激光与光电子学进展 先进成像

结合多尺度注意力和边缘监督的遥感图像建筑物分割模型

杨潇宇，汪西莉*

陕西师范大学计算机科学学院，陕西 西安 710119

摘要 遥感图像分割是遥感图像处理领域的一项重要应用。针对卷积神经网络在遥感图像分割时存在建筑物错分、漏分，建筑物轮廓分割不准确的问题，基于深度学习网络U-Net，提出了一种结合多尺度注意力和边缘监督的遥感图像分割网络MAE-Net。首先，在编码阶段每层引入多尺度注意力模块，该模块对输入特征图进行通道均等分组，每组使用大小不同的卷积核进行特征提取，之后对每组施加通道注意力机制，通过自学习的方式获得更为有效的特征，解决大小不一建筑物特征提取不准确的问题；其次，在解码阶段引入边缘提取模块，构造边缘监督网络，通过损失函数计算边缘标签和预测边缘的误差，帮助分割网络更好地学习建筑物边缘特征，使建筑物边界的分割结果更为连续、平滑。实验结果表明，MAE-Net能够从背景复杂多样、尺度变化较大的遥感图像中完整地分割出建筑物，且分割精度较高。

关键词 图像处理；神经网络；遥感图像；多尺度特征提取；注意力机制；边缘监督

中图分类号 O436 **文献标志码** A **DOI**：10.3788/LOP202259.2228004

Building Segmentation Model of Remote Sensing Image Combining Multiscale Attention and Edge Supervision

Yang Xiaoyu, Wang Xili*

School of Computer Science, Shaanxi Normal University, Xi'an 710119, *Shaanxi, China*

Abstract Remote sensing image segmentation is a crucial application in the field of remote sensing image processing. A semantic segmentation network MAE-Net combining multiscale attention and edge supervision is proposed based on the deep learning network U-Net to address the phenomena of building missing classification, missing segmentation, and inaccurate building contour segmentation in the building segmentation of remote sensing images via convolution neural network. First, a multiscale attention module is introduced into each layer during the coding stage. The module separates the input feature map into equal channels and employs the convolution kernels of various sizes for feature extraction in each group. Thereafter, the channel attention mechanism is used in each group to gain more efficient features through self-learning to solve the problem of inaccurate feature extraction of buildings of various sizes. Second, in the decoding stage, the edge extraction module is introduced to build the edge supervision network. The error between the learning edge label and expected edge is supervised by the loss function to aid the segmentation network in better learning the building edge features and make the building boundary's segmentation result more continuous and smoother. The experimental findings show that MAE-Net can completely segment buildings from remote sensing images with complicated and diverse backgrounds and large-scale changes, and the segmentation accuracy is higher.

Key words image processing; neural network; remote sensing image; multiscale feature extraction; attention mechanism; edge supervision

1 引 言

随着遥感技术的快速发展，从遥感图像中提取各种人工目标成为研究热点。建筑物作为地物之一，在城市规划、基础设施建设、变化检测、地表覆盖分类及地理信息数据库更新等方面都具有重要的作用，准确

收稿日期：2021-08-17；**修回日期**：2021-09-17；**录用日期**：2021-10-13

基金项目：第二次青藏高原综合科学考察研究项目(2019QZKK0405)

通信作者：*wangxili@snnu.edu.cn

提取遥感图像中的建筑物是非常必要的[1-2]。在传统的遥感图像特征提取任务中，通常采取人工设计的方式提取物体特征[3]，再使用传统分割算法实现对遥感图像中物体的分割。但这类分割算法在提取目标特征时具有一定的局限性，很难提高分割的准确率。

随着深度学习的发展，卷积神经网络在图像处理领域表现出色[4]，具有强有力地捕获丰富的空间特征和多尺度信息的能力，其提取到的特征比之前人工构造的特征效果更好，因此被众多学者广泛研究，使得许多基于卷积神经网络的语义分割方法相继被提出。2015年，Long等[5]提出了全卷积网络(FCN)，该网络在去掉全连接层的基础上使用反卷积操作得到预测图，并进行逐像素分类，实现对输入图像的分割。同年，Ronneberger等[6]提出了U-Net，此网络基于编码解码结构实现，网络解码器中的每层在上采样过程中与同层编码器的特征图进行拼接，保留了更多的细节信息，提升了分割效果。2017年，Zhao等[7]提出了PSPNet，此网络是在FCN上的改进，对深层特征图使用金字塔池化模块获得不同尺度的语义信息并融合，将浅层细节信息和深层语义信息综合考虑，使得分割结果更加精细。2018年，Chen等[8]提出了DeepLab方法，该方法使用多尺度空洞卷积来扩大感受野，系统地聚合不同尺度的上下文信息，减少了下采样操作导致的信息损失，进一步提升了分割精度。

以上基于卷积神经网络的图像分割方法，能很好地提取图像的空间特征和多尺度信息，但依然存在一些问题，比如只关注图像的空间信息而忽略了通道信息的重要性、不能很好获取图像中的边缘细节信息和一些复杂的特征信息等。为了解决这些问题，相关学者提出了一些有效的解决方案。2020年，Hu等[9]提出了squeeze-and-excitation networks，其通过学习通道之间的相关性，筛选出针对通道的注意力，增强重要通道的权重，减弱不重要通道的权重，提升了在通道维度上特征信息的表达能力。2019年，Fu等[10]提出了双重注意力网络(DANet)，该网络通过引入自注意力机制在特征的空间维度和通道维度分别抓取特征之间的全局依赖关系，增强特征的表达能力。除了对通道信息相关性的研究，还有一些是对边缘信息提取的研究，Qin等[11]隐式地将精确的边界预测目标注入混合损失中，以减少在边界和图像的其他区域中学习到的信息在交叉传播时的误差，增强了对边界信息的提取精度。Yu等[12]通过类间差异获得尽可能多的特征，再通过明确的语义边界去指导特征学习，从而达到细化边界结果的效果。

上述研究在一定程度上提高了复杂场景中图像分割精度，但由于遥感图像中建筑物类别多样、纹理信息复杂、空间分布不规则、尺度多变、背景复杂等特殊原因，对遥感图像建筑物的分割仍存在很多问题，如建筑物及其边界信息丢失严重、复杂结构建筑物无法完整识别和建筑物漏分、错分等。针对这些问题，本文在原始U-Net上进行改进，通过引入多尺度特征提取结构、通道注意力机制和边缘监督机制增强网络对目标特征的提取能力，提出了一种结合多尺度注意力和边缘监督的遥感图像建筑物分割网络MAE-Net。在建筑物数据集上测试了网络性能，并与其他语义分割常用网络进行了对比。

