ICIC Express Letters
Part B: Applications
Volume **13**, Number **10**, October **2022**

ICIC International ©2022 ISSN 2185-2766
pp. **1009–1018**

COMPARISON OF DEEP LEARNING MODELS FOR DENSE BUILDING SEGMENTATION IN THE CITY USING AERIAL IMAGERY DATA

Dewa Ayu Defina Audrey Nathania, Calvin Surya
Alexander Agung Santoso Gunawan and Edy Irwansyah

Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ dewa.nathania; calvin.surya }@binus.ac.id; { aagung; eirwansyah }@binus.edu

Received February 2022; accepted May 2022

Abstract. *The process of automatic building detection has become a concern for city governments in many developing nations as it is the foundation for urban planning and other purposes, particularly in high-density cities. Deep learning has attracted significant interest in recent years as the most appealing technique for problem resolution in the field of remote sensing. The semantic image segmentation approach, which tries to categorize each pixel in an image into a collection of specified labels, is one application of deep learning. This research proposed recognizing buildings in aerial photographs using deep learning models, specifically U-Net, PSPNet, and DeepLab v3. The model employed is semantic image segmentation, with each image pixel assigned a building and a backdrop class. According to the findings of this study, DeepLab v3 with the ResNet 101 backbone provides very good precision results, with 94.5% in the training set and 88.1% in the test set. DeepLab ResNet 152 produced no significant modifications, indicating that DeepLab ResNet 101 was enough for detecting buildings.*
Keywords: Building detection, Semantic image segmentation, Aerial imagery, DeepLab

1. **Introduction.** Automated building extraction from high resolution satellite imagery is a significant research problem that currently faces various challenges owing to the wide range of variables that must be taken into consideration. Deep learning, a game-changing method used in many remote sensing techniques, has a startling ability to detect structures in satellite or aerial images. There are already a variety of strategies and algorithms available to increase building detection performance. Digital image processing is the computational processing, modification, and interpretation of visual information using computer technology in computer science. With the assistance of current technology, several techniques for efficiently processing visual information have been developed.

Several methods [1] for detecting building structures from standard-contrast high-resolution satellite data images have been presented. The literature research revealed image processing discoveries that have been frequently used in a range of fields, including remote sensing [2,3], object recognition [3,4], image cropping and segmentation [5,6], and others. Zhang et al. [7] proposed a technique for detecting buildings in high-resolution remote sensing data images with characteristic contrast on a global scale. The suggested method generates a population size map for building extraction.

Semantic segmentation is a deep learning approach for labelling each pixel of raster geographic data with a collection of semantic labels such as highways, rivers, vacant land, and buildings. Wu et al. [8] suggested a demarcation system made up of a modified U-Net and a multitasking framework for generating segmentation maps and constructing

DOI: 10.24507/icicelb.13.10.1009

contours while taking boundary control into consideration. To address the issue of unclear object borders, Marmanis et al. [9] suggested a deep convolutional neural model for the segmentation of high-resolution aerial photographs that clearly includes category borders in the segmentation stage.

The state of the art in building extraction with satellite or aerial imagery with a deep learning model using a semantic image segmentation method is as has been done by [10-13]. Despite several studies, detecting buildings in developing-country metropolitan areas with high-resolution satellite data remains a difficult challenge. Chaurasia et al. [14] developed a PSPNet model to show authorities the semantic segmentation process for smart city planning and land use mapping. Liu et al. [15] used DeepLab v3+ to accomplish building extraction from high-resolution remote sensing data. The main challenge is, but is not limited to, variations in the shape of taller buildings, especially in many parts of large dense urban areas such as the city of Jakarta, Indonesia. [16] compared to less dense urban areas such as Palu city, Central Sulawesi, and residential buildings located in new housing complexes which tend to be more organized in structure.

