ICIC Express Letters
Volume **14**, Number **1**, January **2020**

 ISSN 1881-803X
pp. **43–51**

SEMANTIC SEGMENTATION OF AERIAL IMAGERY FOR ROAD AND BUILDING EXTRACTION WITH DEEP LEARNING

Alexander Agung Santoso Gunawan*, Ilma Arifiany and Edy Irwansyah

Department of Computer Science
School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
*Corresponding author: aagung@binus.edu

Received July 2019; accepted October 2019

Abstract. *Road and building extraction from aerial images has many applications in various fields, for example in urban planning, automatic map making, and land use analysis. However, commonly this extraction task is still done by human experts manually, so the process is very expensive and time-consuming. On the other hand, semantic segmentation is the process of classifying each pixel of an image as a class label to understand the image at a pixel level. This paper proposed an algorithm of semantic segmentation based on deep learning, namely Segmentation Networks (SatNet), for automatically segmenting buildings and roads. Our SatNet model is based on Residual Networks (ResNet) for image feature extraction. Finally, we compared the proposed algorithm to the state-of-the-art and our segmentation results have improved significantly.*
Keywords: Road and building extraction, Aerial imagery, Semantic segmentation, Deep learning, SatNet

1. **Introduction.** Satellite imagery and aerial imagery have many applications in agriculture, landscaping, cartography and regional planning. These images can be in visible colors and other spectra. There is also an elevation imagery, usually made with radar or lidar images. Together with the development of aerial and satellite imagery, a field called remote sensing is also developing. Remote sensing is the data acquisition and analysis from remote object by a device that does not physically contact the object [1]. Recently, to perform data analysis and interpretation is still relying on human experts.

In our research, we would like to extract the buildings and roads from aerial imagery. The extraction results have many applications in various fields including automatic map making, city planning, and land use analysis. Nevertheless, this task is still done by human experts manually, so the process is a very expensive and time-consuming process. Moreover, accurate pixel labeling on a large scale is a complex task for humans, because the terrestrial objects have many variations in shapes, and can be occluded by other objects such as trees or buildings [2]. Therefore, automatic object extraction of aerial images is highly demanded, and much effort has been put forward in the remote sensing literature. The objective of this research is to develop algorithm for automatically extracting the locations of objects, especially roads and buildings directly from aerial images.

There are many previous works that have attempted to extract terrestrial objects from aerial imagery. Approaches utilizing Neural Networks (NNs) to solve aerial imagery extraction problems have already been proposed and have achieved good performance. The earliest work [2] proposed a road extraction by using Restricted Boltzmann Machines (RBMs). It formulated the problem of road extraction from aerial images as a patch-based semantic segmentation on pixel level. Furthermore, the authors [3] improved the

DOI: 10.24507/icicel.14.01.43

neural networks for dealing with label noises by using Convolutional Neural Networks (CNNs) and modified loss functions. Finally, in his Ph.D. thesis [4], he concluded that label noise has a negative effect, but its effect of predictions by using neural networks is relatively small because of CNNs strong representation ability.

Recently, there are several works related to road and building extraction from aerial imagery. Xu et al. [5] proposed a segmentation model based on Densely Connected Convolutional Networks (DenseNet) and elaborated the local and global attention units. The aim of the paper is just focused on road extraction method from aerial imagery. Its results for road extraction are comparably same with ours. Furthermore, Liang et al. [6] design a model for road boundary extraction from LiDAR and camera imagery. They used FCNs for extracting deep features of road boundaries and then, CNNs for producing a polyline representation of road boundary. For building extraction, Ji et al. [7] proposed a model based on Mask R-CNN for object-based instance segmentation, and a multi-scale full convolutional network for pixel-based semantic segmentation. Its results for building extraction are slightly higher than ours, but the computational complexity is much higher. Furthermore, Feng et al. used building extraction for cartographic generalization, which is the process of producing small-scale representations of large-scale spatial data [8].

