Object detection of aerial image using mask-region convolutional neural network (mask R-CNN)

Musyarofah¹, Valentina Schmidt², Martin Kada²

¹Remote Sensing Technology and Data Center, National Institute of Aeronautics and Space (LAPAN), 70 LAPAN Street, Pekayon, Pasar Rebo, 13710 Jakarta Timur, Indonesia

²Institute of Geodesy and Geoinformation Science, Technische Universität Berlin, Main Building Room H 5123, Sekretariat H 12, Straße des 17. Juni 135, 10623 Berlin, Germany

Email: musyarofah@lapan.go.id

Abstract. The most fundamental task in remote sensing data processing and analysis is object detection. It plays an important role in classification and very useful for various applications such as forestry, urban planning, agriculture, land use and land cover mapping, etc. However, it has many challenges to find an appropriate method due to many variations in the appearance of the object in image. The object may have occlusion, illumination, viewpoint variation, shadow, etc. Many object detection method has been researched and developed. Recently, the development of various machine learning-based methods for object detection has been increasing. Among of them are methods based on artificial neural network, deep learning and its derivatives. In this research, object detection method of aerial image by using mask-region convolutional neural network (mask-R CNN) is developed. The result shows that this method gives a significant accuracy by increasing the image training and epoch time.

1. Introduction
The use of remote sensing images has been increasing in recent years due to the advanced development of remote sensing technology that offers a better quality and higher resolution of image. The amount of availability of image data is going up, as well. Many various applications such as forestry, urban planning, agriculture, land use and land cover mapping, etc. can be applied by using remote sensing image. Therefore, the demand of remote sensing image data is also growing lately. In order to obtain a good result and get benefit from remote sensing image data, the data should be processed and analyzed carefully.

The most fundamental task in remote sensing data processing and analysis is object detection. It plays an important role in classification. The result and accuracy of object detection can influence the accuracy of image classification and gives a significant effect in image analysis. However, object detection is challenging because the object may has many variation of appearance in image caused by scale variation, occlusion, illumination condition, viewpoint variation, deformation, shadow, background clutter, etc [1].

In the last decades, many study and research have been conducted to find and develop an appropriate method to obtain higher accuracy and better result in object detection. Generally, there are four main categories in object detection method include template matching-based object detection methods,
knowledge-based object detection methods, object-based image analysis (OBIA)-based object methods, and machine learning-based object detection [2]. Nowadays, the development of various machine learning-based methods for object detection is growing. The techniques using machine learning are more advanced and powerful. Therefore, the technique becomes popular and widely used in many applications including some applications that use remote sensing image and analysis. Among of them are methods based on artificial neural network, deep learning and its derivatives.

In remote sensing community, artificial neural network or neural networks, which is the basis of deep learning (DL) algorithm, was popular in 1980s. And then some DL based methods like support vector machine (SVM), random forest (RF), etc. were proposed to do classification and change detection tasks. SVM has an ability to handle high dimensionality data and perform well with limited training samples [3], while RF is easy to use and can obtain high accuracy [4]. However, since 2014, DL as the derivative of neural networks has returned the interest of the remote sensing community to neural networks. In image analysis tasks including object detection, land use and land cover (LULC) classification, etc., DL algorithms give significant achievements [5]–[10].

The advanced development of graphical processing unit (GPU) increases computer performance in parallel processing that is beneficial for DL, especially convolutional neural network (CNN) that computes images parallelly in object recognition and object detection tasks. In fact, deep learning is robust, fit for complex problem and has an ability to solve higher computational task compared to another machine learning-based method [11]. CNN, one of DL fundamental network, uses convolution to learn higher-level features of the object from low-level feature composition in image data. It imitates how human brain works. CNN has succeed in object recognition, detection and classification tasks. Since 2012, DL always wins object detection and image classification competition every year. Based on the previous achievement, the DL framework can be applied in many application fields, particularly in object detection and classification task for 2D image and even 3D image [1], [12].

Mask-region convolutional neural network (mask R-CNN) is an object detection algorithm method based on CNN that has extra feature for instance segmentation and extra mask head. Mask R-CNN is an extension of Faster R-CNN, its predecessor, by adding an additional branch for predicting segmentation masks on each Region of Interest (RoI) [13]. In this research, we implement framework for object (roof) detection and extraction which applies Mask R-CNN model trained to detect and report instances of roof segments within aerial image.

**Neural Networks**

Neural networks are one type of machine learning model. Fundamentally, machine learning is using algorithms to extract information from raw data and represent it in some type of model [14] that can be used to infer things about other data. Neural networks or artificial neural networks (ANN) are a computational model used in machine learning both for classification or regression and are inspired by the function of the human brain. ANNs are composed by several layers of nodes that are connected by links with a weight attached to them. In the fully connected version of ANNs, every node is connected to all nodes in the layers behind and in front of it and to no node in its own layer, meaning nodes in one layer are completely independent of each other. It is a supervised learning algorithm where first the data is forward propagated in the network the error is being calculated and then back-propagated through the network while adjusting the weights [11].

