Instance Segmentation of Water Body from Aerial Image using Mask Region-based Convolutional Neural Network

Sangdaow Noppitak
PhD Student, Multi-agent Intelligent Simulation Laboratory, Department of Information Technology, Faculty of Informatics, Mahasarakham University, Thailand
61011261007@msu.ac.th

Sarayut Gonwirat
PhD Student, Multi-agent Intelligent Simulation Laboratory, Department of Information Technology, Faculty of Informatics, Mahasarakham University, Thailand
61011262003@msu.ac.th

Olarik Surinta
Multi-agent Intelligent Simulation Laboratory, Department of Information Technology, Faculty of Informatics, Mahasarakham University, Thailand
Olarik.s@msu.ac.th

ABSTRACT

Land use is constantly changing, and water plays a critical role in the process. If changes are noticed quickly or are predictable, land use planning and policies can be devised to mitigate almost any problem. Accordingly, researchers present a mask region-based convolutional neural network (Mask R-CNN) for water body segmentation from aerial images. The system's Aerial image water resources dataset (AIWR) was tested. The AIWR areas were agricultural and lowland areas that require rainwater for farming. Many wells were spotted throughout the agricultural areas. The AIWR dataset presents two types of data: natural water bodies and artificial water bodies. The two different areas appear as aerial area images that are different in color, shape, size, and similarity. A pre-trained model of Mask R-CNN was used to reduce network learning time. ResNet-101 was used as backbone architecture. The information gathered in the learning process is limited, and only 720 pictures were produced, Researchers used data augmentation to increase the amount of information for training by using affine image transformation, including scale, translation, rotation, and shear. The experiment found that mask R-CNN architecture can specify the position of the water surface. Measuring method in this case is mAP value. The mAP value is at 0.30 without data augmentation. However, if using the R-CNN mask with data augmentation, the mAP value increased to 0.59.

CCS Concepts

• Computing methodologies → Image segmentation
• Computing methodologies → Neural networks.

Keywords

Instance Segmentation; Water Body; Aerial Image; Mask R-CNN; Transfer Learning.

1. INTRODUCTION

The two terms "land cover" and "land use" are typically used together [12]. Over the past ten years the difference between land cover and land use has attracted many researchers [2] prompted by a change in land cover to accommodate changes in land use. As such, if land use data are accurate and up-to-date, we can apply that information to many objectives, such as city planning, environmental audit or evaluation, and national policy [14].

Elagouz et al. [3] tracked land use in the Nile River, Egypt with RS technology to determine the impact of land changes in urban areas during and after the year 2011. The land changed because of the unplanned expansion of a nearby city. Jazouli et al. [7] said that soil erosion was the most important cause of land degradation throughout the world. Jazouli et al. has predicted the impact of land use changes, which affect soil erosion, in the Oum Er Rbia basin, Morocco. They studied the mountainous areas with steep, slopes, and clay soil where places are higher risk for soil erosion. Soil erosion is sometimes caused by human activities and local weather. Further research would be beneficial for generating land use prediction maps, detecting land use changes, and creating yearly mapping for soil erosion.

Nowadays, deep learning research is very popular, For example, land cover analysis research [13]. The research used deep neural networks for analysis of Landsat 5/7 satellite images to show land cover maps for agriculture, including agriculture areas, water, grass, mixed wood, and border. Kussul et al. [8], used convolutional neural network (CNN), which is the method for classification of recorded images, in remote sensing work. The CNN classified recorded images in category of optical and synthetic aperture radar (SAR) derived from Landsat-8 and Sentinel-1A using CNN type one-dimensional (1-D) and 2-D. The results from CNN were compared with the random forest method and the ensemble neural networks technique. 2D-CNN got the highest score with a 94.6% accuracy rate. However, 2D-CNN still has some problems distinguishing small objects. Spatial resolution of the satellite images is 30 meters, which is low resolution.

Miao et al. [9] presented water body segmentation using restricted receptive field deconvolution network (RRF DeconvNet) for extraction of water body from high-resolution spectrum images. This method did not require infrared spectrum images, and this method also decreased blurring boundaries problem by using a new loss function called edges weighting loss (EWLoss). The researchers tested with the dataset collected from Google Earth. The images from Google earth were in the visible spectrum at 50 meters spatial resolution of the rural area at Suzhou and Wuhan, China. The experiments showed that RRF DeconvNet method using EWLoss had 96.9% accuracy rate.

