P-LINKNET :LINKNET WITH SPATIAL PYRAMID POOLING FOR HIGH-RESOLUTION SATELLITE IMAGERY

Yi Ding[1,*], Muyu Wu[2], Yongshu Xu[1], Songjiang Duan[1]

[1] Chongqing Geomatics and Remote Sensing Center,Chongqing,404100,China - dy@dl023.net,13508397853@163.com, dsj@dl023.net
[2] School of Remote Sensing and Information Engineering,Wuhan University,Wuhan,430079,China - 2935094110@qq.com

Commission III, WG III/1

KEY WORDS: Deep learning, high-resolution remote sensing imagery, building extraction, semantic segmentation, encoder-decoder, fully convolutional networks

ABSTRACT:

Automatic extraction of buildings from high-resolution remote sensing imagery is very useful in many applications such as city management, mapping, urban planning and geographic information updating. Although extensively studied in the past years, due to the general texture of the building and the complexity of the image background, high-precision building segmentation from high-resolution sensing image is still a challenging task. Repeated pooling and striding operations used in CNNs reduce feature resolutions and cause the loss of detail information. In order to solve this problem, we proposed a deep learning model with a spatial pyramid pooling module based on the LinkNet. The proposed model called P-LinkNet that takes advantage of a spatial pyramid pooling module to capture and aggregate multi-scale contextual information. We tested it on Inria Building dataset. Experimental results show that the proposed P-LinkNet is superior to the LinkNet.

1. INTRODUCTION

Automatic extraction of buildings from remote sensing imagery is used in many applications, including urban planning, navigation, and disaster management [Panboonyuen et al., 2019, Wu et al., 2018, Li et al., 2018, Liu et al., 2019]. In recent years, the capability of remote sensing technology has been greatly improved, which leads to the availability and accessibility of high-resolution remote sensing images [Hui et al., 2019, Huang et al., 2016, Guo et al., 2016]. With the use of quality data in large spatially areas, it is possible to perform accurate image segmentation targeting the extraction of buildings.

In the past few decades, various methods of extracting features from images have been widely developed. In traditional classical methods, the spatial and texture features of images are extracted by mathematical descriptors, such as Haar spaces [Viola, Jones, 2001], Scale-invariant Feature Transform(SIFT) [Lowe, 2004], Local Binary Patterns(LBP) [Ojala et al., 2002] and Grey Level Co-occurence Matrix(GLCM) [Gomez et al., 2012]. Since entering the new world, pixel by pixel prediction has been introduced on the basis of feature extraction by classifier such as Support Vector Machines(SVM) [Inglada, 2007], Adaptive Boosting(AdaBoost) [Aytekin et al., 2013], Random Forests [Dong et al., 2015], K-Means [Cheng et al., 2013], and Conditional Random Fields(CRF) [Li et al., 2015]. However, these methods rely heavily on manual design and implementation, which change as the application domain changes. As a result, they are prone to introduce biases and poor generalization, and are time-consuming and labor-intensive.

Fortunately, alongside advancements in computational capabilities and the availability of large volumes of data, deep learning is on the rise, especially the convolutional neural network(CNN), bringing us new solutions, as it can automatically learn effective classification features. In recent years, with the development of semantic segmentation technology, building extraction from high-resolution satellite imagery has been continuously improved.

In the early studies, the semantic tags were determined independently by pixel by pixel by the CNN model based on patch, which only relied on a small patch around the target pixel to predict the tags and ignored the internal relationship between patches. The CNN model based on patch has achieved remarkable results in the extraction of buildings, but it cannot guarantee the spatial continuity and integrity of the building structure [Lin, Saripalli, 2012, Vakalopoulou et al., 2015]. In addition, the CNN method based on patch is very time-consuming.

