Building Segmentation from VHR Aerial Imagery using DeepLabv3+ Architecture

Nuran Aslantaş[1], Bülent Bayram[1], Tolga Bakırman[1]

[1]Yıldız Technical University, Department of Geomatics Engineering, Davutpaşa, Istanbul, Turkey
Email: f5018009@std.yildiz.edu.tr, bayram@yildiz.edu.tr, bakirman@yildiz.edu.tr

KEY WORDS: Building extraction; deep learning; high resolution aerial images; deeplabv3+, WHU dataset

ABSTRACT

Up-to-date building footprint maps are of high demand for geographical applications such as sustainable urban planning and management, smart city applications, urban sprawl monitoring, population estimation and disaster management. Unplanned growth and settlement cause many problems such as deterioration of ecological balance, increase in damage from natural disasters, destruction of fertile lands, and drought. Analysis and monitoring of urban growth is an important issue for urban planning, environmental management, and sustainable development in areas of rapid urbanization. Recent advancements on remote sensing and artificial intelligence technologies provide great opportunities to obtain rapid, reliable and accurate building footprint maps. Very high-resolution satellite images and aerial images are rich data sources of spatial information for obtaining building footprints with automated approaches for cities with a variety of building types. Convolutional neural networks (CNN) have recently been used to successfully recover building footprints from satellite images. However, building segmentation from high resolution data is still a challenging task due to complex backgrounds and heterogeneous data structure. To overcome these problems, deep learning (DL) techniques became useful approaches. Different DL models have been proposed for building segmentation such as U-Net with VGG11 encoder pre-trained on ImageNet, Conditional Random Fields (CRF) with FCN, end-to-end self-cascaded network approach and generative adversarial networks (GAN). Additionally, there are existing open datasets for building segmentation issues such as Massachusetts building, ISPRS Vaihingen and Potsdam, Inria and WHU dataset. In this study, we aimed to investigate the performance of the DeepLabv3+ architecture for building segmentation. The hyper parameters of the architecture have been tested and selected empirically to obtain accurate results. The Wuhan University (WHU) Aerial Building Dataset with a spatial resolution of 0.075 to 0.3 m is used for training and testing. The dataset is split as 80%, 10% and 10% for train, validation and test, respectively. Input images are cropped to 512 x 512 pixels. The accuracy assessment results on the test dataset show that mean intersection over union (IoU) reached 98.23%. The obtained results show that DeepLabv3+ architecture is highly capable for building segmentation from very high resolution aerial imagery.

1. INTRODUCTION

Building footprint extraction is important issue for analysis of the existing settlement pattern which plays an important role in various fields such as urban planning, construction, sustainable crisis management (Erdem and Avdan, 2020). In recent years, aerial image data with a resolution of 5-30 cm can be obtained thanks to high resolution sensors (Marmanis et al., 2016). Thus, building extraction from high resolution images has become an important field of study in remote sensing.

Traditional methods such as manual digitization and fieldwork for building footprint extraction are costly and labor-intensive (Pan et al., 2020). For this reason, autonomous methods have taken the place of traditional methods. Recently, deep learning has become state-of-the-art method for many fields such as land use and land cover classification (Zhang et al., 2019), object detection (Liu et al., 2020) segmentation (Wang et al.,2020) scene classification (Cheng et al., 2020) and road extraction (Lian and Huang, 2020). Due to the success of the method, building extraction from high resolution satellite images with deep learning has become a prominent method for the analysis of existing settlement texture.

High-resolution imageries are important data source for building segmentation studies due to the rich information they provied. Although high resolution imageries are rich data sources, some problems are encountered during the studies such as complex scene (Huang et al., 2019), the need for more memory due to the size of the data and their complex structure (Liu et al., 2019), and the different sizes of the buildings (multi scale problem) (Deng et al., 2021) which are important problems affecting the training phase.

DeepLabv3+ architecture has solved multi-scale problem that arises during the extraction of buildings by the atrous convolution method (Lui et al., 2021). Depthwise separable convolution which is used in architecture provides computation cost reduction (Lui et al., 2021). The high success of the method has revealed that high resolution remote sensing images are an accurate data source for urban mapping, and the Deeeplabv3+ method is a fast and

computationally cost effective method. Therefore, in this study, we aimed to evaluate the performance of DeepLabv3+ architecture for automatic building footprint extraction.

