An Ensemble Architecture of Deep Convolutional Segnet and Unet Networks for Building Semantic Segmentation from High-resolution Aerial Images

Abstract

Building objects is one of the principal features that are essential for updating the geospatial database. Extracting building features from high-resolution imagery automatically and accurately is challenging because of the existence of some obstacles in these images, such as shadows, trees, and cars. Although deep learning approaches have shown significant improvements in the results of image segmentation in recent years, most deep neural networks still cannot achieve highly accurate results with correct segmentation map when processing high-resolution remote sensing images. Therefore, we implemented a new deep neural network named Seg–Unet method, which is a composition of Segnet and Unet techniques, to exploit building objects from high-resolution aerial imagery. Results obtained 92.73% accuracy carried on the Massachusetts building dataset. The proposed technique improved the performance to 0.44%, 1.17%, and 0.14% compared with fully convolutional neural network (FCN), Segnet, and Unet methods, respectively. Results also confirmed the superiority of the proposed method in building extraction.

Keywords: building extraction; image segmentation; remote sensing; Seg–Unet approach

1. Introduction

Highly accurate feature extraction from high-resolution remote sensing imagery produces reliable information for various applications (Shrestha and Vanneschi 2018). The extraction of small ground objects, such as building objects from the imagery of the surface of the earth, can be a potential application (Krizhevsky *et al.* 2012). High-precision building extraction from high-resolution satellite images can perform an essential task in several applications, such as disaster management, geospatial database updating, urban planning, and navigation (Mayer 1999). Raw data should be converted into sensible information by using geospatial information system (GIS) to enable the quantification process. The time-consuming and labor-intensive data interpretation and digitization are often required for this transformation. Although Yuan (2017) introduced a source called volunteered geographic information (VGI) as an alternative option, its availability is restricted due to the differences in positional and completeness accuracy. Participation inequality,

in terms of varying impressions, cultures, and judgments, can be the principal reasons for the aforementioned issue (Shrestha and Vanneschi 2018), thereby restricting the accessibility of dependable and up-to-date building maps. Automatic building extraction using remote sensing imagery needs a promising approach that remains underdeveloped in spite of a decade of research in this field (Marcu and Leordeanu 2016). The main elements that make this process challenging are the wide changes in building appearances in images because of various building features, such as shadows, cars, structures, various roofing materials, and illumination statuses, which are formed by buildings (Yuan and Cheriyadat 2014). Traditional methods have been mixed with genetic algorithms (Sumer and Turker 2013) and support vector machine (SVM) method (Inglada 2007) to detect buildings. Other characteristics, such as multi-spectral features, textures (Levitt and Aghdasi 1998), and shadow properties (Peng and Liu 2005); local structures, such as corners, lines, and edges (Huertas and Nevatia 1988) of remote sensing images, have been utilized as main factors for extracting building objects. The efficiency of these types of approaches is restricted due to the dependence of the method performance on low-level local characteristics. Thus, to well distinguish the features, the utilization and exploitation of representative high-level features that play a principal role in image segmentation are favorable.

In recent studies, feature-based deep convolutional approaches, such as convolutional neural network (CNN), have demonstrated that they can achieve reliable results in image classification for computer vision (He *et al.* 2015, Szegedy *et al.* 2015) and feature semantic segmentation (Vakalopoulou *et al.* 2015, Alshehhi *et al.* 2017, Abdollahi *et al.* 2020). The CNN model is efficient in image processing because of its capability to learn from raw images without following pre-processing steps. In addition, deep convolutional network (DCNN) has become a promising technique in image processing because of its ability to efficiently mix spatial and spectral features on the basis of raw input data without preprocessing (Alshehhi *et al.* 2017). Recent works have revealed that different kinds of deep learning approaches, which are based on CNNs, such as deep convolutional encode–decoder architecture and fully convolutional network (FCN), have shown significant improvements in the remote sensing field. In terms of computational proficiency and accuracy, FCN is the most proficient approach for pixel-wise semantic segmentation. However, several problems restrict model performance in detection, leading to failure in generating inadequate or redundant prediction detection and in identifying numerous objects (Shrestha and Vanneschi 2018, Abdollahi *et al.* 2020). In the next section, previous studies related to applying

promising CNN methods for remote sensing image classification and building semantic segmentation are discussed.

