ATTENTION BASED CONVOLUTIONAL NEURAL NETWORK FOR BUILDING EXTRACTION FROM VERY HIGH RESOLUTION REMOTE SENSING IMAGE

H.R.Hosseinpoor [1]*, F.Samadzadegan [1]

[1] School of Surveying and Geospatial Information Engineering, College of Engineering, University of Tehran - (hosseinpoor, samadz)@ut.ac.ir

Commission VI, WG VI/4

KEY WORDS: Building extraction, Fully convolutional neural networks, Attention mechanism, U-Net

ABSTRACT:

Buildings are a major element in the formation of cities and are essential for urban mapping. The precise extraction of buildings from remote sensing data has become a significant topic and has received much attention in recent years. The recently developed convolutional neural networks have shown effective and superior performance to perform well on learning high-level and discriminative features in extracting buildings because of the outstanding feature learning and end-to-end pixel labelling abilities. However, it is difficult to use the features of different levels with a certain degree of importance that is appropriate to deep learning networks. To tackle this problem, a network based on U-Nets and the attention mechanism block was proposed. The network contains an encoder part and a decoder part and a spatial attention module. The special architecture presented in this article enhances the propagation of features and effectively utilizes the features at various levels to reduce errors. The other remarkable thing is that attention module blocks only lead to a minimal increase in model complexity. We effectively demonstrate an improvement of building extraction accuracy on challenging Potsdam and Vaihingen benchmark datasets. The results of this paper show that the proposed architecture improves building extraction in very high resolution remote sensing images compared to previous models.

1. INTRODUCTION

Remote sensing images with very high resolution (VHR) are widely used in many geoscience applications. Several of the practical applications are based on VHR remote sensing imagery classification at the pixel level, also defined as semantic segmentation. Semantic segmentation is a widely used topic in the field of computer vision and has become increasingly used in remote sensing in recent years. Building extraction from remote sensing images is essentially a problem of segmenting semantic objects and, has important implications in urban planning, population estimation, and topographic map creating and updating (Maggiori et al., 2017; Li et al., 2018; Lu, 2018).

State-of-the-art approaches for semantic image segmentation are built on fully convolutional neural networks (F-CNNs) (Shelhamer, Long and Darrell, 2017). The F-CNNs is based on a pre-trained deep convolutional neural network (DCNN) designed to classify images from, VGG-16 (Simonyan and Zisserman, 2014), ResNet (He et al., 2015), Deeplab (Chen et al., 2018) and DenseNet (Yang et al., 2018). The convolutional layer has been used as the main block in all these architectures, which can take in an input image, assign learnable weights and biases to objects that can be able to differentiate one from the other in images. A variety of FCNs have been proposed, such as SegNet (Badrinarayanan, Kendall and Cipolla, 2017), DeconvNet (Huang et al., 2016), U-net (Ronneberger, Fischer and Brox, 2015). Up to now, a continuously updated network named DeepLab (Chen et al., 2018) has been a new benchmark of semantic segmentation.

The most recent studies in building extraction exclusively utilized FCN-based methods. (Maggiori et al., 2017) designed a two-scale neuron module in an FCN to reduce the trade-off between recognition and precise localization. Studies in (C. Zhang et al., 2018) integrated multiple layers of activation into pixel level prediction based on FCN. Due to the considerable efforts made in recent years in the development of different building extraction algorithms and methods, however, the automatic extraction of buildings from very high spatial resolution remote sensing images is a challenging issue. This is because there are a lot of problems to get an accurate extraction of buildings. There are mainly two difficult issues, For one thing, the shape and scale of buildings are various, which makes it difficult to detect buildings with all scales using a uniform scale model. On the other hand, due to the increased spatial resolution of the sensors used in photogrammetry and remote sensing, objects of various sizes and shapes appear along with variations in color and texture, which makes it impossible to define a pattern that expresses most buildings. Therefore, extracting robust and discriminative of buildings is challenge in VHR remote sensing images.

