ICIC Express Letters
Volume **15**, Number **9**, September **2021**

ICIC International ©2021 ISSN 1881-803X
pp. **941–949**

DEVELOPMENT OF PADDY FIELD MAPPING FROM SATELLITE IMAGE USING SEMANTIC SEGMENTATION METHOD IN CENTRAL BORNEO

Alexander Agung Santoso Gunawan*, Fanny and Edy Irwansyah

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

Received December 2020; accepted March 2021

Abstract. *Paddy fields are one of the main resources in Indonesia, which can be categorized as an agricultural country. Land mapping, paddy field development planning, and the maintenance of the field data for the government are a few things that paddy field detection from satellite images can do. However, these tasks are still done manually by humans, leading to a decrease in the effectiveness and efficiency of the tasks. Furthermore, not only does this method need a huge cost, but there is also a possibility of finding various versions which cause problems when the data needed to be analyzed for reference in making decisions. This research was conducted to present solutions to classify accurately and automatically the paddy fields from satellite imagery. We solved the semantic segmentation problem by using U-Net and U-Net+ResNet-18 models and were evaluated using the dice coefficient metric. Both models were trained using paddy field images of the earth's surface in Central Borneo province. Based on the experiments, U-Net reached 85.32% and U-Net+ResNet-18 reached 89.38% in segmenting paddy fields. The results can be said as good enough in detecting and segmenting paddy field object.*

Keywords: Paddy fields, Semantic segmentations, Classification, Satellite imagery

1. Introduction.

Indonesia is an agrarian country where majority of the population makes a living in the agricultural sector. Unfortunately, many paddy fields have been converted into non-paddy fields and it is increasing from year to year, becoming a threat to the development of Indonesia's agricultural sector. The threat is the reason for the issuance of Presidential Regulation Number 59 of 2019 concerning Control of the Transfer of Function of Paddy Fields. In addition, the government also starts to create a spatial data that can be analyzed to determine which paddy fields can be saved. Therefore, the government needs a single, accurate data that can be used as a reference for development policy, resulting in the One Map Policy. To produce the One Map Policy, the Geospatial Information Agency (GIA) has carried out verification of national paddy fields in 15 largest rice-producing in Indonesia. However, the high cost and effort incurred in producing one version of spatial data for field mapping by employing human experts manually. In here, technological developments by automating land detection can help in producing field mapping for the Indonesian agriculture sector.

Semantic segmentation method is an image classification method in pixel level. This method aims to determine each pixel in the image as an object which can then be classified and segmented per object. In this problem, semantic segmentation can be applied to detecting paddy field in the image of the earth's surface [1]. The main problem is to search the meaningful texture descriptor [2]. The development of robust feature descriptors for viewpoint variation, scale change, and illumination remains a challenge for researchers.

DOI: 10.24507/icicel.15.09.941

Several studies have discussed semantic segmentation. Our previous research [3] has combined the Fully Convolutional Network (FCN) architecture with the Residual Network (ResNet) to segment roads and buildings in the city of Massachusetts. Segmentation is intended to automatically extract buildings and roads from aerial imagery so that it can help map creation, urban planning, and so on. This study aims to exploit a convolutional neural network which has succeeded in image classification with good result [4]. Furthermore, ResNet from He et al. [5] was conducted to provide a solution to the degradation problem where the deeper the layers used in the network, the higher the error rate generated in the training process. The combined model called SatNet has succeeded in achieving 82.26% for building segmentation and 96.76% for road segmentation.

On the other hand, the U-Net architecture designed for semantic segmentation by Ronneberger et al. [6] got much attention. Initially U-Net is designed for problems in biomedical field which needs to know the name of the object being detected and their location of the object on the image. The U-Net architecture can be considered as development of the FCN architecture which uses the concept of skip connection with the level of up-sampling adjusted to the down-sampling level so that it can produce better segmentation than the FCN method which only uses two times the up-sampling level.