Based on the facts regarding the best performance of various deep learning models that have been separately conducted in previous studies as well as the special problems faced in building segmentation, in this paper conduct to detect buildings in dense areas is performed by comparing three distinct deep learning models, including DeepLab (one of the models) utilizing four different backbones, particularly DeepLab (with ResNet 34, 50, 101, and 152), PSPNet (with ResNet 50), and U-Net (ResNet 34). The structure of this paper is as follows. The first section discusses the purpose of the paper. Section 2 discusses similar work on semantic segmentation for building detection. Following the presentation of our technique in Section 3, Section 4 presents the experimental results. In Section 5, we conclude with a discussion.

2. Semantic Segmentation for Building Detection.

2.1. Semantic segmentation using deep learning.

Deep learning algorithms have outperformed several classic computer vision applications in the last decade, including object classification [17,18], detection [19], and semantic segmentation [20,21]. Deep learning studies on semantic segmentation have advanced significantly in recent years. Semantic image segmentation is a basic process in feature extraction that assigns a label to every component, often known as image pixels.

Semantic image segmentation gives a more detailed understanding of images than image classification. To establish the limits of each item in an image classification job, researchers must allocate each pixel to a specific classifier or background in addition to recognizing the items. Two instances of semantic segmentation of remote sensing data are building classification and land use analysis in urban regions [22].

Deep learning in geospatial analysis research has been claimed to have generated effective results through the construction of learning models to solve different tasks using object identification, instance segmentation, or semantic segmentation techniques. According to [23], semantic segmentation task can be formulated as follows. Given an image as a pixel set $S = \{(x_i, y_i), i = 1, 2, \ldots, N\}$ where x_i is a pixel value, N is the number of pixels, $y_i \in \{C_1, \ldots, C_m\}$ is a pixel label, and m is the number of pixel classes. Given θ as a model paramater with activation function $f(x_i; \theta)$, semantic segmentation is a representation learning to minimize crossentropy as an objective function which can be represented as

$$\mathcal{L}(u, y) = -\sum_k y_k \log u_k \tag{1}$$

where u is the groundtruth and $\mathcal{L}(u, y)$ is the crossentropy of y from u.

Optimization process of the objective function can be implemented using supervised learning algorithm, e.g., gradient descent, to predict the most optimum model parameters so that

$$\theta^* = \arg\min_{\theta} \sum_{i=1,\ldots,N} \mathcal{L}\left(\left(f\left(x^i;\theta\right)\right), y^i\right) \tag{2}$$

In remote sensing, semantic segmentation seeks to precisely categorize each pixel in an aerial picture by assigning it to a certain class, such as vegetation, buildings, cars, or roads [24]. This is a critical activity that enables a wide range of applications, ranging from urban planning to change detection and automated map generation. The process of dividing a photo into meaningful pieces, each of which corresponds to one of the pre-specified classes, is known as semantic image segmentation. Extracting things of interest from a scene is one of the problems that semantic segmentation may help with. [25] proposed a method for extracting buildings from aerial data that had previously been explored using different approaches.

Semantic segmentation seeks to simplify visual representation so that it may be more readily studied or comprehended. After the image representations have been segmented and labeled, it can be condensed into a more easily analyzed or comprehended format. Given a digital image as input, semantic image segmentation produces a collection of segments that indicate the semantic class of each pixel in the segment. When comparing pixels from different classes, each pixel in an object class shares some characteristics, such as color, color intensity, or texture.

CNN is used by Saito and Aoki [26] for road and building detection. The study used CNN's typical downsampling architecture, and in the end, a fully connected layer with dropout [27] was added to anticipate the input image. Their method outperforms Mnih's [28] models for both roads and buildings, while using a single model for each class.

2.2. **Building detection using aerial imagery.** Aerial imagery is an important source for land surface analysis, which can yield land use maps. Aerial images provide a larger range of vision than ground search and rescue and can help to avoid the risks associated with ground search and rescue. Because of subjective human variables, substantial false and missed detections may arise when a manually reviewed picture is utilized to analyze a damaged region. As an outcome, it is a difficult task to interpret aerial photos to detect and estimate the level of damage in a region.