ICISS 2020, March 19–22, 2020, Cambridge, United Kingdom
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7725-6/20/03…$15.00
https://doi.org/10.1145/3388176.3388184

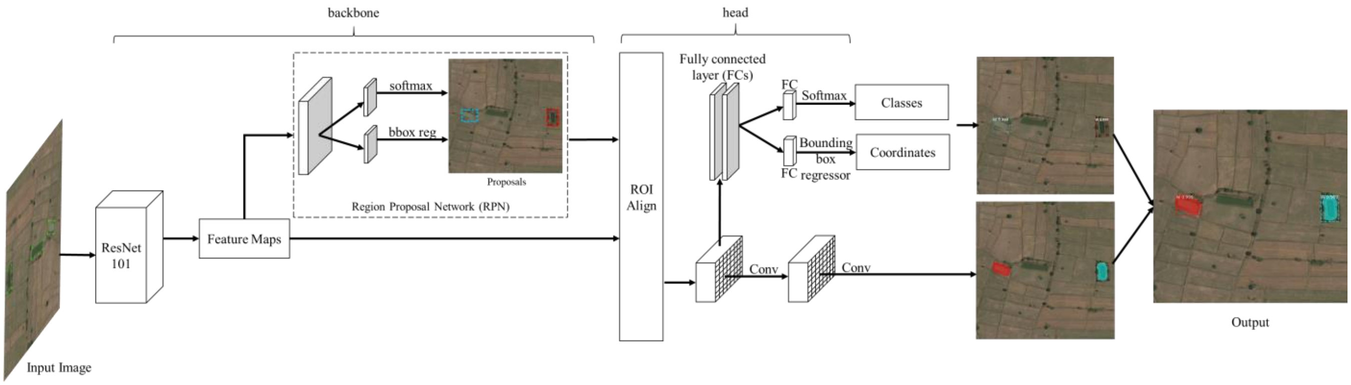


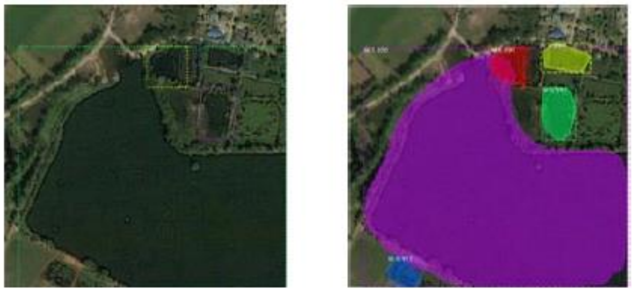
Figure 1. Mask R-CNN Framework.

Wen et al. [15] used Mask R-CNN to segment the building area and the background from Google Earth images. They created a new dataset with 2,000 aerial images in Fujian province, China. The sizes of images used in the experiment range from 1,000x1,000 to 10,000x10,000 pixel. All aerial images were tagged with a label. In the experiment, researcher used pre-trained model of ResNet architecture. All images were resized to 500x500 pixel. The result showed Instance Segmentation using Mask R-CNN resulted in mean Average Precision (mAP) value at 0.9063.

Contribution: this article presents mask region-based convolutional neural network (Mask R-CNN) for water body segmentation from aerial images. This method has been called instance segmentation. ResNet-101 was used as backbone architecture. Mask R-CNN architecture was tested with aerial image water resources dataset (AIWR). The AIWR is the images of agricultural areas in the northeast region of Thailand; these are fertile agricultural areas where people grow rice. The areas require rainwater for farming and many wells can be spotted throughout the agricultural area. Water body data were collected from 2 types, natural water bodies (W1) and artificial water bodies (W2). The aerial images of water bodies were different in color, shape, size, and similarity. This dataset includes 800 images, so AIWR dataset challenges the instance segmentation process.