Deep neural network features have illustrated their ability in semantic segmentation (Long *et al.* 2015, Chen *et al.* 2017), object detection (Girshick *et al.* 2014), and visual identification (Sharif Razavian *et al.* 2014, Audebert *et al.* 2016). Deep convolutional frameworks can be utilized in different remote sensing tasks, such as data merging (Kussul *et al.* 2016), image classification (Yang *et al.* 2018), and detection (Audebert *et al.* 2017). These networks have been successfully utilized to label and classify high-resolution remote sensing images (Penatti *et al.* 2015). Marmanis *et al.* (2018) introduced a deep neural network on the basis of an end-to-end trainable network (DCNN) for detecting boundaries and improving semantic image segmentation. Farabet *et al.* (2012) mixed conditional random fields (CRFs) with multi-scale CNNs to classify dense street scenes. Vakalopoulou *et al.* (2015) implemented a deep convolutional model to identify building features from high-resolution multi-spectral images. Previous works have confirmed that the results of remote sensing imagery classification cannot be decisive (Wilkinson 2005) because improving the resolution of remote sensing images is more useful in the identification and detection of different features on the ground. However, the separation of certain objects with the same spectral values has become difficult due to these improvements, leading to the decrease of the inter-class difference and increase of the intra-class difference of objects such as cars, shadows, streets, and buildings (Paisitkriangkrai *et al.* 2016). That is, extracting sensible spatial features to solve the pixel classification in building extraction has become challenging because various objects may represent similar spectral classes in remote sensing images. Reliable results have been recently achieved by FCN for semantic image segmentation (Fu *et al.* 2017). The method can identify various object classes, including their shapes, such as trees, road objects, and building curves. The model can not only identify the structures of spatial objects but also learn how to categorize pixels and detect what they are (Audebert *et al.* 2016). However, the outcomes are visually degraded during image classification and segmentation when using FCN. The reason is that the model cannot detect objects with multiple borders or small objects because object boundaries are blurred (Maggiori *et al.* 2017). The structures of deep convolutional frameworks have been developed in certain research either by utilizing CRFs mixed with dilated convolution (Chen *et al.* 2014) or by appending skip-layer structure after up-sampling to regenerate high-frequency and comprehensive image information (Marmanis *et al.* 2016), thereby leading to the

performance improvement of semantic segmentation and accuracy improvement of image classification (Sherrah 2016).

Recent works have attempted to boost precision in areas such as pixel labeling; feature extraction from raw data; image encoding, specifically for high-resolution remote sensing imagery on the basis of deep convolutional techniques, such as FCN and CNN (Volpi and Tuia 2016). However, impervious and building objects extracted from high-resolution remote sensing images are difficult to handle due to the presence of various geometric shapes and spatial and spectral features. That is, similar objects in urban areas have various spectral values because high-resolution remote sensing images are usually restricted to three or four channels, and these spectral characteristics may lack the capability to recognize objects. Various objects may also present the same spectral values (e.g., roofs and roads) (Bakhtiari *et al.* 2017, Abdollahi *et al.* 2018).

Although prior scholars have presented helpful insights into different approaches, which can be utilized in pixel labeling, these approaches misclassify certain pixels with the same spectral values and lack the capability to eliminate salt-and-pepper classification noise and to clearly identify object boundary. To solve these issues, we present a new deep neural network called SegUnet, a combination of Segnet and Unet architectures for building objects extraction by using high-resolution aerial imagery. The proposed network is dedicated to restoring pixel position information and produces a high-resolution segmentation map. The model has an encoder–decoder architecture that incorporates index pooling (Segnet) and skip connection (Unet) to generate and disseminate image spatial information. As can be seen in the aforementioned literature review, the proposed method has not been used before, and this study is the first to propose this kind of approach for a given task. The proposed approach is compared with other state-of-the-art deep learning-based techniques, such as FCN (Long et al. 2015), Segnet (Badrinarayanan et al. 2017), and Unet (Ronneberger et al. 2015) on the basis of a similar dataset to demonstrate the ability of the method in building extraction. Such outcomes prove that the new proposed network is efficient in building extraction. The remainder of the paper is organized in the following manner. Section 2 outlines the methodology of the suggested SegUnet approach. Section 3 highlights the results and discussion. Section 4 provides the conclusion.