2. MATERIALS AND METHOD

The DeepLabv3+ architecture is evaluated using public the Wuhan University (WHU) Building Dataset (Ji, et al., 2018) which consists of aerial and satellite imagery. In this study, aerial imagery dataset is used. The aerial dataset consists of more than 220.000 independent buildings and 450 km^2 covering in Christchurch, New Zealand (Figure 1). This area contains countryside, residential, culture, and industrial area. The examples of the dataset shown in Figure 2. The dataset is ideal due to label accuracy and various architecture type of building (Liu et al., 2019). The dataset contains 8188 non-overlapping images. Image size is 512 x 512 pixels. Spatial resolution of aerial images varies between 0.075 and 0.3 m. The dataset was divided into three parts which are the training set (4736 image, containing 130,500 buildings), validation set (1036 image, containing 14,500 buildings) and testing set (2416 image, containing 42,000 buildings).

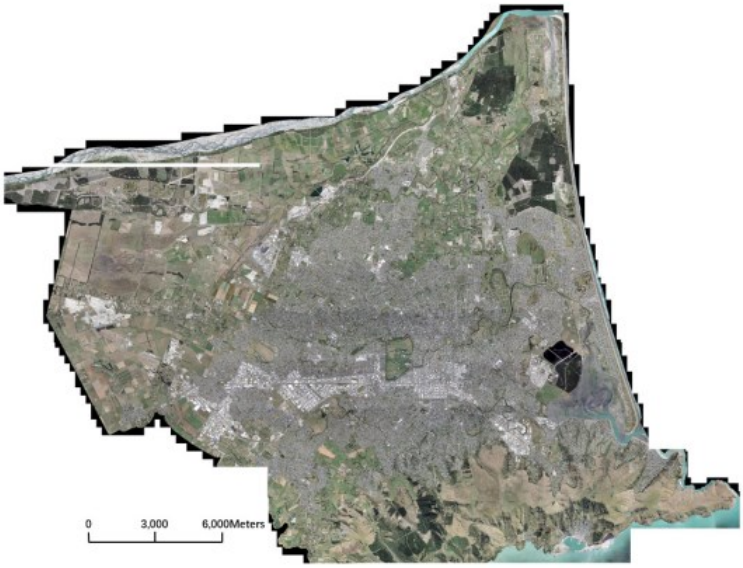


Figure 1: Aerial dataset coverage area copyrighted (Ji, et al., 2018)