This research also attempted to add data augmentation in the category of affine image composed of 4 different methods: scale, translation, rotation, and shear. Augmentation processes were used only in the training process. Data augmentation would be a random parameter value. The images, which trained in each epoch using mask in the R-CNN process, were different. The experiments found that data augmentation had improved the performance of Mask R-CNN in the instance segmentation process when used with AIWR Dataset. The result showed better performance for specifying water bodies. The mAP value increases from 0.30 to 0.59 when researcher used data augmentation.

Paper outline: Section 2 explains the Mask R-CNN architecture used for making instance segmentation. Section 3 explains the data collection process of the aerial image water resources (AIWR) Dataset. Experimental results are explained in Section 4. The final section is the conclusion and future work.



a) b)

Figure 2. Result from a) Faster R-CNN method and b) Mask R-CNN.

2. MASK R-CNN ARCHITECTURE

Mask R-CNN was presented by He et al. [6] in 2017 for improving instance segmentation performance. Mask R-CNN was developed from Faster R-CNN, which was presented by Ren et al. [11] in 2015

Faster R-CNN were designed to use a convolutional network (ConvNet) for feature map extraction of Images. ConvNet can use VGGNet and ZFNet architectures. After that, the region proposal network (RPN) was used for inspecting the object areas. RPN operates location inspection for each object. The reason for running RPN was to create a bounding box of each object, which is called ROI pooling layer. In the ROI pooling layer, ROI in each section would be sent to fully connected layers (FCs) for ROI feature vector calculation before sending the value to the softmax function for consideration of ROI as an object. After that, the function will predict the object type in ROI as shown in Figure 1.

According to the introduction, faster R-CNN [11] is an object detector, so this function can't specify an object in pixel-to-pixel, or we would call it instance segmentation as shown in Figure 2a. Mask R-CNN has been designed to help instance segmentation by using the capabilities of RPN to specify ROI. The next step is to segment the ROI areas to specify the edge of an object as shown in Figure 2b.

2.1 Backbone Architecture

Backbone architecture consists of 2 main networks|; ConvNet and RPN. ConvNet used in this research is ResNet-101 architecture, using a pre-trained model derived from learning of COCO dataset. This architecture can reduce network learning time. The main function of ResNet architecture is to extract feature maps from aerial images, then use the region proposal network (RPN) to find the location of an object using ResNet-101 architecture.

2.2 Head Architecture

An advantage of Mask R-CNN is that it can perform instance segmentation by using the location of any object, derived from RPN, which is another name for the region of interest (ROI). The ROI will be considered whether it is an object or not. If the ROI area is an object, then types of an object would be considered in the next step. This step is similar to Faster R-CNN. After that, areas will be calculated for intersection over union (IoU). Shown in Equation a) The IoU values are assigned to be greater or equal to 0.5.

Any area with an IoU value greater or equal too 0.5 is required to find the perimeter of an image. Sometimes, this method is also known as a segmentation mask. This process is an additional process from Faster R-CNN. In each ROI area, there is only one class. Then, the semantic segmentation model is created. It is as same as binary classification to distinguish an object from background.

3. AERIAL IMAGE WATER RESOURCES DATASET

According to the standard of land use code by fundamental geographic data set: FGDS), Thailand [5] land use classification requires an analysis and transformation of satellite images data together with field survey data. In this article, researchers studied only land use in water bodies. The water bodies in this research can be divided into 2 levels: natural body of water (W1) artificial body of (W2) water.

The deep learning method was used in this research for aerial image data analysis. The aerial images were derived from Bing map by collecting only data in the northeastern region of Thailand. The northeast of Thailand is lowland area mainly used for growing rice, There are also agricultural areas that rely on rainwater for agriculture. As such, there are many ponds in and around the agricultural areas.

The experiments in the study used the Mask R-CNN algorithm which is a suitable method for performing instance segmentation. The model in this experiment can be further developed and applied to water management tasks. Farmers in the northeastern region of Thailand can also create water management plans.

The aerial image data used in this research was 1:50 meters. Every aerial image had 650x650 pixels. Those images included water bodies type W1 and W2 as shown in Figure 3a. Ground truth of all aerial images was set for before sending it to be analyzed and interpreted by remote sensing experts. This assured that the water bodies groupings were correct. An example of ground truth, which has been checked by experts as shown in Figure 3b. Ground truth has been used in learning the algorithm in deep learning mode and also used in further evaluation.