Building detection from aerial and satellite images has been a prominent research area for decades and is of significant interest since it plays an important role in building model development, map updating, urban planning, and reconstruction. Remote sensing frequently acquires photographs of specific locations and serves as effective tools for these sorts of jobs. The ease of access to high-resolution remote-sensing imagery has increased significantly in recent years because of technology breakthroughs in various applications and new platforms such as unmanned aerial vehicles [29].

Aerial or satellite photos have been used in several studies to identify damage [30]. Pre-and post-event satellite photos were utilized in their research to determine alterations caused by natural/man-made disasters. Accurate pixel classification of a wide aerial photograph is a difficult attention challenge for a person to perform because ground things vary greatly in shape, and an object might be obscured by other objects such as trees and building shadows. Saito and Aoki [26] proposed a method for training CNNs for multi-label semantic segmentation of aerial data, as well as a new output function channel-wise inhibited SoftMax (CIS) to train CNNs in such a task. Support vector machines [17], random forests [31], and conditional random fields (CRF) are examples of classifiers [32] which are used to forecast each pixel based on the retrieved characteristics. However, owing to the complexity of building structures, as well as significant similarities with the other categories (e.g., road segments), the prediction results are significantly reliant

on manual feature generation and adaptation, which frequently leads to bias and poor generality.

3. **Research Methodology.** This section describes the methods employed in this study, which consist mostly of a dataset and a deep learning model for creating segmentations from aerial photography. Research flow of this study begins with locating the dataset utilized for the training dataset, and then the dataset is carried out by Data Augmentation with horizontal flip and vertical flip (every 90 degrees becomes a new image). After the training dataset is prepared, it is separated into 80% for training dataset and 20% for validation dataset. Deep learning training procedure is used in conjunction with the training dataset and validation dataset to track the incidence of overfits. The deep learning models used are U-Net, PSPNet, and DeepLab (ResNet 34, 50, 101, and 152). The outcomes of each deep learning model (F1 Score, precision and recall score from validation dataset) are tracked. The test dataset was constructed using a subset of the training dataset. Tassehe test dataset is run with data augmentation every 30 degrees of rotation into a new picture. Evaluating six distinct trained models using the test dataset, the test dataset's F1 Score, recall, and precision values are displayed.

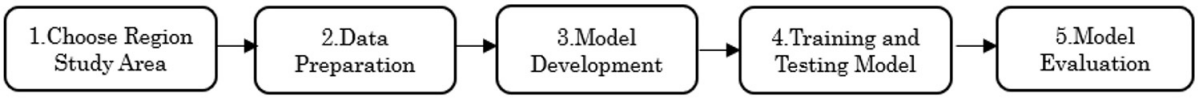


FIGURE 1. Research methodology

3.1. **Dataset.** The dataset used in this research is an aerial imagery of buildings in DKI Jakarta, Indonesia with spatial resolution 25 cm. In this research, the dataset was selected in several sub-regions such as Pasar Minggu, Kramat Jati, and Palmerah (Figure 2).

