Comparison of Semantic Segmentation Deep Learning Methods For Building Extraction

1st Anisa Aizatin
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
2322002@std.stei.itb.ac.id

2nd I Gusti Bagus Baskara Nugraha
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
baskara@itb.ac.id

2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM) | 978-1-6654-7650-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/CENIM56801.2022.10037426

***Abstract*— Urban planners use building extraction on satellite imagery to support government policies. However, the complex depiction of buildings in satellite imagery makes building extraction difficult. One way to extract buildings in satellite imagery is by semantic segmentation deep learning. This study aims to find a suitable deep learning semantic segmentation method by comparing the performance of UNet, UNet++, DeepLabV3, and DeepLabV3+ that combined with ResNet-101 and ResNet-50 as feature extraction algorithms and trained on two public datasets with different characteristics. UNet++ produces the highest performance for predicting both datasets, but with different feature extraction algorithms. MBD feature extraction is more suitable using ResNet-101 while AICrowd uses ResNet-50.However, if we consider time-consuming, DeepLabV3+ and UNet are more efficient for training building datasets because of consuming less time with quietly performance**

Keywords*— building extraction, satellite imagery, deep learning, semantic segmentation

I. INTRODUCTION

Buildings are objects in spatial planning that can produce information for population estimation, urban planning, and management. However, in the real situation, manually building mapping takes time and a lot of cost [1], which affects out-of-date building information. With the development of remote sensing technology, satellite imagery has begun to be used to detect, identify, and analyze building objects in satellite imagery through building extraction.

However, until now, the accuracy and reliability of building extraction results are still a challenge [2]because the depiction of buildings in satellite imagery has various characteristics, such as the shape, size, color, and similarity to the background and other objects [3]. In addition, the significant number of building objects in an area and overlapping depictions of buildings also create difficulties in the extraction, especially in urban areas. Therefore, research for the extraction of building objects is an important research problem in remote sensing [4].

For decades, researchers have carried out study related to building extraction that focuses on its methods. Traditionally, the building extraction method is based on color, spectrum, edge, shape, texture, and semantics [5]. Some traditional building extraction methods are template matching, knowledge-based, region-based, edge-based, and traditional machine learning. Reference [6] used Histogram of Oriented Gradient (HOG), Local Binary Pattern (LBP) and Support Vector Machine (SVM that one of the algorithms in traditional machine learning to extract building features that had high recall values but bad F1-score and precision value. Based on some research [6], [7], [1], [8] that used the traditional method, the performance of the traditional method is not very good, and false detection. Because of the limitation of the traditional method, recently, the building extraction task performs deep learning to produce quite good performance [2].

Deep learning is one of machine learning subsets based on artificial neural networks (ANN). The main idea of deep learning is to imitate the function of the human brain so the machine can learn from data and filter information like the human brain does. In building extraction, deep learning can be used to feature extraction, segmentation, and building detection. The segmentation algorithm consists of two categories based on object class division; semantic segmentation [9] and instant segmentation [10]. Some research used deep learning methods based on semantic segmentation, such as UNet, SegNet, FCN, and DeepLabV3 [10], where those methods do not divide the building objects into certain classes. In addition, based on [11], semantic segmentation deep learning, especially FCN, have better performance than traditional machine learning i.e Naïve Bayes, SVM, and Random Forest (RF). Deep learning semantic segmentation can produce better good performance to extract buildings because it learn and classifies pixels into building and non-building.

A lot of building extraction research have used deep learning semantic segmentation and tried to enhance it. Research in [12] performed the extraction of buildings through semantic segmentation on very high satellite imagery with the derivative method from FCN. Meanwhile, [13] extracted buildings using UNet and FPN that applied SE-ResNeXt-101 EfficientNet-B0 and EfficientNet-B1 as feature extraction algorithms on very high satellite imagery. To determine the performance of DeepLabV3+ which is the extension of DeepLabV3, [14] compared the results of building extraction using DeepLabV3 and DeeplabV3+ with ResNet encoder at different depths.

Most of the architectures of semantic segmentation deep learning use FCN as the base architecture, but FCN has the weakness of losing information during training due to downsampling. Several methods use Atrous Spatial Pyramid Pooling (ASPP) to address the weakness of FCN. Therefore, this paper will compare the performance and complexity of FCN-based algorithms such as UNet, UNet++, and ASPP such as deepLabV3 and deepLabV3+ on satellite image datasets with different characteristics. The algorithm with the best performance and complexity will be used as the basis for further research.