A lot of recent work has aimed at improving the joint encoding of spatial and channel information (Hu *et al.*, 2017; Yang *et al.*, 2018). In order to achieve better results in segmentation of building in VHR remote sensing images, it has been proposed to encoding of channel-wise patterns independently which known as the attention mechanism. In this paper, it attempts to explicitly test this issue by modelling the dependencies between feature maps in order to

* Corresponding author

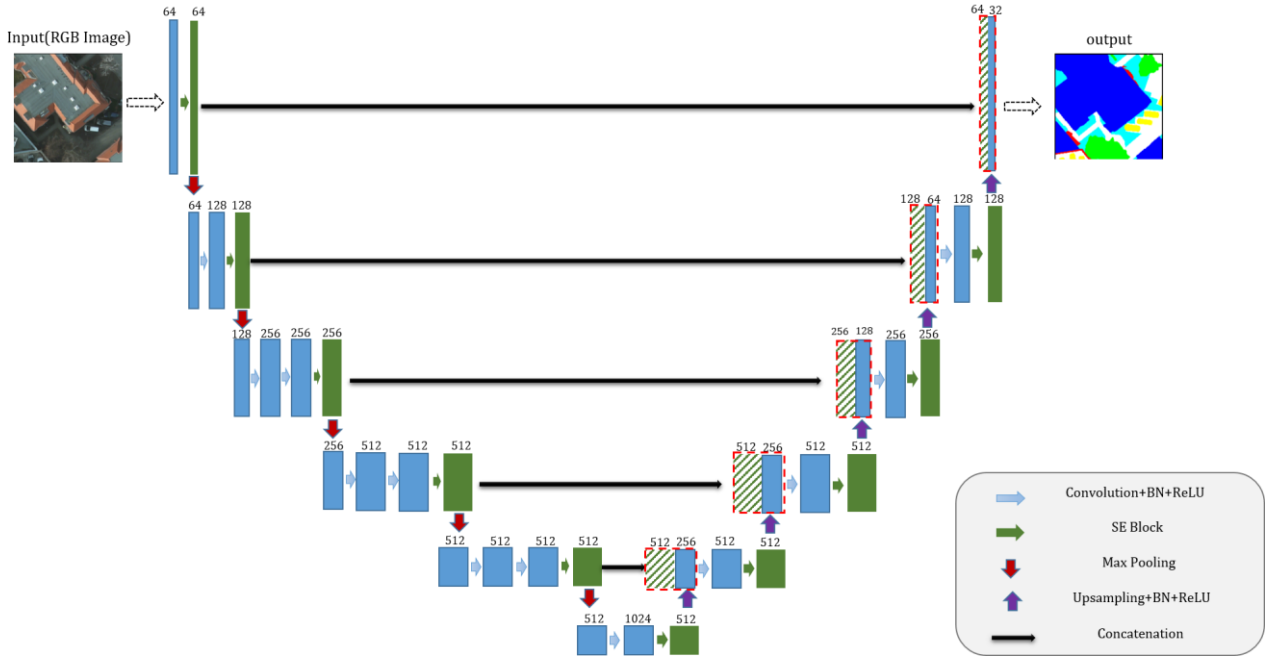

Figure1. Diagram of overall architecture of attention networks

improve its representation. This is done by an architectural component called squeeze & excitation (SE) block (Hu *et al.*, 2017), which can be integrated as an add-on within a CNN. We aim at exploitation the performance of SE blocks to semantic segmentation for extracting buildings by integrating them within the famous UNet network (Ronneberger, Fischer and Brox, 2015)

We use two challenging VHR remote sensing datasets with different spatial image resolutions and different conditions. In the following, this article is organized as follows: In the second section we will outline the proposed method for extracting buildings from VHR images. Section 3 describes the experimental results. Conclusions are given in Sections 4, respectively.

2. METHODOLOGY

This paper presents a deep learning based approach to extract buildings from urban areas in VHR remote sensing images. In recent years, deep learning has become a powerful tool to extract and detect objects in remote sensing. Fully convolutional neural networks (F-CNNs) are adapted as effective tools for the semantic labelling of high-resolution remote sensing data. The encoder–decoder architecture is widely used in F-CNNs. In the encoder, the feature extraction is performed at different levels of the input data scale. The decoder part aims to recovers the spatial resolution of feature maps and to extract target using these feature maps. In order to improve the extraction of buildings from VHR remote sensing imageries, one of the primitive F-CNN models, U-Net network architecture (Ronneberger, Fischer and Brox, 2015) has been used with the idea of using the attention mechanism technique through the use of the SE block. Its overall architecture is shown in Figure 1.

