GENERAL DEEP LEARNING SEGMENTATION PROCESS USED IN REMOTE SENSING IMAGES

Hsuan-Chung Wang[1],

[1] Thinktron ltd., Department of Research and development, 10087 Taipei, Taiwan - jeremywang@thinktronltd.com

Commission II, WG II/III

KEY WORDS: Remote Sensing, Artificial Intelligence, Deep Learning, Segmentation, Satellite, Machine Learning

ABSTRACT:

In the present research, we aim at constructing a general segmentation process for different kinds of remote sensing images and various use cases. We focus on the differences in characteristics of the remote sensing and ordinary images, such as irregular shape, lack of labeled images, and normalization issues. The process includes labeling, preprocessing, augmentation, test data sampling, model building, as well as prediction and merging steps. Labeling serves to identify target objects represented in remote sensing images efficiently. The preprocessing step can be applied to reshape an image aiming to fit the requirements of the general artificial intelligence (AI) model and to accelerate steps. Augmentation mitigates the shortage of labeled images. Test data sampling is performed to evaluate the model performance. Finally, prediction and merging are applied to output a full-sized remote sensing image prediction result. In this research, the landslide segmentation, crop farmland segmentation, and cloud segmentation tasks are considered to evaluate the process. Intersection of union (IOU) is employed as evaluation metric. Eventually, we achieve the performance of 72% IOU in the landslide segmentation task, 83% IOU in the crop farmland recognition task, and the 86% IOU in cloud segmentation task by using the proposed process. This supports that the developed process can by further applied considering different remote sensing images and use cases.

INTRODUCTION

Deep learning techniques have been widely employed in the field of remote sensing in recent years. They have been applied considering various practical use cases, such as agriculture, smart cities, forest management, as well as surveying and mapping. However, only a limited number of the related papers focused on defining a general segmentation process for different kinds of remote sensing images and various use cases. In the present study, we aim to construct a generalized segmentation process.

Remote sensing images have their own characteristics different from ordinary images, which hinder the possibility of directly applying deep learning algorithms in this field. First, remote sensing images usually have specific shapes. Particularly, ordinary images have only RGB bands (also referred to as channels in the deep learning field), while remote sensing images have multiple bands, and the numbers may be flexible. Ordinary images may have the square of hundreds of pixels, while remote sensing images may have the square of tens of thousands of pixels. As a result, remote sensing images should be first preprocessed to fit a particular shape, and the developed artificial intelligence (AI) models should be modified accordingly to be applicable to the images with flexible bands.

Second, the number of labeled remote sensing images is generally rather small as a consequence of the considerable difficulties and costs associated with the labeling task. Therefore, identifying the way to generate a sufficient number of training images is also an important issue. Third, remote sensing images are not normalized to 0 to 255 in general. Therefore, a model can be used only to predict the images already seen by a model instead of being capable of predicting new unknown images, unless the number of images for training is sufficient.

In the present study, we aim to design a general process to deal with all these issue and to successfully apply a deep learning segmentation algorithm to different remote sensing images and various use cases.

1. RELATED WORKS

Zhu et al., (2017) and Ma et al., (2019) have summarized main deep learning applications in the remote sensing field, such as image preprocessing, image registration, object detection, and segmentation. Fully connected network (FCN) model, as explained in Section 3.5.2, has been utilized in their reviewed paper for the purpose of segmentation. Li et al., (2018) have investigated three kinds of tasks in remote sensing image classification: spectral feature classification, spatial feature classification, and spectral-spatial one. These related papers have suggested utilizing convolutional neural networks (CNN) and autoencoder for spectral-spatial classification, which is the target of the segmentation process in the present research. The aforementioned research works have analyzed a large number of related papers and have provided a comprehensive summary in this field.

Kemker et al., (1860) have proposed a new way to generate the sufficient amount of training data. The have considered constructing a virtual scene using the digital image and remote sensing image generation (DIRSIG) method so as to produce a sufficient number of synthetic images as training data. Foivos et al., (2019) have applied U-Net (explained in Section 3.5.1) with a residual block to improve the model performance in the case of land cover classification. Stoian et al., (2019) have applied U-Net combined with a recurrent neural network (RNN) to perform multitemporal satellite images classification. These papers have been presented to address different specific prob-

lems arising while applying deep learning to remote sensing images.

2. RESEARCH METHOD

In the present research, to evaluate the proposed general segmentation process, we consider three cases with different kinds of remote sensing images: landslide segmentation, crop farmland segmentation, and cloud segmentation.

