第 58 卷 第 24 期/2021 年 12 月/激光与光电子学进展 研究论文

激光与光电子学进展

先进成像

基于尺度自适应全卷积网络的遥感影像建筑物提取

冯凡[1]，王双亭[1]，张津[1]，王春阳[1*]，刘冰[2]

[1]河南理工大学测绘与国土信息工程学院，河南 焦作 454000；

[2]中国人民解放军战略支援部队信息工程大学，河南 郑州 450001

摘要 为了提高网络对高空间分辨率遥感影像多尺度建筑物的提取效果，在编码-解码网络的基础上提出了一种高效的尺度自适应全卷积网络。首先，构建多输入多输出结构，实现多尺度特征融合和跨尺度特征聚合。然后，用残差金字塔池化模块学习深层自适应多尺度特征。最后，用基于残差密集连接的聚合特征精化模块进一步处理初始聚合特征，利用不同尺度特征图的像素依赖关系提升分类精度。在差异较大的 WHU 航空数据集和 Massachusetts 数据集上的实验结果表明，相比其他方法，本方法对建筑物的提取效果较好，且训练时间和内存占用情况适中，具有较高的实用价值。

关键词 遥感；图像处理；建筑物提取；全卷积网络；残差金字塔池化；聚合特征精化

中图分类号 TP751.2 **文献标志码** A **doi**：10.3788/LOP202158.2428006

Building Extraction from Remote Sensing Imagery Based on Scale-Adaptive Fully Convolutional Network

Feng Fan[1], Wang Shuangting[1], Zhang Jin[1], Wang Chunyang[1*], Liu Bing[2]

[1] *School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo, Henan* 454000, *China*;

[2] *PLA Strategic Support Force Information Engineering University, Zhengzhou, Henan* 450001, *China*

Abstract The present research proposes an efficient scale-adaptive and fully convolutional network based on an encoder-decoder network, which represents a crucial innovation aimed at improving buildings extraction with various scales from remote sensing imagery with high spatial resolution. First, a multiple-input multiple-output structure is proposed to obtain multiscale features fusion and cross-scale aggregation. Then, a residual pyramid pooling module is deployed to learn deep adaptive multiscale features. Finally, the initial aggregated features are further processed using a residual dense-connected aggregated-feature refinement module. Pixel dependencies of different feature maps are investigated to improve the classification accuracy. Experimental results on the WHU aviation and the Massachusetts datasets show that compared with other methods, the method has a better extraction effect on buildings, and the training time and memory usage are moderate, which has high practical value.

Key words remote sensing; image processing; building extraction; fully convolutional networks; residual spatial pyramid pooling; aggregation feature refinement

OCIS codes 280.4788; 100.2000; 100.4996

收稿日期：2021-01-06；**修回日期**：2021-01-12；**录用日期**：2021-01-20

基金项目：河南省自然基金(182300410111)、河南省高校重点研发项目基金(18A420001)、河南理工大学博士基金(B2016-13)

通信作者：*wcy@hpu.edu.cn

1 引　言

遥感传感器技术的快速发展,推进了大规模高空间分辨率(下文简称为高分辨率)遥感影像的获取和应用。从高分辨率遥感影像中自动提取建筑物是土地利用调查、城市规划、环境监测等重要遥感技术应用的基石,也是遥感领域中长期研究的热点[1]。作为像素级图像分类问题,建筑物提取的核心是特征提取。近年来,以卷积神经网络(CNN)为代表的深度学习方法为遥感图像分析的特征提取提供了端到端的解决方案[2]。其中,网络的静态结构在一定程度上决定了特征提取的模式。但高分辨率遥感影像的复杂语义特征以及建筑物表现出的多尺度特性,导致有针对性的网络设计依然具有一定的挑战性[3]。

高分辨率遥感影像建筑物提取是像素级分类任务,早期研究大多基于建筑物在遥感影像中的特点设计特征,包括光谱特性、颜色、纹理、高度、阴影等[4]。多特征综合和复杂机器学习分类器是建筑物提取领域的重要研究方向,常用的分类器包括支持向量机(SVM)、随机森林(RF)、条件随机场(CRF)等。但人工设计特征的方法一方面费时费力,另一方面缺乏对不同遥感影像的泛化能力[5]。高分辨率遥感影像具有丰富的空间信息,但包含的波段数较少。为了提取具有足够判别力和健壮性的抽象特征,近年来,人们大多通过学习方式从图像数据中自动提取深层次特征。常用于遥感图像分类的 CNN 可分为基于图块的 CNN(Patch-based CNN)和全卷积神经网络(FCNN)[6-8]。基于图块的 CNN 可以有效学习待分类像素及其邻域的空谱联合特征,在高光谱分类领域得到了广泛应用[9]。但该网络存在大量重复计算,限制了其在大规模高分辨率遥感影像任务中的应用。而训练好的 FCNN 能通过一次前向传播分类输入图像的所有像素,相比基于图块的 CNN 效率更优[10]。因此,FCNN 被广泛用于大规模高分辨率遥感影像建筑物提取任务中[11]。CNN 对图像特征的学习通过最优化网络中的各层卷积核实现,其中,网络的静态结构决定了特征学习的模式,而具体的特征提取结果由数据决定,从而表现出一定的健壮性。残差网络(ResNet)和密集网络(DenseNet)中使用的特征融合方式,即特征图相加和特征图连接方式,对 CNN 的优化研究产生了深远影响。基于 CNN 的建筑物特征提取,面临全局语义特征和局部细节特征不能兼顾的基本矛盾[12]。随着卷积的进行,特征抽象程度不断提高,感受野也不断增大,不可避免地会导致空间细节信息的损失。用于建筑物提取的 FCNN 大多使用编码-解码结构,不仅具有逐级解码的特点,还具有恢复空间信息的能力。U-Net 通过融合编码段和对应解码段的特征图有效恢复空间信息,在建筑物提取任务中表现出较大潜力[13]。此外,高分辨率图像中的建筑物具有多尺度特性,遥感图像垂直成像的特点使其语义特征相当复杂,且图像中有大量与建筑物屋顶颜色、纹理相似的地物。为了提高对复杂地物的分析能力,人们基于编码-解码结构进一步探索了特征融合方式。季顺平等[14]构建了 S-UNet (Scale robust U-Net),通过连接降维和上采样后不同解码段的输入,实现了跨尺度特征聚合,增强了网络对多尺度建筑物的分析能力。崔卫红等[15]提出的 VPU (VGG16 pyramid upsample)通过对原始影像的下采样和对不同编码段的特征融合增强了深度特征中建筑物的细节信息,提升了建筑物的边缘提取效果。田青林等[16]使用注意力模块在特征融合前对浅层特征进行了精化,同时通过密集连接解码段特征金字塔,有效提升了网络对建筑物的检测精度。Liu 等[17]在 U-Net 编码段和解码段中间引入了空间金字塔池化(SPP),构建了 USPP 网络,在深度特征学习中增强了对局部特征和全局特征的学习。Shao 等[18]引入了残差空洞卷积模块,可同时增大网络的感受野和提升网络的初始分类结果。Zhang 等[19]在每个编码段和解码段引入残差结构,构建了 Res-UNet,增强了网络中的信息传递。Ibtehaz 等[20]提出的 MultiResUNet 极大提升了医学图像的分割效果,通过大量使用 MultiResBlock 实现了浅层特征的重用,同时也增强了信息传递。Pleiss 等[21]的研究表明,现有深度学习框架中频繁的特征图连接操作会造成过大的显存占用,从而限制算法的实用性。因此,如何针对高分辨率遥感影像的特点优化网络结构,有效学习多尺度特征,还需进一步的深入研究。