This research would like to conduct research on the detection of paddy fields using semantic segmentation. For experiments, the dataset is image of the earth's surface of the Central Borneo province together with manual annotation as ground truth. Furthermore, we wish to combine the U-Net algorithm with ResNet to improve the accuracy of our previous algorithm. This combination is our main research contribution, and it can reach accuracy 89.38% in segmenting paddy fields. Following section will be research methods, the experiment results and finally the conclusion of our research.

2. Research Method.

The research can be split into some phases as follows: (a) requirement, (b) design and development, (c) testing, (d) implementation.

(a) *Requirement.* Project initiated with identifying the problems to get the problem statement, namely paddy field semantic segmentation. The selection of satellite data that is used performed at the end of this phase. In this study, we used the SPOT 6 satellite image dataset with paddy field ground truth that has been verified by GIA. This dataset takes the form of a satellite image covering the province of Central Borneo with polygons as the labels.

(b) *Design and Development.* The research continues with the selection of network architecture models that will be used. The architecture, chosen for semantic segmentation in this study, is the U-Net architecture. This architecture provides a skip connection that exists at each feature extraction stage connecting to the corresponding up-sampling stage. Residual network or ResNet was chosen as the comparison of segmentation performance with U-Net basic architecture. ResNet provides skip connection in a shorter stage, namely in each feature extraction stage. This skip connection serves to help the model avoid degradation problem, where the deeper the convolution layer is, the higher training error rate resulted. After the architecture has been chosen, the selected satellite data will be processed first until it becomes an appropriate dataset. ArcGIS Pro is used to perform image processing. The satellite image is cut into small sizes and the ground truth is processed to become a target label for the image segmentation that is ready to be trained. Each piece of image is ensured to have the same size and coordinates as the ground truth which has been cut into the target label. The data, which has been preprocessed, then needs to be selected and cleaned until finally there are 1100 images as dataset that is ready to be trained with 1000 training data and 100 validation data.

Furthermore, the construction of previously selected network architecture is carried out. The training process is using a model that has been built and the previously processed dataset. The first model (see Figure 1) used the U-Net architecture that refers to [6].