Landslide segmentation task is aimed to detect landslide areas after disasters, such as earthquakes, torrential rains, and typhoons. In this task, 28 UAV images and 348 landslide polygons have been provided by Chinese society of photogrammetry and remote sensing (CSPRS). The data are registered in a region located in Liugui Kaohsiung in the south of Taiwan. There are 12,261 columns, 11,461 rows, and 4 bands corresponding to each image with the cell size of 0.25 meters. The dates when the photos were taken are listed in Table 1. The total area corresponding to all considered polygons covers 6,224,063.52 m^2. The labeled polygons have been generated by CSPRS, and all of them have been labeled by human experts.

Date	Count
03 Apr 2017	10
04 Apr 2017	6
13 Dec 2015	8
19 Dec 2015	3
26 Jun 2016	1

Table 1. Images used in the landslide segmentation task.

The crop farmland segmentation task helps Taiwan Agriculture and Food Agency council (AFA) to estimate the total area of all crop farmlands in Yilan. They seek to employ the information about the labeled farmlands in the sampled area combined with satellite images to recognize the crop farmlands not in the sampled area. Using this technique, they can do the local investigation only on the sample area and identify crop farmlands in the whole Yilan Taiwan. In this task, four Sentinel-2 satellite images and 87,900 labeled farmlands polygons are considered. There are 12,261 columns, 11,461 rows, and 4 bands associated with each image with the cell size of 10 meters. The photographs are taken on 27 Mar 2019, 6 Apr 2019, 26 Apr 2019, and 5 Jun 2019. The high quality labeled data have been provided by CSPRS, including all farmlands in the Yilan Taiwan. CSPRS uses the high-resolution UAV images, and human experts are involved to label the data.

The cloud segmentation task is an essential preprocessing step before applying any kinds of analysis to different remote sensing images, unless a project implies excluding the images with clouds from the training data. In this task, one SPOT5 and one SPOT6 images with the resolution of 10 meters are used, and the target data have been labeled by Thinktron Ltd. The detailed information is provided in Table 2.

ID	Columns	Rows	Bands	Resolution
P0015913_SP5	10,400	45,766	4	10
P0016267_SP6	13,400	39,200	4	10

Table 2. Image information used in the cloud segmentation task.

The evaluation function utilized in this research is intersection of union (IOU) that is one of the most popular evaluation functions used in segmentation tasks. IOU means the intersection of union, implying that the intersection part is used to calculate the intersection of a predicted positive area and a labeled positive area, while the union part is considered to calculate the union of a predicted positive area and a labeled positive area. Therefore, the intersection of union represents the overlapping rate of the labeled positive and predicted positive areas, as illustrated in Figure 1.

Finally, to evaluate the performance of the proposed method, we apply the evaluation function to the testing data sampled in the whole dataset, considering each of three tasks.

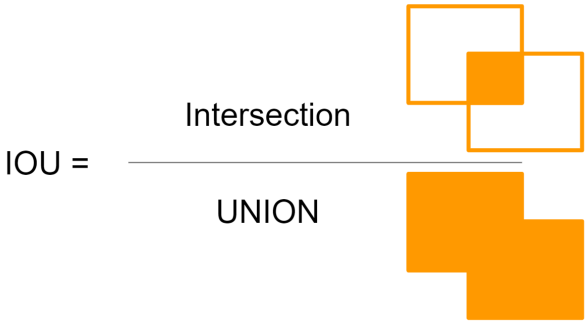


Figure 1. Intersection of union (IOU).

3. MAIN PROCESS

Figure 2 represents the main segmentation process including labeling, preprocessing, augmentation, test data sampling, model building, as well as prediction and merging steps. All these steps are described in detail in the following paragraph.

3.1 Labeling

The accurately labeled data is the most valuable asset to achieve better performance in the AI field, as a model can only learn from these data. Therefore, the sufficient number of labeled images is an essential prerequisite to achieve success in any AI project. Nevertheless, labeling irregular polygons on remote sensing images is rather a difficult task. The following clustering and unsupervised segmentation algorithms can be used to accelerate the process of image labeling. The comparison between them is illustrated in Figure 3. It should be noted that to obtain the ground truth data, it is better not to use the outputs of aforementioned algorithms directly as the labeled data, but apply them as a tool to accelerate the selection process by human experts.

3.1.1 Clustering Clustering algorithms, such as k-means and density-based scan (DBSCAN), can be used to classify each pixel in an image into independent clusters depending on its color digital value. However, such algorithms do not consider the spatial relationship between pixels. Therefore, the pixels far from each other may be classified into the same cluster. Image (b) in Figure 3 represents the output of the k-means algorithm. The areas with the same color are classified as belonging to the same cluster. As a result, if the target has a special color that is different from other pixels, such as a cloud or a landslide, the algorithm may work accurately and usually allows efficiently labeling the target data.