To overcome the problems of patch-based CNNs, Long et al. proposed the fully convolutional networks(FCNs) [Long et al., 2015], which have become a mainstream paradigm for semantic segmentation. FCNs replace the fully connected layers in traditional CNNs with convolutional layers and upsampling layers, which are applicable to the segmentation of images of any size. In the FCN, feature maps with high-level semantics but low resolutions are genreated by down-sampling features using multiple pooling or convolutions with strides [Chen et al., 2018]. Based on the basic FCN8 model, most excellent semantic segmentation networks have been proposed. For example,SegNet [Badrinarayanan et al., 2017] and U-Net [Ronneberger et al., 2015] used the encoder-decoder structure to improve the segmentation accuracy and DeepLab [Chen et al., 2015] used the dilated convolution to enlarge model receptive field. However, yet there is still much room to exploit necessary information in complex scenes. To make good use of global image-level prior for diverse scene understanding, method of [Lazebnik et al., 2006, Lucchi et al., 2011] extracted flobal context information with traditional feature not from deep neural networks. Similar improvement was mader under object detection frame-

*Corresponding author

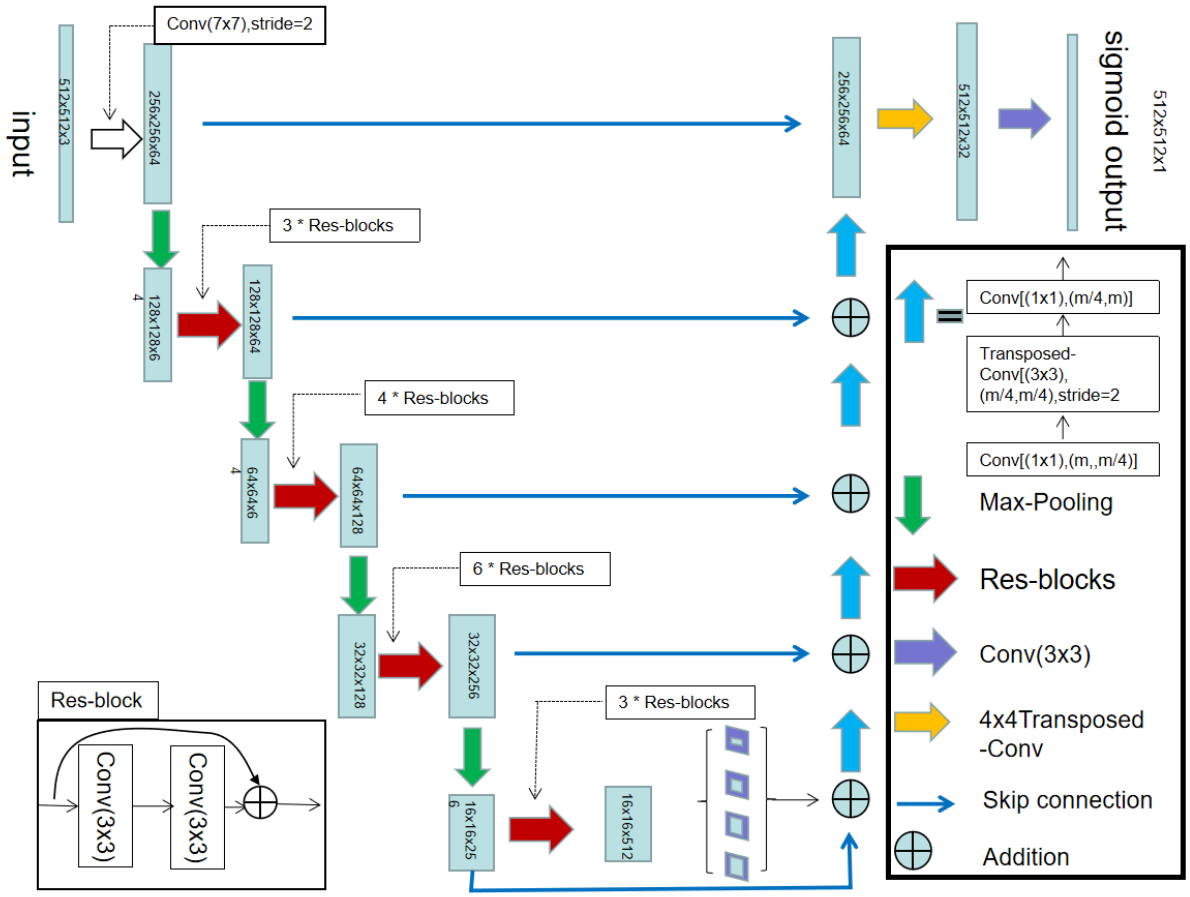


Figure 1. P-LinkNet architecture. Each gem green rectangular block represents a multi-channel feature map. The left part of the diagram represents the encoder. P-LinkNet uses ResNet34 as encoder. The right part is the decoder of P-LinkNet, it is set the same as LinkNet decoder. Compared with LinkNet, P-LinkNet has an additional center part which can enlarge the receptive field and as well as preserve the detailed information of different scales. Each convolution layer is followed by a ReLU activation except the last convolution layer which use sigmoid activation.

works [Szegedy et al., 2014].

