

基于 HED—UNet 遥感图像建筑物语义分割

李松宇

(渤海大学 信息科学与技术学院, 辽宁 锦州 121000)

摘要:作为计算机视觉领域的关键问题之一,语义分割通过对每个像素进行密集的预测并推断标签来实现细粒度的推理。而针对遥感图像中建筑物的语义分割能够节省大量人力和时间成本。但现有语义分割技术,往往经过多次下采样,使得目标细节特征损失严重。该文将最新的 HED—UNet 网络迁移到针对遥感图像建筑物的语义分割中来。HED 网络是一个多尺度融合网络,带有深度监督机制。将 HED 网络嵌入到 U—Net 网络中,通过在多分辨率下对侧输出预测进行深度监督,使训练更为有效。并引入一种层级注意机制,将这些多尺度预测自适应地合并到最终的模型输出中。最后,在武汉大学的遥感建筑物数据集上进行了实验研究,结果表明 HED—UNet 相比于传统语义分割网络性能提升明显,PA 达到了 0.966,MIoU 达到了 0.922。

关键词:语义分割;遥感图像;深度学习;U—Net 网络;HED—UNet

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2022)0058-06

Semantic Segmentation of Buildings in Remote Sensing Images Based on HED—UNet

LI Song-yu

(School of Information Science and Technology, Bohai University, Jinzhou 121000, China)

Abstract:As one of the key issues in the field of computer vision, semantic segmentation implements fine-grained reasoning by intensively predicting each pixel and inferring labels. The semantic segmentation of buildings in remote sensing images can save a lot of manpower and time costs. However, the existing semantic segmentation technology often undergoes multiple downsampling, resulting in serious loss of target detail features. This article migrates the latest HED—UNet network to the semantic segmentation of remote sensing image buildings. The HED network is a multi-scale fusion network with a deep supervision mechanism. The HED network is embedded in the U—Net network, and the training is more effective through deep supervision of the side output prediction under multi-resolution. And introduce a hierarchical attention mechanism to adaptively incorporate these multi-scale predictions into the final model output. Finally, an experimental study was conducted on the remote sensing building data set of Wuhan University. The results showed that compared with traditional semantic segmentation network, the performance of HED—UNet was significantly improved, with PA reaching 96.6% and MIoU reaching 92.2%.

Key words:semantic segmentation; remote sensing image; deep learning; U—Net; HED—UNet

0 引言

目前,建筑对象是重要的地物之一,在许多应用中发挥着重要作用,如地理信息系统、城市规划等地理空间相关应用^[1]。随着传感器技术的迅速发展,伴随着每天都有大量的遥感数据产生。随着高分辨率遥感数据变得更加实用和经济合理,因此,从快速更新、价格合理的遥感影像中提取建筑对象已成为一个重要的现实利益。由于手动描绘建筑对象需要花费大量的精力和时间。遥感技术和新颖的数据科学为自动建筑检测提供了可能,从而有助于动态绘制城市地图,并消除大

量手工工作。语义分割是计算机视觉领域的一个中心课题。语义分割可以理解图像的像素级别,即对图像中的每个像素进行分类,将图像分割成几个感兴趣的对象,并为每个对象分配特定的标签类型^[2]。传统的遥感图像分割方法主要通过提取图像的灰度、颜色、空间纹理、几何特征等底层特征,将图像分割为几个互不连通的区域。低水平特征在同一区域表现出一致性或相似性,但在不同区域存在显著差异。这导致精度差,效率低,泛化能力差。例如,阈值的分割算法 Otsu^[3]根据预定义的阈值将图像划分为不同的区域。但只考

收稿日期:2022-01-04

作者简介:李松宇(1998-),男,硕士生,研究方向为深度学习、图像分割。

虑像素点灰度值本身的特征,一般不考虑空间特征,因此对噪声比较敏感,鲁棒性不高;分水岭变换(WT)算法^[4]是一种基于拓扑理论的数学形态学分割方法;归一化切割(N-Cut)算法^[5]通过提取图像的全局印象来分割图像。