**Deep Learning (DL)**

One definition says that deep learning is a neural network with more than two layers. The problematic aspect to this definition is that it sounds and makes deep learning created since the 1980s. The fact is DL comes several years after neural networks and makes many changes in neural networks architecturally in network styles. DL shows a spectacular result that transcends the earlier generation of neural networks. The evolution of neural networks in DL includes the following facets [14]:

- More neurons than previous networks
• More complex ways of connecting layers/neurons in neural networks
• Explosion in the amount of computing power available to train
• Automatic feature extraction

By the evolution of neural networks, DL can be defined as neural networks with a large number of parameters and layers. There are four fundamental network architectures in DL [14]:

• Unsupervised pre-trained networks
• Convolutional neural networks (CNN)
• Recurrent neural networks (RNN)
• Recursive neural networks

One of the great advantages of DL compared to another traditional machine learning algorithms is automatic feature extraction. By feature extraction, the networks process of deciding which characteristics of a dataset can be used as indicator to label that data reliably [14]. DL had reduced human effort in feature extraction which is an input of classification. Therefore, the accuracy gained by DL is higher than another conventional machine learning algorithms for almost every data type with minimal tuning and human role.

Convolutional Neural Networks (CNN)
Convolutional neural networks (CNN) are one of fundamental network architectures of deep learning. CNNs are a specialized kind of neural network for processing data that has a known, grid-like topology [15]. The network uses a mathematical operation, convolution, which the name of CNN comes from. Therefore, CNNs can be defined as neural networks that employ convolution in place of general matrix multiplication in at least one of their layers. A convolution is a powerful concept for helping to build a more robust feature space based on a signal. So, by using convolution, CNNs obtain the goal to learn higher-order features in the data.

Since CNNs perform well in learning feature of the data, then CNNs are suitable for object recognition, object detection, and classification. In fact, the successful of CNNs in image recognition makes the power of deep learning is recognized. CNNs can identify faces, individuals, street signs, platypuses, and many other aspects of visual data and CNNs are good at building position and rotation invariant features from raw image data, as well [14]. CNNs are very advantageous for the input that has structure, repeating patterns and spatially distributed value like images and audio data. Besides that, CNNs also have been used in natural language translation and sentiment analysis.

CNNs transform the input data from the input layer through all connected layers into a set of class scores given by the output layer [14]. There are many variations of the CNN architecture, but they are based on the pattern of layers, as demonstrated in the following figure.

![Figure 1. High-level general CNN architecture](image-url)
The figure depicts the high-level general CNN architecture that consists of three major groups of layer, namely input layer, feature-extraction (learning) layers and classification layers. Generally, in object detection the input is an image that has spatial information and depth representing the colour channels. Feature extraction layers have convolution layer, activation function layer and pooling layer. In Figure 2, the activation function layer is represented by ReLU as an example that widely used. These three layers find a number of features in the images and progressively construct higher-order features [14]. After extracting feature the next step is classification in classification layer that produces class probabilities or scores.

**Convolution**

A convolution is defined as a mathematical operation describing a rule for how to merge two sets of information. It is important in both physics and mathematics, defines a bridge between the space (time) domain and the frequency domain through the use of Fourier transforms. It takes input, applies a convolution kernel, and gives us a feature map as output. The convolution operation is known as the feature detector of a CNN. The input to a convolution can be raw data or a feature map output from another convolution. It is often interpreted as a filter in which the kernel filters input data for certain kinds of information.

![Figure 2. The convolution operation Source: [14]](image)

Figure 2 illustrates the convolution operation. The input data is convolved by the kernel filter to obtain the convoluted feature.

**Mask Region Convolutional Neural Networks (Mask R-CNN)**

Mask region convolutional neural networks (Mask R-CNN) is a general framework that conceptually simple and flexible for object detection and object instance segmentation. Mask R-CNN efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance [13]. The method of Mask R-CNN is an extension of earlier method, Faster Region Convolutional Neural Network (Faster R-CNN). In Mask R-CNN, a third branch is added for predicting segmentation masks on each Region of Interest (RoI), in parallel with the two existing branches for classification and bounding box regression. The mask branch which is the third branch, is a small fully convolutional network (FCN) applied to each RoI, predicting a segmentation mask in a pixel-to-pixel manner. Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN [16].
2. Methodology

Figure 3 shows the mask R-CNN framework that is used in our implementation. Mask R-CNN adopts the method from Faster R-CNN, which consists of two stages. The first stage, called a Region Proposal Network (RPN), proposes candidate object bounding boxes. The second stage is extracting features using RoIPool from each candidate box and performs classification and bounding box regression. Mask R-CNN is a two stage framework. The first stage scans the image and generates proposals (areas likely to contain an object). And the second stage classifies the proposals and generates bounding boxes and masks.