2 MAE-Net模型

2.1 MAE-Net结构

所提MAE-Net基于编码器、解码器结构实现，网络结构如图1所示。编码器用于获取包含遥感图像建筑物丰富上下文信息的特征图。解码器分为两部分：一部分是类别预测网络，采用上采样操作逐步恢复特征图分辨率，并与编码器同层多尺度注意力(MSA)模块输出的多尺度注意力特征图进行拼接，融合获得更多的特征信息，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；另一部分是边缘监督网络，使用边缘提取(EEB)模块来预测边缘，再通过建立边缘标签和预测边缘之间的损失来指导监督目标分割网络，以进一步细化图像分割网络的特征图，提升网络的细节恢复能力。同时为了提升网络对不同尺度目标信息的获取，在编码阶段每层引入了MSA，首先对通道进行4等分，然后每组使用大小为3、5、7、9的4个卷积核并行在每个分组上进行卷积操作，使网络能够捕捉并利用不同尺度特征图的空间信息。每一层都在不同尺度上获得上下文特征，与不分组的卷积方式相比，每一层都融合了细节和语义信息，提取的特征表达能力更强。之后对每组施加通道注意力机制，建立起通道维度上的依赖关系，丰富特征空间，增强有用信息的表达。

主干网络U-Net包含卷积层、池化层、上采样层和输出层：每个卷积层包括两个3×3卷积，用来提取目标信息；两个批归一化(BN)层[13]用于加速网络的学习速度；并使用两个ReLU修正线性单元[14]作为激活函数；池化层使特征图的尺寸减小为原来的一半；上采样层使用反卷积扩大特征图的尺寸为原来的2倍，恢复特征图的分辨率；输出层包括一个1×1卷积和sigmoid激活函数，进行像素类别预测，输出最终的分割图像。

2.2 多尺度注意力模块

多尺度注意力模块对应图1中的MSA层，其结构如图2所示。在主干网络U-Net中，其编码阶段的卷积核大小固定为3×3，只能提取对应感受野中的局部特征，即提取到单一尺度的特征。为了得到不同尺度更丰富的特征信息，采用了多尺度注意力模块，它通过以下4个步骤实现：首先，在通道维度上通过split convolution Concat(SCC)模块获得多尺度的特征图；其次，通过SE权重模块学习组内各通道间权重；然后，