The aerial images used in this experiment consists of water body: types W1 (see, Figure 3, Column 1, 2, and 3) and W2 (see, Figure 3 Column 4). Aerial image water resources dataset, AIWR has 800 images. Data were chosen at random and divided into 3 sections: training, validation, and test set with ratio 8:1:1. Therefore, 640 aerial images were used for learning and creating the model, 80 images were used for validation, and the remaining 80 images were used for test.

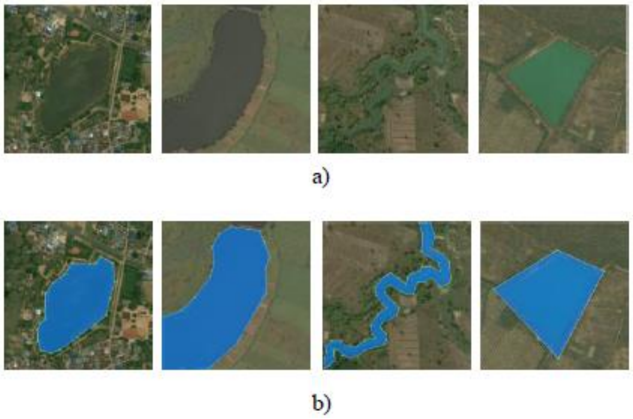


Figure 3. Example of aerial images. a) Water bodies W1 and W2 b) ground truth of water resources.

This dataset challenges for instance segmentation process because the water body are W1 and W2 types. There are 4 challenging objectives: color, shape, size, and similarity as follows:

- **Color**: Figure 4a shows that water bodies have different color, for example white, blue, gray and black. Some areas are covered by unwanted flora, so the images are seen as dark green and black.
- **Shape**: The shape of the areas have different characteristics such as triangles, squares, curves, U-shaped and zigzag as shown Figure 4b.
- **Size**: the water body sizes are different. Size measurement in Bing maps found that the water bodies sizes range from 10, 20, 30, 60 and 120 meters as shown in Figure 4c. When researchers observe 10 meter wide water sources, only a small point can be seen.
- **Similarity**: Aerial images of some water bodies are similar to other types of land use, for example flooded areas, water areas that are obscured by trees, or buildings on water areas etc. Figure 4d uses the dotted lines to show areas that have the characteristics as mentioned above

4. EXPERIMENT AND DISCUSSION

A deep learning algorithm was used in this research for instance segmentation. This method can identify the areas in pixel-to-pixel by using Mask R-CNN architecture. The method is suitable for water body segmentation because it can analyze both natural water bodies and artificial water bodies. The data were collected from aerial image data from agricultural areas in the northeastern region of Thailand There was a total of 800 aerial image data. Those images were divided by the 10-fold cross-validation method. There were 720 images for training, and 80 images were used for test. All aerial images were resized to 512x512 pixel.

In this research, TensorFlow platform was used for training and testing the Mask R-CNN algorithm which runs on GPU GeForce GTX 1070 Ti, Intel(R) Core-i5, 7400CPU @ 3.00GHz, 8GB RAM, Linux Operating system. ResNet architecture is backbone architecture for learning aerial imagery learning. This research used transfer learning [1] to reduce learning time of ResNet architecture. Pre-trained model of ResNet-101 architecture, which derived from the learning process of COCO dataset, was also used. Then, researchers used the mentioned model to perform Fine-Tune for adjusting the parameters in order to make it become suitable for the AIWR dataset.