3.1.2 Unsupervised Segmentation Unsupervised segmentation algorithms, such as watershed segmentation and superpixel segmentation, can be used to classify each pixel in an image into independent clusters. The algorithms considers both a

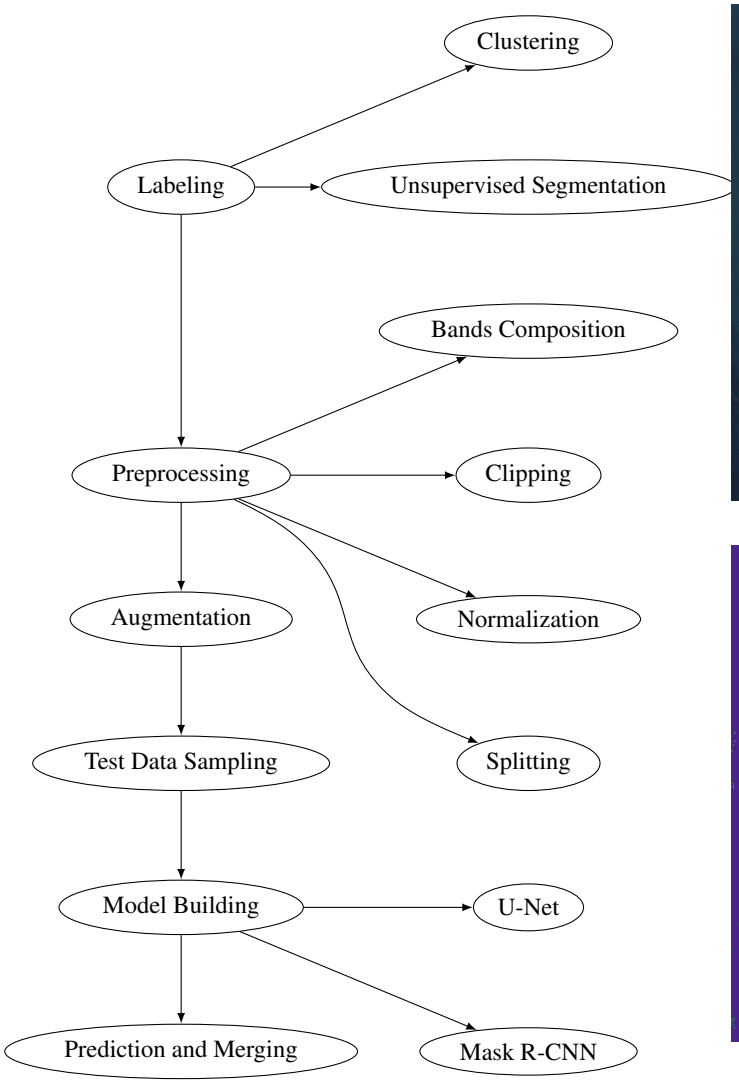
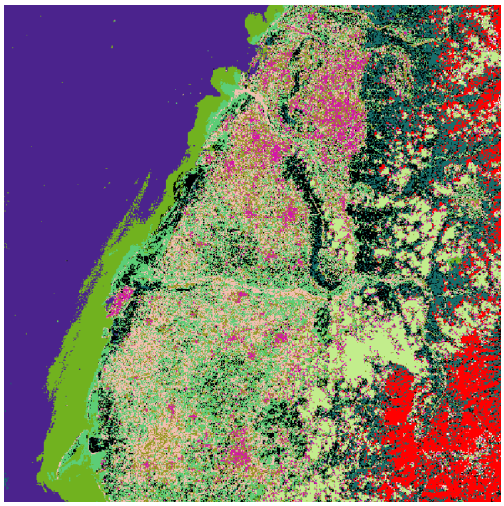
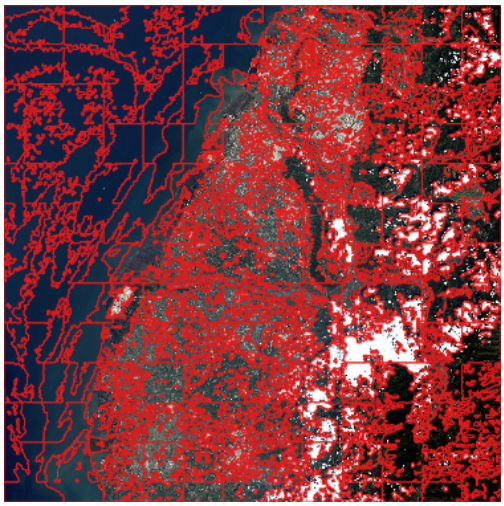


Figure 2. Main Process.

(a) raw image.



(b) k-means clustering.



(c) superpixel segmentation.

Figure 3. Labeling techniques.

color digital value and a spatial index of a pixel. In other words, only neighbor pixels are classified into the same cluster. Image (c) in Figure 3 represents the output of the superpixel segmentation algorithm. One polygon denotes an independent cluster so that a researcher can label the data by selecting the target polygon rather than drawing a polygon by themselves from scratch.