**Backbone CNN**

Backbone is a standard convolutional neural network that serves as a feature extractor. The early layers detect low level features (edges and corners), and later layers successively detect higher level features (object like car, person, sky, etc.). In the original paper of Mask R-CNN, the authors use ResNet for the backbone architecture. We also use ResNet in this research. The process in backbone network started with converting of RGB input image into a feature map of shape when it passes through the backbone network. This feature map will be the input for the next step.

**Region Proposal Network (RPN)**

The RPN is a light-weight neural network that scans the image in a sliding-window fashion and find areas that might contain objects. The regions scanned by RPN are called anchors. Anchors are used to detect multiple objects with different scales and overlapping objects in image. Anchors are boxes with a set of predefined bounding boxes of a certain height and width, so they can capture specific object...
classes with different sizes and aspect ratios. Anchor boxes are distributed over the image area and might be overlap each other to cover as much of image as possible.

The convolutional nature of the RPN handles the sliding-window, so RPN can scan all regions in parallel (on a GPU). RPN scans over the backbone feature map, not over the image directly. Therefore, RPN can reuse the extracted features efficiently and avoid duplicate calculations. The outputs for each anchor that generated by RPN are anchor class and bounding box refinement.

**Proposed Region of Interest (RoI)**

![Proposed Region of Interest](image)

Figure 4. Illustration of anchor boxes in image

Figure 4 shows the illustration of anchor boxes within the image. Anchors are the regions that the RPN scans over, which are boxes which are distributed over the image area. Proposed Regions of Interest (RoI) are generated by RPN which are the regions that contain an object of the class to classify. As illustrated in Figure 4, the proposed RoI contains objects, in this case are the roofs, that bounded by boxes. There are two types of box in each object, the one is formed by connected lines and the other is by dotted line. The box with dotted lines is the anchor box and the box with solid lines is its refinement.

**RoI Classifier and Bounding Box Regressor**

The output from RPN becomes an input for the next step, RoI classifier and bounding box regressor. This step runs on the RoI proposed by RPN and generates two outputs for each RoI, the class of the object in the RoI and bounding box refinement to refine the location and size of the bounding box to encapsulate object. The class that generated by this step is more specific than the class in RPN. In RPN the classes are only foreground and background. In this step, the network is deeper so it can classify a
region to specific classes like person, car, etc. It also generated background class that makes the RoI to be discarded. The bounding box refinement in this stage is similar to the previous.

**Segmentation Masks**

The mask network is the addition that the Mask R-CNN paper introduced, by extending Fast/Faster CNN method. The mask branch is also a convolutional network that takes the positive regions selected by the ROI classifier and generates masks for them. The generated masks are low resolution. But they are *soft* masks, represented by float numbers, so they hold more details than binary masks. The small mask size helps keep the mask branch light. The scaling down of the ground-truth masks is conducted during training to compute the loss. And during inferencing, the predicted masks are scaled up to the size of RoI bounding box.

**3. Results and Discussion**

The data used in the experiment is an aerial image of Everswinkel city, Germany. First, we made a dataset that consists of training set, validation set and testing set from the data. The dataset generation was conducted by using Fishnet tool in ArcGIS.

![Figure 5. Illustration of fishnet in dataset generation](image)

From Figure 5 we can see that the data is divided into several clip images. Every object (roof) in each clip image will be annotated, digitized polygon with class labels of roof (gable, hip and flat). From image that has been annotated, a training set, validation set and testing set were generated with proportion (7:2:1). And then the next step is checking the dataset to see that data has been successfully read according to the masking of each roof types. After all the dataset were checked, the next step is training or learning process by using training set data. The model of mask R-CNN framework that we build in this research has to learn every roof type of the classes from training set. After training process, the next step is validation to tune all hyperparameters in the model. After all hyperparameters were tuned, then we conducted the testing to see if the model is able to detect and segment the object (roof) in image.
Figure 6 shows the result of the experiment conducted in our research. From the result shown in Figure 6, it can be concluded that the network implemented, Mask R-CNN, is able to detect an object in the image (roof top of building) with a good accuracy.