2. Materials and methodology

In this section, we explain the overall framework of Unet, Segnet, and SegUnet models (Figure 1). Subsequently, the prepared high-resolution remote sensing aerial dataset for applying the proposed approach is explained. Finally, the common metrics for calculating the performance of state-of-the-art techniques applied for building extraction are described.

Figure 1. around here

2.1. Unet architecture

The Unet model is an elegant DCNN that can yield accurate image segmentations. The main concept of the Unet model is the replacement of pooling layers with up-sampling operators to complete a typical contracting network by continuous layers, followed by the enhancement of output layer resolution. For localization, the high-resolution features of the contracting part are mixed with up-sampled output. Finally, continuous convolution layer can be used to assemble an accurate outcome on the basis of this information (Long *et al.* 2015). One significant factor in the Unet model is the several feature channels in the up-sampling section where the network can spread context information to layers with high resolution. The Unet deep learning model comprises two principal sections: expansive part (right side) and contracting part (left side). Given that the contracting and expansive parts are symmetric, a U-shaped CNN is formed. The model only utilizes the right part of every convolution and does not have any fully connected layers. For example, considering that the segmentation map only contains pixels, the entire context is accessible in the input image. Therefore, an overlap-tile strategy is utilized to provide a monolithic and random segmentation of large images. For extrapolating the missing context and foretelling pixels in the border section of images, input image mirroring is also utilized. The resolution can be restricted by the GPU memory unless the tiling strategy implements the network to extensive images.

The generic framework of a Unet model is followed by a contracting path that includes two repeated convolution layers of 3×3 window size, followed by a down-sampling layer of 2×2 window size. Activation function (1), which is a kind of transformation function, is used in the convolution process. Assuming that a weight vector is w; a bias vector is b; $x_k(ii, jj)$ is the input of activation function and the output of convolution operation, respectively.

$$Z(x_k(ii,jj)) = f\left(\sum_{k=1}^{k} x_k(ii,jj)\cdot w_k + b_k\right) \Leftrightarrow Z = f(X\cdot W + b). \quad (1)$$

For $f(\cdot)$, activation function (2), that is, rectified linear unit (ReLU) is used in the Unet model. Neurons do not confront the gradient vanishing issue, which arises when the gradient norm declines after sequential updates in the back-propagation process. Neurons also efficiently operate with rectified function because this function encourages sparsity in the hidden layers and prevents saturation during the learning process (Zhou *et al.* 2014). In each down-sampling stage, the number of feature channels is doubled. As previously mentioned, max-pooling layers are utilized to decrease image size, parameter number, and network computing. In the down-sampling method, images are sampled using their principle local correlations. This approach retains efficient information while lessening data processing and allowing the features taken through convolution to have spatial uniformity (Maggiori *et al.* 2017).

An up-sampling, followed by a convolution with a stride of 2×2 that halves the number of feature channels, is used in each step of the expansive path. Two convolution layers of 3×3 kernel size, followed by the ReLU activation function and a concatenation with the correspondingly cropped feature map from the contracting path, are utilized in the expansive path. Eventually, a convolution layer of 1×1 window size and a sigmoid function (3) are utilized for mapping every 32-component feature vector to the desired number of classes (road and non-road) and for mapping the predicted values to probabilities, respectively (Hu *et al.* 2015). The generic framework of the Unet model is illustrated in Figure 2.

$$A(x_k(ii,jj)) = \max(0, Z(x_k(ii,jj))), \quad (2)$$

where $x_k(ii, jj)$ is utilized as the input to the activation function and the output of convolution operation, respectively.

$$S(z) = \frac{1}{1+e^{-z}}, \qquad (3)$$

where S is the output between 0 and 1, and z is the input.