在深度学习算法被广泛应用之前,机器学习方法已经被提出用于遥感图像的语义分割。Mitra 等人^[6]利用改进后的支持向量机(SVM)算法对多光谱遥感图像中的监督像素进行分类。该算法使用少量标记像素对不同的土地覆盖类型进行分割,解决了监督分类中标记像素缺失的问题。Zhang 等人^[7]提出了一种用于合成孔径雷达(SAR)图像分割的改进条件随机场(CRF)^[8]方法。这种方法可以在信息的上下文中捕捉不同尺度的空间结构,从而更好地分割图像的边缘位置。

由于深度学习的快速的发展^[9-10],卷积神经网络(CNNs)在图像分割和图像分类方面取得了不错的进展,并且广泛地应用于语义分割的应用。基于 CNN 基本结构开发的 LeNet^[11]和 AlexNet^[12]等分类模型难以应用于分割,一方面,减少了全连接层参数的数量和维度,导致丢失了物体的详细位置信息;另一方面的限制是输入图像的大小是固定的。为了解决这一问题,Long 等人^[13]提出了一个全卷积网络(FCN),它通过将现有模型的最后一层全连接层更改为 1×1 卷积层来消除输入图像的尺寸限制。此外,生成并保留了像素位置信息的热图,通过转置卷积生成的图像与输入的图像大小相同,其中对所有像素进行分类^[14]。然而,FCN 的分类结果对于大目标是碎片化的,对于小目标的准确率较低。特征图的上采样也导致了分辨率恢复的问题,例如物体边界模糊或描述能力丧失。Badrinarayanan 等人^[15]提出了具有编码器-解码器结构的 SegNet 网络。编码器通过池化层逐渐减小输入特征图的大小,解码器与编码器对称,并通过上采样逐渐恢复图像的细节和特征图的大小。

Ronneberger 等^[16]提出了 U-Net 网络,它结合了特征提取层提取的信息,在上采样过程中恢复了原始图像。但经典的 U-Net 网络,由于多次降采样后对图像进行了大量压缩,导致目标的详细特征信息损失严重以及降低了图像的分辨率,随后的解码网络采取 4 次上采样得到最终的提取结果^[18]。U-Net 只有在最后一层进行图像的像素类别分类,虽然 U-Net 利用了编码部分的收集到的信息,但对于多尺度信息的泛化能力是有限的。针对经典的 U-Net 网络提取地物存在细节特征信息损失的问题以及泛化能力有限问题,Heidler 等人^[17]提出了 HED-UNet 网络模型,对南极海岸线进行语义分割,其中分割的效果比较显著。此

网络,在 U-Net 网络基础上,加入特征金字塔(FPN),并且将 HED 模型迁移在 U-Net 上,使用了具有跳过连接的编码-解码器结构来预测多分辨率的分割掩码。可以很好的解决 U-Net 网络丢失浅层信息的缺点。并且,此网络允许模型在不同分辨率级别输出预测,对这些侧输出添加深度监督可以提高模型的训练效率和泛化性能。而且使用层级注意力机制方法来实现特征图像多尺度自适应的融合。

1 相关工作

1.1 语义分割

图像识别是图像处理的一个重要领域。在图像识别中,图像分类与检测以及图像分割是经常讨论的问题,图像分类与检测可以提供物体识别和位置信息,而图像分割可以确定图像中所有像素的类别。图像分割包括实例分割和语义分割。语义分割是把图像中每一个像素都分配一个类别标签。而实例分割不但要对每个像素进行分类,还要将同一类的物体区分出不同的个体。由于图像的种类多样并且复杂,现有的图像识别任务很难实现,因为图像识别算法需要非常复杂的计算过程和大量的数据和内存。近年来,随着 GPU 等硬件的发展,以及基于 CNN 的深度学习在图像识别领域的出色表现,利用深度学习在遥感领域开展了许多研究。