为了解决高分辨率遥感影像中建筑物提取困难的问题,本文提出了一种尺度自适应网络(SA-Net)。在多输入多输出(MIMO)网络实现多特征融合和跨尺度特征聚合的基础上,进一步融合了残差空间金字塔池化(RSPP)和聚合特征精化(AFR)模块。RSPP 通过残差连接改进金字塔池化操作,以实现自适应多尺度特征图的学习。AFR 对初始聚合特征存在的问题进行优化,提升了网络对不同

类型遥感影像建筑物的提取效果。最后，利用有限的计算资源，在差异较大的高分辨率遥感影像 WHU 航空数据集和 Massachusetts 数据集上进行了建筑物提取实验，验证了本方法的有效性。

2 研究方法

2.1 基于 SA-Net 的建筑物提取

为了提升高分辨率遥感影像复杂场景中多尺度建筑物的提取效果，设计了一系列尺度自适应特征学习模式并提出了 SA-Net。受计算资源的限制，SA-Net 接受切分后的小片影像作为输入，并用线性整流单元(ReLU)作为激活函数。ReLU 会抑制神经元的负值，保证每层的输入非负。输出层包含 2 个通道，用 Sigmoid 函数作为激活函数。Sigmoid 函数将 SA-Net 学到的特征映射到 0～1 范围内，最终完成对输入图像的像素级分类。Sigmoid 函数可表示为

$$X_{\text{Sigmoid}}(P_{x,y})=\frac{1}{1+\exp(-P_{x,y})},\tag{1}$$

式中，$P_{x,y}$ 为特征图中(x,y)位置的像素值。

基于 SA-Net 的建筑物提取流程如图 1 所示。典型监督分类任务的主要步骤有数据预处理、训练和测试。在数据预处理中，传统的数据扩增(Data augmentation)手段，如旋转、镜像和随机位置采样(Random position sampling)策略结合使用。此外，采用随机切片策略(Random cropping strategy)进一步增强了样本的随机性[12]。预测过程中，为了削弱可能存在的边缘不准确问题，根据数据情况探索并有选择地使用了重叠拼接策略(Overlap strategy)。静态的网络结构可被看作特征学习模式，具体的图像特征由数据驱动求解最优化问题获得。

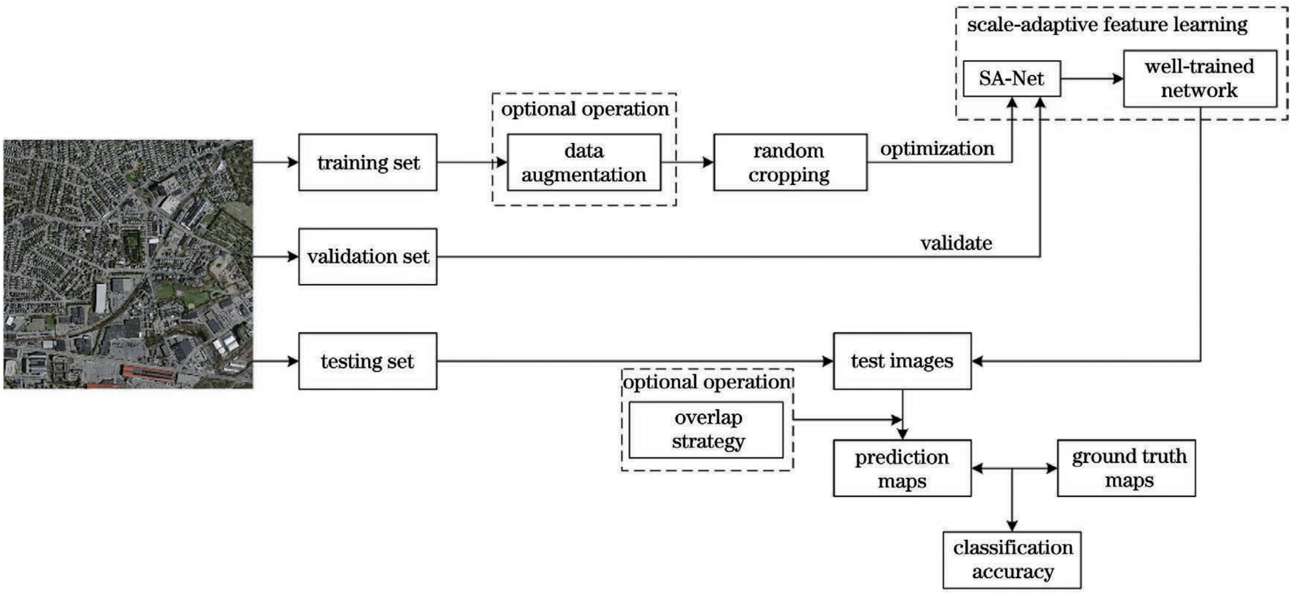

图 1 基于 SA-Net 的建筑物提取流程图

Fig. 1 Flow chart of building extraction based on SA-Net

SA-Net 的结构示意图如图 2 所示。该网络用批归一化(BN)层加快训练，同时控制过拟合。为了降低显存的占用，SA-Net 没有在所有卷积(Conv)后都使用 BN 层。SA-Net 在 U-Net 的基础上，融合了 MIMO、RSPP 和 AFR 等设计。为了增强编码段对多尺度影像的特征提取，将原始影像进行 1/4、1/16 和 1/64 下采样后经过 2 层卷积处理并与对应编码段的特征图进行融合。为了对 U-Net 编码段的最大池化进行补充，在多尺度输入分支中使用了平均池化下采样，通过连接(Concatenation)的方式进行特征融合，以保留不同池化层提取的特征。为了综合利用 4 个解码段的特征，首先，用 1×1 卷积对不同解码段的输出进行降维。然后，用插值操作将调整通道数后的特征图恢复到原始图像的尺寸。最后，将 4 组特征图连接起来，作为后续分析的依据。MIMO 可实现多尺度特征融合和跨尺度特征聚合，从而优化网络的层间关系。在 MIMO 优化 U-Net 的基础上，RSPP 进一步学习了深度多尺度自适应特征图，AFR 则针对多尺度输出部分初始聚合特征存在的问题进行优化。