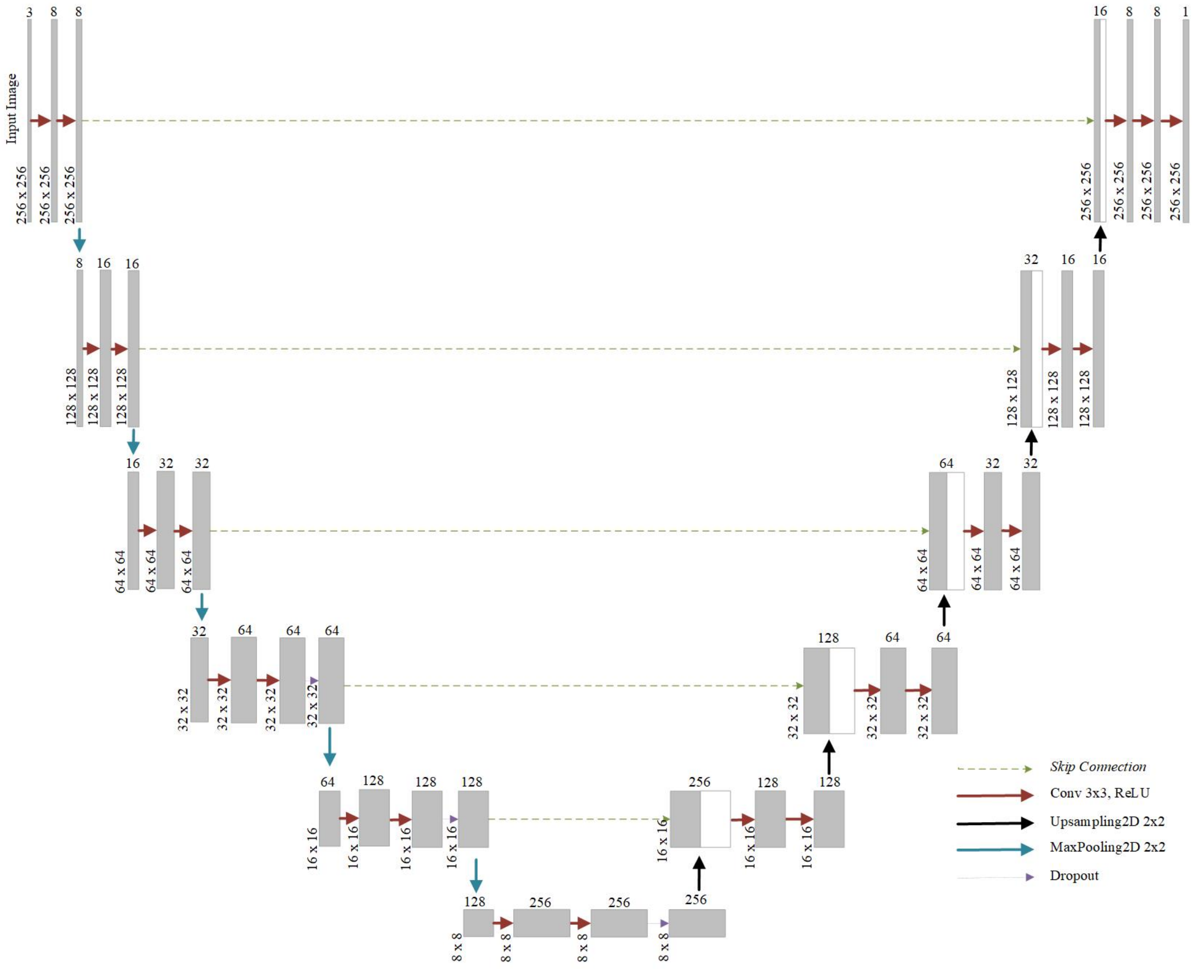


Figure 1. (color online) U-Net model architecture

The second architecture (see Figure 2) is a combination of ResNet-18 with U-Net architecture. Henceforth, this model will be called U-Net+ResNet-18. The convolution layers at feature extraction stages are changed to residual blocks and the up-sampling process is changed according to the residual network. Up-sampling is performed using a transposed convolution which ends with sigmoid activation function. The ResNet-18 network is built upon [5].

In the U-Net+ResNet-18 architecture model, the ResNet 18-layers network is used as feature extraction or down-sampling process. The resulting feature will be deconvolved using the U-Net architecture so that it can be used to segment images. By using the base of U-Net architecture, the ResNet can get the feature of skip connection and encoder-decoder architecture advantages [7]. The residual blocks used in architecture are divided into convolutional blocks and identity blocks (see Figure 3). The convolutional block is used when there is an added stride, while the identity block is used when there is no change in the shape of the input. The difference between the two lies in the identity of the input which is added at the end of the block. In a convolutional block, additional convolution is given to the input shortcut to get the same result so that it can be added. Deconvolution block is a convolutional block that uses transposed convolution [8]. Batch normalization is used in each block to improve the network performance [9].

The architectures for training are U-Net and ResNet-18 combined with U-Net. These two architectures were chosen to compare their performance and evaluation results. Dice coefficient formula is chosen as the method of evaluation of segmentation results. Data processing was performed using the ArcGIS Pro application. The dataset is modeled in