II. DEEP-LEARNING SEMANTIC SEGMENTATION

Semantic segmentation aims to divide objects of images that belong to the same object class. Semantic segmentation in building extraction usually categorizes buildings and non-buildings. Most semantic segmentation deep learning has two networks, the base network, and the segmentation architecture network. The base network or backbone is usually built upon CNN and performs feature extraction, such as ResNet. Residual Network (ResNet) is a deep residual network even very deep residual network architecture, that can have 18, 34, 50, 101,152, or 1202 layers. ResNet has special feature that called skip connection. The skip connection makes direct connection cross-activation by skipping some layer between them. The skip connection can reduce the vanishing gradient problem which is one of the biggest problems in the deep neural network. Different from ResNet 18 and ResNet 34 which use 2 stack layers, ResNet50, ResNet 101, and ResNet1202 use more than 3 stack layers to reduce training time. Currently, for memory reasons, ResNet50 and ResNet101 are more widely used. Meanwhile, thesegmentation architecture is usually built upon fully convolutional network (FCN) and or atrous convolution.

A. UNet

UNet is a symmetric FCN that has U-shape architecture. UNet consists of the contracting path as the encoder, the expansive path as the decoder, and the bridge (bottleneck) that connect the two paths. UNet also has a special feature that is called skip connection. The contracting path is 3x3 convolutional network and 2x2 max pooling. The convolutional network that followed by rectified linear unit (ReLU) is used to capture the context of the input image and the max pooling performs downsampling that the height and width of feature maps are reduced by half. The result of the contracting path is sent to the expansive path by the bridge. The expansive path does upsampling and semantic segmentation that begins with 2x2 transposed convolution. Then, it is combined with the information from the encoder without passing the bridge that is carried by the skip connection. The last step of the expansive path is segmentation based on pixel-wise classification.

B. UNet++

UNet++ or Nested UNet is an extension of the UNet that consist of U-Nets of varying depths. The difference between UNet and UNet++ is the re-design of skip pathways and deep supervision. This skip pathways consist of one or more convolutional layers which keep feature maps between encoder and decoder semantically similar, and dense skip connections which connected between the convolutional layer to keep gradient flow. The deep supervision in UNet++ can be operated in two ways; the accurate mode which computes the average of all segmentation outputs and the fast mode which can determine model pruning to adjust the model complexity and speed gain. The fast mode selects the last segmentation map from one of the segmentation branches.

C. DeepLabV3

DeepLabV3 is the extension of the previous deepLab generation by adding batch normalization in the Atrous Spatial Pyramid Pooling (ASPP) and removing Dense Conditional Random Field (Dense-CRF) in the post-processing step. DeepLabV3 begin with extracting feature at the backbone process. Before the last block of the backbone, atrous convolutional controls the size of the feature map by upsampling the filter by inserting zeros between two spatial dimensions. The last step of the backbone process is to add the ASPP to classify each pixel according to its class. The output from all branches of the ASPP and batch normalization pass a 1x1 convolutional to get the actual size of the segmented image. W

D. DeepLabV3+

The main idea of DeepLabV3+ is to use DeepLabV3 as the encoder that implements Atrous Separable Convolution and simplify decoder architecture to increase segmentation results. DeepLabV3 with modified aligned Xception in encoder extract feature maps by applying atrous separable convolutional. The first step of the proposed decoder is to bilinearly upsample the encoder output and concatenate with the corresponding low-level features. This step applies 3x3 convolutions to refine the feature map and followed by another bilinear upsampling to get the best-predicted image.

TABLE I. ARCHITECTURE FOR SEMANTIC SEGMENTATION

Architecture	Strength and Weakness
U-Net	**Strength** • architecture works well with small training data • easy to scale for multiple classes **Weakness** • disconnected decoder • there is no guarantee that features of the same scale are suitable for feature fusion
UNet++	**Strength** • good segmentation quality with various object sizes • processing speed with pruned schema **Weakness** • not suitable for tiny objects
DeepLabV3	**Strength** • Reduced model computation without increasing parameters **Weakness** • DeepLabV3 consumes much time when training high-resolution image
DeepLabv3+	**Strength** • can encode multi-scale contextual information **Weakness** • because of the large number of parameters, DeepLabv3+ requires a large GPU to train high-resolution remote sensing image

III. EXPERIMENT SETTING

A. Dataset

This study trained the algorithm using two public building datasets, which have different characteristics. The AICrowd Mapping Challenge consists of very high-resolution satellite (VHRS) imagery with a size of 300x300 pixels for each tile, which can be obtained at https://www.aicrowd.com/challenges/mapping-challenge/dataset_files. AICrowd has the characteristics of large building images and few buildings in one tile, besides that the background is not too complex. In addition, this study also used the Massachusetts Building Dataset (MBD) from https://www.cs.toronto.edu/-~vmnih/data/. MBD has high-resolution satellite imagery of the state of Massachusetts with 1500x1500 pixels for each tile and 1-meter spatial resolution. MBD has the characteristics of small building image and there are many buildings in one tile and complex background.