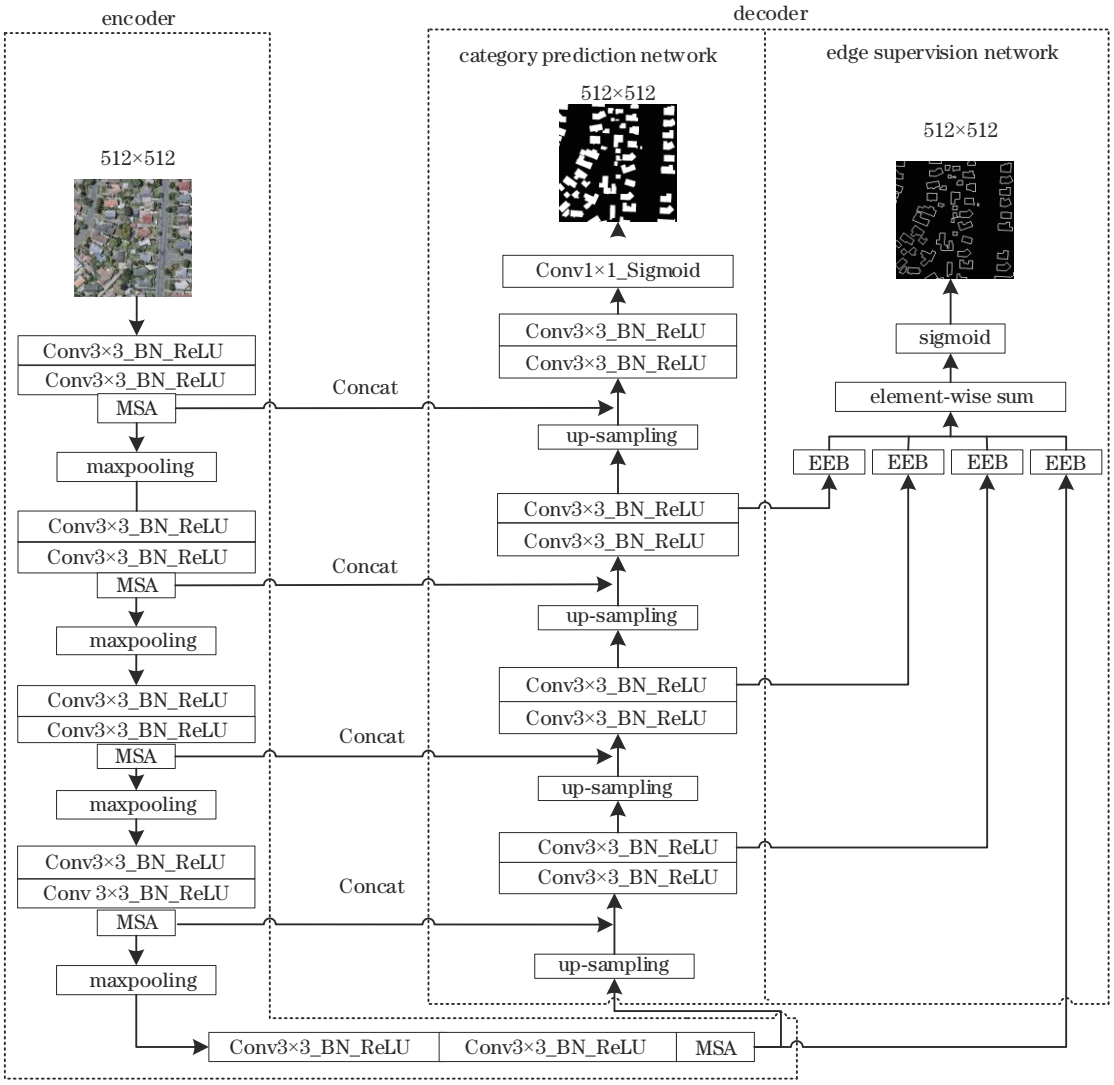

图 1　MAE-Net 结构

Fig. 1　Structure of MAE-Net

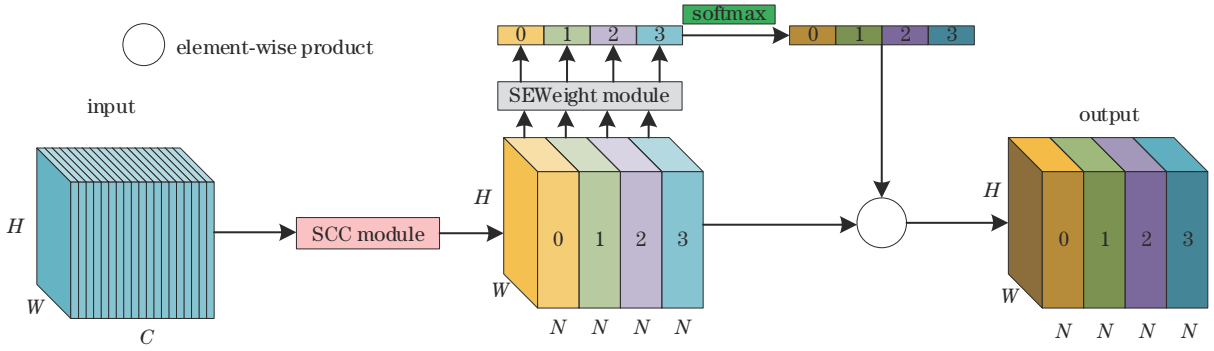

图 2　多尺度注意力模块

Fig. 2　Multiscale attention module

通过 softmax 对组与组间通道权重进行重新标定来自适应学习，充分利用了特征图通道维度上的不同信息来改善网络对建筑物特征的捕捉能力；最后将学习到的通道权重加权到对应的特征图上。

获得多尺度特征图的操作在 SCC 模块中实现，如图 3 所示，对于输入的特征图 $\boldsymbol{X}$，先从通道上将其分成 S 组，其对应的序列为 $\{\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_{S-1}\}$，每组包含的通道数量 $N = C/S - 1$，然后对每组使用不同大小的卷积核进行卷积。SCC 模块中的分组操作可描述为

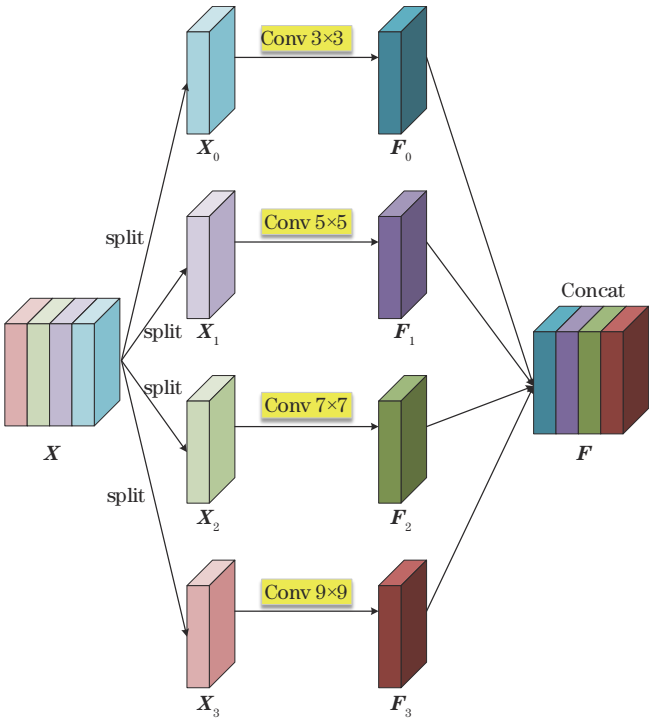

图 3　Split Convolution Concat 模块

Fig. 3　Split convolution concat module

$$\boldsymbol{F}_i=\mathrm{Conv}\left(K_i\times K_i\right)\left(\boldsymbol{X}_i\right),\ i=0,1,2,\cdots,S-1,\tag{1}$$

式中：每个分组使用的卷积核大小 $K_i=2\times(i+1)+1$；$\boldsymbol{X}_i$ 代表每组的输入特征图；$\boldsymbol{F}_i$ 代表不同尺度卷积后输出的特征图。然后将特征图输入 SE 权重模块提取通道注意力权重信息，得到不同尺度的注意力权重，其对应的操作如下：

$$\boldsymbol{Z}_i=\mathrm{SEWeight}\left(\boldsymbol{F}_i\right),\tag{2}$$

式中：$\boldsymbol{Z}_i$ 代表在每个分组施加通道注意力之后的特征图。为了实现组间通道的自适应权重分配，让网络自学习选择不同的空间尺度，采用 softmax 进行组间通道权重的重新标定，具体操作如下：

$$\boldsymbol{w}_{\mathrm{att}i}=\mathrm{softmax}\left(\boldsymbol{Z}_i\right)=\frac{\exp\left(\boldsymbol{Z}_i\right)}{\sum_{i=0}^{S-1}\exp\left(\boldsymbol{Z}_i\right)},\tag{3}$$

式中：$\boldsymbol{w}_{\mathrm{att}i}$ 代表权重重新标定后每个分组对应的全局权重向量；$\boldsymbol{Z}_i$ 代表每个分组不同尺度特征图对应的通道权重向量。最后再对所得到的全局权重向量进行拼接：

$$\boldsymbol{w}_{\mathrm{att}}=\mathrm{Concat}\left[\boldsymbol{w}_{\mathrm{att}\,0}+\boldsymbol{w}_{\mathrm{att}\,1}+\cdots+\boldsymbol{w}_{\mathrm{att}\,i}\right],\tag{4}$$

式中：$\boldsymbol{w}_{\mathrm{att}}$ 代表注意力交互后的多尺度通道权重。最后将得到的权重加权到之前不同尺度卷积后输出的特征图 $\boldsymbol{F}_i$ 中：

$$\boldsymbol{Y}_i=\boldsymbol{F}_i\odot\boldsymbol{w}_{\mathrm{att}\,i},\tag{5}$$

式中：$\odot$ 代表矩阵点乘；$\boldsymbol{Y}_i$ 代表不同尺度卷积后输出的特征图通道加权之后的加权特征图。最后，将得到的所有通道加权特征图拼接成最终的输出特征图：

$$\boldsymbol{F}_{\mathrm{out}}=\mathrm{Concat}\left[\boldsymbol{Y}_0,\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_{S-1}\right],\tag{6}$$

式中：$\boldsymbol{F}_{\mathrm{out}}$ 代表 MSA 模块最终输出的多尺度特征图。

2.3　通道注意力机制

为了更好地对通道信息进行有选择的关注，从而产生更有效的信息输出，在 MSA 中使用通道注意力机制来加强通道信息的获取，该机制对应图 2 中的 SEWeight module。图 4 给出了该机制的结构图，其中 C、H、W 代表特征图的通道数、长、宽。

