的为 U2-Net 网络, 几乎是 U-Net 网络训练时常的两倍. 原始 U-Net 网络训练训练时间为 15 h 3 min, 测试时间为 13.71 s, 而本文方法训练时间为 6 h 49 min, 测试时间为 6.52 s, 相较于原始 U-Net 网络, 训练时间和测试时间都大大减少, 减少可达 50% 以上.

本文方法通过将原始 U-Net 网络的编码网络替换为 EfficientNetV2-S 网络, 同时训练后将分支结构通过结构重参数化方法融合为单路结构, 相比于分支结构进一步减少推理时的参数量和计算量. 实验结果表明, 本文方法相较于原始 U-Net 网络, 参数量、计算量、训练时间和测试时间都大大减少, 网络更加轻量化.

3.3 模型验证

为了验证本文改进的网络以及网络的各个改进模块的有效性, 进行消融实验证明. 消融实验是通过设置基线网络是否使用某一个模块来实现的, 并以 *PA*、*Kappa*、*MIoU* 和 *Macro-F*1 指标来进行衡量, 并统计网络的参数量、计算量、训练时间和测试集测试时间. 实验中, 每个网络训练参数相同, 并且都在同一环境下运行, 得到的测试结果对比如表 4 所示.

表 4 消融实验在测试集上评价结果对比

方法	*PA* (%)	*Kappa*	*MIoU* (%)	*Macro-F*1 (%)
U-Net	86.99	0.809 7	80.67	89.16
U-Net+E	87.18	0.812 8	81.10	89.43
U-Net+E+R	87.30	0.814 4	81.47	89.64
U-Net+E+M	87.35	0.815 3	81.56	89.70
U-Net+E+R+M	87.56	0.818 0	81.72	89.80

表 4 中 E 代表编码网络替换为 EfficientNetV2-S, R 表示在解码网络使用 RepVGG 和 RepVGG-SE 模块, M 表示使用多尺度卷积融合模块. 如表 4 所示, 在编码部分使用 EfficientNetV2-S 进行替换后, 相较于原始 U-Net 网络, *PA* 提高了 0.19%, *Kappa* 提高了 0.003 1, *MIoU* 提高了 0.43%, *Macro-F*1 提高了 0.27%. 在此基础上, 通过添加 RepVGG 模块和 RepVGG-SE 模块后, 比原始 U-Net 网络 *PA* 提高了 0.31%, *Kappa* 提高了 0.004 7, *MIoU* 提高了 0.8%, *Macro-F*1 提高了 0.48%. 通过添加多尺度卷积融合模块, 相比原始 U-Net 网络 *PA* 提高了 0.36%, *Kappa* 提高了 0.005 6, *MIoU* 提高了 0.89%, *Macro-F*1 提高了 0.54%. 在添加全部模块后, 即本文方法, 相比于原始 U-Net 网络, *PA* 提高了 0.57%, *Kappa* 系数提高了 0.008 3, *MIoU* 提高了 1.05%, *Macro-F*1 提高了 0.64%. 表 5 给出了消融实验各网络的参数量和计算量, 以及在数据集上训练和测试的时间.

EfficientNetV2 中大量使用了 DWConv 结构, 使得网络整体的参数量和计算量得以减少, 但 DWConv 在浅层网络运行较慢, 因此, 为了优化网络运行速度, 提

出 Fused-MBConv 模块并在浅层网络部分进行使用, 这一模块通过将 DWConv 替换成普通的 3×3 卷积, 能够明显提升网络速度. 在编码部分, 由于编码网络替换为 EfficientNetV2-S 后, 解码网络的通道数从 256 开始至最后的 24 通道, 相比于原始 U-Net 网络通道数从 1024 开始, 通道数减小很多, 这也是参数量和计算量大大减少的原因, 同时, 训练后通过将分支结构进行融合成一个普通 3×3 卷积, 能够进一步在分支结构的基础上减小推理时的参数量和计算量, 并充分利用底层对 3×3 卷积优化加速的优势, 以更进一步提升网络推理时的运行速度. 多尺度特征融合模块因为卷积核变大, 因此增加了一定的参数量和计算量. 在编码网路下层的 RepVGG-SE 模块中, 由于 identity 分支中使用了通道注意力机制, 该条分支没有和其他分支进行融合合并, 因此增加了一定的参数量和计算量. 表 6 给出了本文方法的分支结构融合前和融合后的对比结果.