Figure 2. around here

177

178 **2.2. Segnet architecture**

179 The Segnet model consists of encoder and corresponding decoder parts, followed by the last layer
180 of pixel-wise classification (Badrinarayanan *et al.* 2017). The overall architecture of deep
181 convolutional Segnet model is illustrated in Figure 3. Each layer in the encoder part has a
182 corresponding layer in the decoder part, and both sections include 13 convolutional layers that
183 correspond to the initial 13 convolutional layers in the model named the VGG16 network
184 (Simonyan and Zisserman 2014), which is outlined for feature classification. A multi-class
185 classifier named Softmax (Equation 4) is fed into the last decoder network to generate independent
186 classification possibilities for individual pixels. Softmax output transforms into possibility
187 dispensation as it always ranges [0–1] and adds up to 1. The *n* channel of image possibility is the
188 output of the Softmax classifier, where *n* presents the number of classes, *x* is the output vector of
189 the model, and index *i* is in the range (0, ..., n-1). $s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_i}}$. (4)

190 For producing and batch normalizing a collection of feature maps, every encoder in the encoder
191 network implements a filter bank with a convolution. Subsequently, ReLU is utilized as an
192 activation function, followed by max-pooling layers with a kernel size of 2×2. Using a factor of 2,
193 the outcoming output is sub-sampled. For achieving translation invariance over tiny spatial
194 changes in the input data, max-pooling layers are utilized. Although additional translation
195 invariance for strong classification can be obtained by multiple map-pooling layers, a
196 corresponding spatial resolution loss of feature maps occurs. Therefore, before implementing sub-
197 sampling, storing, and capturing, boundary information is essential in encoder feature maps. For
198 up-sampling the input feature maps in the decoder network, the memorized sub-sampling indices
199 from the corresponding encoder feature maps are utilized. Dense feature maps are produced by
200 convolving a trainable decoder filter back with these feature maps. Subsequently, a batch
201 normalization step is implemented to each map. The whole feature map in the Unet model
202 (Ronneberger *et al.* 2015) is first transferred to the corresponding decoders, and then is
203 concatenated to up-sample decoder feature maps (using deconvolution), whereas the Segnet model
204 reutilizes pooling indices. In addition, the Segnet model utilizes the whole weights of the pre-

trained convolutional layer from the VGG network as pre-trained weights, whereas no max-pool 5 block and conv 5 exist in the Unet model as in the architecture of the VGG network.

Figure 3. around here

2.3 Seg-Unet architecture

Similar to the Unet architecture, the SegUnet model comprises three sections (Do *et al.* 2019): 1) The encoder or contracting part, which is similar to the VGG network, has four blocks. In every block, two convolution layers are followed by batch normalization and max-pooling layers. After every max-pooling index, the number of features is doubled in the convolutional layer. 2) Bottleneck, which only comprises two convolution layers, is a place for storing sparse feature maps. 3) The decoder or expanding part restores the input image resolution by using up-sampling layers. For transferring local contextual information into the decoder part, each encoder layer is connected to the corresponding decoder layer. Unlike the Unet model, the same padding is utilized instead of valid padding. For classifying each pixel and generating the segmentation map, a 1×1 convolution layer with sigmoid function is utilized at the last decoder block. The loss function of binary cross-entropy is also applied to quantify the contrast between two possibility spreads and assess the efficiency of the technique whose output value may be between 0 and 1. The over-fitting issue can be prevented from using the new network because data normalization uses the batch normalization layer, which is placed after the convolutional layer. Moreover, the sparse feature map can be well restored using an up-sampling layer on the basis of the max-pooling index in the decoder network. The overall framework of the proposed SegUnet deep neural network is demonstrated in Figure 4.