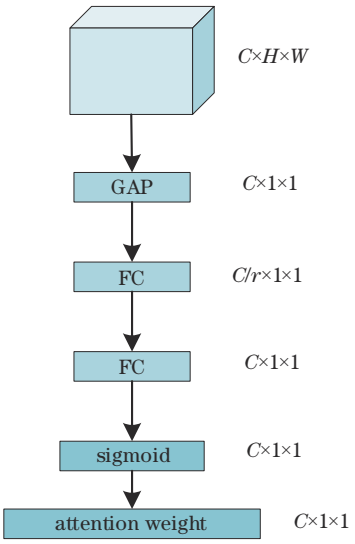

图 4　SEWeight module

Fig. 4　SEWeight module

该机制主要包含两部分，第 1 部分对输入的特征图 $C\times H\times W$ 从空间维度上进行全局平均池化（GAP），在全局感受野上对全局信息进行编码将其嵌入通道中，输出一个 $C\times1\times1$ 维度的编码向量，全局平均池化对应的操作为

$$\boldsymbol{G}_{\mathrm{c}}=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{x}_{\mathrm{c}}(i,j),\tag{7}$$

式中：i,j 表示特征图上每一个像素点的位置；$\boldsymbol{x}_{\mathrm{c}}$ 表示每个具体位置上的像素值。

第 2 部分建立各通道间的依赖关系，使网络可以自己学习更新通道权重。首先，对之前全局平均池化输出的 $C\times1\times1$ 向量，使其经过两个全连接层（FC）建立起代表各通道间的动态权重关系，再通过 sigmoid 激活函数归一化权重信息，最终获得所需要的权重信息。此部分对应的操作可以描述为

$$\boldsymbol{W}_{\mathrm{c}}=\sigma\left\{W_1\delta\left[W_0\left(\boldsymbol{G}_{\mathrm{c}}\right)\right]\right\},\tag{8}$$

式中：W_0、W_1 代表两个全连接层；δ 代表 Relu 激活函数；σ 代表 sigmoid 激活函数；$\boldsymbol{W}_{\mathrm{c}}$ 代表最终得到的通道注意力权重向量。

2.4　边缘提取模块

边缘提取模块对应图 1 中的 EEB 层，其结构如图 5 所示。所提方法通过此模块从特征图中获取预测

边缘，引导分割网络学习更多的图像边缘信息，从而完善图像分割结果，提升边缘分割精度。

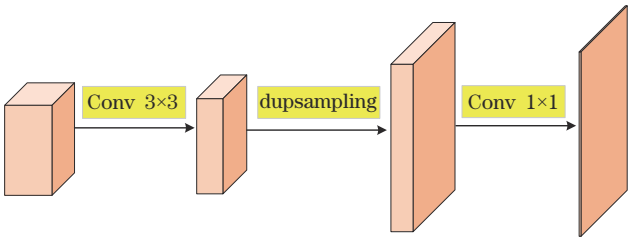

图 5 边缘提取模块

Fig. 5 Edge extraction module

在每个 EEB 模块中，对于输入的特征图，先对其进行 2 次 3×3 卷积，使其特征图在通道维度减半，然后再对其进行数据相关上采样[15]，使其特征图的尺寸大小直接上采样到原输入特征图的大小，最后再通过 1 个 1×1 卷积调整特征图的通道大小为 1，作为 EEB 模块最终的输出预测图。边缘监督网络一共含有 4 个 EEB 模块，将这 4 个模块的预测边缘信息进行 element-wise sum，融合更多的深层语义和浅层细节信息，输出最终的边缘预测图，并于边缘标签建立损失函数关联，监督图像分割网络更好对目标进行分割。

2.5 损失函数

经典的二分类交叉熵损失函数的表达式为

$$L_{\mathrm{BCE}}=\left\{\eta(i,j)\log P(i,j)+\left[1-\eta(i,j)\right]\times\log\left[1-P(i,j)\right]\right\},\tag{9}$$

式中：$\eta(i,j)$代表像素点(i,j)处的真实标签；$P(i,j)$代表像素点(i,j)由分割网络输出的目标物体预测概率；$\eta(i,j)\log P(i,j)$部分对应目标物体区域的损失；$\left[1-\eta(i,j)\right]\log\left[1-P(i,j)\right]$部分对应背景区域的损失。

在网络训练时，编码器中的图像分割网络和边缘监督网络都会进行损失计算，对这两个任务都使用式(9)的二分类交叉熵损失函数来计算损失。建筑物分割从本质上可以看成是建筑物体类与背景类的二分类任务，对应式(9)：$\eta(i,j)$代表像素点(i,j)处的真实标签；$P(i,j)$代表像素点(i,j)由图像分割网络输出的建筑物预测概率，建筑物预测的损失使用L_{BBCE}来表示。边缘分割从本质上可以看成是建筑物边缘类与背景类的二分类任务，对应式(9)：$\eta(i,j)$代表像素点(i,j)处的真实标签；$P(i,j)$代表像素点(i,j)由边缘监督网络输出的建筑物边缘预测概率，边缘预测的损失使用L_{EBCE}来表示。最后，将两个任务的损失进行求和得到最终的损失：

$$L_{\mathrm{all}}=L_{\mathrm{BBCE}}+L_{\mathrm{EBCE}},\tag{10}$$

式中：L_{EBCE}代表边缘预测损失；L_{BBCE}代表建筑物预测损失；L_{all}代表总损失。

3 实验结果与分析

3.1 数据集介绍

实验使用 WHU Building Dataset 和 Satellite Dataset Ⅱ(East Asia)两个数据集[16]对所提 MAE-Net 进行评估。

3.1.1 WHU Building 数据集

WHU Building Dataset 包括从新西兰 Christ Church 地区提取的 22 万栋建筑物，空间分辨率为 0.075 m，覆盖面积为 450 km^2，包含农村、城镇、文化区和工业区，具有不同颜色、大小和用途的多种多样的建筑类型，是评估建筑物提取算法潜力的理想数据集。该数据集共有 8189 幅图像和对应的像素级标签图，分辨率为 512 pixel×512 pixel，其中训练集 4736 幅、验证集 1037 幅、测试集 2416 幅。

3.1.2 Satellite Dataset Ⅱ(East Asia)数据集

Satellite Dataset Ⅱ(East Asia)建筑物数据集覆盖东亚 550 km^2，地面分辨率为 2.7 m，包含训练集和测试图的整体图像(训练集 2 幅图像、测试集 1 幅图像)及这些大规模图像裁剪后的小幅图像。所有图像无缝裁剪成 17388 幅、分辨率为 512 pixel×512 pixel 的小幅图像，用于训练和测试，其中训练集包含 13662 幅图像，测试集包含 3726 幅图像。