2.2 卷积原理及卷积层间关系的优化

对于高分辨率光学遥感影像，常用二维卷积网

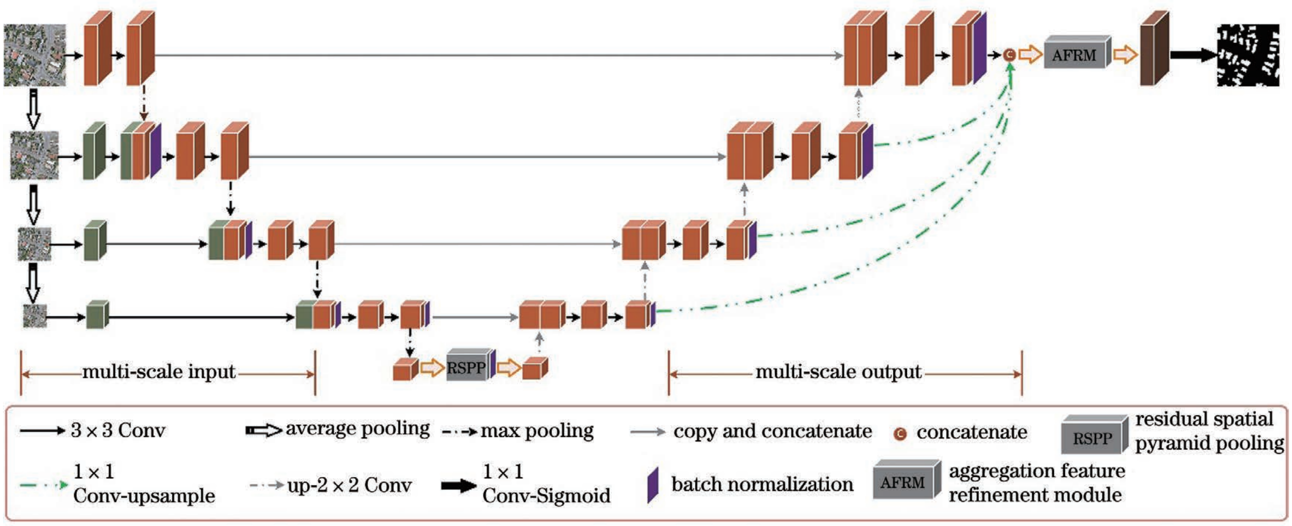

图 2 SA-Net 的结构

Fig. 2 Schematic diagram of the SA-Net

络进行特征提取。卷积层是卷积网络的核心,运算单元是每层包含的若干个卷积核。卷积核通过内积的方式遍历输入数据的每一个通道,最终输出特征图。第 i 层、第 j 个特征图对应位置的值可表示为

$$X_{i,j}^{x,y}=f\left[\sum_{m}\sum_{h=0}^{H_i-1}\sum_{w=0}^{W_i-1}k_{i,j,m}^{h,w}X_{(i-1),m}^{(x+h),(y+w)}+b_{i,j}\right], \quad (2)$$

式中,$k_{i,j,m}^{h,w}$ 为第 i 层、第 j 个卷积核(h,w)位置的值,该卷积核对前一层第 m 个特征图进行卷积,H_i 和 W_i 为卷积核尺寸,$X_{(i-1),m}^{(x+h),(y+w)}$ 为前一层第 m 个特征图$(x+h,y+w)$的值,$X_{i,j}^{x,y}$ 为运算结果,$b_{i,j}$ 为偏置项,$f(\ )$为激活函数。

连续卷积层的数量越多,提取的特征抽象程度就越高,但卷积层间的关系对特征提取效果有很大影响。相比不同层的线性堆叠,类有向无环图网络具有更强的特征提取能力[22]。类图网络指网络中某一层有多个输入或多个输出,主要通过残差连接进行特征融合实现,特征融合的方式有逐像素相加(Pixel-wise addition)和连接两种。设 CNN 第 k 层的输出为 X_k,若使用逐像素相加的方式融合第 j 层$(0<j<k)$输出的特征,则 X_k 可表示为

$$X_k=f_k(X_{k-1})+g_k(X_j), \quad (3)$$

式中,$f_k(\ )$为该层对输入的非线性变换函数,$g_k(\ )$为残差连接中包含对输入处理的函数。若第 k 层的输入为前 $l(l\leqslant k)$层特征图连接的结果,则第 k 层的输出可表示为

$$X_k=f_k([X_{k-1},\cdots,X_{k-l}]), \quad (4)$$

式中,$[\cdot]$为特征图的连接操作。

相比逐像素相加,特征图连接可以保留不同层提取特征的数值结果,为特征重用提供可能。因此,SA-Net 中 MIMO 用连接操作融合最大池化和平均池化提取的特征。残差连接能将网络浅层的特征传递到深层,有效加强了网络中信息的传递。文献[23]的研究结果表明,包含残差连接的网络等价于若干子网络的集成,而带有非线性变换函数的残差连接本身就具有提取特征的能力。因此,基于高分辨率遥感图像特点,在编码-解码网络中研究了卷积层间关系的优化。

2.3 残差金字塔池化

深层 CNN 的实际感受野远远小于理论感受野,导致网络对不同尺度的目标分析能力不足[24]。Zhao 等[25]提出了 SPP 并构建了 PSPNet,对深层 CNN 提取的特征进行了进一步处理,提高了网络的最终分类效果。首先,用不同比例的平均操作获取特征图的局部信息和全局信息。然后,将各个分支的局部特征和全局特征经过 1 层卷积降维后上采样到原始图像的尺寸。最后,将输入特征图和处理后的各个分支特征串联起来,作为 SPP 模块的输出。

SPP 模块位于 PSPNet 末端,当 SPP 模块被置于 U-Net 编码段和解码段之间时,一方面,SPP 模块每个分支仅包含 1 个卷积层,特征提取能力有限。输入特征图直接与输出串联,导致最终的输出包含未被进一步处理的特征。另一方面,SPP 的 4 个分支中卷积层卷积核数量相等,并将池化后的特征图直接连接,即每个分支的通道数相等,这种预先定义好的特征组成不利于遥感图像的尺度自适应性。因此,提出了一种 RSPP 模块,其结构如图 3 所示。RSPP 包含 4 个分支,每个分支包含 2 条路径。每