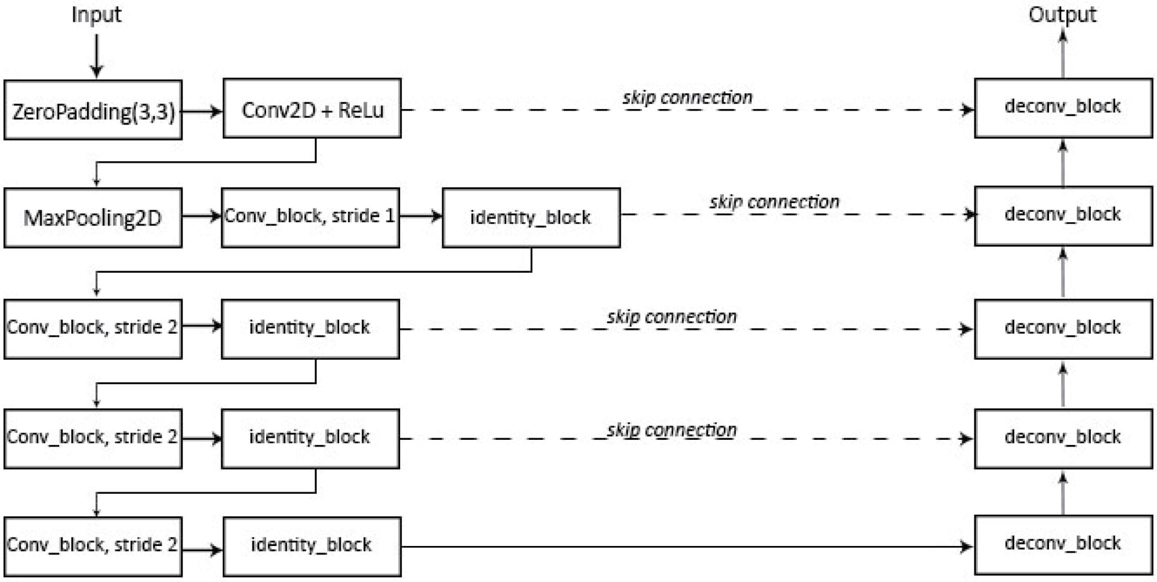


FIGURE 2. U-Net+ResNet-18 model architecture

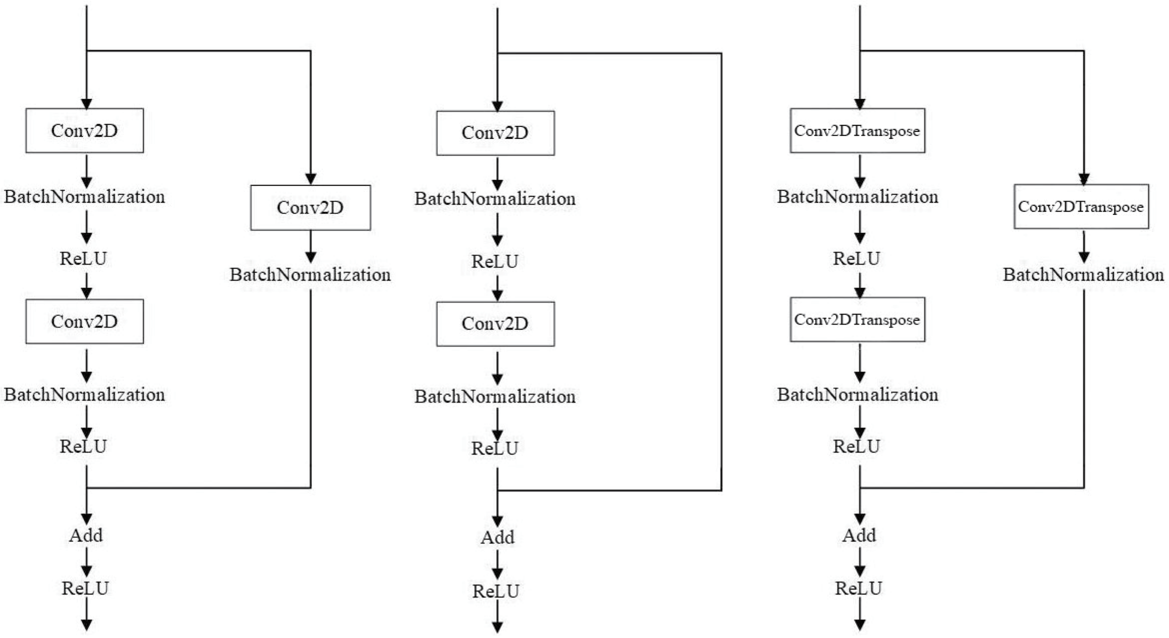


FIGURE 3. Convolutional block (left), identity block (center) and deconvolutional block (right)