1.2 HED 网络

HED 网络^[19](holistically-nested edge detection)是用于边缘检测的方法,HED 网络是一种多尺度多融合的网络结构,如图 1 所示,在 VGG16 网络前五层加入侧输出,将每一组的最后一个卷积层的输出提取出来,但由于每一组得到的图片的大小尺寸不同,所以需要使用转置卷积对每一组得到的尺寸分别扩大 2 至 16 倍,这样在每个尺度的上得到的图片,都是相同的大小。并且在最后一层,通过训练权重来融合上面各层的边缘结果。

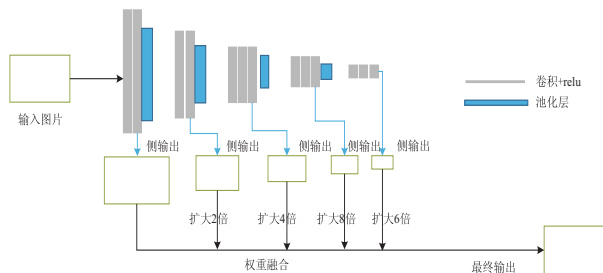


图 1 HED 网络模型

1.3 U-Net 模型

经典的语义分割模型 U-Net 如图 2 所示,它主要分为两个部分,第一个部分是 Encoder,即编码部分,通过卷积操作和下采样操作来减小图片的尺寸,来提取

浅层的特征,通过编码部分可以生成五个有效的特征层。第二部分是 Decoder,即解码部分,利用编码部分生成的 5 个特征层进行上采样,并且进行 copy and crop,即特征融合,得到所有有效特征层的最终融合。最终利用最后一层的有效特征层,通过 softmax 对每一个特征点进行分类。

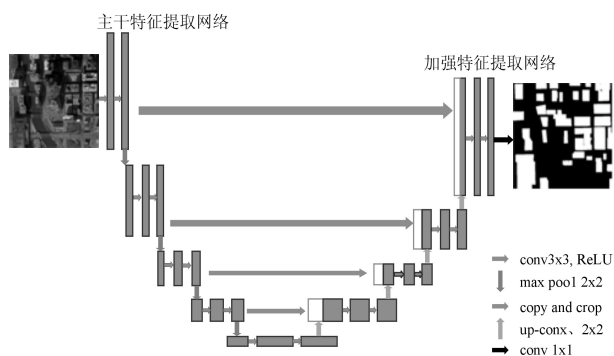


图 2 U-Net 网络模型

U-Net 只有在最后一层进行图像像素类别分类,尽管 U-Net 利用了来自编码部分的收集到的信息,但

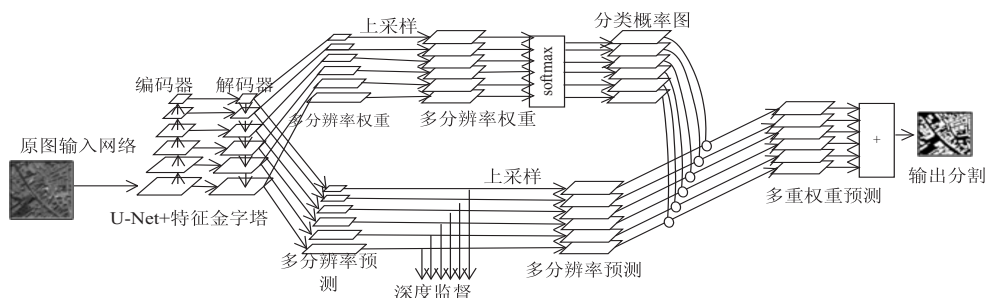


图 3 HED-UNet 网络模型

此模型受到由于 HED 网络模型层次化结构的启发,为每个特征图添加一个用于语义分割的侧输出,总共有 6 个输出,这些多尺度输出以两种不同的方式使用。

2.1.1 深度监督

在构建这个模型的特征金字塔时,模型可能没有太多的动机将有意义和信息丰富的特征编码映射到最低分辨率的特征图上。为了明确提供这种动机,加入深度监督,以便模型能够预测特征金字塔中每个特征图的真实值。