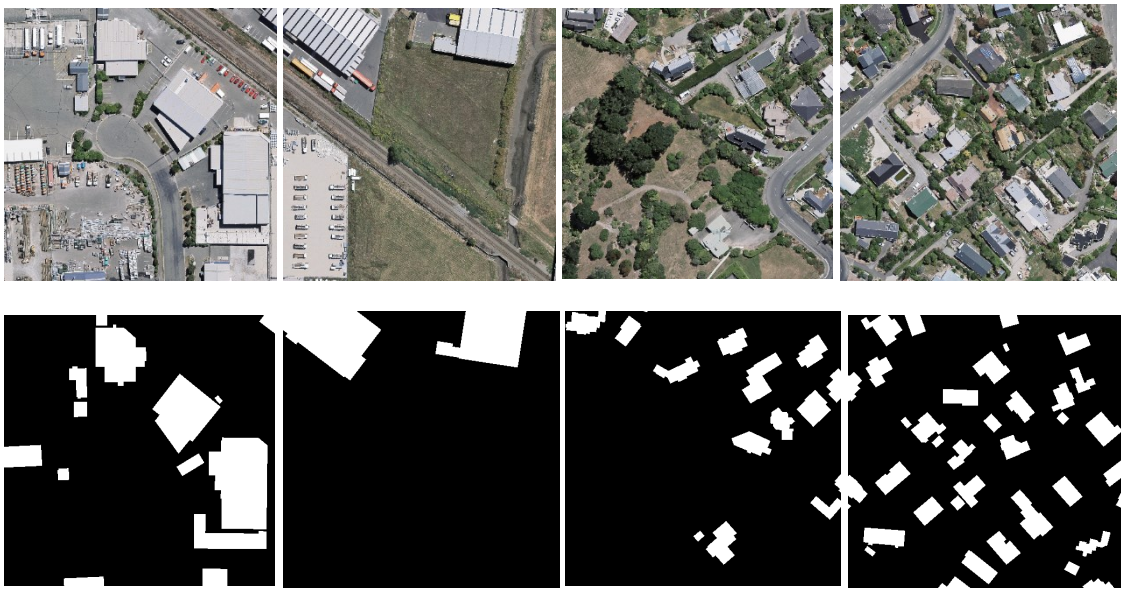


Figure 2: Examples of aerial image dataset

DeepLabv3+ architecture is a encoder-decoder network based on Fully Convolutional Networks (FCNs). It extends the previous version (DeepLabv3) by adding a decoder module to refine the extraction results (Chen et al., 2018). The method includes two concepts namely spatial pyramid pooling and encoder-decoder. Spatial pyramid, which is used in DeepLab (Chen et al., 2017a; Chen et al. 2017b) architecture, present a solution for multi scale problem. In DeepLabv3+ architecture atrous spatial pyramid pooling is used to solve this problem. Encoder module captures multiscale information by atrous convolution and decoder module refines the extraction result.

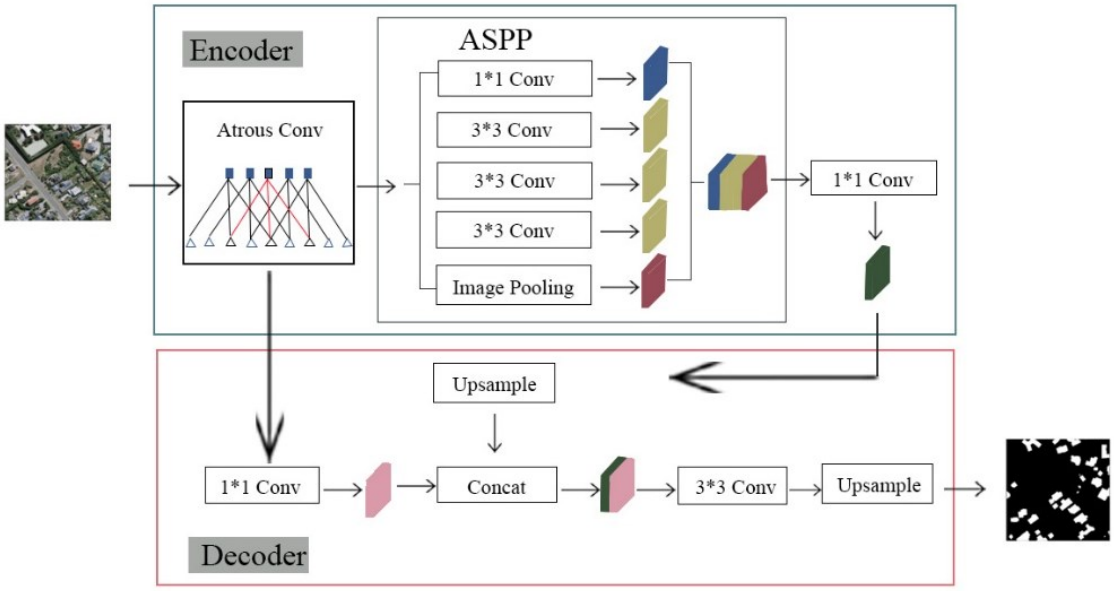


Figure 3: DeepLabv3+ model

Four metrics, that is precision rate, recall rate, F1 score and Intersection-over-Union (IoU) are used to evaluate the performance of the architecture. The formulas are as follows:

$$precision = \frac{TP}{TP+FP} \quad (1)$$

$$recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1 = 2 * \frac{precision*recall}{precision+recall} \quad (3)$$

$$IoU = \frac{TP}{TP+FP+FN} \quad (4)$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative pixels in prediction.