In our research, to solve semantic segmentation, we utilized Fully Convolutional Networks (FCNs) approach [9], which uses CNNs to transform image pixels to pixel labels. Our approach repurposes the renown deep learning architecture, Residual Networks (ResNet) [10] to solve semantic segmentation problems. We trained and tested our architecture, called as Segmentation Networks (SatNet) on a publicly available large aerial imagery dataset. We showed that our architecture can classify all pixels in aerial imagery into buildings and roads more accurately than Semantic Segmentation for Aerial Imagery (SSAI) algorithm [11], which we used as benchmark.

The remainder of the paper is composed as follows. First we discuss semantic segmentation in Section 2, and then is followed by explanation of SatNet, in Section 3. In Section 4, we discuss experiment results of SatNet for road and building extraction. Finally, we concluded our work with suggestions for the future research in Section 5.

2. **Semantic Segmentation.** Semantic segmentation is understanding images at the pixel level where we want to assign each pixel in the image as a label. For example, see the following image from Stanford background dataset [12]. In addition to recognizing the cow and houses behind them in Figure 1, semantic segmentation must also determine the boundaries of each object. Therefore, unlike classification, semantic segmentation needs pixel-wise predictions from images. Before CNNs were used in computer vision, researchers used machine learning approaches such as Random Forest classifiers for semantic

FIGURE 1. Input image (left), its semantic segmentation (right)

segmentation. One of the early CNNs approaches is the patch classification where each pixel is separately classified into labels using image patches around it. The main reason for using image patches is that classification networks usually have a fully connected layer and therefore a fixed size image is required.

In 2015, fully convolutional networks or FCNs [9] by Long et al., promote the CNNs architecture for semantic segmentation without fully connected layers. This approach allows segmentation maps to be created for any size images much faster than the patch classification approach. Almost all recent researches in semantic segmentation adopt this paradigm. The main problem of CNNs for semantic segmentation is the pooling layers. In CNNs architecture, a pooling layer can increase the field of view and cope with the image context by scarifying the exact object position in image. However, semantic segmentation requires the proper alignment between an image and its segmentation map. It means semantic segmentation requires exact position to be preserved. One approach to overcome this problem is encoder-decoder architecture. The encoder gradually decreases the spatial dimensions with the pooling layer. Sometimes the pooling is omitted to preserve the image details. And the decoder gradually increases the spatial dimensions. One of popular encoder-decoder architectures is U-Net [13]. The architecture consists of shrinking layers to capture context and symmetric expanding layers which allow exact object position in image. There is a shortcut connection from the encoder to the decoder to help the decoder recover the object details. For our research, we also use encoder-decoder architecture to extract buildings and roads from aerial images. Our encoder is based on ResNet architecture [10], which has no pooling layers at all. For ResNet, its skip connections enable very deep networks and then can gradually decrease the spatial dimensions without pooling layers.

3. **Segmentation Networks (SatNet).** In the beginning of the research, we explore the idea of transfer learning [14]. Transfer learning is a machine learning method, in which a model developed for certain task is reused as the starting point for a model on another task. The goal of transfer learning is for accelerating the training process. We first tried to reuse pre-trained ImageNet models for ResNet [10] and Xception [15] networks for our semantic segmentation problem. However, this approach failed, and even significant train times after tuning in several ways cannot lead to modest results. It could be that the high-level features in pre-trained models which needed to distinguish pixel-wise labels for semantic segmentation are not available.