2.1 U-Net Architecture

U-Net was proposed for Bio Medical Image Segmentation. The architecture contains two paths. First path called as the encoder which is used to capture the context in the image. This path follows the typical architecture of a convolutional network with alternating convolution and pooling operations and progressively down-samples feature maps, increasing the number of feature maps per layer at the same time. The second path is called as the decoder which is used for upsampling with transposed convolutions to enable precise localization. Thus it is an end-to-end fully convolutional neural network (F-CNN), i.e. it only contains Convolutional layers and does not contain any dense layer because of which it can accept image of any size. The core idea of U-Net is to gradually fuse high-level with low-resolution features from top layers with low-level but high-resolution features from bottom layers, which is helpful for the decoder part to produce high-resolution segmentation results.

In this paper we use the CNN of the VGG family (Simonyan and Zisserman, 2014) as the encoder component in the UNet architecture. According to the UNet architecture (Figure 1), each blue rectangular block represents a multi-channel features map passing through a series of transformations (blue arrow). The height of the rod shows a relative feature map size, while their widths are proportional to the number of channels (the number is explicitly subscribed to the corresponding rod). According to the original VGG family block architecture, to reduce the computation volume, the dimension of the feature maps in the Encoder section are halved using the max pooling operator at each step (red arrow) , While upsampling methods is used (purple arrow) to achieve the dimensions of the input image in the decoder section. The green rectangles represent the SE block, which is explained in the section 2.2. Skip connections (black arrow) are located between the encoder and decoder feature

maps with similar spatial resolution, which aim to provide more contextual information to decode and assist in the aiding flow of gradients in the network.

2.2 Spatial Attention Module (SE block)

Though the great success of U-Net, the working mechanism is still unknown (Z. Zhang *et al.*, 2018). Low-level and high-level features are complementary by nature, where low-level features are rich in spatial details but lack semantic information and vice versa. In extracting low-level features, only concepts such as points, lines or edges are encoded. The skip connection of U-Nets is a common way to help decoders recover object details information from the encoder path by reusing feature maps. The fusion of high-level features with such low-level features helps little, because low-level features are too noisy to provide sufficient high resolution semantic guidance. Experimentally, the semantic and resolution overlap between low-level and high-level features plays an important role in the effectiveness of feature fusion (Z. Zhang *et al.*, 2018).

The mechanism of attention can be weighted high level information using low level features. Inspired by the attention mechanism, the channel attention module is presented to enhance the useful information of low level features and to eliminate noise to avoid overuse of low level features. We use squeeze & excitation (SE) block method presented in (Hu *et al.*, 2017), which can be seamlessly integrated as an add-on within a U-net architecture. This SE block factors out the spatial dependency by global average pooling to learn a channel specific descriptor, which is used to rescale the input feature map to highlight only useful channels. As this component 'squeezes' along spatial domain and 'excites' or reweights along the channels, it is termed as squeeze & excitation block (Roy, Navab and Wachinger, 2019).

For channel attention mechanism, the input feature map U = [u1, u2, …, uc] was considered as a combination of channels $u^i \in R^{H\times W}$. Spatial squeeze is performed by a global average pooling layer, producing vector z ∈ R1×1×C with its kth element.

$$z_k = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} u_k(i, j) \quad (1)$$

This operation set the global information in vector z. This vector is transformed to $\hat{z} = W_1(\delta(W_2 z))$, with $W_1 \in R^{C\times C/2}$, $W_2 \in R^{C/2\times C}$ being weights of two fully-connected layers and the ReLU operator δ(·). This encodes the channel-wise dependencies. The dynamic range of the activations of $\hat{z}$ are brought to the interval [0, 1], passing it through a sigmoid layer $\sigma(\hat{z})$. The resultant vector is used to excite U to:

$$\hat{U}_{SE} = [\sigma(\hat{z}_1)u_1, \sigma(\hat{z}_1)u_1, ..., \sigma(\hat{z}_C)u_C] \quad (2)$$

The activation $\sigma(\hat{z}_i)$ indicates the importance of the i^{th} channel, which are rescaled. As the network learns, these activations are adaptively tuned to ignore less important channels and emphasize the important ones. The architecture of the block is illustrated in Figure 2.