The MBD and AIcrowd Mapping Challenge building datasets were randomly cropped to 256x256 pixels to ensure that all input images had the same size when trained and considering memory performance. In addition, this study adjusted contrast and brightness for AICrowd Mapping Challenge building datasets

B. Training

We performed the training using Google Collab with a 35 GB GPU. The training stage for building extraction uses deep learning based on semantic segmentation or pixel-level classification to categorize objects into buildings and non-buildings and segment each building. This study trained datasets using UNet and UNet++ which perform fully convolutional neural network architecture as baseline. In addition, to compare with the algorithm that performs Atorus Spatial Pyramid Pooling (ASPP) as the baseline, we used DeepLabV3 and DeepLabV3+. In the background, the encoder part did feature extraction. We performed feature extraction using ResNet101 and ResNet50 for each algorithm. We repeat each algorithm in ten epochs and train it using 137 MBD image inputs with 4 images for validation and 8366 image inputs with 1274 images for validation from AICrowd Mapping Challenge Building Dataset.

C. Evaluation

We compare computational complexity, the Intersection of Union (IoU), and model performance between the algorithms. In this study, we just compare time training to see computational complexity. The IoU value shows the intersection between the predicted building shape and the original building shape; model performance can be seen from the accuracy value that shows the proportion of correctly predicted pixel and total pixel. The value of IoU and overall accuracy are in the range of 0 to 1, and the highest value indicates the best segmentation performance. The IoU and overall accuracy can be formulated as follows:

$$Overall\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Intersection\ of\ Union\ (IoU) = \frac{TP}{TP + FN + FP} \quad (2)$$

where true positive (TP) represents the number of correctly labelled pixels as buildings, false positive (FP) represents the number of pixels that are misclassified as buildings, true negative (TN) represents the number of correctly labelled pixels as non-buildings, and false negative (FN) represents the number of pixels that are misclassified as non-buildings In addition, this study performed qualitative evaluation by analyzing the building prediction image resulting from the testing model. We describe the evaluation in descriptive form.

IV. RESULT AND DISCUSSION

A. Training Time

Training time is one of the factors that can represent computational complexity. Table 3 shows the comparison of training time between the algorithm with different backbone depths and datasets. Training on MBD takes less time than AICrowd dataset. On MBD, the depth of the backbone does not impact the time training but in AICrowd, ResNet-101, which has 101 layers, needs approximately 30% more time than ResNet-50. The DeepLabV3 with ResNet-10 architecture is the most time-consuming both in MBD and AICrowd datasets. Meanwhile, The next generation of DeepLabV3, DeepLabV3+ is the fastest architecture. Generally, we can assume that the number of images and image resolutions greatly affect training time-consuming,

TABLE II. TRAINING TIME (SECOND)

Dataset	Algorithm	Training Time	
		ResNet-50	Resnet-101
MBD	UNet	9	8
	UNet++	10	12
	DeepLabV3	10	14
	DeepLabV3+	10	8
AICrowd	UNet	187	380
	UNet++	608	890
	DeepLabV3	1035	1434
	DeepLabV3+	177	250

B. Model Performance

The testing results on the two datasets with different types of satellite imagery using network architectures based on semantic segmentation show accuracy values and IoU scores that are not too different, either applying ResNet-50 or ResNet-101 backbones. However, among the four methods, UNet++ has the best accuracy values and IoU. Table2 shows that testing with UNet++ and ResNet-101 on MBD performs the best IoU score. On the other hand, UNet++ with ResNet-50 performs the highest IoU score. AICrowd Mapping Challenge, which has a large building depiction, has good accuracy for each method, which means it can correctly predict more than 90% of total pixels. From Table 2, we can assume that all models are quite good to predict building and non-building with overall accuracy of more than 80% but from the IoU score, all models are not clear to predict the building shape.