Therefore, we decided to train our architecture without any transfer learning. We design our architecture, which inspired by ResNet and will be referred as Segmentation Network (SatNet) model. This architecture (see Figure 2) is full pre-activation with 18 layers, including the deconvolution layers. Our architecture is based on the FCNs paradigm, so it no longer uses fully connected layers and sigmoid activation functions to make predictions. Pooling layers are also removed and there are only convolution layers here. All convolution layers use Same Padding parameter, which means output size is same to input size when stride parameter equals 1. Because the architecture is a full pre-activation architecture, each layer will pass batch-normalization and rectified linear unit (ReLu) activation function. Furthermore, dropout regularization is only used in encoder part. Before training step, there are several parameters that must be determined, that is learning rate, normalization coefficient and dropout rate. We also performed image augmentation before training step, by rotating each image and its label with random degrees of rotation. By this image augmentation, the resulted model does not favor objects with certain orientations.

The roads and buildings dataset used for our research is the Massachusetts City dataset from Mnih's Ph.D. thesis [4]. The dataset is made from images released officially by the Massachusetts City Government. Labels for these images are made using data from the

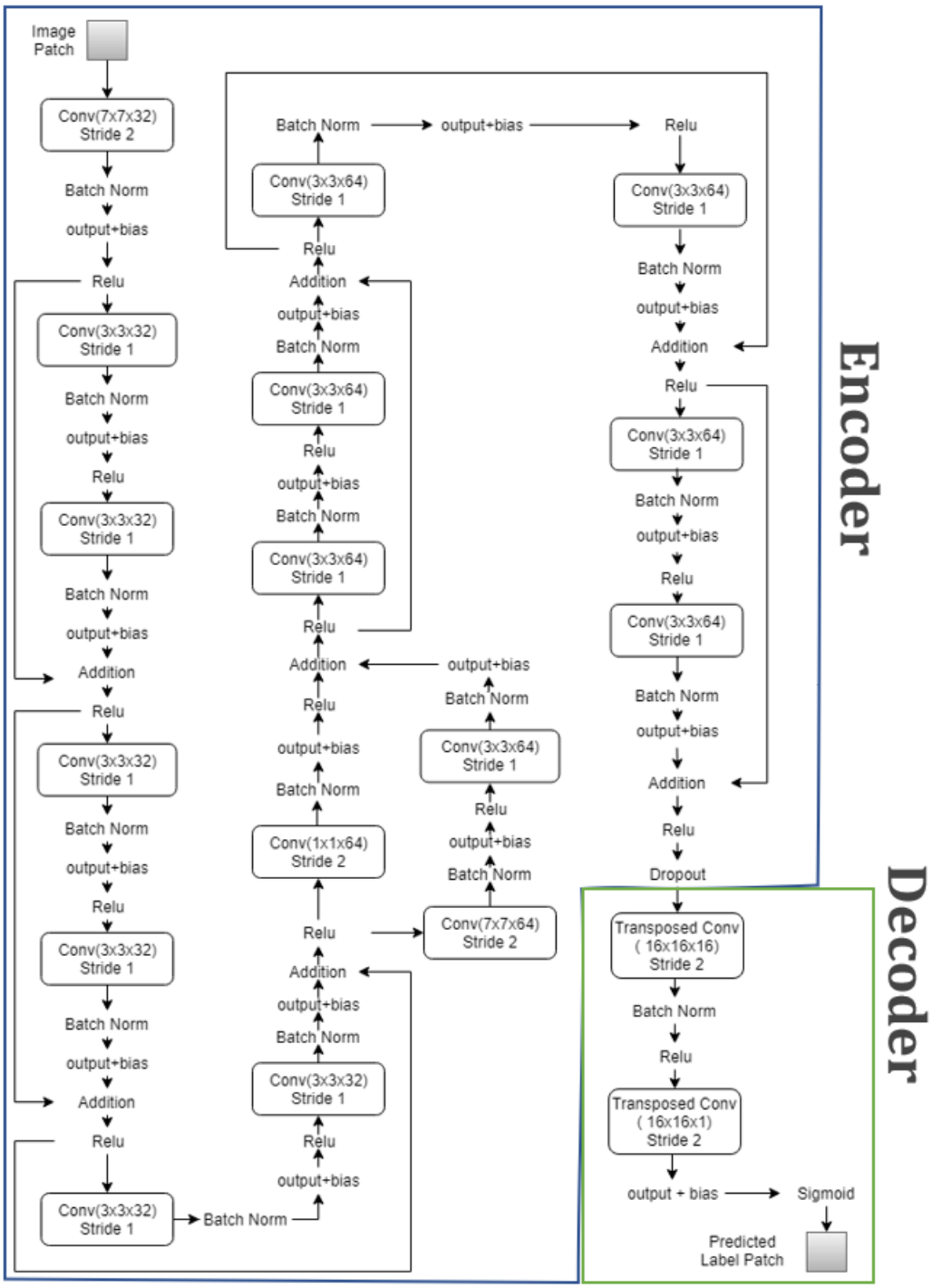


FIGURE 2. SatNet architecture