such a way that the features that can be key to paddy field interpretation can be seen. Image datasets with labels are also adjusted to size and name for ease of the training process. When there are additions or another incremental, it can be started over from this phase [10]. Both architectures will be trained with parameter of batch size = 1 and epochs = 100. Architectural training uses 1000 input images for each epoch. This training process begins with calculating the value during the forward propagation process to produce the predictive value. This value will then be evaluated using the dice coefficient evaluation metric [11] where $|X|$ and $|Y|$ are the cardinalities of the two sets.

$$\textit{Dice Coefficient} = \frac{2|X \cup Y|}{|X| + |Y|}$$

The function used to calculate the loss value is the binary dice loss which is a combination of the binary cross-entropy loss which is commonly used for binary class classification problems [12] and dice loss which is the loss of the dice coefficient metrics.

The next backward propagation is carried out to improve the bias and weight values in the artificial neural network using the Adam optimization function [13]. This process continues for 100 epochs so that a mathematical model with optimal bias values and weights are produced that can be used for the paddy field segmentation problem. The resulting model is implemented in the application to perform semantic segmentation where each pixel in the input image will be predicted. The model will produce a matrix measuring 256×256 with each element having a value range of 0 to 1. The value is the possibility when each matrix element (pixels in the image) is classified as a paddy field in the training process. After obtaining the image segmentation results using each model, the area calculation will be carried out based on the number of pixels classified as paddy fields. The calculation is done using the spatial resolution of SPOT 6 imagery where each pixel measures 1.5 meters × 1.5 meters on the actual scale [14]. With n as the number of pixels, which is classified as paddy field, the equation for obtaining $Area$ can be written as below.

$$Area = n \times 1.5 \times 1.5$$

(c) *Testing.* In this phase, evaluation will be carried out using evaluation metrics and loss function to calculate the percentage of segmentation and loss evaluations generated through training data and validation of each model. Testing is also done by taking the best epoch results which are then tested using images that are not training data or validation data so that the performance of the segmentation results of the two models can be seen and analyzed.

(d) *Implementation.* At this stage the implementation of the application design is carried out. Features and functions that have been designed and the model that has been built will be implemented in an application. The graphic interface is developed based on eight golden rules using the Kivy library in Python. The application is able to simulate and analyze the segmentation results.

3. Results and Analysis.

The dataset used in this study consists of imagery of the surface of the earth in Central Borneo province and the target label used to determine the class of paddy fields or non-paddy fields. The dataset used consists of 1100 data, with 1000 training data and 100 validation data. Based on the experiments conducted, the two models managed to achieve optimum results with the following parameters in Table 1.