3. RESULTS AND DISCUSSION

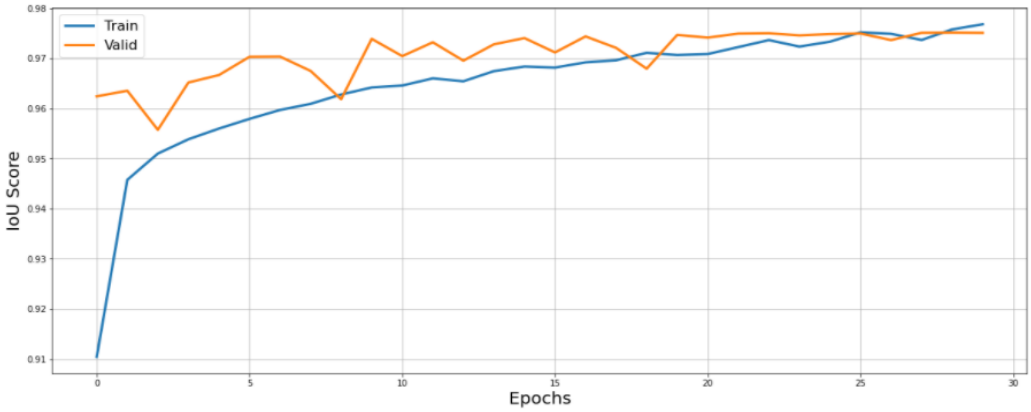
In this experiment, deep residual network ResNet-101 is used as encoder with pre-trained weights on ImageNet. Number of epoch and batch size is set as 30 and 4, respectively. Sigmoid function and dice loss are used as activation function and loss function, respectively. Dataset were randomly divided by 8:1:1 to constitute the training set, validation set, and testing set respectively. We did not apply any data augmentation technique. The training of network is conducted on Pytorch environment. All the experiments are ran on Nvidia Tesla P100 PCI-E 16 GB GPU.

IoU and loss values in training and validation dataset for 30 epochs are shown in Figure 4. The final accuracy result for test dataset shown in Table 1. Although, The WHU dataset contains buildings with different textures, shapes and sizes, the results show that the DeepLabv3+ architecture is an effective method for building extraction. Examples from test, corresponding ground truth and predicted images are shown respectively in Figure 5. DeepLabv3+ results were similar to label in terms of different and amorphous building types (Figure 6). Small structures were generally

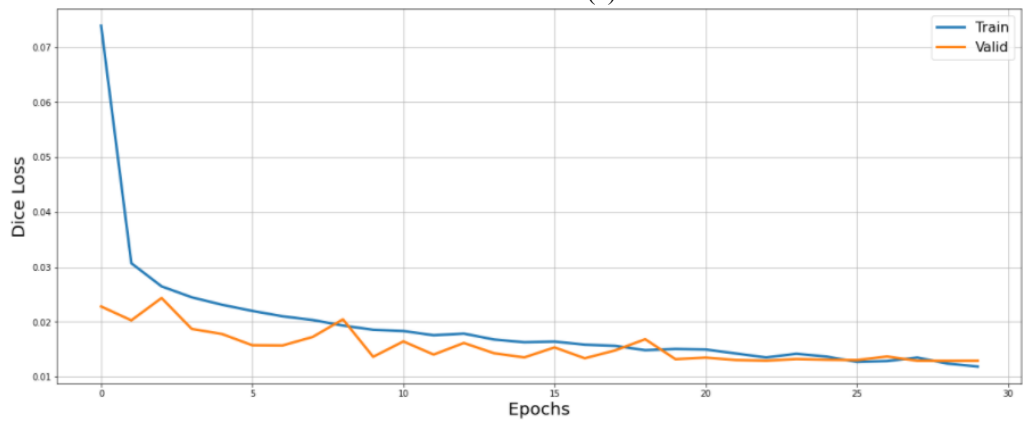
not predicted correctly. Also, small buildings that are very close to each other are predicted as adjacent building (Figure 7).