OpenStreetMap project. This data is limited to areas with average noise level of around 5% or less. The Massachusetts road dataset consists of 1171 aerial photographs from the state of Massachusetts. Each image is 1500 × 1500 pixels, which covers an area of 2.25 square kilometers. The data were randomly divided into a training set of 1108 images, an evaluation set of 14 images, and a test set of 49 images. The dataset covers a variety of urban, suburban, and rural areas and covers an area of more than 2600 square kilometers. It is the largest and most challenging aerial image label data collection.

While the Massachusetts building dataset consists of 151 aerial photos of the Boston area, with each image measuring 1500 × 1500 pixels for area of 2.25 square kilometers. All datasets cover approximately 340 square kilometers. The data were randomized into

137 images for training set, a test set of 10 images, and an evaluation set of 4 images. In the training stage, we conduct the training process on above training datasets and adjust the model parameters to obtain optimum results. The best model in the training stage is then evaluated using the related test dataset.

As the comparison model, we used state of the art in semantic segmentation on aerial imagery called, that is SSAI model [11]. Besides having impressive results, SSAI also provides the implementation code. It is also based on the Fully Convolutional Networks (FCNs) paradigm. The SSAI model (see Figure 3) has six layers, including the pooling layer and deconvolution convolution for up-sampling process. The first convolution layer is done with 2×2 max pooling with stride 1 to speed up the training process. In SSAI model, each layer goes through batch-normalization and ReLu activation function. After ReLu, dropout regularization is performed to get rid of unused neurons.