3.2 实验环境

实验平台为 Intel(R) Xeon(R) Silver 4214 CPU @ 2.20 GHz、128 GB 内存、显存 48 GB 的工作站，软件配置为 Ubuntu18.04 系统、Pytorch 深度学习框架。在训练过程中使用两块 NVIDIA GeForce RTX 3090 24 GB 显存 GPU 进行计算，使用 Adam 优化算法进行优化，初始学习率设置为 0.001，每个训练批次随机读入 14 张分辨率为 512 pixel×512 pixel 的图片，迭代次数设置为 50。

3.3 评价指标

为了对建筑物分割网络性能进行评估，分别采用交并比(IOU)、准确率(P)、召回率(R)和F_1分数这 4 个指标来量化分析所提算法分割结果。IOU 表示标签真实值与预测值的交集和并集的比值，P表示预测正确的正类个数占全部预测为正样本的比例，R表示预测正确的正类个数占全部正样本的比例，F_1分数同时兼顾了分类模型的准确率和召回率，是模型精确率和召回率的一种调和平均，其表达式分别为

$$R_{\mathrm{IOU}}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}},\tag{11}$$

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}},\tag{12}$$

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}},\tag{13}$$

$$S_{F_1}=\frac{2P\times R}{P+R},\tag{14}$$

式中：N_{TP}为目标正确分类的像素数目；N_{FN}为目标分为背景的像素数目；N_{FP}为背景分为目标的像素数目。

3.4 MAE-Net消融实验结果与分析

为了验证所提MAE-Net及网络各个模块的有效性，进行了消融实验。消融实验的设置是通过对网络结构的调整实现的。所提网络模型是基于编码解码结构的U型网络，故选取U-Net作为实验基线模型。

消融实验使用WHU Building训练集训练网络，使用测试集测试评价分割结果，并利用评价指标F_1和IOU来衡量。在训练和测试时，所有模型的输入图像尺寸均为512 pixel×512 pixel，训练时，输出图像是与输入图像大小相同的预测标签图，测试时，输出图像是与输入图像大小相同的分割结果图。实验中，MAE-Net训练参数设置与其他子网络相同且在同一环境下运行。消融实验在WHU Building测试集上的评价结果如表1所示。

表1 消融实验在WHU Building测试集上评价结果对比

Table 1 Comparison of evaluation results of ablation experiment on WHU Building test set

No.	Baseline	MSA	EEB	F_1-score	IOU
1	√	×	×	0.9420	0.8904
2	√	√	×	0.9502	0.9052
3	√	×	√	0.9496	0.9041
4	√	√	√	0.9520	0.9084

从表1可以看到：基线U-Net的IOU和F_1分别为0.8904和0.9420；添加MSA模块后，IOU比基线模型提高了1.48个百分点，F_1提高了0.82个百分点；添加EEB模块后IOU比基线模型提高了1.37个百分点，F_1提高了0.76个百分点；最后将MSA和EEB两个模块都添加到基线网络中，也就是所提MAE-Net，与基线网络结果相比，MAE-Net的IOU和F_1分别提高了1.8个百分点、1个百分点。即MSA模块和EEB模块对建筑物的分割精度都有较好的提升。表2给出了消融实验在WHU Building测试集上训练和测试时间的结果。

从表2可以看到：MAE-Net的训练时间要比基线网络U-Net的训练时间多4 h，增加的这些时间一部分耗费在MSA模块中多尺度特征提取上，另一部分耗费在EEB模块中边缘预测上。MAE-Net在每张测试图像上的平均测试时间比基线网络U-Net多0.012 s，但在基线网络上只添加EEB模块，每张测试图像上的测试时间还要比基线网络U-Net少0.011 s，这更加验证了EEB模块的有效性。综合评价指标结果、训练和测试时间等3个因素，虽然MAE-Net的训练和测试时间要比基线网络U-Net稍微逊色一点，但是在分割准确率上，MAE-Net更加出色，其能在较小的计算量下显著地提高建筑物的分割精度。

表2 消融实验在WHU Building测试集上训练和测试时间对比

Table 2 Comparison of train and test time of ablation experiment on WHU Building test set

No.	Baseline	MSA	EEB	Train time /h	Test time per picture /s
1	√	×	×	5.13	0.077
2	√	√	×	8.16	0.078
3	√	×	√	6.83	0.066
4	√	√	√	9.13	0.089

图6给出了消融实验在WHU Building测试集上的分割结果，从左到右依次为原始图像、标签图、基线U-Net分割结果图、基线和MSA分割结果图、基线和EEB分割结果图、MAE-Net分割结果图。图中黑色代表背景，白色代表建筑物。所给6幅图像建筑物的光照、颜色、大小、形状等均不相同。

在image 1中，4个标记框都含有非常细微的边缘信息：在只有基线网络U-Net的情况下，这4处边缘细节都被U-Net忽略了，出现了建筑物漏分、边界信息模糊、粘连在一起的情况；在只引入MSA模块时，一处建筑物漏分的错误被解决，一处边界信息粘连的错误被改善；在只引入EEB模块时，两处建筑物漏分的错误被解决，一处边界信息粘连的错误被解决；当将两个模块全部引入后，这4处边缘细节都被网络很好地分割，解决了U-Net暴露的建筑物漏分、边界信息模糊、粘连在一起的问题。Image 2和image 3图像中包含一些小目标的建筑物，U-Net对这些小目标的提取能力较差，出现了错分和边界模糊的现象，在分别引入MSA、EEB模块后，网络对这几处的问题都有一定的改善，但还是存在漏分和错分的现象，所提MAE-Net的分割结果则令人满意。Image 5和image 6图像是边界清晰的大目标：U-Net对这些目标的提取能力同样非常差，特别是在image 5图像中出现了大面积的漏分现象，还有其对与地面背景相似的建筑物提取效果也非常不好；在只引入MSA模块时，两幅图像有三处建筑物漏分的现象被解决；在只引入EEB模块时，只有一处边缘细节被改善，其他建筑物漏分的现象都没有很好解决，这也说明边缘监督对边缘分割的指导作用更明显，使其分割的效果更好。Image 4图像在使用MAE-Net分割时，其对建筑物及其边缘的分割效果都非常好，分割结果已经非常接近标签图。

为了定量分析MAE-Net对不同大小建筑物的识别精度和对建筑物边缘的识别精度，按照建筑物的大小进行分类，将像素低于2000的建筑物划分为小尺度(small-scale)建筑物类别，将像素高于2000的建筑物划分为大尺度(large-scale)建筑物类别，在这两个类别上进行对比实验，验证网络对不同大小建筑物的识别精度。还设置了边缘尺度(edge-scale)建筑物类别，通过对网络输出的建筑物分割图进行边缘提取，得到分