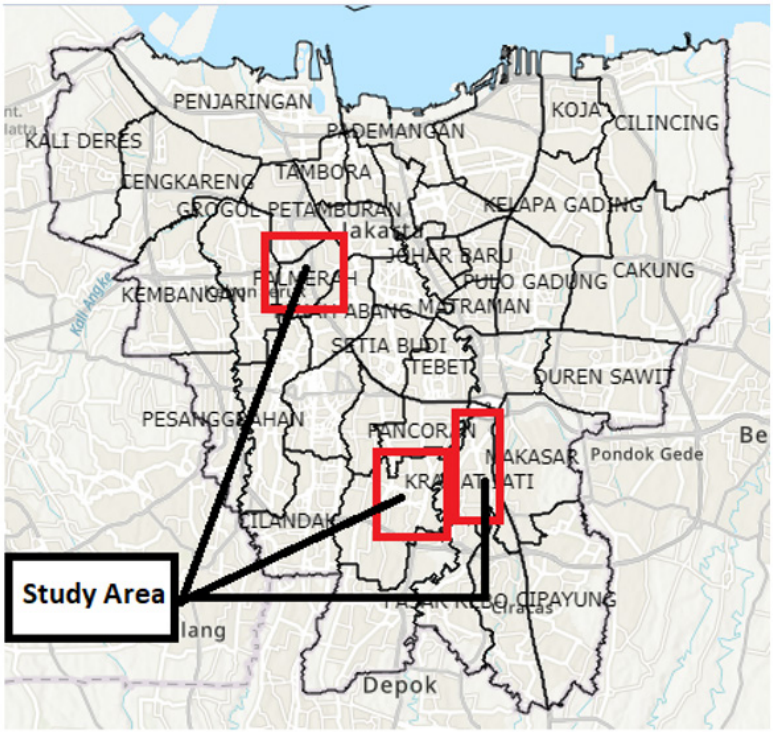


FIGURE 2. Study area in Pasar Minggu, Kramat Jati, and Palmerah sub-district

3.2. **Data preprocessing.** The training dataset consists of 6566 ground patches with $256 \times 256 \times 4$ pixel dimensions. To enhance the sample size, every 90 degrees of the training dataset was recorded, increasing the number of ground patches from 6566 to 25413. The dataset is separated into 2 different parts, which are training dataset and validation dataset, with a proportion of 80% and 20%, respectively. The training dataset has 20331 images, while the validation dataset has 5082 images.

FIGURE 3. Samples of building label of training dataset

The test dataset was created with three distinct regions in mind. The test dataset comprises 279 ground patches with dimensions of $128 \times 128 \times 4$. To extend the test dataset, 2 areas are augmented for every 30 degrees of the test dataset collected, and 1 area is augmented for every 90 degrees of the test dataset captured, bringing the total number of ground patches captured from 279 to 2061. Figure 4 depicts examples of ground truth photographs from the test dataset.

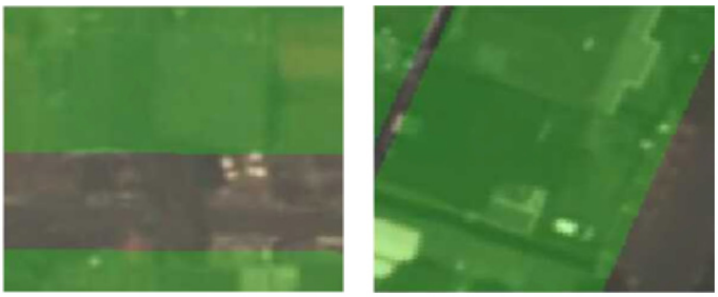


FIGURE 4. Sample of building label of test dataset

Researchers examine all traces of the buildings before creating the dataset (training and testing) to guarantee that the buildings used for train and test are accurately tagged. As demonstrated in Figure 5, researchers improved the annotations that were not adequately marked when seen visually (between the building and the label).

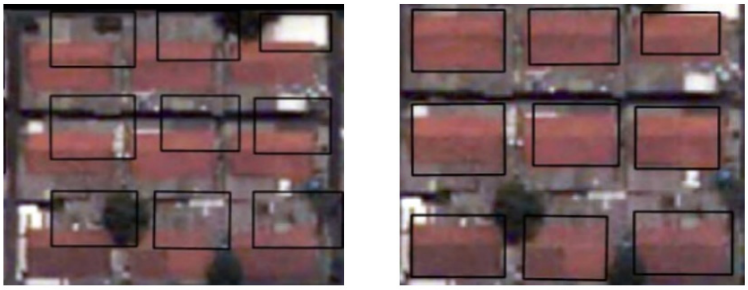


FIGURE 5. Before (left) and after repair annotations (right) on structures that are not correctly designated