Although extensively studied in the past years, due to the general texture of the building and the complexity of the image background, high-precision building segmentation from high-resolution sensing image is still a challenging task. On the one hand, the satellite images are high-resolution, so the network should have large receptive field that can cover the whole image. On the other hand, some buildings in the images are small and with complex shape. In this case, it is of the great significance to preserve the detailed spatial information. Taking these factors into account, we proposed a deep learning model called P-LinkNet, which can properly handle these challenges.

The spatial pyramid pooling structure of the PSPNet [Zhao et al., 2017] model aim to handle the problems of segementing objects at different spatial scales. The approach is well-known for achieving robust and efficient performance for dense semantic labeling. The network structure consists of several branches of dilated convolution operations to enlarge the receptive field. Spatial pyramid pooling shows better performance at pixel-level prediction tasks such as scene parsing and semantic segmentation. However, the PSPNet model only utilizes FCN based on ResNet as the backbone and lacks up-sampling capabilities. LinkNet [Chaurasia, Culurciello, 2017] is an effecient semantic segmentation neural network which takes the advantage of skip connections, residual blocks and encoder-decoder architecture. It combines both lower and higher layer to gerner-ate the final result. At the same the time, it runs fast. Although it achieves better performance, it has no extraction of multi-scale contextual features capability.

P-LinkNet makes full use of the advantages of LinkNet and PSPNet. It uses LinkNet with pretrained encoder as its backbone and has a traditional spatial pyramid pooling module in the center part. By combining the encoder-decoder structure and the spatial pyramid pooling module, the proposed P-LinkNet can capture multi-scale features and effectively restore detailed context information of buildings at all scales.

Transfer learning is a efficient method that can directly improve network performance in most situation, especially when the training data is limitd. In semantic segmentation field, initializing encoders with ImageNet pretrained weights has shown promissing results.

2. METHOD

2.1 Network Architecture

In this section, we introduce each component of our efficient network P-LinkNet. It combines a U-shaped encoder-decoder

structure with spatial pyramid pooling. In the following P-LinkNet is explained in more details.

LinkNet is a typical fully convolutional network which comes from UNet. The main structure of LinkNet is U-shape. Compared to UNet, in order to reduce the computation, LinkNet don't concatate the feature map of encoder and the feature map of decoder directly, but through a convolutional layer, and then add directly.

In a deep neural network, the size of its receptive field can reflect how much information is used in the context. Using pooling layers could multiply increase the receptive field of feature points,but may reduce the resolution of center feature maps and drop spacial information. Global average pooling is a good model as the global contextual prior, which is commonly used in image classification [He et al., 2016]. To further reduce context information loss, we propose a hierarchical global prior, containing information with different scales and varying among different sub-regions. The proposed model called P-LinkNet that takes advantage of a spatial pyramid pooling module to capture and aggregate multi-scale contextual information. As shown in Figure1.

Spatial pyramid pooling is a useful method to use global information at four different scales. Through pyramid pooling, spatial features on four different spatial sacles can be identified. After encoding by the backbone, the following pyramid level separates the feature map into different sub-regions and forms pooled representation for different locations. The output of different levels in the pyramid pooling modules contains the feature map with different sizes. To get the weight of global feature, we use 1x1 convolution layer after each pyramid level to reduce dimension of context representation to 1/N of the original one if the level size of pyramid is N. Then we upsample the low-dimension feature maps to get the same size as the input feature map by bilinear interpolation. Finally, we concatenate the different levels of features and the input feature map. Note that eh number and size of pyramid levels can be adjusted. As shown in Figure2.

To explain our structure, P-LinkNet provides an effective global contextual prior-level scene parsing. The pyramid pooling module can collect levels of information, more representative than global pooling [**?**]. In end-to-end learning, the global pyramid pooling module and the local LinkNet feature can be optimized simultaneously.