表 5 消融实验在网络轻量化方面结果对比

方法	Params (M)	FLOPs (G)	训练时间	测试时间 (s)
U-Net	31.04	46.14	15 h 3 min	13.71
U-Net+E	20.88	4.51	5 h 3 min	5.48
U-Net+E+R	20.89	4.51	5 h 21 min	5.50
U-Net+E+M	21.40	5.68	6 h 33 min	6.63
U-Net+E+R+M	21.40	5.67	6 h 49 min	6.52

表 6 分支结构融合前和融合后对比

融合前后	*PA* (%)	*Kappa*	*MIoU* (%)	*Macro-F*1 (%)	Params (M)	FLOPs (G)
融合前	87.56	0.8180	81.72	89.80	21.50	5.77
融合后	87.56	0.8180	81.72	89.80	21.40	5.76

由表 6 可知, 在分支结构融合前和融合后, 每项指标的结果都一样, 表明了分支结构融合的可行性和有效性, 同时, 参数量和计算量相比未融合前有了减少, 提高了网络推理时的效率, 使模型更加的轻量化. 本文方法通道数最多为 256, 因此融合后参数量和计算量减少相对较少, 如果网络层数更深并且特征图通道数更多, 分支结构融合后减少的参数量和计算量更大.

3.4 高分辨率遥感图像分割验证

遥感图像通常为高分辨率图像, 图像尺寸较大, 由于硬件设备的限制, 无法直接对高分辨率遥感图像进行分割, 因此将高分辨率图像切割成 256×256 大小的图片, 再送入网络进行分割得到分割结果, 最后进行拼接得到最终的高分辨率图像分割结果.

切割和拼接过程中, 如果直接按照普通滑动窗口的方式以 256 为步长进行切割, 可能会破坏目标的完整性, 导致分割结果不够准确和精细, 并且直接拼接后会存在两张图片拼接处细节不足等问题, 因此采用以 256 为步长, 重叠区域为 128 对高分辨率遥感图像进行滑动窗口切割, 切割到图片边缘时不足 256 尺寸则以边缘为起点后退 256 步长切割边缘图片, 横向和纵向都采用此方式进行切割; 拼接复原时, 以两张图片重叠区域的一半为界, 舍去各自的边缘部分后, 再进行拼接, 横向和纵向都采用此方式, 这样就很大程度上避免了上述存在的问题. 实验采用的模型为第 3.1 节所训练的模型, 对比结果如表 7 和图 11 所示.

表 7 不同方法在高分辨率图像分割的各项指标对比

方法	*PA* (%)	*Kappa*	*MIoU* (%)	*Macro-F*1 (%)
FCN-8s	83.40	0.7310	73.44	84.52
SegNet	82.50	0.7183	70.05	82.28
DeeplabV3+	85.81	0.7721	76.21	86.46
U-Net	86.72	0.7858	78.36	87.81
UNet++	84.53	0.7496	74.90	85.56
U2-Net	86.05	0.7756	78.25	87.73
Att U-Net	86.99	0.7904	79.39	88.45
本文	**87.55**	**0.8002**	**80.96**	**89.41**

由表 7 可知, 本文方法的 *PA* 为 87.55%, *Kappa* 为 0.8002, *MIoU* 为 80.96%, *Macro-F*1 为 89.41%, 在所有对比网络中各项指标均取得了最高值. 相较于原始 U-Net, *PA* 提高了 0.83%, *Kappa* 提高了 0.0144, *MIoU* 提高了 2.6%, *Macro-F*1 提高了 1.6%.