TABLE III. COMPARISON IOU SCORE AND ACCURACY (%)

Dataset	**Algorithm**	**ResNet-50**		**ResNet-101**	
		IoU	accuracy	IoU	accuracy
MBD	DepLabV3	75,92	85,09	75,65	85,11
	DeepLabV3+	78,46	87,15	77,04	86,05
	UNet	79,31	87,9	75,98	85,59
	UNet++	79,94	88,06	80,83	88,55
AIcrowd Mapping Challenge Dataset	DepLabV3	76,21	94,79	77,04	95,20
	DeepLabV3+	76,82	95,09	75,86	94,77
	UNet	77,17	95,18	77,15	95,09
	UNet++	78,48	95,51	77,86	95,39

Fig.1 performs the prediction results of MBD testing. DeepLabV3 can predict the presence of the building, but it cannot distinguish between buildings and the background around buildings. Hence, DeepLabV3 cannot detect the distance between buildings. DeepLabV3+ and UNet produce better building predictions than DeepLabV3 and can detect the distance between buildings, but they cannot predict the shape of the building. UNet++ produces the best building prediction results, but it cannot predict small buildings with UNet++ and ResNet101 as the backbone.

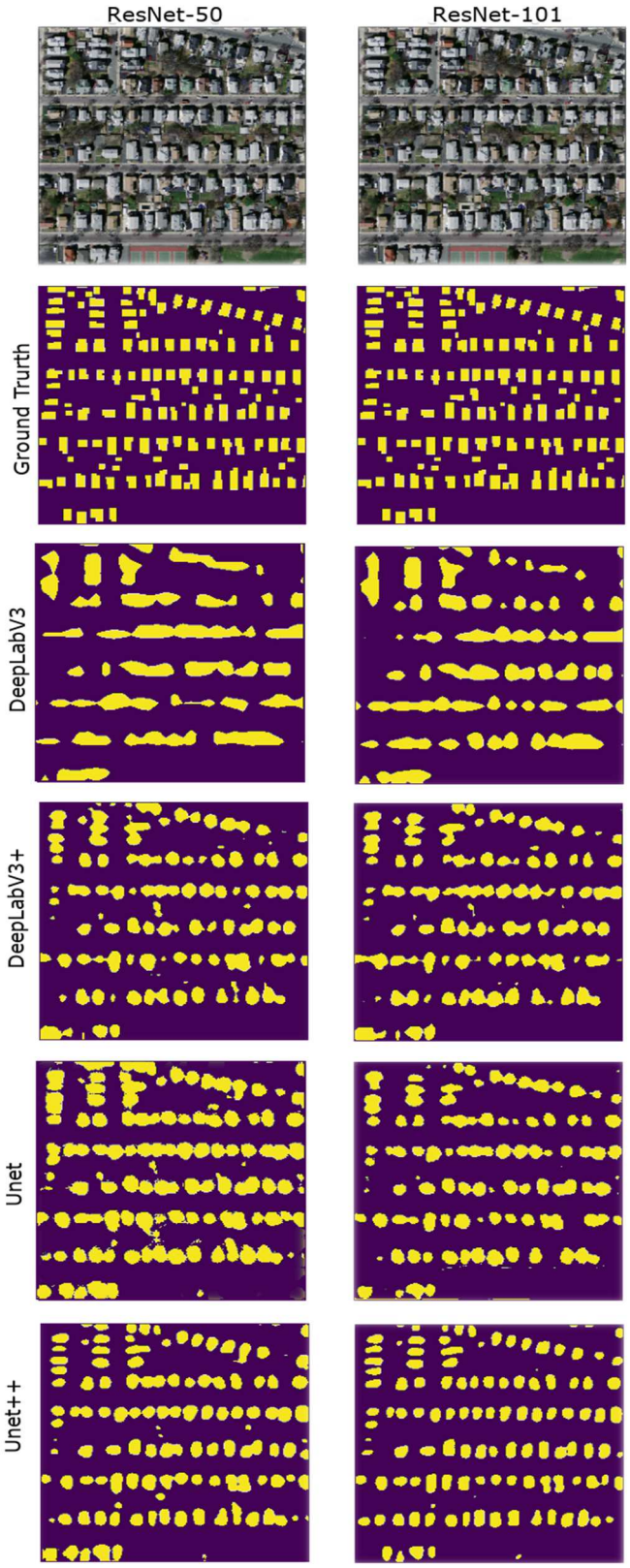

Fig. 1. Prediction Result on Massachusset Building Dataset

The AICrowd dataset with large building depictions delivers good building predictions. Fig.2 shows that each method can predict every building in the tile except buildings with very small sizes, especially with UNet++ and ResNet-50 as the backbone. However, the predicted shape of the building has an edge that does not match the ground truth.