Figure 4. around here

2.4. Dataset

To apply the proposed SegUnet model for building extraction, the Massachusetts building dataset (Mnih 2013) is used. Given the computational restriction, the original dataset that contains 10, 4, and 137 aerial images for test, validation, and training with a spatial dimension of 1500×1500 pixel dimension is divided into the size of 384×384, respectively. The total number of images used in this study is 1,564, where 1,532, 24, and 8 images are considered for training, validation, and test, respectively. Certain samples of a building dataset with various scenes are depicted in Figure 5.

Figure 5. around here

2.5. Evaluation metrics

In this study, four principal calculation measurements, namely, overall accuracy (OA) (5), F1 score (6), recall (7), and precision (8) are utilized on the basis of the confusion matrix (Ghasemkhani *et al.* 2020) with four main factors, such as false negative (FN), false positive (FP), true negative (TN), and true positive (TP), to assess the model performance for extracting building features from high-resolution aerial imagery. OA is specified as the sum of rightly identified pixels divided by the entire number of pixels. Precision is calculated as a percentage of precisely identified pixels among the identified pixels of the building. Meanwhile, F1 score is the combination of recall and precision metrics. Recall is specified as a percentage of correctly predicted pixels among all the actual pixels of building, whereas F1 score is the combination of recall and precision (Wang *et al.* 2020).

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$F1 = \frac{2 \times \mathrm{Pr}ecision \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \quad (6)$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (7)$$

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \quad (8)$$

3. Results and performance evaluation

In this part, the quantitative and visual results of the proposed SegUnet model and other state-of-the-art building extraction approaches, such as the Segnet model for semantic pixel-wise segmentation, the FCN model for image semantic segmentation, and the deep convolutional Unet model, are discussed.

3.1. Experiment results

By considering a representative section for the images with a specific attention on various surroundings and building features, the visual inspection of classification maps achieved by the proposed SegUnet model was implemented for qualitative analysis. For training the proposed model, the ground truth labels, and all the prepared samples were treated as inputs to the model. The parameters and framework of the proposed approach, such as the number of blocks and size of each block, are illustrated in Figure 4. For updating the parameters of the proposed model and minimizing the energy function while training the network, an exceptional optimization algorithm is needed. Therefore, in our network, we utilized one of the most common optimizers called adaptive moment estimation (Adam) to update parameters such as biases and weights and to lessen the losses. We set the learning rate of the SegUnet network to $1e-4$ during training to speed up the processing and achieve an improved performance. In this study, the whole process of the introduced network for extracting building features from aerial imagery was performed on a GPU Nvidia Quadro P5000 with a computation capacity of 6.1 and a memory of 16 GB under the framework of Keras with TensorFlow backend.

Figure 6 depicts the results of two images obtained by the proposed approach for building extraction. The figure is presented in three columns and four rows. The first, second, and third columns respectively represent the original image, the ground truth image, and the building segmentation results, which were obtained by the SegUnet model. Meanwhile, the second and fourth rows represent the zoomed results. Figure 6 shows that the proposed SegUnet model achieves the OA of 92.33% and 91.3% for Image 1 and Image 2, respectively, proving that the model can generally extract buildings from high-resolution aerial images accurately. However, the FN (illustrated as blue pixel) and FP (illustrated as green pixel) of the identified pixels illustrate multiple failures for our suggested approach and show multiple issues with the data. The proposed approach can identify a building where tiny nearby buildings emerge as a joined area, which increases the FP

between the spaces of the building. However, the proposed model cannot make a right building prediction where no building is found in the label image, but one exists in the original image that appears as an FP prediction.

Figure 6. around here

3.2. Discussion

To verify the performance of the proposed SegUnet technique for extracting building objects from high-resolution remotely sensing aerial imagery, we compared the method with other DCNNs. Specifically, we compared the suggested SegUnet model with the deep convolutional encoder–decoder approach called the Segnet model, FCN technique, and deep convolutional Unet model. By comparing the results achieved via Segnet and Unet models with the results of the proposed SegUnet model, the difference in the accuracy for building extraction can be witnessed.