The number of pyramid levels and size of each level can be modified, which are related to the size of feature map that is fed into the pyramid pooling layer. This structure extracts different subregions through pooling kernels of different sizes. Thus, the multi-stage kernels should keep in a reasonable gap. Our pyramid pooling module is four-level one with size of 1x1, 2x2, 3x3, 6x6 respectively. By our four-level pyramid, the pooling kernel cover the whole, half of, and small portions the feature map, thereby gets the contextual information. Note that the number and the size of pyramid levels can be modified, which according to the size of the feature map fed into the pyramid pooling layer.

2.2 Pretrained Encoder and Decoder

Transfer learning is a useful method that can directly improve network performance in most situation and reduce the computation. In the training period, we found that transfer learning

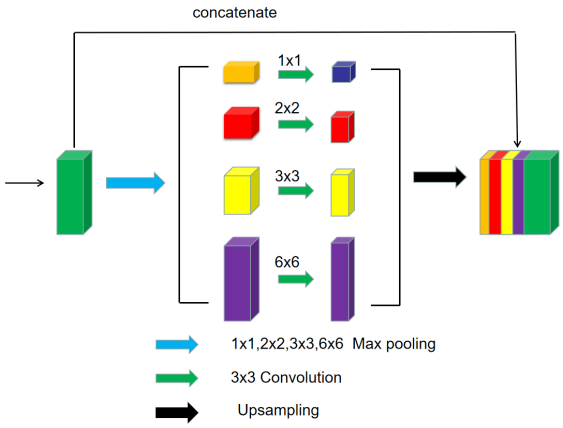


Figure 2. An illustration of the spatial pyramid pooling module structure with four pooling scales in P-LinkNet.

can accelerate our model convergence and make it have better performance without extra cost.

Deep pre-trained networks lead to good performance. However, increasing depth of the network layer may bring additional optimization difficulty. ResNet solves the problem with skip connection in each block. Latter layers of deep ResNet [He et al., 2016] mainly learn residues based on previous ones. P-LinkNet uses ResNet34 pre-trained on ImageNet [Deng et al., 2009] dataset as its encoder, which shows high precision and runs fast.

ResNet34 is originally designed for classification task on mid-resolution images of size 256x256. But in our experiment, the task is to segment buildings, which is usually to resize 512x512. After encoding, the feature map size is 1/32 of the input image. Then we use the pyramid pooling module to obtain context information. The decoder of P-LinkNet remains the same size as the original LinkNet, which has been proven effective. The decoder part uses transposed convolution layers to upsampling, restoring the resolution of feature map from 16x16 to 512x512. We used batch normalization between each transposed convolution layers and which is followed by ReLU non-linearity.

2.3 Loss Function

Cross-entropy loss(CE) is the most commonly used loss function in semantic segmentation tasks. For binary classification, CE loss function can be described as:

$$CE(p,y) = \begin{cases} -\log(p) & if\ y = 1 \\ -\log(1-p) & otherwise. \end{cases}$$

where y$\in$ {0,1} is the ground truth label and p $\in$[0,1] is the prediction result. For notational convenience, here we define the probability p_t as:

$$p_t = \begin{cases} p & if\ y = 1 \\ 1-p & otherwise \end{cases}$$

Then we can get the loss function of CE as: $CE(p_t) = -\log(p_t)$.

For the task of segmentation, IoU is usually used to measure the performance of any segmentation approach. As a result, there exists a lot of its surrogates, and the goal is to minimize the gap between the actual IoU value and its differentiable approximation. Dice loss is proposed in view of the small foreground proportion, which is essentially to measure the overlap between two samples as:

$$L_{Dice} = 1 - 2|A \cap B|/(|A| + |B|).$$

where A is the foreground of target while B is the foreground of prediction.In our task, the error ground truth labels instead of the category imablance are the most serious impact that constrains the effectiveness of the model training. In order to avoid the network convergence, we choose Dice loss + BCE loss as our loss function.