Table 1. Default parameter experiment

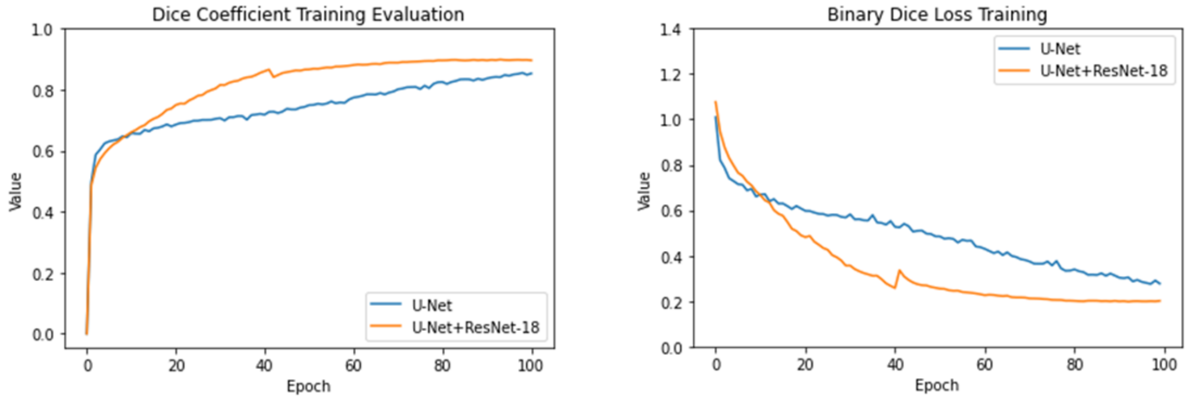
Image Size	256
Batch Size	1
Epoch	100
Optimizer	Adam
Dropout Rate	0.5
Activation Function on Hidden Layer	ReLu
Activation Function on Output Layer	Sigmoid

There are different changes of learning rates for each model. Detail of the changes can be seen as:

$$learning\ rate_{U\text{-}Net} \begin{cases} 0.00035, & epochs > 0 \\ 0.00025, & epochs > 25 \\ 0.00015, & epochs > 50 \end{cases}$$

$$\textit{learning rate}_{\textit{U-Net+ResNet-18}} \begin{cases} 0.00015, & \textit{epochs} > 0 \\ 0.000005, & \textit{epochs} > 40 \\ 0.0000001, & \textit{epochs} > 80 \end{cases}$$

Result Comparison Model U-Net and U-Net+ResNet-18 for Data Training. The type of training carried out is supervised learning. Training is carried out in 100 epochs and results can be seen every 1 step. The U-Net model training process takes approximately 10 hours while the U-Net+ResNet-18 model training process takes approximately 14 hours. Figure 4 shows the comparison charts with dice coefficient (see Figure 4(a)) and loss of the two models for training data (see Figure 4(b)).



(a) (b)

FIGURE 4. (a) Dice coefficient training data comparison chart; (b) binary dice loss training data comparison chart

To see the most accurate model, the result which has the lowest loss and the highest *dice coefficient* was selected. The comparison results can be seen in Table 2 which contains the best results for each model.

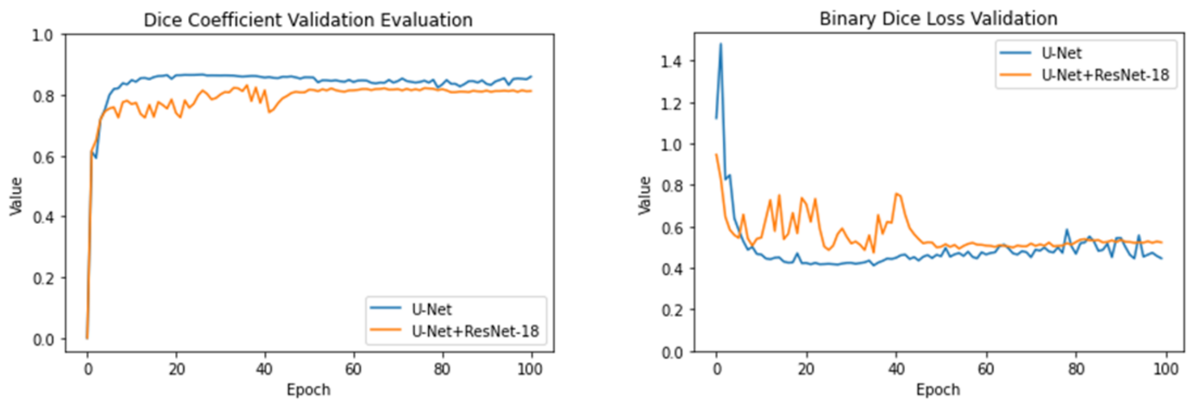
TABLE 2. Comparison of U-Net and U-Net+ResNet-18 training results