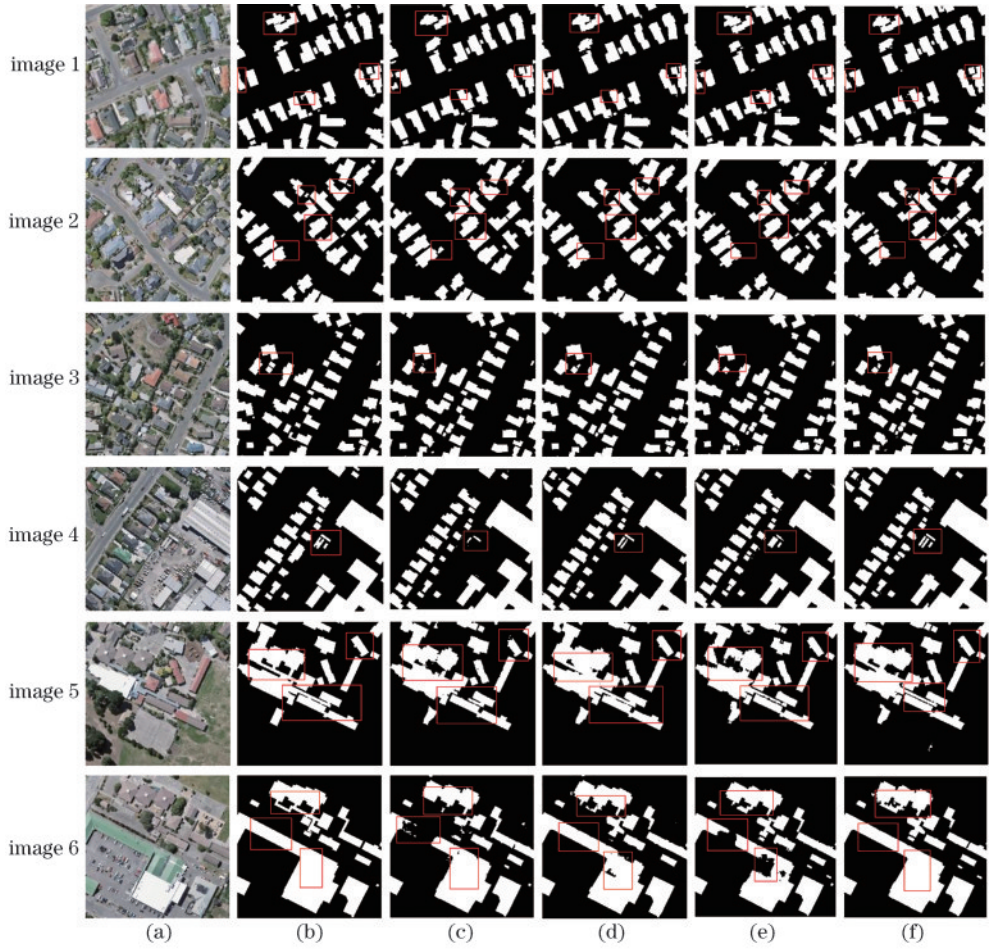

图 6　消融实验在 WHU Building 测试集上的分割结果。(a)原始图像;(b)标签图;(c)基线分割结果;(d)基线和 MSA 分割结果;(e)基线和 EEB 分割结果;(f) MAE-Net 分割结果

Fig. 6　Segmentation results of ablation experiment on WHU Building test set. (a) Original images; (b) image label;(c) baseline processing result; (d) baseline with MSA processing result; (e) baseline with EEB processing result; (f) MAE-Net processing result

割后的建筑物边缘图,并与边缘标签进行对照评价,验证网络对建筑物边缘的识别精度。对图 6 中的 6 幅图像按照建筑物的像素进行分类,将 image 1 和 image 2 归为小尺度建筑物类别,将 image 5 和 image 6 归为大尺度建筑物类别,由于 image 3 和 image 4 既包含大尺度建筑物又包含小尺度建筑物,其包含全面的建筑物大小类别,所以将 image 3 和 image 4 归为边缘尺度建筑物类别,用来对比验证网络对建筑物边缘的识别精度。然后在划分好的 3 个实验组别上,使用 MAE-Net 和基线网络 U-Net 分别进行测试。表 3 给出了 U-Net 和 MAE-Net 对不同大小建筑物及建筑物边缘的识别精度对比结果。

从表 3 可以看出,相比基线网络 U-Net,MAE-Net 对不同大小建筑物都有很好的识别精度,对建筑物边缘的识别更是有大幅度的提升。以上所有消融实验的结果表明,MSA、EEB 两个模块都有很好的特征提取能力:MSA 侧重从多尺度角度和各个通道间建立起的全局联系提取特征信息,能很好对建筑物的全局信息进行分割;EEB 由于有边缘标签的监督,其对边缘信息的感知更加敏感,能很好对边缘细节特征进行提取,对边缘的分割更加准确平滑。上边 6 幅实验结果图都能很好印证这个结论。

表 3　两种模型对建筑物及建筑物边缘识别精度结果

Table 3　Accuracy results of two models for building and building edge recognition

Scale	U-Net		MAE-Net	
	IOU	F_1-score	IOU	F_1-score
Small-scale(image 1+image 2)	0.904	0.949	0.910	0.953
Large-scale(image 5+image 6)	0.918	0.957	0.927	0.963
Edge-scale(image 3+image 4)	0.438	0.609	0.545	0.706

3.5　MAE-Net 模型和其他模型的比较

3.5.1　WHU Building 数据集

在 WHU Building 数据集的验证集上,将 MAE-Net 与 U-Net[6]、SiU-Net[16]、SRINet[17]、Ra-CGAN[1]和 DeNet[18]进行了对比,数据集划分均一致,对比结果如表 4 所示。

U-Net 是基于编码解码结构的网络模型,其在每个解码层添加了跳跃连接,也是所提模型的基线网络结构,此网络对应的 IOU 为 88.13%。SiU-Net 以原始

表 4　不同方法在 WHU Building Dataset 验证集上的各项指标对比

Table 4　Experimental results of different methods on WHU building validation set

Method	F_1-score	P	R	IOU
U-Net[6]	0.9135	0.9542	0.8826	0.8813
SiU-Net[16]	0.9385	0.9380	0.9390	0.8840
SRINet[17]	0.9423	0.9521	0.9328	0.8909
Ra-CGAN[1]	0.9490	0.9510	0.9460	0.8960
DeNet[18]	0.9480	0.9500	0.9460	0.9012
MAE-Net	0.9542	0.9554	0.9530	0.9124