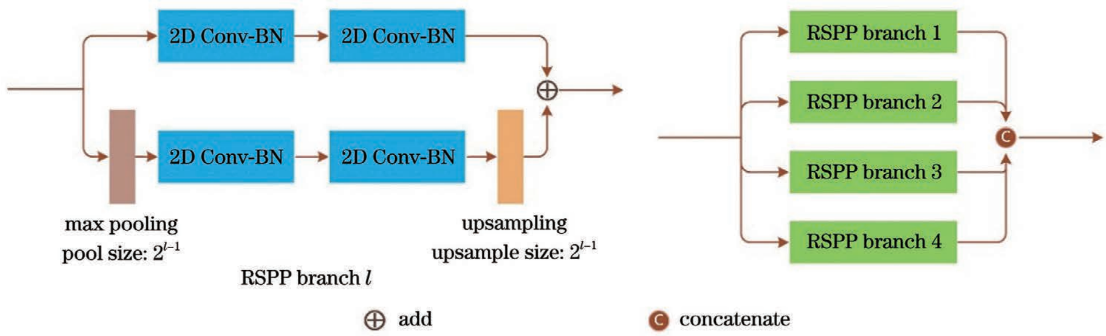

图 3　RSPP 的结构

Fig. 3　Schematic diagram of the RSPP

条路径包含 2 个卷积层。特征图采样仅在每个分支的 1 条路径中采用，2 条路径学习的特征通过逐像素相加进行融合。将 4 个分支的输出相连接，作为 RSPP 最终的输出。RSPP 第 l 个分支的计算过程可表示为

$$X_l^{\mathrm{RSPP}} = H_1(X) + X_l^{\mathrm{Up}}\{H_2[X_l^{\mathrm{Max}}(X)]\}，\tag{5}$$

式中，X 为输入特征图（本模型中为第 4 个编码段的输出），$H_l(\ )$为分支中卷积层对输入的非线性变换，X_l^{Up} 和 X_l^{Max} 为对特征图的上采样和最大池化操作，其对维度的变化比例为 2^l。RSPP 最终的输出可表示为

$$X^{\mathrm{RSPP}} = [X_1^{\mathrm{RSPP}},\cdots,X_l^{\mathrm{RSPP}}]。\tag{6}$$

RSPP 每个分支中的任意 1 条路径，均可看作是另 1 条路径的非恒等残差连接，具有独立提取特征的能力[9]。因此，RSPP 模块每个分支的 2 条路径在网络训练过程中会根据数据调节每条路径的权重，起到集成提取特征的效果。由于其中 1 条路径会对特征图进行缩放，不同比例的缩放对不同尺度建筑物具有一定的自适应能力。

2.4　聚合特征精化

对于多尺度输出的初始聚合特征，由于不同尺度的特征直接通过上采样达到原始图像的尺寸，虽然空间尺寸相同，但特征的精细程度存在较大差别。不同分支的特征图中固定大小区域内的像素，有些是通过学习得到，有些是通过上采样插值得到，直接用初始聚合特征进行分类时缺乏对跨尺度特征的利用。因此，提出了一种 AFR 模块，其结构如图 4 所示。首先，用密集连接的卷积层进行特征精化。其中，密集连接可以重用每一层学到的像素依赖关系。然后，用全局恒等残差连接融合初始聚合特征和处理后的特征，达到补偿空间信息、增强网络中信息传递的目的。设 AFR 包含 m 个卷积层，则第 m 层的输出 X_m 可表示为

$$X_m = A_m([X_{k-1},\cdots,X_1]) + X_0，\tag{7}$$

式中，$A_m()$为第 m 层卷积层的非线性变换函数，X_0 为初始聚合特征。由于每个卷积层对不同通道使用的参数不同，从而实现对精细程度差异较大的初始聚合特征的精化。而文献[16]中用于特征精化的空洞卷积效果并不理想，原因可能是密集连接的卷积层已经覆盖了像素精细程度的差异，不需要空洞卷积额外提升感受野。此外，设置空洞率需要大量的超参数调优，不利于建筑物的提取。

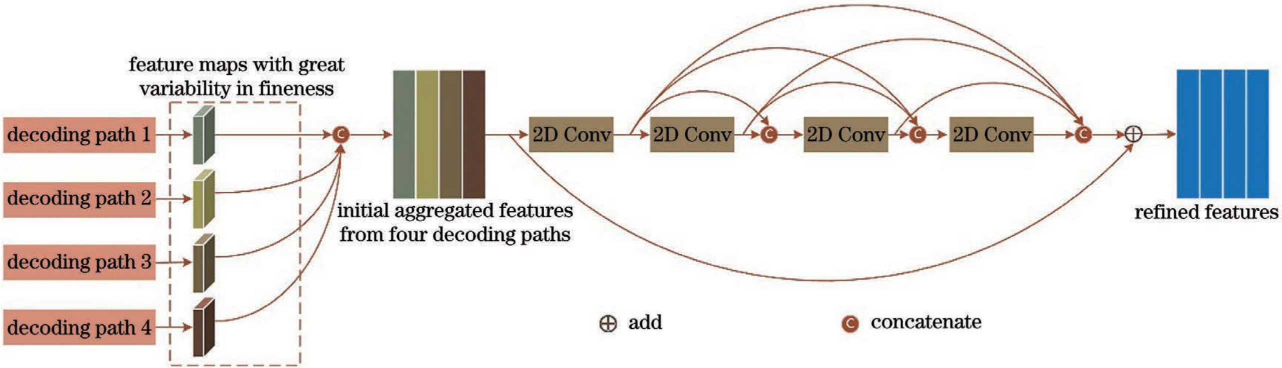

图 4　AFR 模块的结构

Fig. 4　Schematic diagram of the AFR module

3　实验设计

3.1　实验数据

为了充分验证本网络的有效性，用空间分辨率不同且建筑物特征差异较大的 WHU 航空数据集、Massachusetts 建筑物数据集进行实验。两个数据集中的两组影像实例如图 5 所示。可以发现，两个数据集中的建筑物均有多尺度特性，但 WHU 数据集更明显，一些建筑物甚至能占尺寸为 512 pixel×512 pixel 影像的一半以上。WHU 数据集中的建筑物尺寸较大且边缘清晰，而 Massachusetts 数据集中的建筑物较小，难以区分边缘和背景。

图 5　两组实例影像。(a)WHU 数据集；(b)Massachusetts 数据集

Fig. 5　Two sets of example images. (a) WHU dataset; (b) Massachusetts dataset

1) WHU 航空数据集是 WHU 建筑物数据集的子数据集，WHU 建筑物数据集由 Ji 等[26]制作并开源，为建筑物提取研究提供了一套范围大、精度高且覆盖多种数据源的数据集。其中，航空影像采集自新西兰，空间分辨率为 0.3 m，覆盖面积为 450 km^2。数据集中的图像已被切分成尺寸为 512 pixel×512 pixel 的子图，且所有子图均被划分为训练集、验证集、测试集，分别包括 4736 张、1036 张、2416 张图像。

2) Massachusetts 建筑物数据集由 Minh[27]建立，覆盖的地表范围约为 340 km^2，影像的空间分辨率 1 m。Massachusetts 数据集预先被划分为训练集、验证集和测试集，分别包括 137 张、4 张、10 张图像，每张影像的尺寸均为 1500 pixel×1500 pixel。相比 WHU 航空数据集，Massachusetts 影像的空间分辨率较低，建筑物占整张影像的比例较小，且标签存在一定错误，因此对分类网络提出了较高的要求。