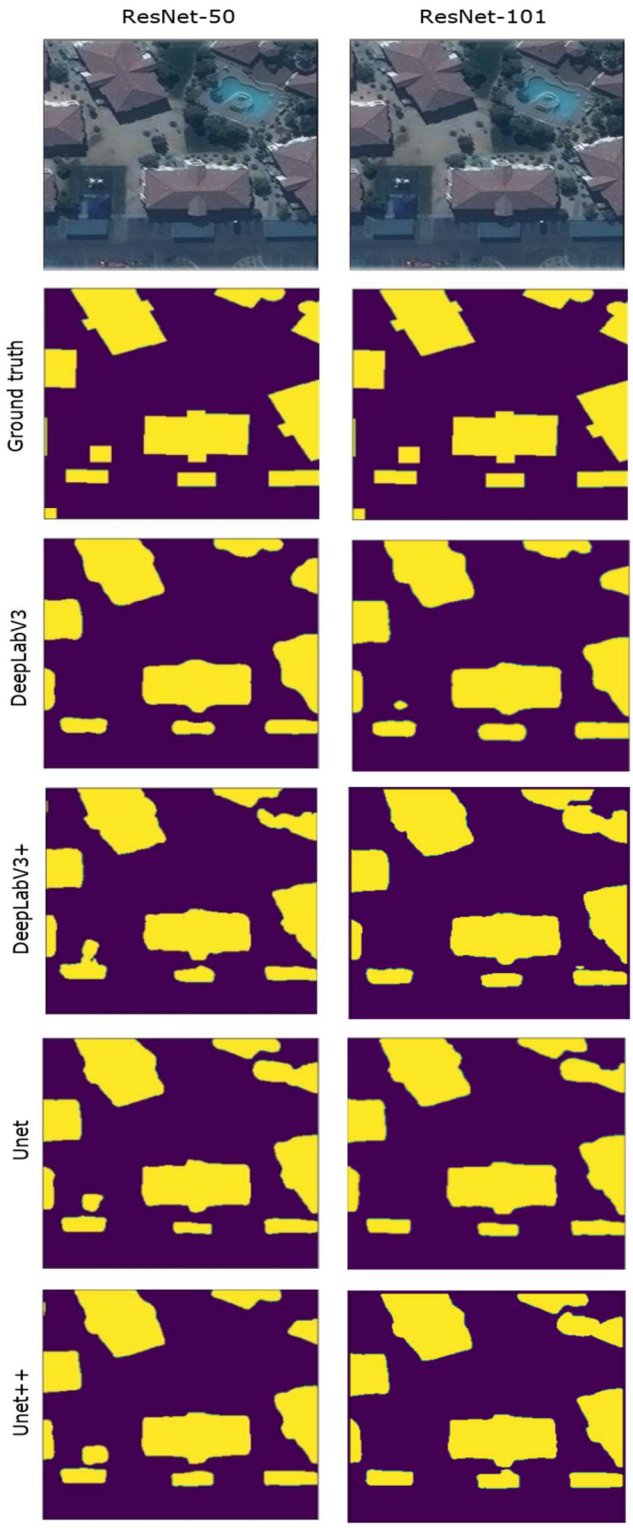

Fig. 2. Prediction Result on AICrowd Mapping Challenge Dataset

UNet++ produces a good prediction for extracting buildings based on image prediction result, IoU score, and accuracy. ResNet-101 is suitable to use in MBD, and ResNet-50 performs better performance in AICrowd. This is because 50 layers in ResNet-50 is sufficient to extract AICrowd dataset feature but needs more layers for extracting MBD dataset feature that more complex. In addition, ResNet-50 as a backbone can predict small buildings better than ResNet-101 both in MBD and AICrowd datasets. However, from the

prediction results, the shape of the predicted buildings, both high-resolution and high-resolution datasets, has imperfect edges. In addition, buildings with exceedingly small sizes are still difficult to detect.