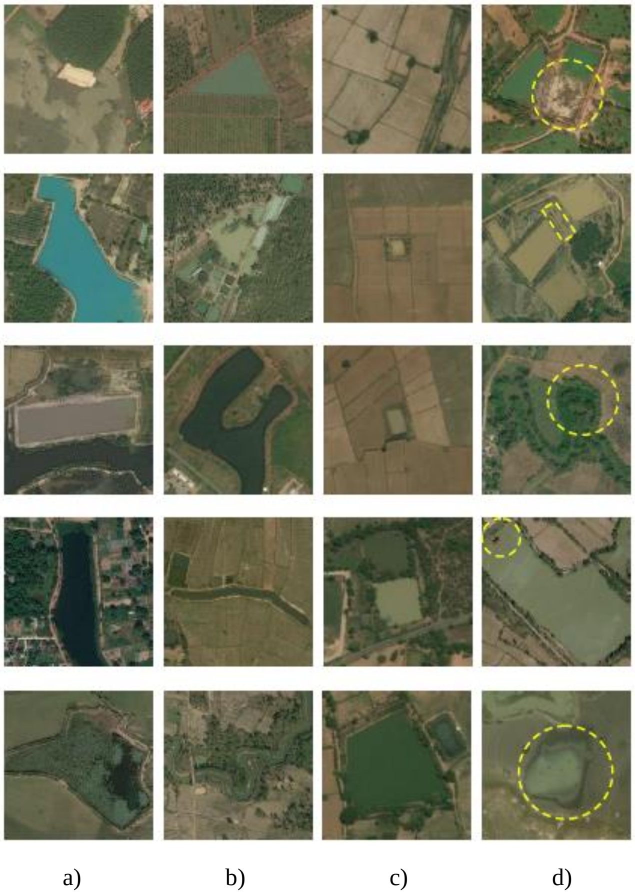


Figure 4. Challenges of instance segmentation of collected data are a) color, b) shape, c) size, and d) similarity.

The parameters used for Fine-tune consist of NUM_CLASSES=2, BATCH_SIZE=4, FPN_CLASSIF_FC_LAYERS_SIZE=512, IMAGES_PER_GPU = 1, IMAGE_MIN_DIM = 512, IMAGE_MAX_DIM = 512, IMAGE_SHAPE=[512, 512, 3], RPN_ANCHOR_SCALES=(8, 16, 32, 64, 128), STEPS_PER_EPOCH=100, TRAIN_ROIS_PER_IMAGE=32, VALIDATION_STEPS=5, and LEARNING_RATE=0.0001

One of the deep learning problems is the amount of training data is too small. A common way to solve the problem is to perform data augmentation, which can be divided into 2 groups including the traditional, white-box method or black-box method. Two common methods for image augmentation in traditional transformations are affine image transformations and color modification [10]

This research used data augmentation, affine image transformations series, which includes scale={"x": (0.8, 1.2), "y": (0.8, 1.2)}, translate_percent={"x": (-0.2, 0.2), "y": (-0.2, 0.2)}, rotate=(-25, 25), and shear=(-8, 8). An example of aerial images obtained after data augmentation are shown in Figure 5.

4.1 Model Evaluation

To Evaluation Mask R-CNN algorithm, researchers used mean average precision (mAP) [4] ,which is a method for evaluating the effectiveness of image retrieval by an intersection over union (IoU) calculation from the following equation.

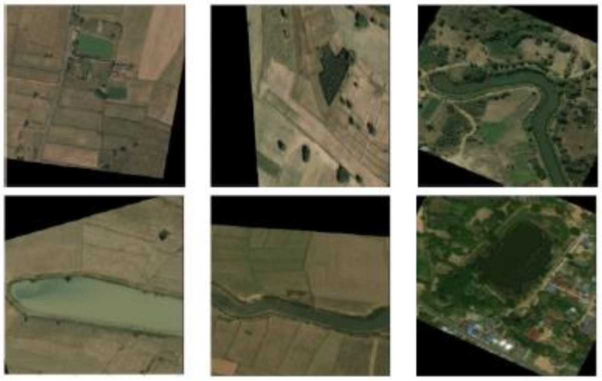


Figure 5. Examples of data augmentation.

$$IoU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \quad (1)$$

where $B_p \cap B_{gt}$are the areas of intersection between the predicted area. Ground truth (gt) is bounding boxes and $B_p \cup B_{gt}$ is the area of union, determined by the value of $IoU \geq 0.5$

After that, true positive (TP), a correct detection, and false positive (FP) (A wrong detection) were calculated. The detection were performed by $values \geq 0.5$, false negative (FN) (A ground truth not detected) and true negative (TN) (corrected misdetection). The TP, FP, FN, TN value are taken to calculate precision (P) and Recall (R) value.