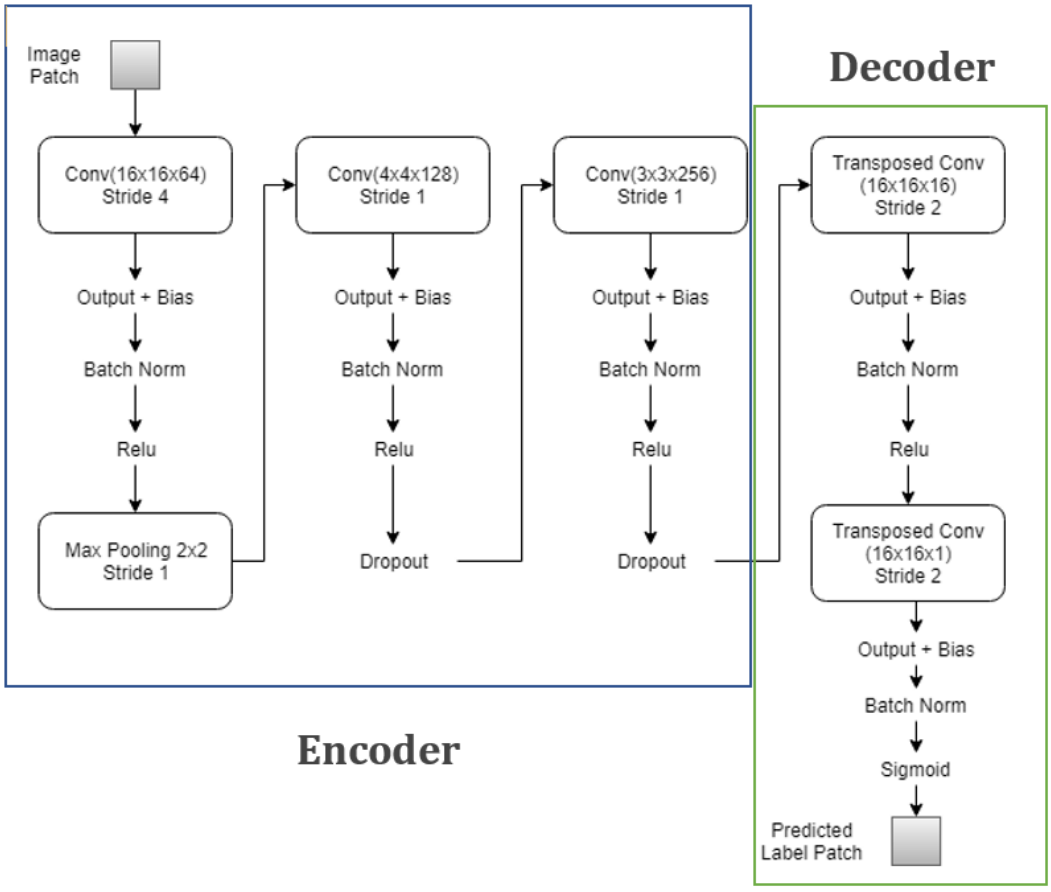


FIGURE 3. SSAI architecture

4. **Experiment Results.** First, we evaluate the training stage. Figure 4 shows the comparison of training loss values in road segmentation using SSAI and SatNet models.

Semantic segmentation can be considered as binary classification on pixel-wise level. To evaluate the model performance of binary classification, we usually use accuracy values in both training and testing stage. By comparing the highest training accuracy and the testing accuracy, it can be detected whether there is overfitting in the developed model. Table 1 shows the comparison of model accuracy of SSAI and SatNet models in training and testing stage for road dataset.

From accuracy results in Table 1, the SatNet has a slightly higher level of accuracy than the SSAI model. Both models can be considered as good models for road extraction problem, because of their high accuracy. There is also no significant difference from the