3.2　对比模型

实验中的模型基于 Keras 2.2.4 实现，后端使用 Tensorflow。实验运行在单张 RTX2070 显卡上，显存为 8 G。用 U-Net 及其若干改进版本作为对比模型，包括 S-UNet[14]、USPP[15]、Res-UNet[17]和 MultiResUNet[18]。为了控制模型参数，U-Net 及其相关模型第一个编码段的卷积核数目均设定为 32。为了对比每个模型的显存占用，在单张输入影像尺寸为 256 pixel×256 pixel 的条件下，测试了每个模型可用的最大批次数。各个模型的参数量和最大批次数如表 1 所示。可以发现，SA-Net 的参数量和显存占用适中；U-Net 的显存占用最低，原因是其结构简洁，且未使用 BN 层；MultiResUNet 的显存占用极大，原因是该模型中的特征图连接操作比较频繁。

表 1　不同模型的参数量和最大批次数

Table 1　Number of parameters and the maximum number of batches of different models

Model	U-Net	USPP	S-UNet	Res-UNet	SA-Net	MultiResUNet
Parameter number /10^6	7.76	4.82	7.97	4.73	7.13	7.26
Max batch size	37	22	23	19	25	6

3.3　评价指标

为了评估本模型在实验数据集上对建筑物的提取效果，用四种常用的评价指标，即查准率(Precision)、查全率(Recall)、交并比(IOU)和 F1 分数(F1 score)对提取效果进行评估，可表示为

$$X_{\mathrm{Precision}}=\frac{X_{\mathrm{TP}}}{X_{\mathrm{TP}}+X_{\mathrm{FP}}},\tag{8}$$

$$X_{\mathrm{Recall}}=\frac{X_{\mathrm{TP}}}{X_{\mathrm{TP}}+X_{\mathrm{FN}}},\tag{9}$$

$$X_{\mathrm{IOU}}=\frac{X_{\mathrm{TP}}}{X_{\mathrm{TP}}+X_{\mathrm{FP}}+X_{\mathrm{FN}}},\tag{10}$$

$$F_1=\frac{2\times X_{\text{Precision}}\times X_{\text{Recall}}}{X_{\text{Precision}}+X_{\text{Recall}}},\tag{11}$$

式中，X_{TP} 为所有图像中被正确分类为建筑物的像素数，X_{FP} 为被错误分类为建筑物的像素数，X_{FN} 为被错误分类为背景的像素数。

3.4 训练策略和实验设置

由于 WHU 航空影像已经被切分成尺寸为 512 pixel×512 pixel 的子图，因此，对 Massachusetts 数据集中的影像进行了相同的处理。Massachusetts 数据集初步处理后仅包含 1233 张影像，因此，用随机旋转、翻转和镜像的手段扩增数据。在生成样本的过程中，采用随机位置采样策略，保证样本之间有一定的重叠和随机性。最终，Massachusetts 数据集的训练样本数量为 8631。受显存限制，直接用尺寸为 512 pixel×512 pixel 的图像进行训练只能选择较小的批次数(Batch size)，且训练样本缺乏足够的随机性。为了解决上述问题，用随机采样子图的方式进行训练。设数据集的训练样本数量为 P，训练批次数为 B，每次随机从 B 张尺寸为 512 pixel×512 pixel 的遥感影像中选取尺寸为 256 pixel×256 pixel 的子图像，并将选出的 B 张子图像作为训练的一个批次(Batch)。为了验证随机选取子图进行训练的有效性，基于 U-Net 在 WHU 数据集进行了建筑物提取实验，并与其他文献中未采用该策略的模型进行对比，结果如表 2 所示。可以发现，随机采样策略能在计算资源有限的条件下有效提升分类效果。

表 2 不同模型在 WHU 数据集的随机采样训练和常规训练结果

Table 2 Random sampling training and regular training results of different models in the WHU dataset

Model	Batch size	Image size(pixel×pixel)	Graphic card	Video memory /G	IOU /%
U-Net (random cropping training)	16	256×256	RTX 2070	8	88.58
U-Net (Ref. [15])	8	512×512	Nvidia P6000	24	84.08
U-Net (Ref. [4])	6	512×512	Nvidia Titan XP	12	86.80

为了验证网络结构对分类精度的影响，用广泛采用的二分类交叉熵(BCE)作为损失函数；优化器可极大影响拟合速度和算法的稳定性，为了公平对比，实验中的模型均采用 Adam 作为优化器，WHU 数据集和 Massachusetts 数据集的学习率分别为 0.0001 和 0.0003。MultiResUNet 模型的批次数为 6，其余模型的批次数均为 16。此外，基于训练数据规模，将 WHU 数据集和 Massachusetts 数据集的训练轮数分别设为 300 轮和 200 轮，确保所有网络能在合理的训练时间内充分拟合。由于原始影像尺寸过大以及采用的随机采样策略，测试时必须基于小块图像，还需对测试结果进行拼接才能获得原始图像的预测结果。文献[28]指出全卷积网络预测结果图的中心部分精度较高，表现出边缘退化现象。针对该问题，采用的重叠拼接策略示意图如图 6 所示。对 Massachusetts 数据集进行了重叠尺寸(Padding size)的对比实验，最终设置的重叠尺寸为 64。对于 WHU 数据集，由于图像已经预先切分为较小的尺寸(512 pixel×512 pixel)，且刚好能被随机采样的切片尺寸(256 pixel×256 pixel)整除，若进行重叠测试会导致图像较细的边缘不能被包含在切片内部，提取精度无法从该策略中受益。此外，WHU 数据集按顺序拼接，不需要对图像进行填充，测试效率较高，因此对该数据集不采用重叠拼接策略，详细的实验设置如表 3 所示。

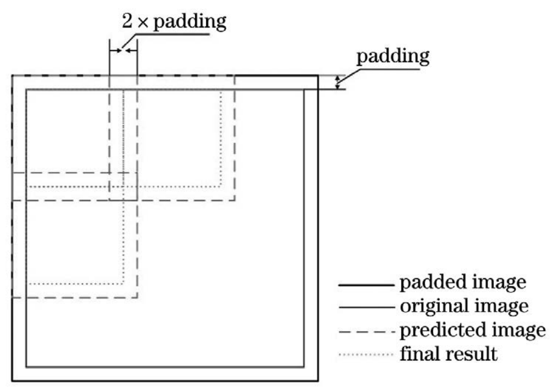

图 6 重叠拼接策略示意图

Fig. 6 Diagram of the overlap strategy

4 实验结果与分析

表 4 为不同模型在两个数据集上的定量评价结果。可以发现，由于采用随机采样策略和充分的训练，基准模型 U-Net 取得了较好的精度。Res-UNet 和 MultiResUNet 在两个数据集上的 IOU 和 F1 分数均低于 U-Net，且 MultiResUNet 和 U-Net 的精度差距较大，这表明全卷积模型在不同类型数据之间的泛化能力有限。DeepLab V3+的分类结果也验证了该结论，这表明模型优化要结合数据特点。USPP和S-UNet的IOU和F1分数均高于U-Net，验