AP value was considered by average of maximum precision at a set of 11-spaced recall levels . The equation is as follows:

$$AP = \frac{1}{11} \sum_{r \in \{0,0.1,..,1\}} P_{inter\ p}(r) \quad (2)$$

with $P_{inter\ p}(r) = \max_{\tilde{r}.\tilde{r} \geq r} p(\tilde{r})$

where $p(\tilde{r})$ is the measured precision at recall $\tilde{r}$
After that, mAP value are calculated as the following equation.

$$m\text{AP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_\text{i} \quad (3)$$

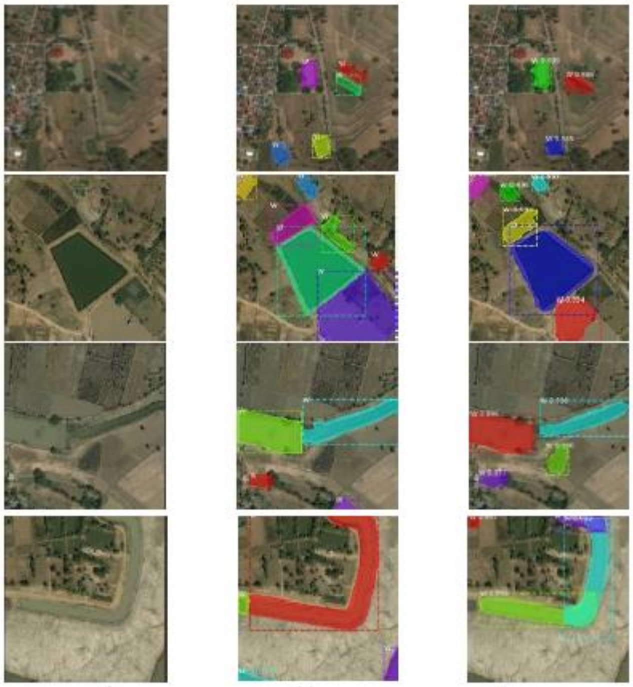
where N is number of query

4.2 Result of Instance Segmentation of Water Body

Table 1 is the result of the experiment of mask R-CNN architecture to segment water bodies from the AIWR dataset. Augmentation data experiments of AIWR dataset were performed by affine image transformations method, including scale, translation, rotation, and shear. The result shows that the loss error value from training processes was up to 1.08. That result in the mAP was as low as 0.30, but when researchers tested again using data augmentation, the loss errors were reduced to only 0.41 and the mAP increased to 0.59, which is almost 2 times higher. However, the data augmentation process takes 12 day and 9 hours to learn.

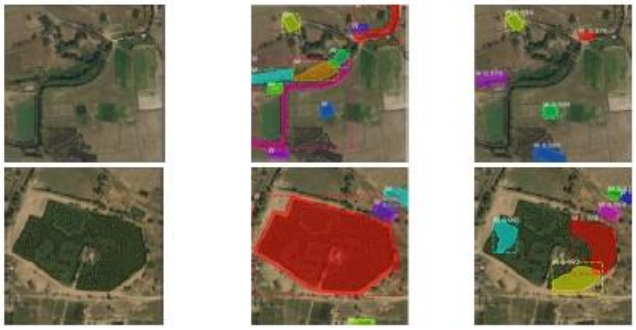
Table 1. The result of the experiment using mask R-CNN with the AIWR Dataset

Augment	Validation loss	mAP	Training Time	Test Time /img
False	1.08	0.30	11d 15h 16min 27s	3 µs
True	0.41	0.59	12d 9h 48min 25s	4 µs



a) b) c)

Figure 6. Result of instance segmentation using mask R-CNN with data augmentation. a) Aerial images b) images with ground truth, and c) instance segmentation.



a) b) c)

Figure 7. Error results in segmentation. a) Aerial images b) images with ground truth and c) error of instance segmentation.

Figure 6 included the data augmentation method in data training in order to create a model. The result shows that data augmentation in data training leads to a better result of segmentation. Figure 7 demonstrates errors from instance segmentation. It is because Figure 7c (Row 1) cannot segment the river areas covered with trees, and Figure 7c (Row 2) is the area covered by unwanted flora.