众所周知,这种所谓的深度监督可以提高神经网络的学习效率以及泛化能力。这是通过训练中间层输出的真实数据,以提供额外的、更直接的训练,反馈给早期的层来实现的。在此模型中,深度监督为每一个多分辨率预测创建一个相应的真实分割,然后将这些多尺度真实样本与预测样本进行比较,以提供额外的损失项。由此产生的深度监督促进网络更好的捕捉更大的结构,并通过在有意义的深层编码的特征来利用可用的感受野。

对于多尺度信息的泛化能力是有限的。

2 本文模型

2.1 模型结构

HED-UNet 网络是一个混合模型,将 HED 迁移在 U-Net 网络上如图 3 所示,此模型在 U-Net 网络的基础上,增加了 HED 的思想。并且引入的深度监督以及注意力机制方法。

从编码器-解码器主干开始,此模型需要一个相对较大的空间上下文窗口来作为其决策的基础,并且该模型添加了一个特征金字塔(FPN),该特征金字塔具有 6 个分辨率级别,对应于 5 个下采样和上采样。在这个特征金字塔中,最精细的特征图是全图像分辨率,最粗糙的特征图是分辨率的 1/32 的特征图。为了覆盖任务所需的足够大的感受野,这里选择 6 个分辨率级别。在解码器部分,融合浅层和深层的信息。浅层的网络关注细节特征,而高层的网络关注于深层语义信息。

2.1.2 多尺度融合的策略

在下一步,此结构是利用了 HED 网络的思想。将中间层的输出结合成一个全分辨率预测。通过这种方式,此模型可以将靠近边缘的细粒度轮廓与远离边界的更强大的高级预测结合起来。HED 网络中使用的合并方式是使用学习到的权重来合并侧输出预测。为了进一步提高合并性能,提出了基于层级注意力的合并机制。

2.2 层级注意力机制

对遥感建筑物分割时,并不需要关注给定场景的所有区域的细节,所以此模型引入层级注意力机制,实现多尺度自适应的融合。

这种注意力合并机制最初是在自然语言处理中作为单词和标记之间的顺序注意进行了探索,后来也被应用到计算机视觉中作为图像中的空间注意。此模型利用层级注意力机制结合多尺度预测。在此模型中,这种注意力使得网络关注于它认为对当前场景的每个像素最有用的特征,而不是为特征融合设置固定的权重。

因此,既不是顺序注意力也不是空间注意力,而是允许模型关注不同的分辨率级别。

对于每个预测级别,都会创建一个权重图。然后对权重图进行上采样以匹配输出分辨率,并通过 softmax 函数将其转换为分类概率图。为了获得最终预测,计算预测和注意力掩码之间的点积。