表 3　WHU 航空数据集和 Massachusetts 数据集的实验设置

Table 3　Experimental settings of WHU ariel dataset and Massachusetts dataset

Parameter	WHU ariel dataset	Massachusetts dataset
Training image number (size)	4736 (512 pixel×512 pixel)	8631 (512 pixel×512 pixel)
Validation image number (size)	4144 (256 pixel×256 pixel)	144 (256 pixel×256 pixel)
Training epoch	300	200
Steps per epoch	296	540
Batch size	16(6 for MultiResUNet)	16(6 for MultiResUNet)
Iteration number	296×300	540×200
Padding size	0	64

表 4　不同模型在 WHU 和 Massachusetts 数据集上的定量评估结果

Table 4　Quantitative evaluation results of different models on the WHU and Massachusetts datasets　　unit: %

Dataset	Model	Precision	Recall	IOU	F1 score
WHU	U-Net	94.37	93.52	88.58	93.94
	USPP	94.50	94.35	89.44	94.42
	MultiResUNet	97.00	90.01	87.57	93.37
	S-UNet (Ref. [14])	95.20	93.00	88.80	94.09
	SR-FCN (Ref. [4])	94.40	93.90	88.90	94.15
	S-UNet	94.74	93.77	89.14	94.25
	DeepLab V3+ (Ref. [4])	91.60	94.60	87.10	93.08
	Res-UNet	92.71	93.90	87.44	93.30
	SA-Net	95.27	93.80	89.62	94.53
Massachusetts	U-Net	85.84	81.18	71.60	83.44
	MultiResUNet	93.22	66.84	63.74	77.86
	USPP	88.50	79.37	71.95	83.69
	S-UNet	86.05	81.50	71.99	83.71
	Res-UNet	87.08	77.66	69.64	82.10
	Res-UNet (Ref. [11])	86.21	80.26	71.14	83.13
	JointNet (Ref. [11])	86.21	81.29	71.99	83.68
	SA-Net	86.78	82.70	73.45	84.69

证了跨尺度特征聚合和增大感受野对建筑提取的有效性。在 WHU 数据集上，USPP 和 S-UNet 的 IOU 分别比 UNet 高 0.86 和 0.56 个百分点，而在 Massachusetts 数据集上，USPP 和 S-UNet 的 IOU 仅有微弱优势，这表明上述模型缺乏对不同影像中建筑物尺度的自适应性。

在 WHU 和 Massachusetts 两个数据集上，相比对比模型，SA-Net 取得了最高的 IOU 和 F1 分数。在 WHU 数据集上，SA-Net 的 IOU 分别比 U-Net、USPP、S-UNet、Res-UNet、MultiResUNet 提升了 1.04、0.18、0.48、2.18、2.05 个百分点。在 Massachusetts 数据集上，SA-Net 的优势更明显，IOU 比 MultiResUNet 和 Res-UNet 分别提升了 9.71 和 3.81 个百分点，相比 U-Net、USPP 和 S-UNet 的提升也超过了 1 个百分点，这表明 SA-Net 对不同数据有较好的适应能力，对复杂场景中的建筑物也有较好的提取效果。SA-Net 和对比模型在 WHU 数据集中的训练时间如表 5 所示，可以发现，MultiResUNet 的训练时间远长于其他模型，这表明频繁的特征图连接对模型效率有负面影响。SA-

表 5　不同模型在 WHU 数据集上的训练时间

Table 5　Training time of different models on the WHU dataset　　unit: h

Model	U-Net	USPP	S-UNet	Res-UNet	SA-Net	MultiResUNet
Training time	10.7	11.8	12.8	13.4	13.3	35.1

Net 的训练时间适中，与 Res-UNet、S-UNet 相近，略长于 U-Net 和 USPP。

图 7 和图 8 为不同模型在两个数据集上得到的一些具有代表性的建筑分割图。可以发现，其他模型预测图中经常出现三种不良分割类型(遗漏小型建筑、大型建筑的不连续提取和易混淆地面物体的

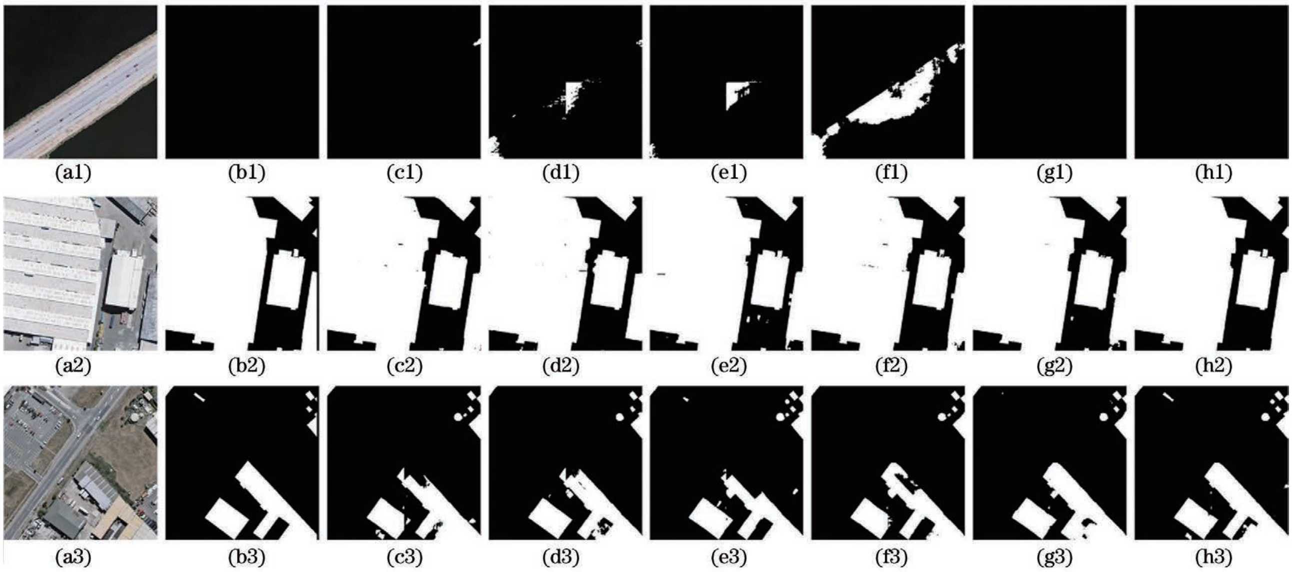

图 7　不同模型对 WHU 数据集的分割结果。(a)图像；(b)标签；(c)U-Net；(d)MultiResUNet；(e)Res-UNet；(f)S-UNet；(g)USPP；(h)SA-Net

Fig. 7　Segmentation results of WHU dataset by different models. (a) Image; (b) label; (c) U-Net; (d) MultiResUNet; (e) Res-UNet; (f) S-UNet; (g) USPP; (h) SA-Net