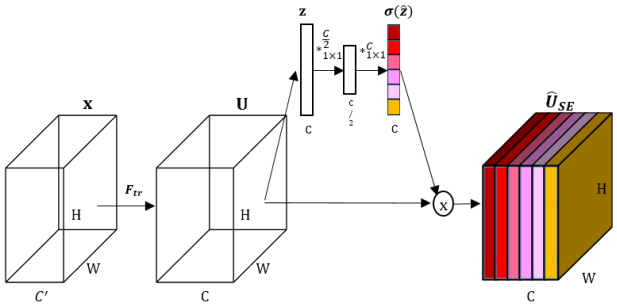

Figure 2. Squeeze & Excitation (SE) block

3. EXPERIMENTAL RESULTS

In this section, the effectiveness of the proposed scheme for building extraction in VHR remote-sensing images was investigated. All networks were trained and tested with Pytorch framework on Tesla K80 GPU device.

3.1 Dataset

The ISPRS 2D semantic labelling VHR remote sensing imageries of urban districts are used in the experiments, including the Vaihingen (Germany) and Potsdam (Germany) datasets, as these are open asset datasets provided online. The Potsdam set contains two sorts of optical images, including near-inferred, red, green bands (NIR-RG) and red, green, blue bands (RGB), respectively. The Potsdam dataset has a blue channel, containing 38 ortho-rectified aerial IRRGB images of ≈ 6000×6000 at 5 cm spatial resolution, which made small details visible. The Vaihingen dataset comprised of 33 large image patches, extracted from a larger orthophoto imagery captured over Vaihingen. The images have a 9 cm (GSD) ground distance, each image consisting of near infrared, red and green channels. In addition, extra digital surface models (DSMs) are supplied. For both dataset, each of the ground truth labels are made up of building and unknown (clutter).

3.2 Dataset pre-processing

Given the limited memory of the GPU and obtaining more training samples, images in both sets were split into smaller patches. We generated a training set using crop out image patches of size 224 × 224 pixels by sliding through each training image without any overlap. In the training phase, about twenty percent of the images were randomly used as the validation set and the remaining labelled images as the training models. Moreover, according to the defined red–green–blue (RGB) values of the six land cover classes, required objects can be extracted such as (255, 255, 255) colour, which means the building type. The optimization is performed with a batch size of 16 on the RGB tiles using a Stochastic Gradient Descent (SGD). In order to train the network at the learning stage, the learning rate was adjusted to 0.01, which decreased with 0.1 after every 10 repetitions. The momentum was set to 0.95. Training was continued till validation loss converged. During testing, each tile is processed by sliding a 224 × 224 window with a stride of 112 (i.e. 50% overlap).