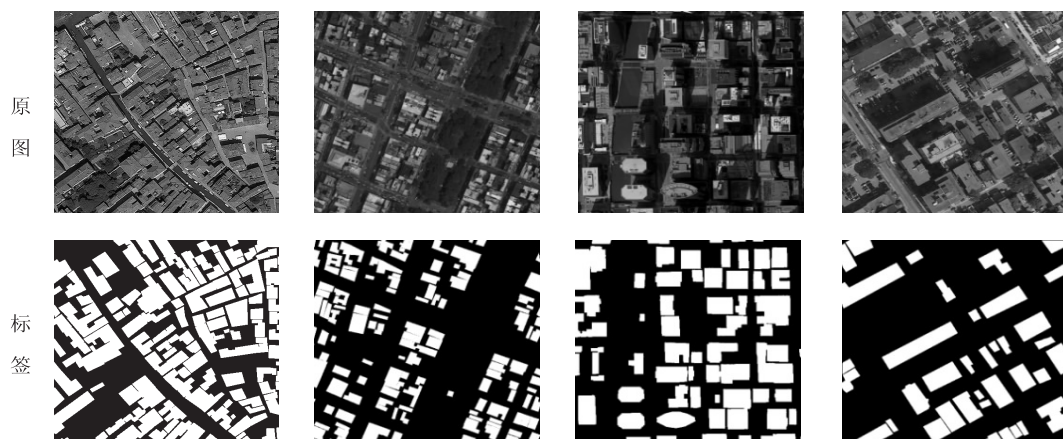


图 A

图 B

图 C

图 D

图 4 数据集部分展示

3.1 图像预处理

图像扩增 (image augmentation) 技术是通过对训练图像做一系列随机改变,来产生相似但有不同的训练样本,从而增加训练数据集的数量,在训练模型时,

样本的数量大,使得训练出的模型具有广泛性。本实验使用的是 Augmentor 库,对图片进行一系列的旋转、缩放等操作来获取更多的训练样本。扩增的图片如图 5 所示。

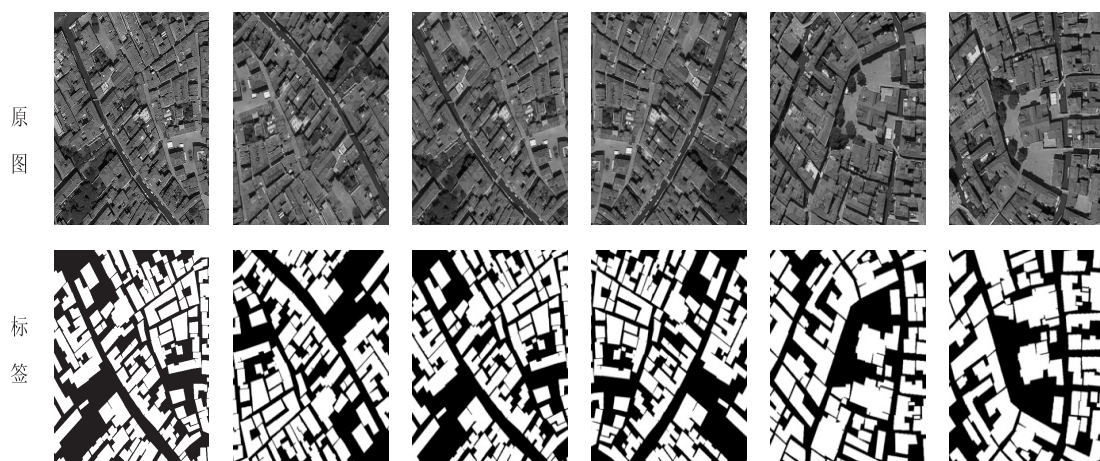


图 5 图像扩增

第 1 列是原始图像和标签图,右数 5 列是通过图像扩增后生成的原图和标签图。

3.2 网络参数的设置

实验环境见表 1。

表 1 实验环境

系统环境	CPU	显卡	编译环境	运行环境
Windows 10	Core (TM) i7-11370H	Nvidia GeForce RTX 3070Ti	python3.7	Pytorch-1.8-GPU

实验采用了遥感图数据集的 500 张图片,其中对图片进行了去噪和增强,选用其中 400 进行图像扩增,生成 1 000 张图片,其中选用 800 张图片作为训练集,200 张图片作为验证集,剩余的 100 张没有进行图像

扩增的图片作为测试集,使学习的结果具有泛化性。

其中输入图片的大小为 512×512 像素, Batch size 大小为 2,迭代次数为 100 次。取训练时选 0.000 1 为初始学习率,为了防止迭代次数增多出现局部最小值,