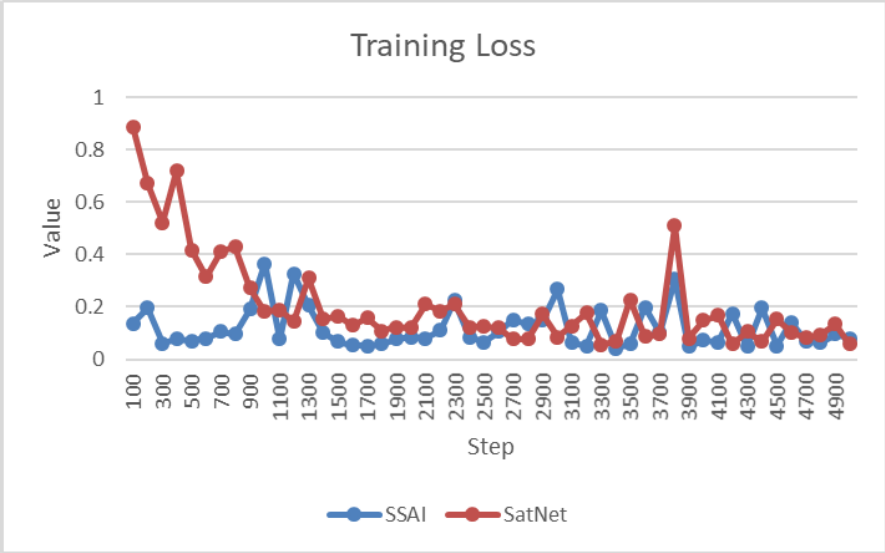
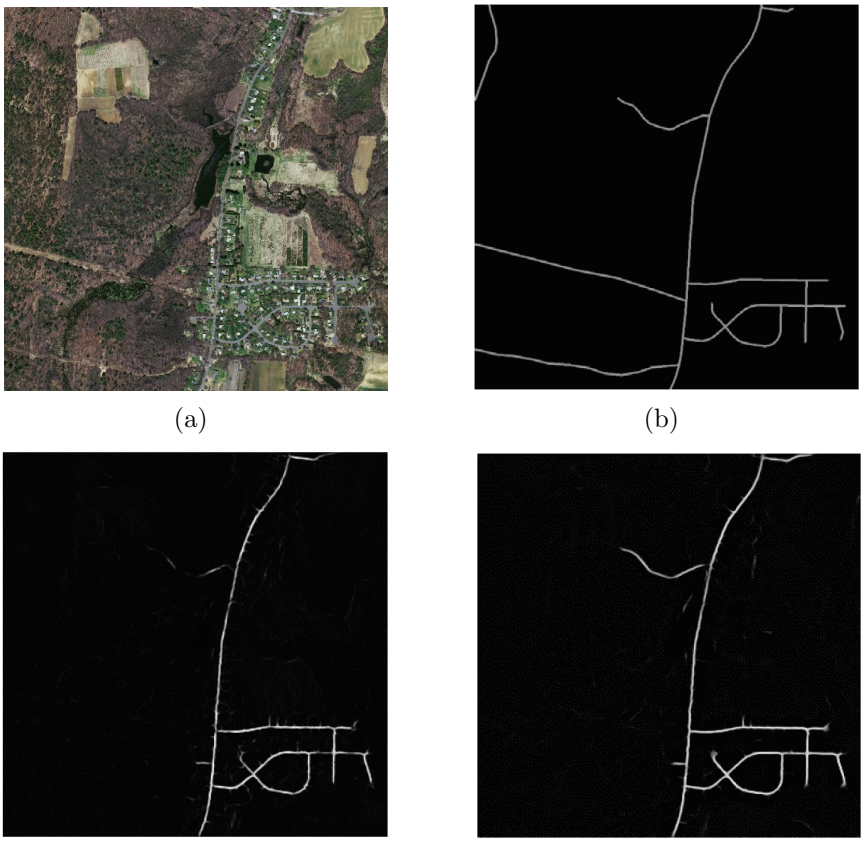


FIGURE 4. (color online) Training loss comparison of SSAI and SatNet models on road dataset

TABLE 1. Model accuracy of SSAI and SatNet in road dataset

Model	Step	Learning Rate	Highest Training Accuracy	Testing Accuracy
SSAI	3400	0.0125	96.66%	96.33%
SatNet	3300	0.0125	97.06%	96.76%



(a) (b)

(c) (d)

FIGURE 5. (a) Road image in dataset, (b) ground truth, (c) SSAI result, (d) SatNet result

accuracy of training and testing stage. It means there is no overfitting issue in developed models. As illustration, Figure 5 displays one of road images in dataset, together with its label ground-truth, the SSAI and SatNet result for the road image.

Next, Figure 6 shows the comparison of training loss values in building segmentation using SSAI and SatNet models. Furthermore, Table 2 shows the comparison of model accuracy of SSAI and SatNet models in training and testing stage for building dataset.