In order to measure the performance of classification and object detection model, the metric of measurements are needed to evaluate the model. There are evaluation metrics that show how well a model is doing in terms of real world performance. By these evaluation metrics, the quality of a model can be revealed and compared to different models on the same tasks. Average precision (AP) is a popular metric in measuring the accuracy of object detectors like Faster R-CNN, Mask-R CNN, etc. Mean average precision (mAP) is the average of AP. In some context, the AP is computed for each class and averages them. But in some context, they mean the same thing. IoU measure the overlap between 2 boundaries, that is used to measure how much the predicted boundary overlap with the ground truth (the real object boundary). IoU threshold can be predefined in some dataset. Precision measures how accurate is your prediction (the percentage of accurate prediction). Recall measures how good you find all the positives. IoU measures the overlap between 2 boundaries.

\[
\text{precision} = \frac{TP}{TP + FP} \\
\text{recall} = \frac{TP}{TP + FN}
\]

From the experiment in our research, we obtained a good performance of the model we build in term of accuracy of the object detection. The mAP resulted is 0.9014186807449848 (mAP @ IoU=0.50). And another metrics are the precisions and recalls with value of 0.8871293140410786 and 0.530022520443246 consecutively.
Figure 7 shows the result of object (roof) segmentation by using Mask R-CNN model we build. From the result we can see that the segmentation mask does not really fit on the object (roof). Therefore, improvement of the model by changing some parameters and adding more architectural variations should be conducted to make the segmentation better and increasing the accuracy. In domain of architecture used in instance segmentation task, a bigger version of ResNet or another Feature Pyramid Network (FPN) approach should be tried and explored to improve the object detection and segmentation result. The different RPN architecture also should be explored more.

Adding more features, changing the training strategy and increasing the amount of training set data can also improve the result. Another issue that can improve the result is environmental conditions. It means that we should consider environmental conditions in image data such downtowns with closely standing high-rises, three shaded buildings, waterbodies nearby, etc. The environmental conditions might affect the detection and segmentation step during training and testing.

The term of balanced dataset should also be considered to reduce the domination of certain task that sometime influence to missclassification. Class imbalance is another issue with real-world training data. The number of flat roof buildings compared to another roof type should be considered, since it will lead to a problem for training a neural network classifier.

4. Conclusion
Mask-R CNN can be implemented in object (roof) detection of aerial image with a good accuracy (mAP 90.14% & precision 88.71%). RGB (spectral of image) can be used as a feature. The result of segmentation mask does not really fit on the object (roof). By using Mask-R CNN, the object is not only detected, but we can also know which pixels are belonged to object.
5. References

[1] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, “Object Detection with Deep Learning: A Review,” IEEE Transactions on Neural Networks and Learning Systems. 2019, doi: 10.1109/TNNLS.2018.2876865.

[2] G. Cheng and J. Han, “A survey on object detection in optical remote sensing images,” ISPRS Journal of Photogrammetry and Remote Sensing. 2016, doi: 10.1016/j.isprsjprs.2016.03.014.

[3] F. Melgani and L. Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” IEEE Trans. Geosci. Remote Sens., 2004, doi: 10.1109/TGRS.2004.831865.

[4] M. Belgiu and L. Drăgu, “Random forest in remote sensing: A review of applications and future directions,” ISPRS Journal of Photogrammetry and Remote Sensing. 2016, doi: 10.1016/j.isprsjprs.2016.01.011.

[5] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, “Deep learning in remote sensing applications: A meta-analysis and review,” ISPRS Journal of Photogrammetry and Remote Sensing. 2019, doi: 10.1016/j.isprsjprs.2019.04.015.

[6] X. X. Zhu et al., “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources,” IEEE Geoscience and Remote Sensing Magazine. 2017, doi: 10.1109/MGRS.2017.2762307.

[7] E. Maggiori, Y. Tarabalka, G. Charpiait, and P. Alliez, “Fully convolutional neural networks for remote sensing image classification,” in International Geoscience and Remote Sensing Symposium (IGARSS), 2016, doi: 10.1109/IGARSS.2016.7730322.

[8] B. Huang, B. Zhao, and Y. Song, “Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery,” Remote Sens. Environ., 2018, doi: 10.1016/j.rse.2018.04.050.

[9] M. Kampffmeyer, A. B. Salberg, and R. Jenssen, “Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2016, doi: 10.1109/CVPRW.2016.90.

[10] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, “Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data,” IEEE Geosci. Remote Sens. Lett., 2017, doi: 10.1109/LGRS.2017.2681128.

[11] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature. 2015, doi: 10.1038/nature14539.

[12] L. Liu et al., “Deep Learning for Generic Object Detection: A Survey,” Int. J. Comput. Vis., 2019, doi: 10.1007/s11263-019-01247-4.

[13] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, doi: 10.1109/ICCV.2017.322.

[14] J. Patterson and A. Gibson, Deep learning: A Pracionar Approach. 2017.

[15] A. C. Ian Goodfellow, Yoshua Bengio, “Deep Learning Book,” Deep Learn., 2015, doi: 10.1016/B978-0-12-391420-0.09987-X.

[16] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.