Table 1: Quantitative evaluation results on the testing set of the WHU Building dataset by DeepLabv3+ model in terms of precision, recall, F1-Score, and IoU

Model	Precision	Recall	F1-Score	IoU
DeepLabV3+	99.08%	99.09%	99.09%	98.23%



(a)



(b)

Figure 4: IoU (a) and loss values (b) for 30 epochs

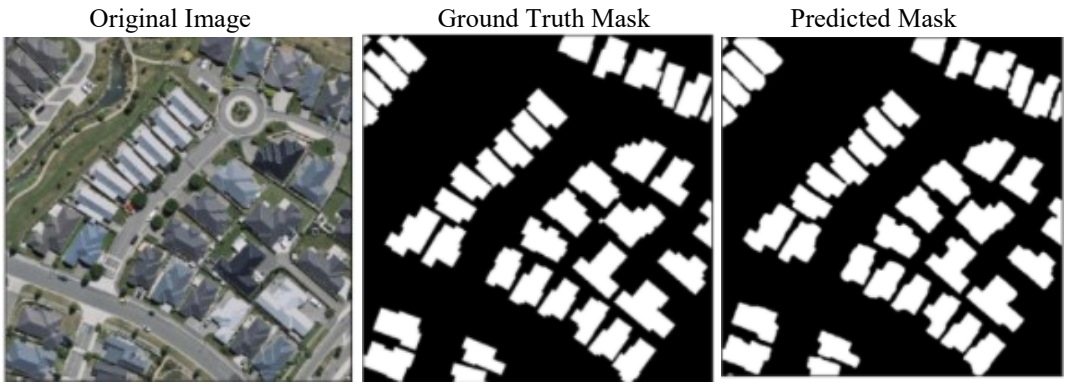


Figure 5: Building footprint extraction sample result

Original Image Ground Truth Mask Predicted Mask

Figure 6: Footprint extraction results in various building types

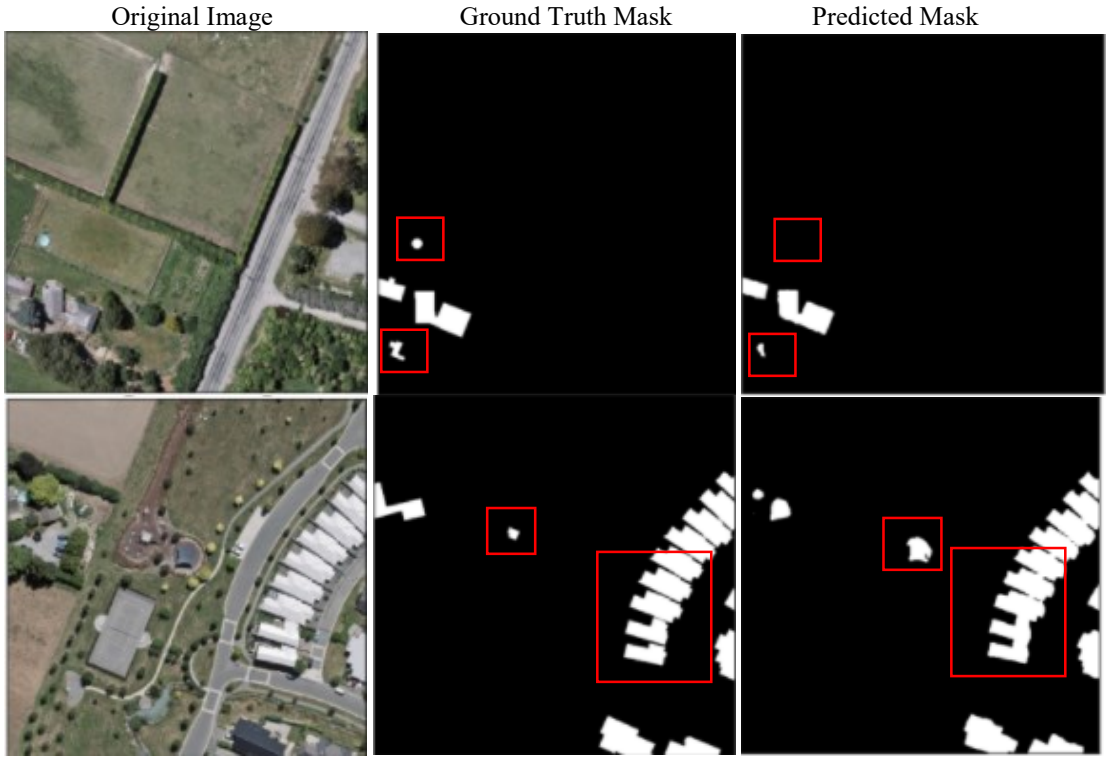


Figure 7: Footprint extraction results for small buildings and buildings close to each other