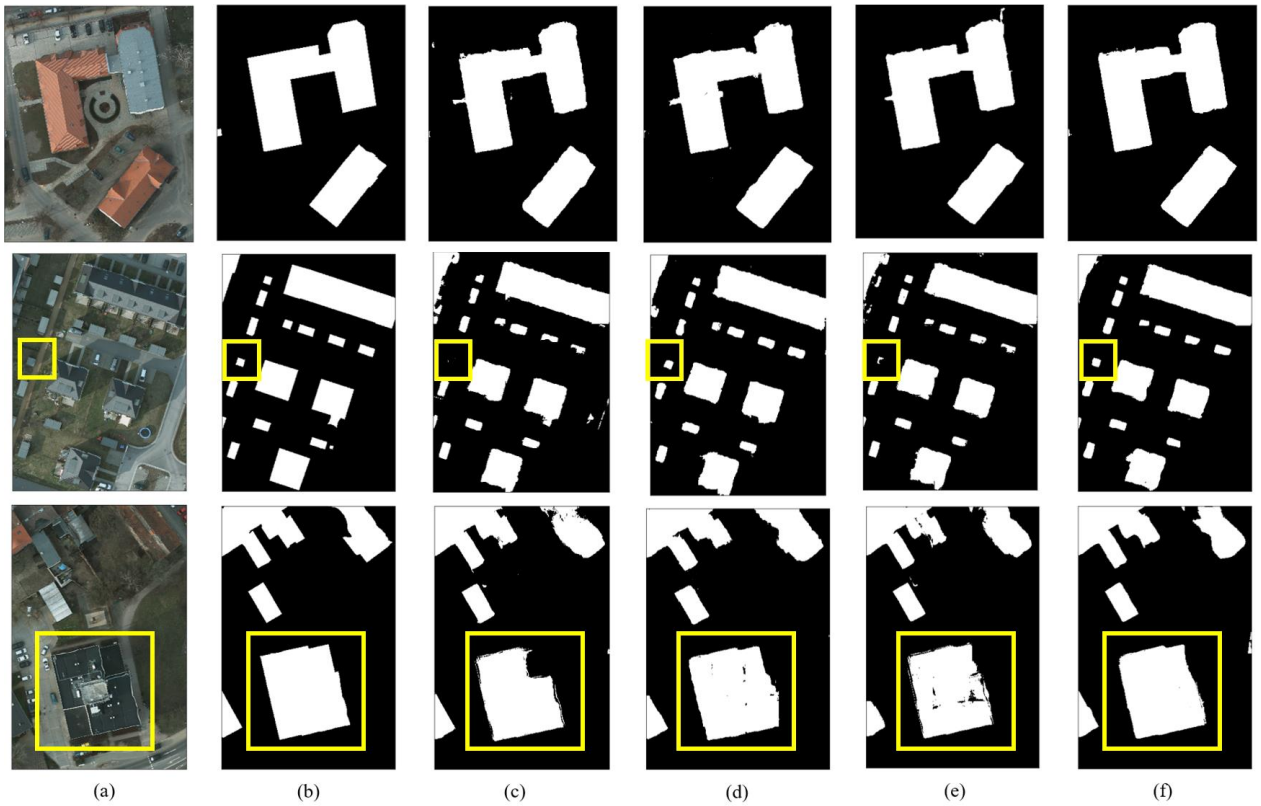

Figure3. The building extraction maps of Images from Potsdam data. (a) Original image. (b) Ground truth. Extracted building map from (c) UNet11. (d) UNet11 with SE block. (e) UNet16. (f) UNet16with SE block.