学习率采取指数衰减型,动量因子设置为 0.5。训练过程中使用监测器对测试集损失值进行监测,当损失值连续 50 次没有下降,则认为模型训练完毕,训练提前停止。

3.3 损失函数

使用了对二元交叉熵损失的自适应平衡修正,对于将地面真实划分为正像素 Y_+ 和负像素 Y_- 和预测 $\hat{p}$ 的单一图像,给出为:

$$L(\hat{p}) = -\frac{|Y_-| \sum_{j \in Y_-} \log \hat{p}_j}{|Y_+ \cup Y_-|} - \frac{|Y_+| \sum_{j \in Y_+} \log(1 - \hat{p}_j)}{|Y_+ \cup Y_-|} \quad (1)$$

无论两个类大小之间的比率如何,此损失函数都会对正类和负类赋予相同的权重。由于这个特性于语义分割,因此,它被用作两个任务的损失函数。

3.4 实验评价指标

实验结果评价指标采用精准率 (Pixel Accuracy) 和平均交并比^[2] (MIoU)。

像素精度 (PA): 它是语义分割的度量标准,计算正确分类的像素数量与其总数之间的比率。

$$PA = \frac{\sum_{i=0}^k P_{ii}}{\sum_{i=0}^k \sum_{j=0}^k P_{ij}} \quad (2)$$

MIoU: 这是用于分割目的的标准度量。它计算两个集合的交集和并集之间的比率,该比率可以转换为真阳性 (交集) 的数量与真阳性、假阴性和假阳性 (并集) 的总和。该 IoU 每个类别是在每个类的基础上计算的,然后取平均值。

$$MIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{P_{ii}}{\sum_{j=0}^k P_{ij} + \sum_{j=0}^k P_{ji} - P_{ii}} \quad (3)$$

其中, k 为类别总数; P_{ij} 为类别 i 的像素被推断为 j 类别像素的总数; P_{ji} 为类别 j 的像素被推断为 i 类别像素的总数; P_{ii} 为类别 i 的像素被推断为类别 i 像素的总数; P_{ij} 和 P_{ji} 属于被分错的像素数量; P_{ii} 属于被分对的像素数量。

3.5 实验结果与分析

首先,将数据集分为训练集、验证集和测试集。然后在训练集上训练不同的网络模型,并在验证集上进行验证。最后,在测试集上对训练后的模型进行评估。

图 6 和表 2 分别是各实验模型建筑物分割效果对比图和模型测试结果的汇集。实验分别比较了 HED-UNet、U-Net 以及传统方法 FCN8s、FCN16s 以及 FCN32s, HED-UNet 表现出了非常好的性能,其 PA 指标比起 U-Net 网络上升 5.2%, MIoU 指标达到 0.922,效果提高的明显。

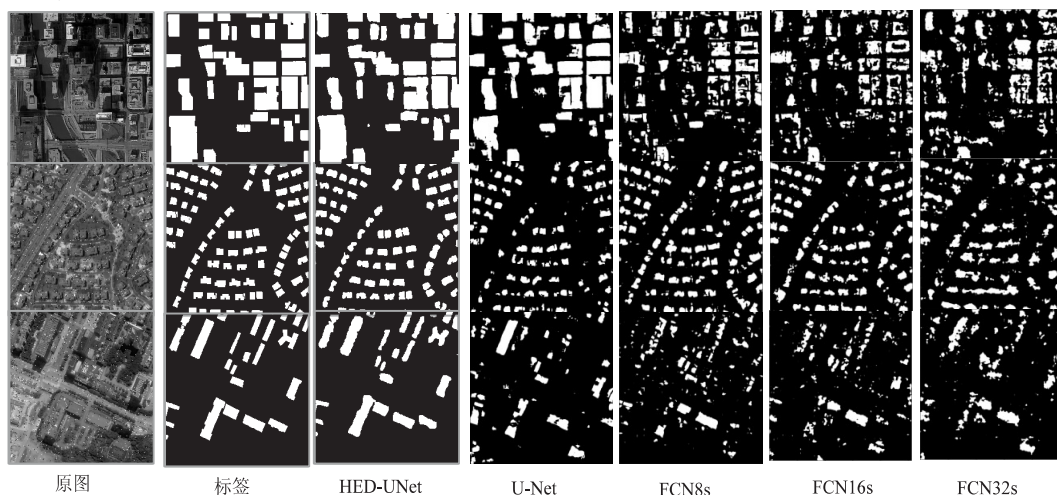


图 6 各网络模型预测

表 2 实验结果

实验名称	Pixel Accuracy	MIoU
HED-UNet	0.966	0.922
U-Net	0.914	0.844
FCN-8s	0.877	0.761
FCN-16s	0.878	0.728
FCN-32s	0.861	0.662