4. CONCLUSION

The dramatic increase in population has led to the rapid and unplanned construction of urban areas. The analysis of the existing settlement pattern is an important issue in the studies to be carried out for sustainable growth, disaster management and protection of ecological balance. Obtaining the settlement pattern of urban areas with traditional methods is a time-consuming and costly process. Thanks to the developments in sensor technology, it has been possible to obtain high-resolution and cost-effective data. In the study, the success of the DeepLabv3+ method for

automatic building footprint extraction from high resolution aerial images was examined. DeepLabv3+ was chosen for its high success in multi-scale data. Although high resolution imageries are rich data sources, they cause some problems in inference and classification processes due to their complex structure. Multi-scale problem is one of the prominent problems. Based on 98.23% IoU success achieved as a result of the study, it has been shown that DeepLabv3+ architecture can be effiently used in mapping the existing settlement pattern in urban areas.

REFERENCES

Cheng, G., Xie, X., Han, J., Guo, L., and Xia, G., 2020. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, pp. 3735-3756.

Chen, L.C., Zhu, Y., Papandreou, G., Schrof, F., and Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation: Proceedings of the European conference on computer vision (ECCV), pp. 801–818

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40, pp. 834-848

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017b. Rethinking atrous convolution for semantic image segmentation: Proceedings of the European conference on computer vision (ECCV)

Deng, W., Shi, Q., and Li, J., 2021. Attention-Gate-Based Encoder–Decoder Network for Automatical Building Extraction: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, pp. 2611 – 2620.

Erdem, F., and Avdan, U., 2020. Comparison of Different U-Net Models for Building Extraction from High-Resolution Aerial Imagery: International Journal of Environment and Geoinformatics 7(3), pp. 221-227.

Huang J., Zhang, X., Xin, Q., Sun, Y., and Zhang, P., 2019. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network: ISPRS Journal of Photogrammetry and Remote Sensing, 151, pp. 91-105.

Ji, S., Wei, S., and Lu, M., 2018. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set: IEEE Transactions on Geoscience and Remote Sensing, 57, pp. 574 – 586.

Lian, R., and Huang, L., 2020. DeepWindow: Sliding Window Based on Deep Learning for Road Extraction From Remote Sensing Images: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 1905 – 1916.

Liu, H., Luo, J., Huang, Bo., Hu, X., Sun, Yang, Y., Xu, N., and Zhou, N., 2019. DE-Net: Deep Encoding Network for Building Extraction from High-Resolution Remote Sensing Imagery: Remote Sensing, 11(20), 2380.

Liu, J., Wang, S., Hou, X., and Song, W., 2019. A deep residual learning serial segmentation network for extracting buildings from remote sensing imagery: International Journal of Remote Sensing, 41, pp. 5573-5587.

Liu, M., Fu, B., Xie, S., He, H., Lan, F., Li, Y., Lou, P., and Fan, D., 2021. Comparison of multi-source satellite images for classifying marsh vegetation using DeepLabV3 Plus deep learning algorithm: Ecological Indicators, 125,

Liu, Q., Xiang, X., Wang, Y., Luo, Z., and Fang, F. 2020. Aircraft detection in remote sensing image based on corner clustering and deep learning: Engineering Applications of Artificial Intelligence, 87.

Marmanis, D., Schindler, K., Wegner, J.D., Gallaiani, S., Datcu, M., and Stilla, U., 2016. Classification With an Edge: Improving Semantic Image Segmentation with Boundary Detection: ISPRS Journal of Photogrammetry and Remote Sensing, 135, pp. 158-172.

Pan, Z., Xu, J., Guo, Y., Hu, Y., and Wang, G., 2020. Deep Learning Segmentation and Classification for Urban Village Using a Worldview Satellite Image Based on U-Net: Remote Sensing, 12(10), 1574.

Wang, S., Chen, W., Xie, S.M., Azzari, G., and Lobell, D.B., 2020. Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery: Remote Sensing, 12(2), 207.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., and Atkinson, P.M., 2019. Joint Deep Learning for land cover and land use classification: Remote Sensing of Environment, 221, pp. 173-187.