3.3. **Deep learning model for building segmentation.** This research implemented three different models, with one model being used with four different depths of layer of ResNet (Residual Network), which are, first, U-Net, which was proposed in [33], and which was used with four different depths of layer of ResNet. The U-Net model is made of from 2 parts which are symmetric, and skip connection, each of which is explained as follows. a) Contracting path: This is made up of several convolution and max pooling layers that steadily reduce the dimension of the image by 2 while increasing the number of channels (depth). The U-Net is created with the ResNet 34 (34 Layers) model as the backbone for the input in the contracting path. The purpose of contracting path is to generate features by studying the objects contained in image data. b) The expansive path is made up of some transpose convolution layers and standard convolution layers that the dimension of the picture data provided by the contracting path increased by 2 while decreasing the channels in the image by 2. Expansive path seeks to recover the position of an item in image data using up-sampling that develops features carrying information about the position of an object in the image data progressively. c) Object position information in the image is retrieved by using a skip connection in conjunction with concatenation of two layers, between output of transposed convolution layer in expansive path and features created in the contracting path at the same level. The purpose of last layer of expansive path is restoring the dimensions of the output in previous layer to original image data dimensions.

Second, the PSPNet (Pyramid Scene Parsing Network) model was proposed in [34]. The PSPNet model is made up of three major components, the functions of which are as follows. a) To construct a feature map from an input image, use ResNet 50 as the backbone. b) The pyramid parsing module uses the first module's (ResNet 50) features map to extract representations of four distinct sub-representations (with 1×1, 2×2, 3×3, and 6×6 convolution). The outcome of the representation recovered in the previous layer is up-sampled and concatenated with all of the up-sampled representations as well as the features mapped in the first module. c) To acquire final semantic segmentation results, the convolutional layer employs a representation of the input picture obtained by the second module.

Third, [35] suggested the DeepLab v3 model. DeepLab is an image segmentation tool that was created to help with semantic segmentation challenges. DeepLab employs spatial pyramid pooling (SPP) at a variety of grid sizes or uses multiple parallel atrous convolutions at variable speeds (a technique known as Atrous Spatial Pyramid Pooling, or ASPP). The model architecture makes use of ResNet and Atrous convolution to extract image properties while minimizing image size, as well as improved Atrous Spatial Pyramid Pooling (ASPP). Following ResNet blocks, there are additional four layers, the first of which is a 1×1 convolution, and the final three of which are convolutions utilizing a 3×3 kernel at varied rates (rate = 6, rate = 12, and rate = 18).

4. **Results and Discussion.** Tables 1 and 2 summarize the findings of this study. It can be shown that PSPNet obtains the greatest precision in the test set, but when examined from the recall, PSPNet, as well as DeepLab with the backbone ResNet 34 and 50, cannot forecast effectively. This is due to the fact that the F1 Score and recall result from validation set and test set are significantly different. According to these results, DeepLab ResNet 34 and 50 cannot generalize the building because the ResNet layer is less deep (less complex), but the recall from DeepLab increases with the increase in the ResNet until the optimum point is reached in ResNet 101, because there is no significant difference between ResNet 101 and 152, so DeepLab with ResNet 101 is good for detecting buildings.

DeepLab ResNet 50 outperformed DeepLab ResNet 152 with a precision of 93.98 percent, outperforming it by only 87 percent. According to the test results in Table 2, PSPNet,