3. EXPERIMENTS

3.1 Dataset

In this paper, we ecaluate the proposed P-LinkNet on Inrial Building dadataset [Maggiori et al., 2017], which inclueds 180 images with public labels and 180 images without public labels. For quantitative analysis, we only use the former in this paper, which was cropped to over 10000 training pictures and more than 2000 test pictures. There are five dissimilar urban settlements (Austin, Chicago, Kitsap County, Western Tyrol and Vienna) with 36 images respectively, ranging form densely populated areas to alpine towns. The dataset is formulated as a binary segmentation problem, in which buildings are labeled as foreground and other objects are labeled as background. As shown in Figure3.

(a) image (b) label

Figure 3. Examples of Inria dataset. (a) Original image. (b) Mask label.

3.2 Implementation Details

The whole experimental process is shown in the Figure 4. To avoid overfitting, we did data augmentation in ambitious way, including horizontal flip, vertical flip, transformation during the training period. The dataset is formulated as a binary segmentation problem, in which buildings are labeled as foreground and other objects are labeled as background. We did not do any augmentation in test time and used 0.5 as our prediction to generate binary outputs.

For our best model, we choose SGD as our optimizer. The original learning rate was set 0.01, and we adjust learning rate by MultiStepLR. During the course of experiments, we notice

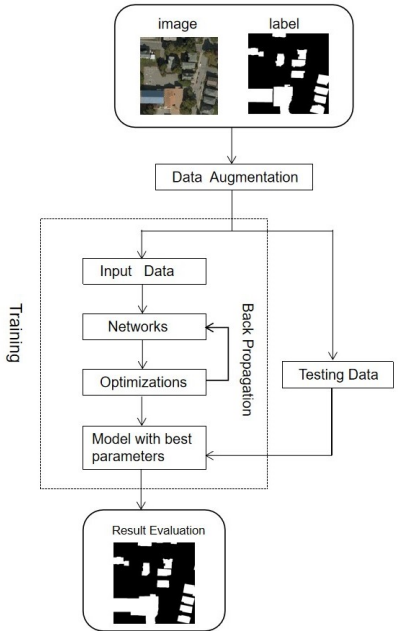
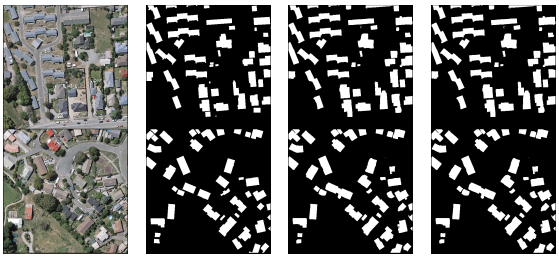


Figure 4. The schematic workflow of this study.

that appropriately large "cropsize" can yield good performance and "batchsize" in the batch normalization layer is of good performance. Due to limited physical memory on GPU cards, the batch size during training phase we fixed as 16. It took about 30 epochs for our network to converge.

3.3 Results

We propose to use spatial pyramid pooling to acquire multi-scale features, which can improve the extraction accuracy of buildings. The model aggregates the spatial context information form the low convolutional layers and multi-scale features to alleviate the problem of spatial information loss. For evaluation, both pixel-wise(Pixel Acc.) and intersection over union(IoU) are used. We found our model is better than LinkNet34. LinkNet34 had pretrained encoder which made the network has good performance. But the IoU of our model is 84.48 while the LinkNet34 is 84.26. And the precision of our model is 91.50 while the LinkNet34 is 91.04. Figure 5 shows some of the experimental results. By adding spatial pyramid pooling, P-LinkNet can obtain larger receptive field and multi-scale information at the same time, and thus alleviated the global infromation loss occurred in LinkNet34.



(a) image (b) label (c) P-LinkNet (d) LinkNet

Figure 5. Examples of building extraction results produced by two models on Inria dataset. (a) Original image. (b) Mask label. (c) prediction maps from P-LinkNet. (d) prediction maps from LinkNet.

4. CONCLUSION

In recent years, deep learning, especially convolutional neural networks, have been widely applied in computer vision and semantic segementation. However, automatic building extraction from high-resolution remote sensing imagery is still a challenging task due to a large variety of appearing patterns and its spatial scale. To address this issues, we have proposed a semantic segmentation model called P-LinkNet based on LinkNet for high resolution satellite imagery. By multi-scale pooling in the center part, P-LinkNet can get better performance in building extraction, which demonstrating that eht encoder-decoder and spatial pyramid pooling module are two powerful tools that need to be merged to take effect for builing segmentation.