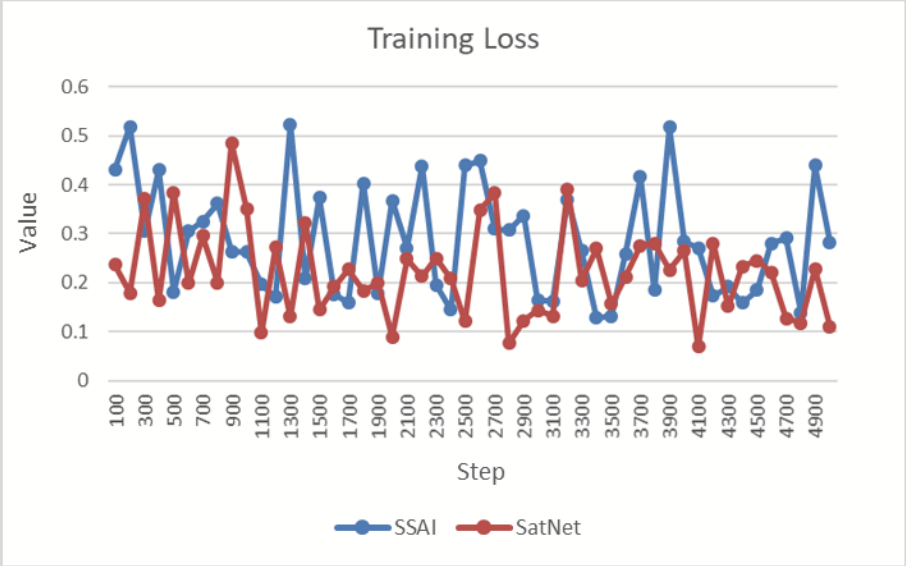


FIGURE 6. (color online) Training loss comparison of SSAI and SatNet models on building dataset

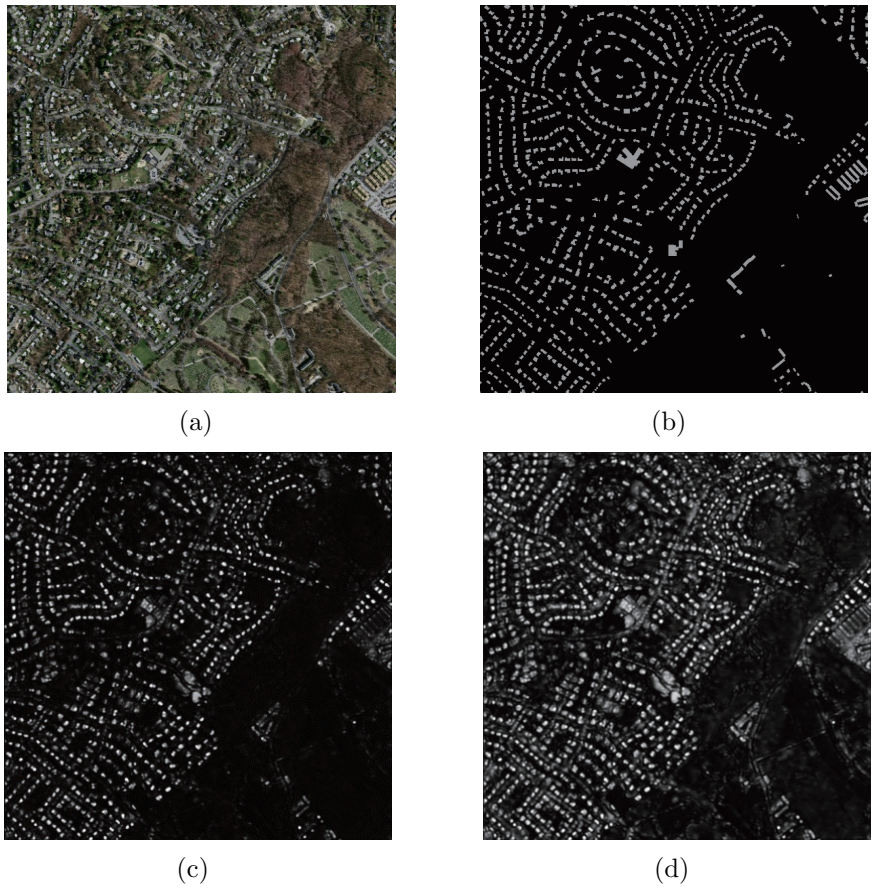
TABLE 2. Model accuracy of SSAI and SatNet in building dataset