The visual outcomes of building extraction by using the suggested SegUnet model and other comparative techniques for calculating the efficiency of the SegUnet approach in building extraction are illustrated in Figure 7. The obtained outcomes demonstrate that the influence of shortcomings can be reduced to a specific degree by using the proposed methods because these methods consider the spatial information for semantic segmentation. However, FCN and Segnet approaches predict additional FNs and FPs, which are depicted by blue and green colors, respectively. Thus, these methods cannot precisely preserve and achieve boundary information, leading to the detection of FNs and FPs and production of a low-resolution segmentation map. The Unet model, which utilizes deconvolution layers and skip connection, can also achieve and preserve boundary information with higher accuracy than FCN and Segnet methods, thus obtaining a correct segmentation map. By contrast, the proposed SegUnet model, which utilizes skip connection (Unet) and index pooling (Segnet), can predict fewer FNs and FPs, preserve boundary information, and produce a correct segmentation map compared to other comparative approaches.

Figure 7. around here

To test the efficiency of the introduced SegUnet approach for building extraction in comparison with other DCNNs, we demonstrated the quantitative results of the techniques in Table 1. The first eight rows of Table 1 present the quantitative accuracy of the four main metrics achieved by the comparative approaches for the eight images, whereas the last row presents the average accuracy of the metrics. As shown in Table 1, the FCN model can obtain higher accuracy for the recall factor than other methods because the model predicts many FNs. By contrast, the Unet method can obtain higher accuracy for precision and OA factors than FCN and Segnet methods. Moreover, the Unet method is the second-best approach in building extraction and can obtain a correct segmentation map. Finally, the average accuracy for F1 score and OA factors achieved by the proposed SegUnet model is higher than those by other techniques with almost 0.14%, 1.17%, and 0.44% higher than the Unet, Segnet, and FCN approaches, respectively. These results indicate that the proposed model can improve the results and exceed other state-of-the-art techniques in building extraction from high-resolution remote sensing imagery. Figure 8 plots the clear differences between the introduced SegUnet model and other deep learning approaches for building object segmentation. Figure 8 also illustrates that the proposed SegUnet network achieves higher precision for the OA factor than other techniques.

Figure 8. around here

Table 1 around here

4. Conclusion

For extracting building objects from high-resolution aerial imagery, we presented a new deep neural network called the SegUnet model, which is a combination of Segnet and Unet techniques, in this work. We applied the proposed model on the Massachusetts building dataset. After training and validating the method, we utilized four accuracy metrics to assess the efficiency of the indicated technique in building extraction, which achieved a 92.73% accuracy on average for OA. This result indicated that the proposed model can produce a correct segmentation map and can accurately extract building objects. Furthermore, we compared the visual and quantitative results of the proposed SegUnet model with those of other deep learning techniques, such as Segnet, FCN, and Unet models, to show its effectiveness. The results confirmed that the proposed method obtained the best

quantitative and visual performances and outperformed other DCNNs in building extraction from high-resolution aerial imagery.

Author Contributions: Conceptualization, A.A. and B.P.; methodology and formal analysis, A.A.; data curation, A.A.; writing—original draft preparation, A.A.; writing—review and editing, B.P.; supervision, B.P.; funding, B.P. and A.A.

Funding: This research is supported by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, the University of Technology Sydney (UTS). This research is also supported by Researchers Supporting Project (RSP) number RSP-2020/14, King Saud University, Riyadh, Saudi Arabia.

Conflict of Interest: The authors declare no conflict of interest.