V. CONCLUSION

This research compared several semantic segmentation deep learning methods to train high and very high-resolution satellite imagery. The best method in this study will become the primary method used for future research. We compared the methods based on the accuracy, IoU values, and image prediction of the testing results. From the result analysis, we can assume that:

- for large and very high-resolution datasets, we suggest using DeepLabV3+ or UNet because both algorithms perform quietly performance with less time-consuming.
- for small and medium to high-resolution datasets, we suggest using UNet++ which performs the best performance. If we consider the building shape, UNet++ with ResNet-50 as the backbone gives more clear edge and shape, but if we consider classification building and nonbuilding UNet++ with ResNet gives the best accuracy.
- DeepLabV3 does not suitable to extract building especially for training large and very-high resolution datasets because it is too complex for DeepLabv3

However, both the algorithms have not been able to predict the edges of buildings and exceedingly small buildings accurately. Therefore, further research will improve the prediction of the edges of each building that using an algorithm according to the dataset and resource with adding polygonization in post-processing.

REFERENCES

[1] G. S. Xia, J. Huang, N. Xue, Q. Lu, and X. Zhu, "GeoSay: A geometric saliency for extracting buildings in remote sensing images," *Computer Vision and Image Understanding*, vol. 186, pp. 37–47, Sep. 2019, doi: 10.1016/j.cviu.2019.06.001.

[2] Y. Wang, L. Gu, X. Li, and R. Ren, "Building Extraction in Multitemporal High-Resolution Remote Sensing Imagery Using a Multifeature LSTM Network," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 9, pp. 1645–1649, Jul. 2020, doi: 10.1109/lgrs.2020.3005018.

[3] L. Li, J. Liang, M. Weng, and H. Zhu, "A multiple-feature reuse network to extract buildings from remote sensing imagery," *Remote Sens (Basel)*, vol. 10, no. 9, Sep. 2018, doi: 10.3390/rs10091350.

[4] V. Ostankovich and Vi. Afanasye, "Illegal Buildings Detection from Satellite Images using GoogLeNet and Cadastral Map," in *2018 International Conference on Intelligent Systems (IS)*, 2018, pp. 616–623.

[5] S. Ji, S. Wei, and M. Lu, "Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, Jan. 2019, doi: 10.1109/TGRS.2018.2858817.

[6] D. Konstantinidis, T. Stathaki, V. Argyriou, and N. Grammalidis, "Building Detection Using Enhanced HOG-LBP Features and Region Refinement Processes," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 10, no. 3, pp. 888–905, Mar. 2017, doi: 10.1109/JSTARS.2016.2602439.

[7] Y. Wang, Q. Meng, Q. Qi, J. Yang, and Y. Liu, "Region merging considering within- and between-segment heterogeneity: An improved hybrid remote-sensing image segmentation method," *Remote Sens (Basel)*, vol. 10, no. 5, May 2018, doi: 10.3390/rs10050781.

[8] F. Dornaika, A. Moujahid, Y. el Merabet, and Y. Ruichek, "Building detection from orthophotos using a machine learning approach: An empirical study on image segmentation and descriptors," *Expert Syst Appl*, vol. 58, pp. 130–142, Oct. 2016, doi: 10.1016/j.eswa.2016.03.024.

[9] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169. Elsevier Ltd, May 01, 2021. doi: 10.1016/j.eswa.2020.114417.

[10] L. P. Osco *et al.*, "A review on deep learning in UAV remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102456, Oct. 2021, doi: 10.1016/j.jag.2021.102456.

[11] Y. Li, B. He, T. Long, and X. Bai, "Evaluation the performance of fully convolutional networks for building extraction compared with shallow models," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Jul. 2017, pp. 850–853.

[12] K. Moghalles, H. C. Li, Z. Al-Huda, and E. A. Hezzam, "Multi-Task Deep Network for Semantic Segmentation of Building in Very High Resolution Imagery," in *2021 International Conference of Technology, Science and Administration, ICTSA 2021*, Mar. 2021. doi: 10.1109/ICTSA52017.2021.9406538.

[13] P. Borba, F. de Carvalho Diniz, N. C. da Silva, and E. de Souza Bias, "Building Footprint Extraction Using Deep Learning Semantic Segmentation Techniques: Experiments and Results," in *IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposiu*, Oct. 2021, pp. 4708–4711. doi: 10.1109/igarss47720.2021.9553855.

[14] L. Yang, H. Wang, K. Yan, X. Yu, J. Li, and D. Man, "Building Extraction of Multi-source Data Based on Deep Learning," in *2019 IEEE 4th International Conference on Image, Vision and Computing*, 2019, pp. 296–300.