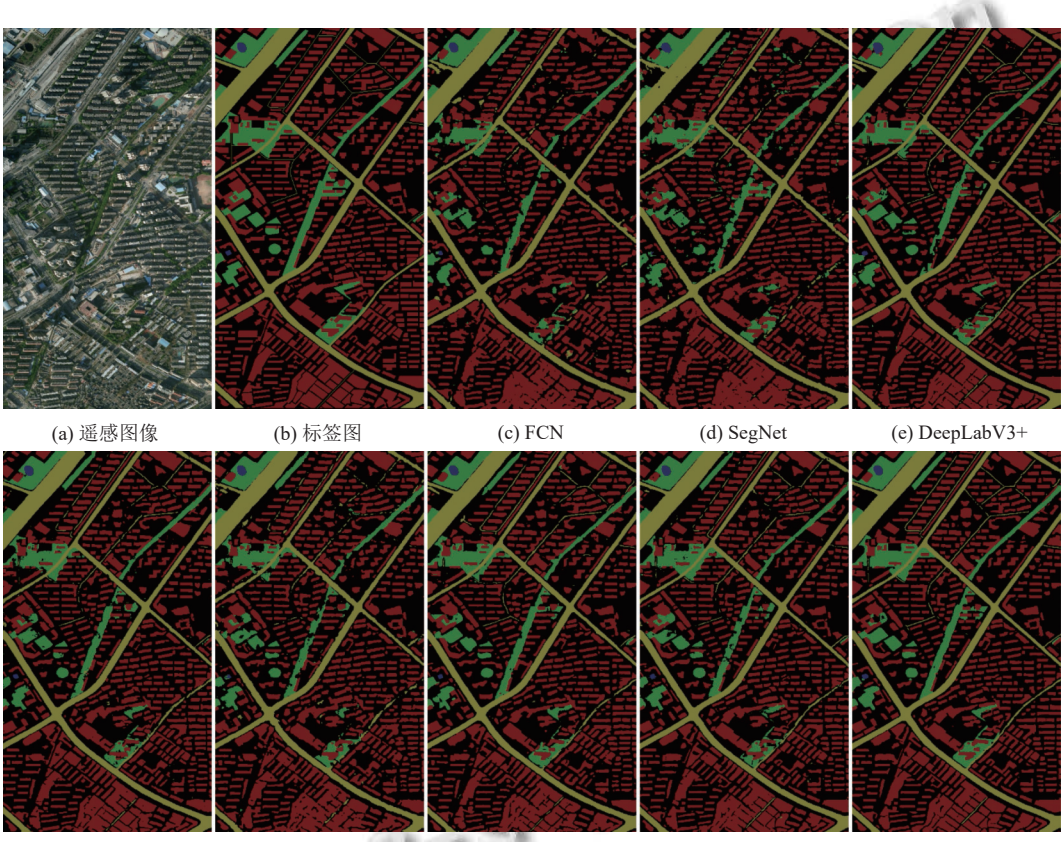
从图 11 中也可以看出, 本文方法在大场景环境下分割结果更加准确, 目标分割更加完整, 分割结果更加接近真实标签图片.

如表 8 所示, 在使用本文改进模型进行高分辨率遥感图像分割时, 通过使用改进的切割和拼接方法后, 各项指标均得到大幅提升, 表明本文提出的方法对于大场景高分辨率图像的分割具有先进性.

4 结论语

本文基于 U-Net 提出了一种用于遥感图像分割的改进网络, 该网络使用 EfficientNetV2-S 作为编码网络, 提高特征提取能力并且更加轻量化, 使用卷积结构重参数化方法, 训练时多分支结构, 增强网络训练性能, 推理时将多分支结构融合为单路结构, 提高推理效率,