图像及其下采样图像作为并行网络的输入，并行网络的两个分支共享相同的 U 型网络结构与权重，然后将分支的输出串联起来作为最终输出，其最终分割结果的 IOU 为 88.40%。SRINet 通过连续融合多级别特征来捕获和聚合用于语义理解的多尺度上下文，从而提升了多尺度目标的分割精度，其分割结果的 IOU 为 89.09%。Ra-CGAN 采用对抗训练的方法，通过一个具有多级通道注意力机制的生成模型和一个判别网络模型，来矫正标签图和分割图之间的差异，改善分割结果，其最终的 IOU 为 89.60%。DeNet 遵循 Inception 思想，通过改进下采样操作来减小浅层细节信息的丢失率，设计了 densely upsampling convolution 上采样模块，最大程度利用深层语义信息和浅层空间信息来生成对应的建筑物分割结果图，其最终的 IOU 为 90.12%。MAE-Net 模型在此数据集上的评价结果最好，其 IOU 值最高，为 91.24%。图 7 给出了 MAE-Net 与基线网络 U-Net 的实验对比结果。

所提算法利用多尺度特征提取结构、通道注意力机制和边缘监督机制，增强了网络的特征提取能力，使网络对不同尺度目标的感知能力更强，对有用通道的关注力更高，对边缘细节信息更加敏感。因此，所提算法在每项指标上都取得最高的精度，表明了其有效性。

3.5.2　Satellite Dataset Ⅱ (East Asia) 数据集

在 Satellite Dataset Ⅱ (East Asia) 数据集的测试集上，将 MAE-Net 与 U-Net[6]、SiU-Net[16]、AugUst-Net[19]、RSIS-MLCA[20] 和 Ra-CGAN[1] 进行了比较，数据集划分均一致，对比结果如表 5 所示。

AugU-Net 通过对输入图像进行光谱增强操作，扩充光谱维度的样本空间，将原始图像重新采样作为新的输入样本。RSIS-MLCA 是一种基于编码解码结构的网络结构，其在网络编码阶段加入了通道注意力机制，能比较充分地对通道信息进行筛选，提高了分割精度。从表 5 可以看出，相较于其他网络，MAE-Net 在 IOU、R 和 P 方面均有提高，建筑物分割效果更好。但在 R 上，对比方法 U-Net 要比 MAE-Net 高一些，这是因为 R 表示预测正确的建筑物像素数占全部建筑物像素数的比例，高的 R 是因为预测正确的建筑物像素数多，但这也会使网络将更多背景分为建筑物，带来更多的误判，从而降低 P。对于对比方法 U-Net，虽然其 R 比所提网络高，但是由于其网络对背景的更高误检率，导致其 P 要比所提网络低 17.5 个百分点。对于对比方法 Ra-CGAN，其将网络 RSIS-MLCA 作为生成网

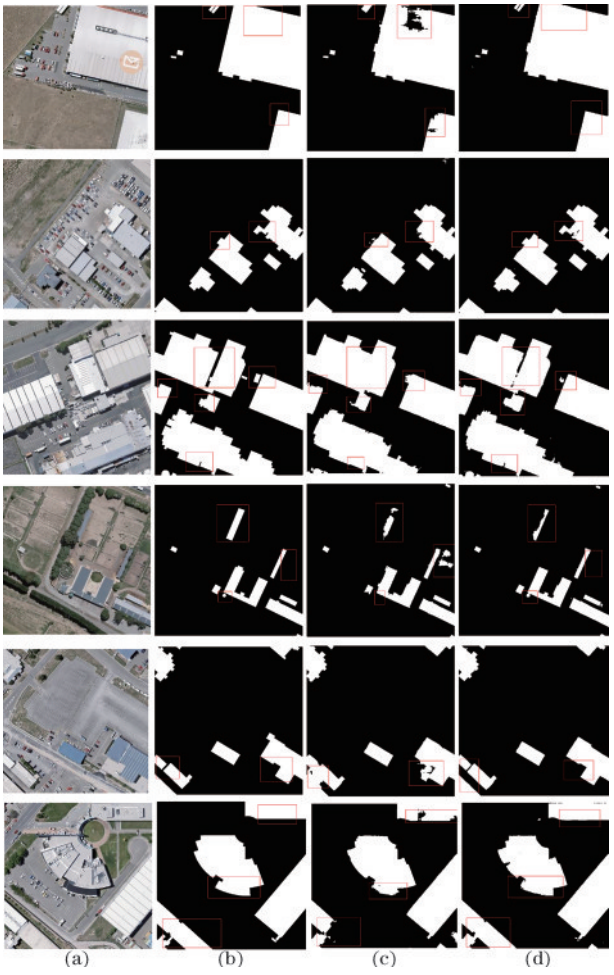

图 7　WHU Building 验证集中深度模型分割对比结果。(a) 原始图像；(b) 标签图；(c) U-Net 分割结果；(d) MAE-Net 分割结果

Fig. 7　Segmentation results of comparative experiment on WHU Building validation set. (a) Original images; (b) image label; (c) U-Net processing result; (d) MAE-Net processing result

表 5　不同方法在 Satellite Dataset Ⅱ (East Asia) 测试集上的各项指标对比

Table 5　Experimental results of different methods with the Satellite Dataset Ⅱ (East Asia) test set

Method	F_1-score	P	R	IOU
U-Net[6]	0.746	0.653	0.869	0.594
SiU-Net[16]	0.758	0.725	0.796	0.611
AugU-Net[19]	0.780			0.640
RSIS-MLCA[20]	0.796	0.826	0.768	0.661
Ra-CGAN[1]	0.812	0.852	0.775	0.677
MAE-Net	0.802	0.828	0.805	0.690

络，再使用一个判别网络模型，采用对抗训练的方式对生成网络输出的分割结果图进行矫正，以此来生成最终的分割结果图。此网络在关注通道注意力的基础之上，还使用了对抗训练，加强了网络对高阶数据分布特征的学习，使其在数据样本不均衡的条件下，改善了分割效果，所以其分割结果在指标 P 上要比所提网络高，这是对抗训练在数据样本不充足条件下的优势。