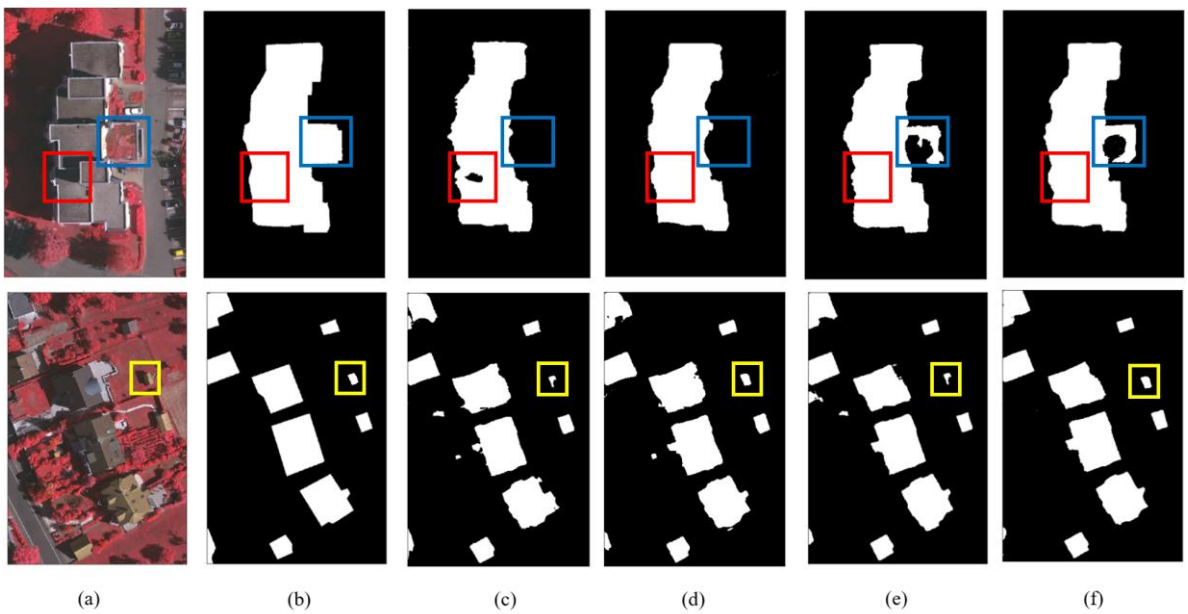

Figure4. The building extraction maps of Images from Vaihingen data. (a) Original image. (b) Ground truth. Extracted building map from (c) UNet11. (d) UNet11 with SE block. (e) UNet16. (f) UNet16with SE block.

3.3 Result

The table 2 shows the results of the proposed method using the attention mechanism in the UNet network structure with different encoders from the VGG family. In this study, VGG11 and VGG16 encoders were used in UNet model to demonstrate the ability of the attention mechanism structure for improving the buildings extraction of VHR remote sensing imagery. These models are named unet11 and unet16, respectively. In addition, when using the SE block in

attention mechanism mode, the models were named unetatt11 and unetatt16. In order to quantitatively compare the proposed network, we used three popular criteria, named Recall, Precision, and F-measure to evaluate the performance of the proposed algorithm. They are defined as follows.

$$\mathrm{Re}call = \frac{TP}{TP + FP} \tag{3}$$

$$\mathrm{Pr}ecision = \frac{TP}{TP + FN} \tag{4}$$

$$F - measure = \frac{2 \times \mathrm{Pr}ecision \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \tag{5}$$

Where TP (true positive) represents the total number of building pixels correctly classified in the reference maps; FP (false positive) represents the number of background pixels misclassified as buildings; FN (False Negative) represents the number of pixels of a building object selected as background pixels. Generally, recall and precision can express the accuracy of a building's extraction, and F-measure criterion is a fictitious measure of recall and precision.

Data	Criteria	Method			
		Unet 11	**Unet 16**	**Unetatt 11**	**Unetatt 16**
Potsdam	Recall	0.951	0.958	0.962	0.946
	Precision	0.929	0.951	0.936	0.957
	F-measure	0.939	0.954	**0.949**	**0.958**
Vaihingen	Recall	0.91	0.915	0.913	0.914
	Precision	0.927	0.931	0.935	0.939
	F-measure	0.918	0.922	**0.923**	**0.926**

Table 1. Accuracy measures of the three algorithms for the building extraction results using Recall, Precision, and F-measure in test images of two dataset.

Figures 3 and 4 show the building extraction results by the attention mechanism architectures and traditional UNet network with Vgg11 and Vgg16 pretrained encoders. Generally, architectures that use VGG16 encoders will produce better results than VGG11 because of the deeper layers and more feature maps.

The better performance of our proposed algorithm is clearly perceivable when we use SE block with VGG16 encoder in UNet model structure. This model can extract buildings with various scales and the extracted buildings are complete and continuous under complex building environments. The proposed algorithm can recognize buildings at both small and large scales, as shown in the labelled yellow ellipses in Figure 3 and 4.

Especially, as illustrated in the red rectangles corresponding to Figure 4, the architecture of UNet with VGG16 encoder and SE block can detect buildings completely, while the others performed poorly in detecting buildings on the dark side, which indicates the superiority of attention mechanism method in extracting buildings with different appearances. However, Roof of buildings covered with vegetation, are difficult to be extracted completely using the proposed algorithm. For example, as shown in the blue rectangles in Figure 4, there are part of buildings that are covered with vegetation, and all the algorithms cannot effectively detect it.