5. CONCLUSION

In this paper evaluated the accuracy of instance segmentation by Mask R-CNN together with data augmentation. The mAP values were used as the measuring method. This research tested with aerial images of water resources dataset (AIWR). The areas are the lowlands which require rainwater for farming. The challenges of AIWR dataset the collection of 2 types of water bodies: natural water bodies and artificial water bodies. The two types of data are different in color, shape, size, and similarity. This paper used a pre-trained model to reduce learning time of the Mask R-CNN. This research has shown that the mask R-CNN architecture combined with data augmentation can identify the water surface using the mAP value for measurement. The value was up to 0.59. It is almost two times greater than not using data augmentation method.

In future work, because the data tested is aerial photography obtained from Bing map, only RGB colors can be evaluated. If other research can use data from satellites, such as Landsat, which has a band specifically for water analysis, the result of an analysis of water bodies with different color might give higher accuracy. Any new architecture suitable for water body analysis might be used to expect an even higher accuracy rate.

6. REFERENCES

[1] Bunrit, S., Nittaya, K. and Kerdprasop, K. 2019. Evaluating on the Transfer Learning of CNN Architectures to a Construction Material Image Classification Task. *International Journal of Machine Learning and Computing*. 9, 2 (2019), 201–207.

[2] Caldas, M.M., Goodin, D., Sherwood, S., Campos Krauer, J.M. and Wisely, S.M. 2015. Land-cover change in the Paraguayan Chaco: 2000–2011. *Journal of Land Use Science*. 10, 1 (Jan. 2015), 1–18.

[3] Elagouz, M.H., Abou-Shleel, S.M., Belal, A.A. and El-Mohandes, M.A.O. 2019. Detection of land use/cover change in Egyptian Nile Delta using remote sensing. *Egyptian Journal of Remote Sensing and Space Science*. (Jan. 2019), 0–5.

[4] Everingham, M., Gool, L. Van, Williams, C.K.I., Winn, J. and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision. 88, 2 (2010), 303–338.

[5] GISTDA 2013. *Fundamental Geographic Data Set (FGDS)*.

[6] He, K., Gkioxari, G., Dollár, P. and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)* (Venice, Italy, Oct 22-29, 2017), 2961–2969.

[7] Jazouli, A. El, Barakat, A., Khellouk, R., Rais, J. and Baghdadi, M. El 2019. Remote sensing and GIS techniques for prediction of land use land cover change effects on soil erosion in the high basin of the Oum Er Rbia River (Morocco). *Remote Sensing Applications: Society and Environment*. 13, (Jan. 2019), 361–374.

[8] Kussul, N., Lavreniuk, M., Skakun, S. and Shelestov, A. 2017. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters*. 14, 5 (May 2017), 778–782.

[9] Miao, Z., Fu, K., Sun, H., Sun, X. and Yan, M. 2018. Automatic Water-Body Segmentation from High-Resolution Satellite Images via Deep Networks. *IEEE Geoscience and Remote Sensing Letters*. 15, 4 (2018), 602–606.

[10] Mikołajczyk, A. and Grochowski, M. 2018. Data augmentation for improving deep learning in image classification problem. In *Proceedings of the International Interdisciplinary PhD Workshop (IIPhDW)* (swinoujście, Poland, May 09 - 12, 2018), 117–122.

[11] Ren, S., He, K., Girshick, R. and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal

Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 39, 6 (2017), 1137–1149.

[12] Rujoiu-Mare, M.-R. and Mihai, B.-A. 2016. Mapping Land Cover Using Remote Sensing Data and GIS Techniques: A Case Study of Prahova Subcarpathians. *Procedia Environmental Sciences*. 32, (2016), 244–255.

[13] Storie, C.D. and Henry, C.J. 2018. Deep learning neural networks for land use land cover mapping. In *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS)* (Valencia, Spain, July 22-27, 2018), 3445–3448.

[14] Treitz, P. and Rogan, J. 2004. Remote sensing for mapping and monitoring land-cover and land-use change-an introduction. *Progress in Planning*. 61, 4 (2004), 269–279.

[15] Wen, Q., Jiang, K., Wang, W., Liu, Q., Guo, Q., Li, L. and Wang, P. 2019. Automatic Building Extraction from Google Earth Images under Complex Backgrounds Based on Deep Instance Segmentation Network. *Sensors*. 19, 2 (Jan. 2019), 333.