为了更客观地对不同方法的性能进行综合评估，降低由于数据集正负样本严重不匹配所造成的影响，使用整体评价指标IOU来进行网络性能评价。从表5可以看出，MAE-Net在此数据集上IOU评价值最高为69.0%，相比于对比方法U-Net、SiU-Net、AugUst-Net、RSIS-MLCA和Ra-CGAN分别提高了9.6个百分点、7.9个百分点、5.0个百分点、2.9个百分点和1.3个百分点。图8给出了MAE-Net与基线网络U-Net的实验对比结果。从图8可以看出，U-Net模型存在建筑物无法完整识别和建筑物漏分的现象，而MAE-Net模型表现较好，能很好分割出不同尺度的建筑物，改善建筑物无法完整识别和建筑物漏分的现象。这表明MAE-Net中的MSA和EEB是有效的，可以提升对遥感图像建筑物的分割效果。

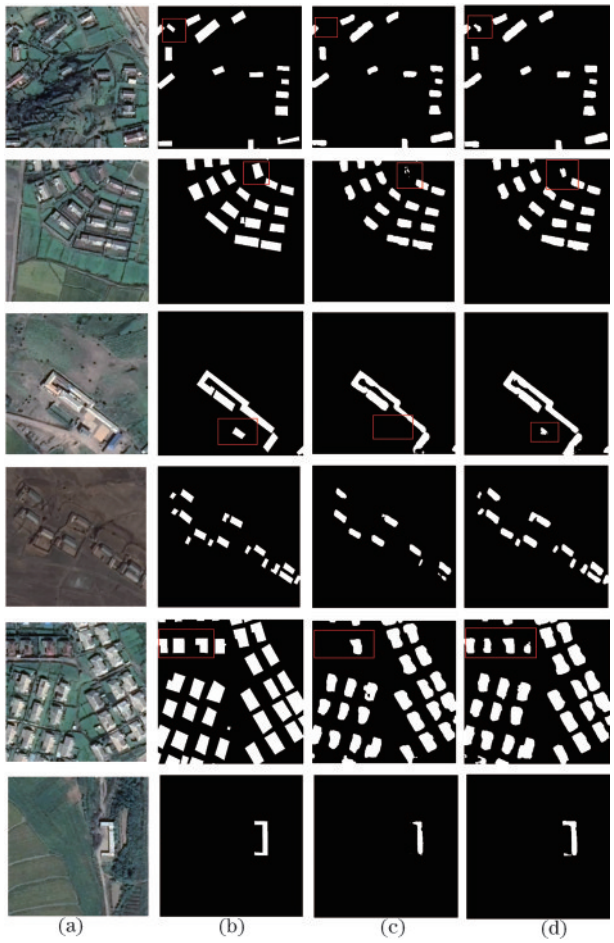

图8 Satellite Dataset Ⅱ(East Asia)测试集中深度模型分割对比结果。(a)原始图像；(b)标签图；(c) U-Net分割结果；(d) MAE-Net分割结果

Fig. 8 Segmentation results of comparative experiment on Satellite Dataset Ⅱ (East Asia) test set. (a) Original images; (b) image label; (c) U-Net processing result; (d) MAE-Net processing result

4 结 论

基于U-Net提出了结合多尺度注意力和边缘监督的分割网络，它是端到端卷积神经网络模型，可用于遥感图像建筑物的分割。在编码器中添加多尺度注意力模块，增加了网络对不同大小建筑物特征的提取能力，在保持空间信息的情况下，提取更多不同尺度建筑物局部细节信息。通道注意力机制显式地建立不同尺度特征通道间的依赖关系，强化了有用特征，提升了网络对目标特征的获取能力。在解码器中添加边缘监督网络，通过计算预测边缘和边缘标签的损失，指导监督图像分割网络，提升了分割网络对建筑物边缘的感知能力，改善了建筑物错分、漏分，建筑物轮廓不准确的问题。实验结果表明，较原始U-Net模型以及其他施加注意力和特征融合的分割模型，MAE-Net对遥感图像中建筑物细节特征信息保留更加完整，在建筑物边缘处具有更高的分割准确度，对尺度变化较大的建筑物分割更加准确，增强了建筑物以及建筑物边缘的完整性，提升了分割精确度。

参 考 文 献

[1] 余帅，汪西莉. 含多级通道注意力机制的CGAN遥感图像建筑物分割[J]. 中国图象图形学报，2021，26(3)：686-699.
Yu S，Wang X L. Remote sensing building segmentation by CGAN with multilevel channel attention mechanism [J]. Journal of Image and Graphics，2021，26(3)：686-699.

[2] Yang G，Zhang Q，Zhang G X. EANet：edge-aware network for the extraction of buildings from aerial images [J]. Remote Sensing，2020，12(13)：2161.

[3] 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.
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.

[4] 陈小龙，赵骥，陈思溢. 基于注意力编码的轻量化语义分割网络[J]. 激光与光电子学进展，2021，58(14)：1410012.
Chen X L，Zhao J，Chen S Y. Lightweight semantic segmentation network based on attention coding[J]. Laser & Optoelectronics Progress，2021，58(14)：1410012.

[5] Long J，Shelhamer E，Darrell T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition，June 7-12，2015，Boston，MA，USA. New York：IEEE Press，2015：3431-3440.

[6] Ronneberger O，Fischer P，Brox T. U-net：convolutional networks for biomedical image segmentation[M]//Navab N，

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

[7] 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, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[8] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[9] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[10] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 3141-3149.

[11] Qin X B, Zhang Z C, Huang C Y, et al. BASNet: boundary-aware salient object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 7471-7481.

[12] Yu C Q, Wang J B, Peng C, et al. Learning a discriminative feature network for semantic segmentation [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1857-1866.

[13] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]//International Conference on Machine Learning, July 6-11, 2015, Lille, France. Cambridge: The MIT Press, 2015: 448-456.

[14] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines[C]//proceedings of the 27th International Conference on International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel. Madison: Omnipress, 2010: 807-814.

[15] Tian Z, He T, Shen C H, et al. Decoders matter for semantic segmentation: data-dependent decoding enables flexible feature aggregation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 3121-3130.

[16] 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.

[17] Liu P H, Liu X P, Liu M X, et al. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network[J]. Remote Sensing, 2019, 11(7): 830.

[18] Liu H, Luo J, Huang B, et al. DE-net: deep encoding network for building extraction from high-resolution remote sensing imagery[J]. Remote Sensing, 2019, 11 (20): 2380.

[19] Maggiori E, Tarabalka Y, Charpiat G, et al. High-resolution aerial image labeling with convolutional neural networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(12): 7092-7103.

[20] 余帅, 汪西莉. 基于多级通道注意力的遥感图像分割方法[J]. 激光与光电子学进展, 2020, 57(4): 041012.
Yu S, Wang X L. Remote sensing image segmentation method based on multi-level channel attention[J]. Laser & Optoelectronics Progress, 2020, 57(4): 041012.