图 6 是分别在上述网络中进行测试,左一是原图,左二是标签图,左三是 HED-UNet 的测试结果,左四

是 U-Net 的测试结果,右三至右一次分别是 FCN8s、FCN16s 以及 FCN32s 的测试结果从测试图上也可以明显看出 HED-UNet 预测的图接近标签图,提取的建筑物基本上完整,而且边界信息相比于其他传统网络预测也有明显的提升,内部空洞较少。

4 结束语

研究了对遥感建筑图像的语义分割,采用比较新颖的 HED-UNet 模型,针对传统模型目标的细节特征

信息损失严重。此模型引入了 HED 思想、多尺度融合的网络结构、深度监督以及注意力合并机制,大多数性能的提高来自于自适应的训练程序和一些额外的层,与已经存在的层相比,它们不需要大量的计算资源。该算法提取建筑物的像素准确率可达到 0.966,MIoU 达到 0.922,实验效果表明了该算法在遥感建筑物语义分割方面的优越性。

参考文献:

- [1] VAKALOPOULOU M, KARANTZALOS K, KOMODAKIS N, et al. Building detection in very high resolution multispectral data with deep learning features[C]//2015 IEEE international geoscience and remote sensing symposium (IGARSS). [s. l.]: IEEE, 2015: 1873–1876.
- [2] GARCIA-GARCIA A, ORTIZ-ESCOLANO S, OPREA S, et al. A review on deep learning techniques applied to semantic segmentation[J]. arXiv:1704.06857, 2017.
- [3] OTSU N. A threshold selection method from gray-level histograms[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62–66.
- [4] LIU Y, MOORE J C, CHENG X, et al. Ocean-driven thinning enhances iceberg calving and retreat of Antarctic ice shelves[J]. Proceedings of the National Academy of Sciences, 2015, 112(11): 3263–3268.
- [5] SHI J, MALIK J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888–905.
- [6] MITRA P, SHANKAR B U, PAL S K. Segmentation of multispectral remote sensing images using active support vector machines[J]. Pattern Recognition Letters, 2004, 25(9): 1067–1074.
- [7] ZHANG P, LI M, WU Y, et al. Hierarchical conditional random fields model for semisupervised SAR image segmentation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(9): 4933–4951.
- [8] KRÄHENBÜHL P, KOLTUN V. Efficient inference in fully connected crfs with gaussian edge potentials[J]. Advances in Neural Information Processing Systems, 2011, 24: 109–117.
- [9] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504–507.
- [10] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436–444.
- [11] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.
- [12] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25: 1097–1105.
- [13] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4): 640–651.
- [14] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). Boston, Massachusetts, USA: IEEE, 2015: 1–9.
- [15] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481–2495.
- [16] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//International conference on medical image computing and computer-assisted intervention. [s. l.]: Springer, 2015: 234–241.
- [17] HEIDLER K, MOU L, BAUMHOER C, et al. HED-UNet: combined segmentation and edge detection for monitoring the antarctic coastline[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1–14.
- [18] 徐昭洪, 刘宇, 全吉成, 等. 基于 VGG16 预编码的遥感图像建筑物语义分割[J]. 科学技术与工程, 2019, 19(17): 250–255.
- [19] XIE Saining, TU Zhuowen. Holistically-nested edge detection[C]//Proceedings of the 2015 IEEE international conference on computer vision. Santiago: IEEE, 2015: 1395–1403.