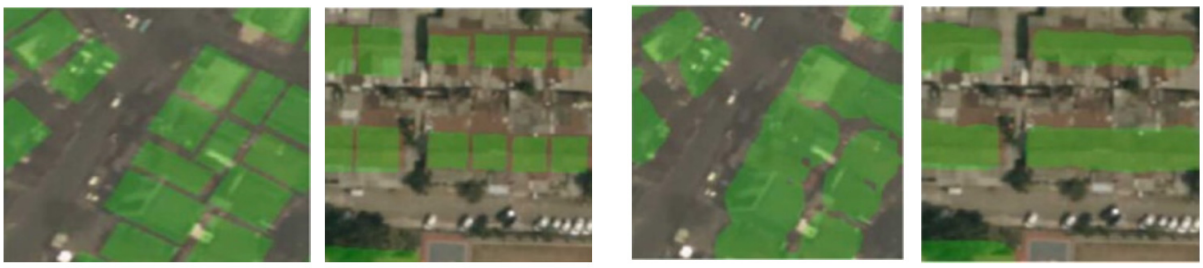
Table 1. The validation results

	Model list					
	Unet ResNet 34	PSPNet ResNet 50	DeepLab ResNet 34	DeepLab ResNet 50	DeepLab ResNet 101	DeepLab ResNet 152
F1 Score	87.1106%	84.769%	85.6986%	94.9005%	**95.0925%**	89.8088%
Precision	86.2302%	82.6462%	85.924%	93.9827%	**94.5239%**	87.6973%
Recall	88.0092%	87.0038%	85.4744%	**95.8364%**	95.668%	92.0245%

Table 2. The test results

	Model list					
	Unet ResNet 34	PSPNet ResNet 50	DeepLab ResNet 34	DeepLab ResNet 50	DeepLab ResNet 101	DeepLab ResNet 152
F1 Score	86.9625%	78.3771%	69.1274%	81.7926%	**89.1978%**	89.1619%
Precision	88.548%	**91.107%**	90.041%	89.4901%	88.1997%	89.0823%
Recall	85.4328%	68.7685%	56.0977%	75.3144%	**90.2188%**	89.2416%

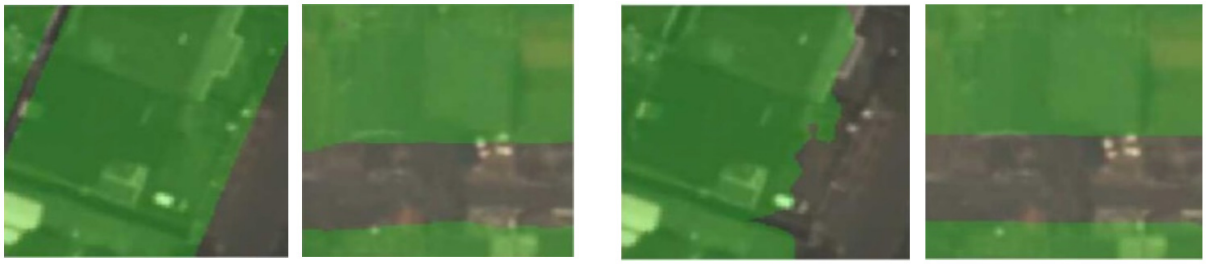
DeepLab ResNet 34, and ResNet 50 all have a high precision of 91 percent, 90 percent and 89 percent, respectively, but a low recall of roughly 68 percent, 56 percent, and 75 percent. This suggests that both models can detect a small number of structures, yet each makes an accurate prediction. In the case of the DeepLab model, the researcher hypothesizes that the model requires a deeper ResNet, as evidenced by the decreasing difference in recall and F1 Score between validation and test results for each DeepLab model, starting with recall and F1 Score of 29.3767 percent and 16.5712 percent, respectively, and ending with recall and F1 Score of 20.522 percent and 1 percent, respectively, on DeepLab ResNet 50. When researchers employ DeepLab ResNet 101, they observe a significant decline in recall and F1 Score compared to ResNet 34 and 50, specifically a ratio of 5.4492 percent for recall and 5.8947 percent for F1 Score. The best findings indicate that DeepLab ResNet 152 is the best model for detecting buildings, with a recall ratio of only 2.7829 percent and an F1 Score of 0.6469 percent. Although DeepLab ResNet 152 generated the best results, DeepLab ResNet 101 also produced good segmentation with an accuracy of 88.1 percent and F1 Score of 89.197 percent on the test set. The label and prediction of deep learning models on validation dataset and test dataset on DeepLab with ResNet 101 are shown in Figure 6 and Figure 7.



(a) (b)

Figure 6. Label (a) and prediction (b) on validation dataset using DeepLab ResNet 101

The researcher concluded two things based on the findings in Figure 6 and Figure 7. To begin with, the deep learning models that perform well at detecting buildings (DeepLab) continue to struggle with determining the minimum distance between adjacent buildings, lowering the precision, recall, and F1 Score of the two models. Second, as illustrated



(a) (b)

FIGURE 7. Label (a) and prediction (b) on test dataset using DeepLab-ResNet 101

in Figure 7, there are building annotations that pass through the roof of the building, lowering the model's performance on the test dataset (accuracy, recall, and F1 Score).