References

1. Abdollahi, A., Bakhtiari, H.R.R. & Nejad, M.P., 2018. Investigation of svm and level set interactive methods for road extraction from google earth images. *Journal of the Indian Society of Remote Sensing*, 46 (3), 423-430.
2. Abdollahi, A., Pradhan, B., Shukla, N., Chakraborty, S. & Alamri, A., 2020. Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review. *Remote Sensing*, (12), 1444.
3. Alshehhi, R., Marpu, P.R., Woon, W.L. & Dalla Mura, M., 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry Remote Sensing*, 130, 139-149.
4. Audebert, N., Boulch, A., Lagrange, A., Le Saux, B. & Lefevre, S., 2016. Deep learning for remote sensing. Technical Report. DOI: 10.1109/JURSE.2017.7924536.
5. Audebert, N., Boulch, A., Randrianarivo, H., Le Saux, B., Ferecatu, M., Lefèvre, S. & Marlet, R., 2017. Deep learning for urban remote sensing. *Joint Urban Remote Sensing Event (JURSE)*IEEE, 1-4. DOI: 10.1109/JURSE.2017.7924536.
6. Audebert, N., Le Saux, B. & Lefèvre, S., 2017. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. *Asian Conference on Computer Vision*, 180-196. DOI: 10.1007/978-3-319-54181-5_12.
7. Badrinarayanan, V., Kendall, A. & Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 39 (12), 2481-2495.
8. Bakhtiari, H.R.R., Abdollahi, A. & Rezaeian, H., 2017. Semi automatic road extraction from digital images. *The Egyptian Journal of Remote Sensing and Space Science*, 20 (1), 117-123 Available from: http://www.sciencedirect.com/science/article/pii/S1110982317300820.

9. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A.L., 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. 834 - 848. DOI: 10.1109/TPAMI.2017.2699184.
10. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.J.I.T.O.P.A. & Intelligence, M., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. 40 (4), 834-848.
11. Do, N.-T., Joo, S.-D., Yang, H.-J., Jung, S.T. & Kim, S.-H., 2019. Knee bone tumor segmentation from radiographs using seg-unet with dice loss. *25th International Workshop on Frontiers of Computer Vision (IW-FCV), Gangneung, South Korea*.
12. Farabet, C., Couprie, C., Najman, L. & Lecun, Y., 2012. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis Machine Intelligence,* 35 (8), 1915-1929.
13. Fu, G., Liu, C., Zhou, R., Sun, T. & Zhang, Q., 2017. Classification for high resolution remote sensing imagery using a fully convolutional network. *Remote Sensing,* 9 (5), 498.
14. Ghasemkhani, N., Vayghan, S.S., Abdollahi, A., Pradhan, B. & Alamri, A., 2020. Urban development modeling using integrated fuzzy systems, ordered weighted averaging (owa), and geospatial techniques. *Sustainability,* 12 (3), 809.
15. Girshick, R., Donahue, J., Darrell, T. & Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580-587. Available from: https://arxiv.org/abs/1311.2524.
16. He, K., Zhang, X., Ren, S. & Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision*, 1026-1034. Available from: https://arxiv.org/abs/1502.01852.
17. Hu, F., Xia, G.-S., Hu, J. & Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing,* 7 (11), 14680-14707.
18. Huertas, A. & Nevatia, R., 1988. Detecting buildings in aerial images. *Computer Vision, Graphics, Image Processing,* 41 (2), 131-152.
19. 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 Remote Sensing,* 62 (3), 236-248.
20. Krizhevsky, A., Sutskever, I. & Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097-1105. DOI: 10.1145/3065386.
21. Kussul, N., Shelestov, A., Lavreniuk, M., Butko, I. & Skakun, S., 2016. Deep learning approach for large scale land cover mapping based on remote sensing data fusion. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 198-201. DOI: 10.1109/IGARSS.2016.7729043.