Model	Step	Learning Rate	Highest Training Accuracy	Testing Accuracy
SSAI	3500	0.0125	82.51%	78.56%
SatNet	4100	0.00625	89.20%	82.26%

From accuracy results in Table 2, SatNet accuracy is much higher than SSAI. Nevertheless, both models can be considered just modest models for building extraction problem, because their accuracy is just around 85%. Furthermore, the accuracy of training has significant difference to the testing accuracy. It means there is overfitting issue here. We can handle this issue by adding new amounts of training data. As illustration, Figure 7 shows one of building images in dataset, together with its label ground-truth, the SSAI and SatNet result for the road image.

5. **Conclusions.** As conclusions, both SSAI and SatNet models can be used for semantic segmentation of aerial and satellite imagery. Overall, the SatNet model is better than the SSAI model to extract road and building from aerial imagery due to its higher accuracy. Based on our analyses, we can suggest for future research to add new amounts of training building data in order to overcome overfitting issue and to improve accuracy of the SatNet model.

Acknowledgment. This work is supported by Directorate General of Research and Development Strengthening, Indonesian Ministry of Research, Technology, and Higher Education, as a part of Penelitian Terapan Unggulan Perguruan Tinggi (PTUPT) Research Grant to Bina Nusantara University with contract number: 12/AKM/PNT/2019 and contract date: 27 March 2019.



(a) (b)

(c) (d)

FIGURE 7. (a) Building image in dataset, (b) ground truth, (c) SSAI result, (d) SatNet result

REFERENCES

[1] T. Lillesand, R. W. Kiefer and J. Chipman, *Remote Sensing and Image Interpretation*, 7th Edition, Wiley, 2015.

[2] V. Mnih and G. E. Hinton, Learning to detect roads in high-resolution aerial, *European Conference on Computer Vision (ECCV)*, Crete, Greece, 2010.

[3] V. Mnih and G. Hinton, Learning to label aerial images from noisy data, *Proc. of the 29th International Conference on Machine Learning (ICML'12)*, Edinburgh, Scotland, 2012.

[4] V. Mnih, *Machine Learning for Aerial Image Labeling*, Ph.D. Thesis, Graduate Department of Computer Science, University of Toronto, 2013.

[5] Y. Xu, Z. Xie, Y Feng and Z. Chen, Road extraction from high-resolution remotesensing imagery using deep learning, *Remote Sensing Journal*, vol.10, no.1461, pp.1-16, 2018.

[6] J. Liang, N. Homayounfar, W. C. Ma, S. Wang and R. Urtasun, Convolutional recurrent network for road boundary extraction, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, California, USA, 2019.

[7] S. Ji, Y. Shen, M. Lu and Y. Zhang, Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples, *Remote Sensing Journal*, vol.11, no.1343, pp.1-20, 2019.

[8] Y. Feng, F. Thiemann and M. Sester, Learning cartographic building generalization with deep convolutional neural networks, *ISPRS International Journal of Geospatial-Information*, vol.8, no.258, pp.1-20, 2019.

[9] J. Long, E. Shelhamer and T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA, 2015.

[10] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016.

[11] S. Saito, T. Yamashita and Y. Aoki, Multiple object extraction from aerial imagery with convolutional neural networks, *Journal of Imaging Science and Technology*, vol.60, no.1, pp.1-9, 2016.

[12] S. Gould, R. Fulton and D. Koller, Decomposing a scene into geometric and semantically consistent regions, *Proc. of International Conference on Computer Vision (ICCV)*, Kyoto, 2009.

[13] 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 (MICCAI)*, Munich, Germany, 2015.

[14] M. Hussain, J. J. Bird and D. R. Faria, A study on CNN transfer learning for image classification, *UKCI 2018: Advances in Computational Intelligence Systems*, Nottingham, UK, 2018.

[15] F. Chollet, Xception: Deep learning with depthwise separable convolutions, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, Honolulu, Hawaii, 2017.