Model	Epoch	Learning Rate	Loss	Dice Coefficient
U-Net	98	0.00015	0.2755	0.8552
U-Net+ResNet-18	93	0.0000001	0.1976	0.8988

Based on the table, it can be seen that the U-Net+ResNet-18 model has better evaluation results than the U-Net model, both in the comparison of the dice coefficient value and the loss value of the two models. To determine the most accurate model must also be through the evaluation results of the validation data. Validation data is not used in the training process so that the model must be able to make predictions without first studying the data, which is different from the training data used by the model for learning. Figure 5 consists of the comparison of the dice coefficient (see Figure 5(a)) and loss evaluation results of data validation using the U-Net model and the U-Net+ResNet-18 model (see Figure 5(b)).

From the graph, the dice coefficient and loss values of each model using validation data can be seen. The comparison results can be seen in Table 3 which contains the best results from each of the following models as such.

From the results of the validation data, it can be seen that the U-Net model has a better level of accuracy than the U-Net+ResNet-18 model but worse in the training data. From



(a) (b)

Figure 5. (a) Dice coefficient validation data comparison chart; (b) binary dice loss validating data comparison chart

Table 3. Comparison of U-Net and U-Net+ResNet-18 validating results

Model	**Epoch**	**Learning Rate**	**Loss**	**Dice Coefficient**
U-Net	36	0.00025	0.4122	0.8613
U-Net+ResNet-18	36	0.00015	0.4735	0.8322

the charts it can be deduced that the training of combination model should be stopped early for avoiding the overfitting. Therefore, to obtain the best results from each model based on training data and validation data, evaluation is carried out by taking the largest average of the difference between the dice coefficient and loss values. In Table 4 it can be seen that the combination model is stopped in earlier epoch. The averaging equation is as follows:

$$Average = Max\left(\frac{(Dice\ Coefficient_{training} - loss_{training}) + (Dice\ Coefficient_{validation} - loss_{validation})}{2}\right)$$

Based on the above equation, the best average results are described in Table 4 below.

Table 4. Comparison of U-Net and U-Net+ResNet-18 average results

Model	**Epoch**	**Learning Rate**	**Training**		**Validation**		**Average**
			Loss	**Dice Coefficient**	**Loss**	**Dice Coefficient**	
U-Net	**100**	0.00015	0.2771	0.8532	0.4466	0.8601	**0.4948**
U-Net+ResNet-18	**76**	0.000005	0.2055	0.8938	0.5046	0.8223	**0.5030**

Based on the results of the table above, the U-Net+ResNet-18 model has the best segmentation results with an average value of 50.30% achieved on the 76th epoch. Figure 6 shows the example results of the two models on their best epoch.

The results of the segmentation area calculation can be seen in Table 5. The result of the combination model is closer to the ground truth.

4. Conclusions.

Based on the experiments, the following conclusions can be drawn.

- This research has produced an application that can detect paddy fields using semantic segmentation methods with U-Net and U-Net+ResNet-18 models which are