22. Levitt, S. & Aghdasi, F., 1998. An investigation into the use of wavelets and scaling for the extraction of buildings in aerial imagesed. *Proceedings of the 1998 South African Symposium on Communications and Signal Processing-COMSIG'98 (Cat. No. 98EX214)*, 133-138. DOI: 10.1109/COMSIG.1998.736936.
23. Long, J., Shelhamer, E. & Darrell, T., 2016. Fully convolutional networks for semantic segmentationed. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431-3440. DOI: 10.1109/TPAMI.2016.2572683.
24. Maggiori, E., Tarabalka, Y., Charpiat, G. & Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience Remote Sensing,* 55 (2), 645-657.
25. Marcu, A. & Leordeanu, M., 2016. Dual local-global contextual pathways for recognition in aerial imagery. *Available from: https://arxiv.org/abs/1605.05462.*
26. Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M. & Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry Remote Sensing,* 135, 158-172.
27. Marmanis, D., Wegner, J.D., Galliani, S., Schindler, K., Datcu, M. & Stilla, U., 2016. Semantic segmentation of aerial images with an ensemble of cnss. *ISPRS Annals of the Photogrammetry, Remote Sensing Spatial Information Sciences,* 3, 473-480.
28. Mayer, H., 1999. Automatic object extraction from aerial imagery—a survey focusing on buildings. *Computer Vision Image Understanding,* 74 (2), 138-149.
29. Mnih, V., 2013. *Machine learning for aerial image labeling, ph.D. Dissertation, dept. Comput. Sci., univ. Toronto, Canada*.
30. Paisitkriangkrai, S., Sherrah, J., Janney, P. & Van Den Hengel, A., 2016. Semantic labeling of aerial and satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations Remote Sensing,* 9(7), 2868-2881.
31. Penatti, O.A., Nogueira, K. & Dos Santos, J.A., 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 44-51. DOI: 10.1109/CVPRW.2015.7301382.
32. Peng, J. & Liu, Y., 2005. Model and context-driven building extraction in dense urban aerial images. *International Journal of Remote Sensing,* 26 (7), 1289-1307.
33. Ronneberger, O., Fischer, P. & Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentationed. *International Conference on Medical Image Computing and Computer-assisted Intervention*, 234-241. DOI: 10.1007/978-3-319-24574-4_28.
34. Sharif Razavian, A., Azizpour, H., Sullivan, J. & Carlsson, S., 2015. Cnn features off-the-shelf: An astounding baseline for recognitioned. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 806-813. DOI: 10.1109/CVPRW.2014.131.
35. Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *Available from: https://arxiv.org/abs/1606.02585.*

36. Shrestha, S. & Vanneschi, L., 2018. Improved fully convolutional network with conditional random fields for building extraction. *Remote Sensing,* 10 (7), 1135.
37. Simonyan, K. & Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. 1-14. *Available from: https://arxiv.org/abs/1409.1556.*
38. Sumer, E. & Turker, M., 2013. An adaptive fuzzy-genetic algorithm approach for building detection using high-resolution satellite images. *Computers, Environment Urban Systems,* 39, 48-62.
39. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A., 2015. Going deeper with convolutionsed. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1-9. DOI: 10.1109/CVPR.2015.7298594.
40. Vakalopoulou, M., Karantzalos, K., Komodakis, N. & Paragios, N., 2015. Building detection in very high resolution multispectral data with deep learning features. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*IEEE, 1873-1876. DOI: 10.1109/IGARSS.2015.7326158.
41. Volpi, M. & Tuia, D., 2016. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience Remote Sensing,* 55 (2), 881-893.
42. Wang, S., Hou, X. & Zhao, X., 2020. Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block. *IEEE Access,* 8, 7313-7322.
43. Wilkinson, G.G., 2005. Results and implications of a study of fifteen years of satellite image classification experiments. *IEEE Transactions on Geoscience Remote Sensing,* 43 (3), 433-440.
44. Yang, X., Ye, Y., Li, X., Lau, R.Y., Zhang, X. & Huang, X., 2018. Hyperspectral image classification with deep learning models. *IEEE Transactions on Geoscience Remote Sensing,* 56 (9), 5408-5423.
45. Yuan, J., 2017. Learning building extraction in aerial scenes with convolutional networks. *IEEE Transactions on Pattern Analysis Machine Intelligence,* 40 (11), 2793-2798.
46. Yuan, J. & Cheriyadat, A.M., 2014. Learning to count buildings in diverse aerial scenes. *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 271-280. DOI: 10.1145/2666310.2666389.
47. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A. & Oliva, A., 2014. Learning deep features for scene recognition using places database. *27th International Conference on Neural Information Processing Systems, Montreal, Canada, 1, 487-495.*