Although our model has achived satisfactory result, P-LinkNet still has some problems in boundary accuracy, we plan to improve by introducing edge attention module and modifying the loss function to improve the extraction accuracy in the future.

REFERENCES

Aytekin, O., Zongur, U., Halici, U., 2013. Texture-Based Airport Runway Detection. *IEEE Geoscience and Remote Sensing Letters*, 10(3), 471–475.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481–2495.

Chaurasia, A., Culurciello, E., 2017. Linknet: Exploiting encoder representations for efficient semantic segmentation. *2017 IEEE Visual Communications and Image Processing (VCIP)*, 1–4.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2015. Semantic image segmentation with deep convolutional nets and fully connected crfs. *International Conference on Learning Representations*.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 833–851.

Cheng, Y.-Q., Li, H.-C., Celik, T., Zhang, F., 2013. Frft-based improved algorithm of unsupervised change detection in sar images via pca and k-means clustering. *2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS*, 1952–1955.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dong, Y., Du, B., Zhang, L., 2015. Target Detection Based on Random Forest Metric Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(4), 1830–1838.

Gomez, W., Pereira, W. C. A., Infantosi, A. F. C., 2012. Analysis of Co-Occurrence Texture Statistics as a Function of Gray-Level Quantization for Classifying Breast Ultrasound. *IEEE Transactions on Medical Imaging*, 31(10), 1889–1899.

Guo, Z., Shao, X., Xu, Y., Miyazaki, H., Ohira, W., Shibasaki, R., 2016. Identification of Village Building via Google Earth Images and Supervised Machine Learning Methods. *Remote Sensing*, 8(4), 271.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Huang, Z., Cheng, G., Wang, H., Li, H., Shi, L., Pan, C., 2016. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 1835–1838.

Hui, J., Du, M., Ye, X., Qin, Q., Sui, J., 2019. Effective Building Extraction From High-Resolution Remote Sensing Images With Multitask Driven Deep Neural Network. *IEEE Geoscience and Remote Sensing Letters*, 16(5), 786–790.

Inglada, J., 2007. Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. *Isprs Journal of Photogrammetry and Remote Sensing*, 62(3), 236–248.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2, 2169–2178.

Li, E., Femiani, J., Xu, S., Zhang, X., Wonka, P., 2015. Robust rooftop extraction from visible band images using higher order crf.

Li, X., Li, Z., Yang, J., Liu, Y., Fu, B., Qi, W., Fan, X., 2018. Spatiotemporal characteristics of earthquake disaster losses in China from 1993 to 2016. *Natural Hazards*, 94(2), 843–865.

Lin, Y., Saripalli, S., 2012. Road Detection and Tracking from Aerial Desert Imagery. *Journal of Intelligent and Robotic Systems*, 65(1), 345–359.

Liu, Y., Li, Z., Wei, B., Li, X., Fu, B., 2019. Seismic vulnerability assessment at urban scale using data mining and GIS-science technology: application to Urumqi (China). *Geomatics, Natural Hazards and Risk*, 10(1), 958–985.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

Lucchi, A., Li, Y., Boix, X., Smith, K., Fua, P., 2011. Are spatial and global constraints really necessary for segmentation. *2011 International Conference on Computer Vision*, 9–16.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 3226–3229.

Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 971–987.

Panboonyuen, T., Jitkajornwanich, K., Lawawirojwong, S., Srestasathiern, P., Vateekul, P., 2019. Semantic Segmentation on Remotely Sensed Images Using an Enhanced Global Convolutional Network with Channel Attention and Domain Specific Transfer Learning. *Remote Sensing*, 11(1), 83.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Szegedy, C., Reed, S. E., Erhan, D., Anguelov, D., 2014. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*.

Vakalopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N., 2015. Building detection in very high resolution multispectral data with deep learning features. *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 1873–1876.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 1, 511–518.

Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., Xu, Y., Shibasaki, R., 2018. Automatic Building Segmentation of Aerial Imagery Using Multi-Constraint Fully Convolutional Networks. *Remote Sensing*, 10(3), 407.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6230–6239.