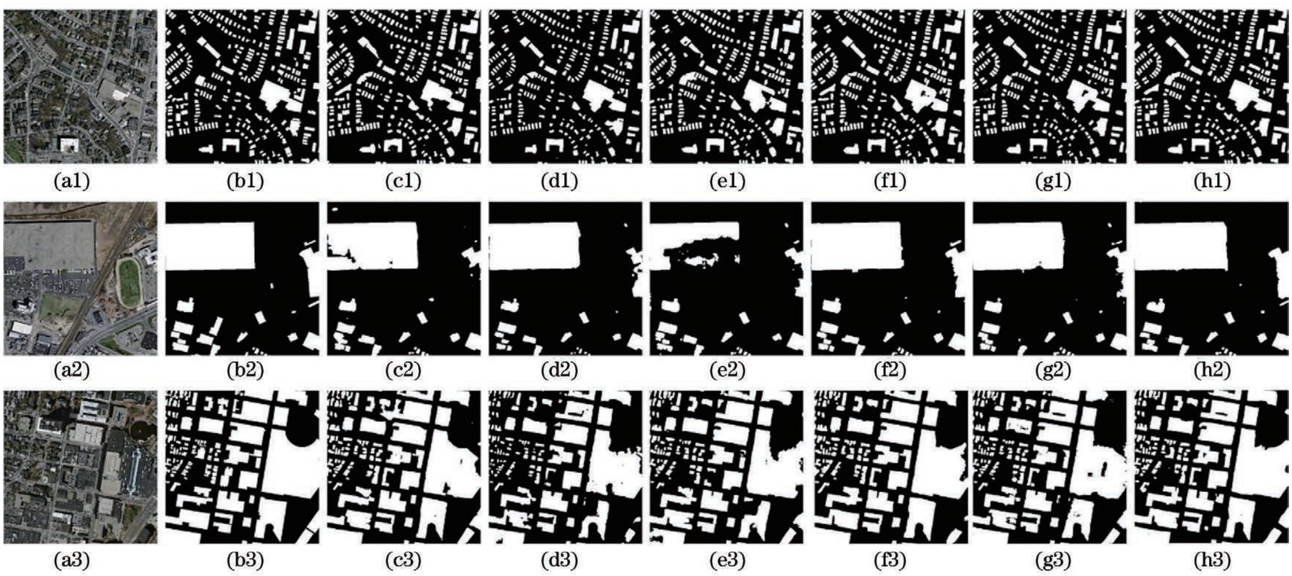

图 8　不同模型对 Massachusetts 数据集的分割结果。(a)图像；(b)标签；(c)U-Net；(d)MultiResUNet；(e)Res-UNet；(f)S-UNet；(g)USPP；(h)SA-Net

Fig. 8　Segmentation results of the Massachusetts dataset by different models. (a) Image; (b) label; (c) U-Net; (d) MultiResUNet; (e) Res-UNet; (f) S-UNet; (g) USPP; (h) SA-Net

错误分类),而 SA-Net 较少出现上述三类不理想的分类结果,这表明 SA-Net 可适应不同类型的高分辨率遥感影像,学习稳健的多尺度特征。

为了验证本方法中 MIMO、RSPP 和 AFR 的有效性,进行了模型消融实验,结果如表 6 所示。可以发现,加入 MIMO 后网络对两个数据集的分割精度均有较大提升,相比 U-Net,在 WHU 数据集上的 IOU 和 F1 分数分别提升了 0.79 和 0.44 个百分点,在 Massachusetts 数据集上分别提升了 1.42 和 0.97 个百分点。RSPP 和 AFR 在不同数据集上对网络的影响差异较大。其中,RSPP 对网络在 WHU 数据集中精度的提升较大,对 Massachusetts 数据集则不明显。相反,AFR 对 WHU 数据集中网络的精度没有提升,在 Massachusetts 数据集中的提升比较明显。这表明 RSPP 和 AFR 对于建筑物尺度差异较大的数据集有各自的适应性,如 WHU 数据集建筑物多尺度现象明显,RSPP 中带有采样的分支能学到更大感受野的深度特征;而 Massachusetts 数据集中建筑物的尺寸普遍较小,RSPP 不带采样的分支起主要作用,模块退化为普通的两层卷积。此外,对于建筑物和背景较难区分的 Massachusetts 数据集,AFR 的特征精化也起了积极的作用。综上所述,MIMO、RSPP 和 AFR 三个模块的有效性得到了充分验证,也验证了 SA-Net 良好的尺度自适应能力,在 WHU 和 Massachusetts 两个数据集上均取得了较好的建筑物提取精度。

表 6 消融实验的评估结果

Table 6 Evaluation results of ablation experiments unit: %

Dataset	Index	U-Net (base-line)	MIMO	RSPP	AFR	IOU	F1 score
WHU	1	√				88.58	93.94
	2	√	√			89.37	94.38
	3	√	√	√		89.67	94.55
	4	√	√	√	√	89.62	94.53
Massachusetts	1	√				71.60	83.44
	2	√	√			73.02	84.41
	3	√	√	√		73.06	84.44
	4	√	√	√	√	73.45	84.69

遥感影像的标记费时费力,为了进一步检验本方法的实用价值,进行了小样本条件下的建筑物提取实验。选取 WHU 数据集初始训练集的前 1000 张影像作为训练样本,选取 Massachusetts 数据集未扩增前的 1233 张影像进行训练。小样本条件下,将 WHU 数据集和 Massachusetts 数据集的训练样本量分别降低至 3.4 小节中的 21%和 14%,实验结果如表 7 所示。可以发现,相比 U-Net、USPP 和 S-UNet,SA-Net 在小样本条件下的精度有明显提升,原因是本方法的特征提取能力较强。

表 7 小样本条件的实验结果

Table 7 Experimental results of small sample conditions unit: %

Dataset	Model	Precision	Recall	IOU	F1 score
WHU	U-Net	83.03	85.87	73.05	84.43
	USPP	87.42	86.60	77.00	87.01
	S-UNet	87.11	86.64	76.80	86.87
	SA-Net	88.92	86.23	77.86	87.55
Massachusetts	U-Net	86.30	73.46	65.79	79.36
	USPP	86.64	75.79	67.86	80.85
	S-UNet	84.56	79.21	69.20	81.80
	SA-Net	87.49	79.28	71.21	83.18

此外，相比表 4 中的实验结果，小样本条件下各个模型的分类精度均有大幅度下降，且各模型在 WHU 数据集中的精度降幅较大，这也验证了初始样本对于分类的重要性；数据扩增得到的新样本对于分类精度的提升有限，也表明单一网络的优化设计无法从根本上解决建筑物提取面临的困难。作为综合性工程问题，必须针对高分辨率遥感图像的特点，探索更多有效的分类精度提升手段，如半监督学习和 SA-Net 的结合以及针对遥感图像分类训练过程引入适用性更强的损失函数或激活函数。