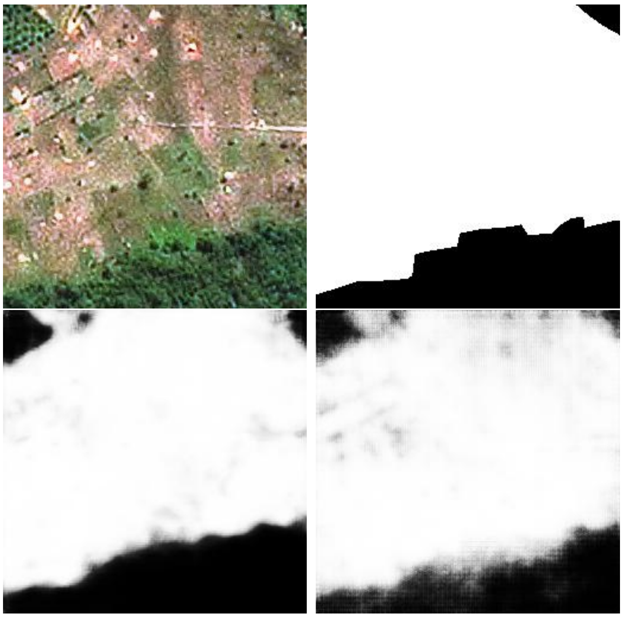


FIGURE 6. Image, ground truth, U-Net and U-Net+ResNet-18 segmentation

TABLE 5. Results of ground truth, U-Net and U-Net+ResNet-18 segmentation areas

Ground truth	15.17200 ha
U-Net	11.38118 ha
U-Net+ResNet-18	12.12570 ha

trained with the SPOT 6 satellite image dataset in Central Borneo provence. The dataset is divided into 1000 training data and 100 validation data.

- The evaluation results with the dice coefficient using the best U-Net model can reach 85.32% with a loss of 27.71% in training data and reach 86.01% with a loss of 44.66% for validation data achieved at the 100th epoch. The training process is carried out for approximately 10 hours.
- The evaluation results with the dice coefficient using the best model U-Net+ResNet-18 can reach 89.38% with a loss of 20.55% in training data and up to 82.23% with a loss of 50.46% for validation data achieved on the 76th epoch. The training process is carried out for approximately 14 hours.
- The experimental results suggest that the training of combination model should be stopped early for avoiding the overfitting. Furthermore, the best average evaluation was achieved by the U-Net+ResNet-18 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.

REFERENCES

[1] P. Kaiser, Learning aerial image segmentation from online maps, *IEEE Trans. Geoscience and Remote Sensing*, 2017.

[2] S. Thewsuwan and K. Horio, Local spatial information with bag-of-visual-words model via graph-based representation for texture classification, *International Journal of Innovative Computing, Information and Control*, vol.16, no.5, pp.1611-1621, 2020.

[3] A. A. S. Gunawan, I. Arifiany and E. Irwansyah, Semantic segmentation of aerial imagery for road and building extraction with deep learning, *ICIC Express Letters*, vol.14, no.1, pp.43-52, 2020.

[4] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *The 3rd International Conference on Learning Representations (ICLR)*, San Diego, 2015.

[5] K. He, X. Zhang, S. Ren and J. Sun, *Deep Residual Learning for Image Recognition*, Microsoft Research, 2015.

[6] O. Ronneberger, P. Fischer and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, Computer Science Department and BIOSS Centre for Biological Signalling Studies, University of Freiburg, Germany, 2015.

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, *European Conference on Computer Vision (ECCV2018)*, Munich, 2018.

[8] V. Dumoulin and F. Visin, A guide to convolution arithmetic for deep learning, *arXiv.org*, arXiv: 1603.07285, 2018.

[9] S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, *The 32nd International Conference on Machine Learning (ICML)*, Lille, 2015.

[10] A. Singh and P. J. Kaur, A simulation model for incremental software development life cycle model, *International Journal of Advanced Research in Computer Science*, vol.8, no.7, pp.126-132, 2017.

[11] R. R. Shamir, Y. Duchin, J. Kim, G. Sapiro and N. Harel, Continuous dice coefficient: A method for evaluating probabilistic segmentations, *arXiv.org*, arXiv: 1906.11031, 2016.

[12] Z. Zhang and M. R. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, *The 32nd Conference on Neural Information Processing Systems*, Montréal, Canada, 2018.

[13] S. Gandhi and B. Sarkar, Remote sensing techniques, *Essentials of Mineral Exploration and Evaluation*, pp.81-95, 2016.

[14] D. P. Kingma and J. L. Ba, Adam: A method for stochastic optimization, *The 3rd International Conference for Learning Representations (ICLR)*, San Diego, 2015.