并在支路使用通道注意力机制, 强化有用特征, 抑制无效信息, 在网络编码部分结合多尺度卷积模块融合多尺度特征, 使用下采样模块将浅层信息传递到深层网络, 增强网络对不同尺度目标特征的提取能力, 也更好地结合上下文信息. 实验中, 通过裁剪和过采样高分辨率遥感图像生成实验数据, 使用 warmup 和余弦退火组合的训练策略, 使用标签平滑的交叉熵损失函数和 Dice 损失函数的组合缓减过拟合和类别不均衡的问题, 实验结果表明, 本文改进网络在所有对比的网络中各项指标均为最高, 并且通过消融实验验证了各模块对于网络分割性能的有效性, 改进的网络相较于原始 U-Net 参数量和计算量都大大减少, 网络性能更高的同时更加轻量化, 在大场景高分辨率图像中也取得了最好的效果. 本文方法整体改善了各目标错分和漏分的现象, 加强了各尺度目标的完整识别, 提高了遥感图像分割性能, 证明了本文方法的有效性.



(a) 遥感图像 (b) 标签图 (c) FCN (d) SegNet (e) DeepLabV3+

(f) U-Net (g) UNet++ (h) U2-Net (i) Attention U-Net (j) 本文

图 11 各网络模型高分辨率遥感图像分割结果

表 8 不同切割拼接方法的各项指标对比

方法	*PA* (%)	*Kappa*	*MIoU* (%)	*Macro-F*1 (%)
普通方法	84.84	0.755 1	75.74	86.12
本文	87.55	0.800 2	80.96	89.41

参考文献

1 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述. 计算机学报, 2017, 40(6): 1229–1251. [doi: 10.11897/SP.J.1016.2017.01229]

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

3 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.

4 Zhou ZW, Siddiquee MR, Tajbakhsh N, *et al*. Unet++: A nested U-Net architecture for medical image segmentation. Proceedings of the 4th International Workshop on Deep Learning in Medical Image Analysis and Multimodal

Learning for Clinical Decision Support. Granada: Springer, 2018. 3–11.
5 Oktay O, Schlemper J, Le Folgoc L, *et al*. Attention U-Net: Learning where to look for the pancreas. arXiv:1804.03999, 2018.
6 Qin XB, Zhang ZC, Huang CY, *et al*. U^2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition, 2020, 106: 107404. [doi: 10.1016/j.patcog.2020.107404]
7 苏健民, 杨岚心, 景维鹏. 基于 U-Net 的高分辨率遥感图像语义分割方法. 计算机工程与应用, 2019, 55(7): 207–213. [doi: 10.3778/j.issn.1002-8331.1806-0024]
8 范自柱, 王松, 张泓, 等. 基于 W-Net 的高分辨率遥感卫星图像分割. 华南理工大学学报 (自然科学版), 2020, 48(12): 114–124.
9 Cao Y, Xu JR, Lin S, *et al*. GCNet: Non-local networks meet squeeze-excitation networks and beyond. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul: IEEE, 2019. 1971–1980.
10 王曦, 于鸣, 任洪娥. UNET 与 FPN 相结合的遥感图像语义分割. 液晶与显示, 2021, 36(3): 475–483. [doi: 10.37188/CJLCD.2020-0116]
11 Zhu Y, Sapra K, Reda FA, *et al*. Improving semantic segmentation via video propagation and label relaxation. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 8848–8857.
12 Tan MX, Le QV. Efficientnetv2: Smaller models and faster training. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 10096–10106.
13 Ding XH, Zhang XY, Ma NN, *et al*. RepVGG: Making VGG-style convNets great again. Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13728–13737.
14 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.
15 Tan MX, Le QV. Efficientnet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 6105–6114.
16 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.
17 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.
18 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1–9.
19 Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269.
20 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille: JMLR.org, 2015. 448–456.
21 Goyal A, Bochkovskiy A, Deng J, *et al*. Non-deep networks. arXiv:2110.07641, 2021.
22 Müller R, Kornblith S, Hinton G. When does label smoothing help? Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 422.
23 Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv:1711.05101, 2018.
24 Kingma DP, Ba J. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations. San Diego: 2014.
25 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]
26 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.

(校对责编: 牛欣悦)