5. **Conclusion.** In this research, the detection of buildings in dense locations is accomplished by approach method called semantic segmentation using aerial pictures, by comparing 3 different models of deep learning, with DeepLab (one of the model) using 4 different backbones, notably DeepLab (with backbone ResNet 34, 50, 101, and 152), PSPNet (backbone ResNet 50) and U-Net (backbone ResNet 34). The suggested approach was taught and evaluated using aerial photographic images of Pasar Minggu, Kramat Jati, and Palmerah Subdistricts, DKI Jakarta Province, Indonesia. The DeepLab v3 model with the backbone ResNet 101 produced high precision, recall, and F1 Scores in detecting buildings, namely 88.1%, 90.2%, and 89.1% in the test set, respectively. Unet ResNet 34 is also good at detecting buildings with precision, recall and F1 Scores in the test set as 88.5%, 85.4%, and 86.9%. The suggestion for future work is that, given the limitations of this research, it can be expanded by using new models such as DeepLab V3+ and PSPNet + UNet Decoder, or by adding a ResNet layer depth to PSPNet to identify model flaws. Additionally, the model's accurate building detection is critical in research on building damage assessment and also in the preparation of large-scale building maps for detailed urban spatial planning.

REFERENCES

[1] M. Aamir, Y. Pu, Z. Rahman, M. Tahir, H. Naeem and Q. Dai, A framework for automatic building detection from low-contrast satellite images, *Symmetry*, vol.11, no.1, DOI: 10.3390/sym11010003, 2018.

[2] F. Chen, R. Ren, T. Van de Voorde, W. Xu, G. Zhou and Y. Zhou, Fast automatic airport detection in remote sensing images using convolutional neural networks, *Remote Sensing*, vol.10, no.3, DOI: 10.3390/rs10030443, 2018.

[3] M. Aamir, Y. Pu, W. Abro, H. Naeem and Z. Rahman, A hybrid approach for object proposal generation, *International Conference on Sensing and Imaging*, vol.506, pp.251-259, 2017.

[4] N. Ayoub, Z. Gao, B. Chen and M. Jian, A synthetic fusion rule for salient region detection under the framework of DS-evidence theory, *Symmetry*, vol.10, no.6, DOI: 10.3390/sym10060183, 2018.

[5] Z. Rahman, Y. Pu, M. Aamir and F. Ullah, A framework for fast automatic image cropping based on deep saliency map detection and Gaussian filter, *International Journal of Computers and Applications*, vol.41, no.3, pp.207-217, 2018.

[6] H. Naeem, G. Bing, M. Naeem, M. Aamir and M. Javed, A new approach for image detection based on refined Bag of Words algorithm, *Optik*, vol.140, pp.823-832, 2017.

[7] A. Zhang, X. Liu, A. Gros and T. Tiecke, Building detection from satellite images on a global scale, *arXiv Preprint*, arXiv: 1707.08952, 2017.

[8] G. Wu et al., A boundary regulated network for accurate roof segmentation and outline extraction, *Remote Sensing*, vol.10, no.8, DOI: 10.3390/rs10081195, 2018.