5 结 论

针对高分辨率遥感影像建筑物的提取，为了获得尺度自适应特征，提出了三个网络优化设计模块，即 MIMO、RSPP 和 AFR，同时构建了尺度自适应网络 SA-Net。首先，基于 MIMO 在编码-解码网络的基础上进行多尺度特征融合和跨尺度特征聚合。然后，针对初始聚合特征存在的问题，使用 AFR 进行特征精化，提升聚合特征的判别力。最后，利用 RSPP 扩大网络感受野的同时，提升网络对多尺度建筑物的适应性。基于有限的计算资源，在差异较大的两个高分辨率遥感影像上进行了建筑物提取实验，验证了本方法的有效性。通过模型消融实验和小样本实验，进一步验证了 MIMO、RSPP 和 AFR 等优化设计模块的有效性。实验结果表明，SA-Net 的参数量少于 U-Net，显存占用适中，继承了 U-Net 的轻量特性，实用性较好。但本方法中的网络优化设计不能从根本上解决高分辨率建筑物提取困难的问题。为了利用有限的标记样本取得更好的分类精度，还需进一步研究本方法和半监督学习的结合。同时，还需将损失函数和激活函数与遥感影像的特点相结合，在 SA-Net 的基础上进一步研究探索。此外，本方法提出的模块具有即插即用的特点，可被改进或应用在其他地物提取任务中，有望将 SA-Net 蕴含的设计思想进一步扩展。

参 考 文 献

[1] Liu C, Huang X, Zhu Z, et al. Automatic extraction of built-up area from ZY3 multi-view satellite imagery: analysis of 45 global cities[J]. Remote Sensing of Environment, 2019, 226: 51-73.

[2] Ma L, Liu Y, Zhang X L, et al. Deep learning in remote sensing applications: a meta-analysis and review[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2019, 152: 166-177.

[3] Li W J, He C H, Fang J R, et al. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data[J]. Remote Sensing, 2019, 11(4): 403.

[4] Ji S P, Wei S Q, Lu M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery[J]. International Journal of Remote Sensing, 2019, 40(9): 3308-3322.

[5] Wu Z H, Gao Y M, Li L, et al. Fully convolutional network method of semantic segmentation of class imbalance remote sensing images[J]. Acta Optica Sinica, 2019, 39(4): 0428004.
吴止锾, 高永明, 李磊, 等. 类别非均衡遥感图像语义分割的全卷积网络方法[J]. 光学学报, 2019, 39(4): 0428004.

[6] Yuan Q Q, Shen H F, Li T W, et al. Deep learning in environmental remote sensing: achievements and challenges[J]. Remote Sensing of Environment, 2020, 241: 111716.

[7] Li Y, Zhang H K, Shen Q. Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network[J]. Remote Sensing, 2017, 9(1): 67.

[8] Wang J C, Shen L, Qiao W F, et al. Deep feature fusion with integration of residual connection and attention model for classification of VHR remote sensing images[J]. Remote Sensing, 2019, 11(13): 1617.

[9] Feng F, Wang S T, Zhang J, et al. Hyperspectral images classification based on multi-feature fusion and hybrid convolutional neural networks[J]. Laser & Optoelectronics Progress, 2021, 58(8): 0810010.
冯凡, 王双亭, 张津, 等. 基于多特征融合和混合卷积网络的高光谱图像分类[J]. 激光与光电子学进展, 2021, 58(8): 0810010.

[10] Maggiori E, Tarabalka Y, Charpiat G, et al. Convolutional neural networks for large-scale remote-sensing image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(2): 645-657.

[11] Zhang Z X, Wang Y H. JointNet: a common neural network for road and building extraction[J]. Remote Sensing, 2019, 11(6): 696.

[12] Kang W C, Xiang Y M, Wang F, et al. EU-Net: an efficient fully convolutional network for building extraction from optical remote sensing images[J]. Remote Sensing, 2019, 11(23): 2813.

[13] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W

M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.

[14] Ji S P, Wei S Q. Building extraction via convolutional neural networks from an open remote sensing building dataset[J]. Acta Geodaetica et Cartographica Sinica, 2019, 48(4): 448-459.
季顺平, 魏世清. 遥感影像建筑物提取的卷积神经元网络与开源数据集方法[J]. 测绘学报, 2019, 48(4): 448-459.

[15] Cui W H, Xiong B Y, Zhang L Y. Multi-scale fully convolutional neural network for building extraction[J]. Acta Geodaetica et Cartographica Sinica, 2019, 48(5): 597-608.
崔卫红, 熊宝玉, 张丽瑶. 多尺度全卷积神经网络建筑物提取[J]. 测绘学报, 2019, 48(5): 597-608.

[16] Tian Q L, Qin K, Chen J, et al. Building change detection for aerial images based on attention pyramid network[J]. Acta Optica Sinica, 2020, 40(21): 2110002.
田青林, 秦凯, 陈俊, 等. 基于注意力金字塔网络的航空影像建筑物变化检测[J]. 光学学报, 2020, 40(21): 2110002.

[17] Liu Y H, Gross L, Li Z Q, et al. Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling[J]. IEEE Access, 2019, 7: 128774-128786.

[18] Shao Z F, Tang P H, Wang Z Y, et al. BRRNet: a fully convolutional neural network for automatic building extraction from high-resolution remote sensing images[J]. Remote Sensing, 2020, 12(6): 1050.

[19] Zhang Z X, Liu Q J, Wang Y H. Road extraction by deep residual U-Net[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(5): 749-753.

[20] Ibtehaz N, Rahman M S. MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation[J]. Neural Networks, 2020, 121: 74-87.

[21] Pleiss G, Chen D L, Huang G, et al. Memory-efficient implementation of DenseNets[EB/OL]. (2017-7-21)[2021-01-01]. https://arxiv.org/abs/1707.06990.

[22] Feng F, Wang S, Wang C, et al. Learning deep hierarchical spatial-spectral features for hyperspectral image classification based on residual 3D-2D CNN[J]. Sensors, 2019, 19(23): 5276.

[23] Veit A, Wilber M J, Belongie S J, et al. Residual networks behave like ensembles of relatively shallow networks[C]//Advances in Neural Information Processing Systems, December 5-10, 2016, Barcelona, Spain. [S.l.: s.n.], 2016: 550-558.

[24] Liu Y G, Yu J Z, Han Y H. Understanding the effective receptive field in semantic image segmentation[J]. Multimedia Tools and Applications, 2018, 77(17): 22159-22171.

[25] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[26] Ji S P, Wei S Q, Lu M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(1): 574-586.

[27] Mnih V. Machine learning for aerial image labeling[D]. Toronto: University of Toronto, 2013: 84-88.

[28] Sun Y, Tian Y, Xu Y P. Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: structural stereotype and insufficient learning[J]. Neurocomputing, 2019, 330: 297-304.