4. CONCLUSION

Extracting buildings from VHR remote sensing images has been a popular topic for the past two decades. However, the large variety of scales and the appearance of buildings make the task particularly challenging in VHR imagery. In this paper, an attention-based network was proposed for 2D building extraction in VHR images. The attention-based network contained an encoder part and a decoder part, which can guide to recalibrate intermediate feature maps, for image segmentation. In our experiment on three different F-CNN architectures for building extraction task we demonstrate that our proposed model yield a consistent improvement in segmentation performance.

REFRENCE

Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017) 'SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), pp. 2481–2495. doi: 10.1109/TPAMI.2016.2644615.

Chen, L. C. *et al.* (2018) 'DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), pp. 834–848. doi: 10.1109/TPAMI.2017.2699184.

He, K. *et al.* (2015) 'Deep Residual Learning for Image Recognition', ((ed.), Oxford, U.K., Pergamon Press PLC, 1989, Section 3, pp.111-120. (ISBN 0-08-036148-X)), pp. 1–9. doi: 10.3389/fpsyg.2013.00124.

Hu, J. *et al.* (2017) 'Squeeze-and-Excitation Networks', pp. 7132–7141. Available at: http://arxiv.org/abs/1709.01507.

Huang, Z. *et al.* (2016) 'Building extraction from multi-source remote sensing images via deep deconvolution neural networks', *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016-Novem, pp. 1835–1838. doi: 10.1109/IGARSS.2016.7729471.

Kemker, R., Kanan, C. and Carlson, C. F. (2012) 'Deep Neural Networks for Semantic Segmentation of Multispectral Remote Sensing Imagery', *International Journal of Applied Earth Observation and Geoinformation*, 15(1), pp. 38–48. doi: 10.1016/j.jag.2011.07.002.

Li, L. *et al.* (2018) 'A Multiple-Feature Reuse Network to Extract Buildings from Remote Sensing Imagery', *Remote Sensing*, 10(9), p. 1350. doi: 10.3390/rs10091350.

Lu, K. (2018) 'Extraction of Buildings From Aerial Images', *PQDT - Global*, p. 64. Available at: https://search.proquest.com/docview/2087742600

Maggiori, E. *et al.* (2017) 'Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification', *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 55(2), pp. 645–657. doi: 10.1109/TGRS.2016.2612821.

Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-Net: Convolutional Networks for Biomedical Image Segmentation', *Computer Vision and Pattern Recognition*, pp. 1–8. Available at: http://arxiv.org/abs/1505.04597.

Roy, A. G., Navab, N. and Wachinger, C. (2019) 'Recalibrating Fully Convolutional Networks With Spatial and Channel "Squeeze and Excitation" Blocks', *IEEE Transactions on Medical Imaging*, 38(2), pp. 540–549. doi: 10.1109/TMI.2018.2867261.

Shelhamer, E., Long, J. and Darrell, T. (2017) 'Fully Convolutional Networks for Semantic Segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), pp. 640–651. doi: 10.1109/TPAMI.2016.2572683.

Simonyan, K. and Zisserman, A. (2014) 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *American Journal of Health-System Pharmacy*, 75(6), pp. 398–406. doi: 10.2146/ajhp170251.

Wu, G. *et al.* (2018) 'Automatic building segmentation of aerial imagery usingmulti-constraint fully convolutional networks', *Remote Sensing*, 10(3), pp. 1–18. doi: 10.3390/rs10030407.

Yang, H. *et al.* (2018) 'Building Extraction in Very High Resolution Imagery by Dense-Attention Networks', *Remote Sensing*, 10(11), p. 1768. doi: 10.3390/rs10111768.

Zhang, C. *et al.* (2018) 'Convolutional Neural Network-Based Remote Sensing Images Segmentation Method for Extracting Winter Wheat Spatial Distribution', *Applied Sciences*, 8(10), p. 1981. doi: 10.3390/app8101981.

Zhang, Z. *et al.* (2018) 'ExFuse: Enhancing Feature Fusion for Semantic Segmentation', pp. 1–17. Available at: http://arxiv.org/abs/1804.03821.