[9] D. Marmanis, K. Schindler, J. Wegner, S. Galliani, M. Datcu and U. Stilla, Classification with an edge: Improving semantic image segmentation with boundary detection, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol.135, pp.158-172, 2018.
[10] R. Hamaguchi and S. Hikosaka, Building detection from satellite imagery using ensemble of size-specific detectors, *CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp.187-191, 2018.
[11] G. Prathap and I. Afanasyev, Deep learning approach for building detection in satellite multispectral imagery, *IEEE International Conference on Intelligent Systems (IS)*, pp.461-465, 2018.
[12] W. Li, C. He, J. Fang, J. Zheng, H. Fu and L. Yu, Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data, *Remote Sensing*, vol.11, no.4, DOI: 10.3390/rs11040403, 2019.
[13] Z. Pan, J. Xu, Y. Guo, Y. Hu and G. Wang, Deep learning segmentation and classification for urban village using a worldview satellite image based on U-Net, *Remote Sensing*, vol.12, no.10, DOI: 10.3390/rs12101574, 2020.
[14] K. Chaurasia, R. Nandy, O. Pawar, R. Singh and M. Ahire, Semantic segmentation of high-resolution satellite images using deep learning, *Earth Science Informatics*, vol.14, no.4, pp.2161-2170, 2021.
[15] H. Liu, J. Luo, B. Huang et al., DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery, *Remote Sensing*, vol.11, no.20, DOI: 10.3390/rs11202380, 2019.
[16] E. Irwansyah, Y. Heryadi and A. Gunawan, Semantic image segmentation for building detection in urban area with aerial photograph image using U-Net models, *IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS)*, pp.48-51, 2020.
[17] X. Liu, Z. Deng and Y. Yang, Recent progress in semantic image segmentation, *Artificial Intelligence Review*, vol.52, no.2, pp.1089-1106, 2018.
[18] J. Shi, X. Yuan and M. Elhoseny, Weakly supervised deep learning for objects detection from images, *Urban Intelligence and Applications*, pp.231-242, 2020.
[19] X. Yuan, L. Xie and M. Abouelenien, A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data, *Pattern Recognition*, vol.77, pp.160-172, 2018.
[20] V. Badrinarayanan, A. Kendall and R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.39, no.12, pp.2481-2495, 2017.
[21] L. Chen, G. Papandreou, F. Schroff and H. Adam, Rethinking atrous convolution for semantic image segmentation, *arXiv Preprint*, arXiv: 1706.05587, 2017.
[22] F. Fang, X. Yuan, L. Wang, Y. Liu and Z. Luo, Urban land-use classification from photographs, *IEEE Geoscience and Remote Sensing Letters*, vol.15, no.12, pp.1927-1931, 2018.
[23] A. Valada, W. Dhall and Burgard, Convoluted mixture of deep experts for robust semantic segmentation, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Workshop, State Estimation and Terrain Perception for all Terrain Mobile Robots*, 2016.
[24] Y. Heryadi and E. Irwansyah, *Deep Learning: Applications within the Geospatial Sector*, AWI Technology Press, 2020.
[25] A. Khalel and M. El-Saban, Automatic pixelwise object labeling for aerial imagery using stacked U-Nets, *arXiv Preprint*, arXiv: 1803.04953, 2018.
[26] S. Saito and Y. Aoki, Building and road detection from large aerial imagery, *Image Processing: Machine Vision Applications VIII*, vol.9405, DOI: 10.1117/12.2083273, 2015.
[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *The Journal of Machine Learning Research*, vol.15, no.1, pp.1929-1958, 2014.
[28] V. Mnih, *Machine Learning for Aerial Image Labeling*, Ph.D. Thesis, University of Toronto, 2013.
[29] L. Ma, M. Li, X. Ma, L. Cheng, P. Du and Y. Liu, A review of supervised object-based land-cover image classification, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol.130, pp.277-293, 2017.
[30] A. J. Cooner, Y. Shao and J. B. Campbell, Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 Haiti Earthquake, *Remote Sensing*, vol.8, no.10, DOI: 10.3390/rs8100868, 2016.
[31] Y. Dong, B. Du and L. Zhang, Target detection based on random forest metric learning, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.8, no.4, pp.1830-1838, 2015.
[32] E. Li, J. Femiani, S. Xu, X. Zhang and P. Wonka, Robust rooftop extraction from visible band images using higher order CRF, *IEEE Trans. Geoscience and Remote Sensing*, vol.53, no.8, pp.4483-4495, 2015.

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

[34] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, Pyramid scene parsing network, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2881-2890, 2017.

[35] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.40, no.4, pp.834-848, 2018.