3.2 Preprocessing

In the preprocessing step, the main purpose is to transform an image to fit the requirements of an AI model and to accelerate the execution of the subsequent steps. The following four steps: band composition, clipping, normalization, and splitting, are explained in this section.

3.2.1 Band Composition Combining all bands of a satellite image can be used to transform shapes of training images to those acceptable to be inputted into a model. In general, source satellite images are split into independent images according to their bands (red, green, and near infrared ones). However, it is required to consider them as a single image to train a deep learning model. Therefore, CNN (the most important computer vision algorithm in the deep learning field that is explained in Section 3.5) can be used to capture the spatial relationship between neighbor pixels across multiple bands in parallel.

3.2.2 Clipping Clipping an image according to an area in the labeled data and a source image can be used to enable a model to learn from the accurate information and to reduce the computational resources. In particular cases, the images used

for labeling are different from those used to train the model, and they may not be fully overlapping. Once non-overlapping areas are included into the training set, the model will learn based on the incorrect information. In other words, if the target exists in the area that is not covered by the labeled data, the model will learn this pattern as a non-target one. As a result, the image outside of the overlapping area should be clipped. This step should be done earlier considering the computing efficiency and the time cost of the following steps. For example, the plain area is deemed as the only valuable area in the farmland recognition task, and therefore, the mountain areas should be clipped, as illustrated in Figure 4.

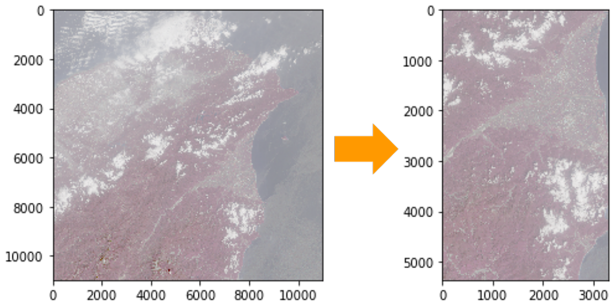


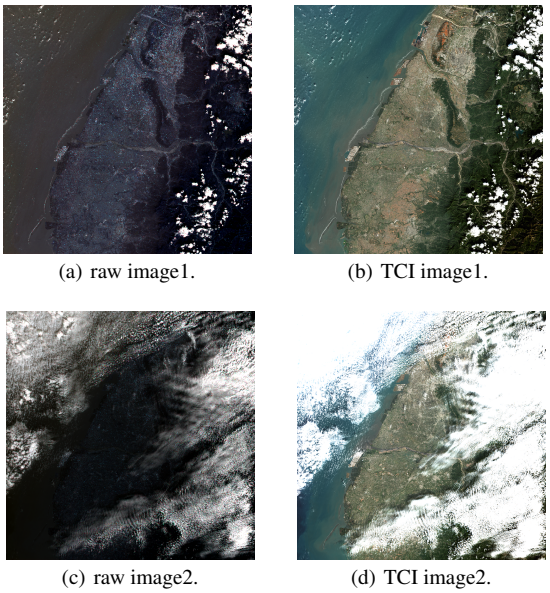
Figure 4. Clipping.

3.2.3 Normalization An appropriate normalization method can facilitate improving the model generalization ability, and therefore, the model will achieve better performance on unknown images. It is recommend to use the true color images (TCI) that have pixel value range from 0 to 255 for RGB bands as the training data. The difference in the color digital values between different TCIs are more comparable. Figure 5 represents TCI provided by European Space Agency (ESA) that have been registered by the Sentinel-2 2A product. It is evident that TCIs on the right side are clearer and can be compared with each other, while raw images on the left side may cause difficulties to perform this task.

However, TCIs only work in RGB bands rather than all bands. Other bands can only be calibrated to ground reflectance. However, the comparability is still questionable. It should be noted that this process cannot ensure the perfect generalization ability even if TCIs are used. Using a large number of labeled images with the great extent of variety is always recommend.

3.2.4 Splitting Splitting is used to generate a large amount of training data. In general, remote sensing images are difficult to obtain, and therefore, the number of such images is limited. Moreover, the difficulties associated with the labeling task also decrease the number of training images. However, even if the number of such images is limited, they are usually high-quality ones. Remote sensing images have generally rather high resolution. Unlike ordinary photographic images with a small number of pixels, remote sensing images have the square of tens of thousands of pixels. As a result, splitting an image into pixels and generating the larger amount of the training data is the best strategy to fit the requirements of AI models.

The way to split an image can be considered as capturing a moving window. A window is put in the left-top of an image, and then, is gradually shifted right and down to capture every area in the image. The window size defines the size of split images, while the step size sets the overlapping rate between them. The size of split images depends on the size of a target.



(a) raw image1. (b) TCI image1.

(c) raw image2. (d) TCI image2.

Figure 5. TCI image comparison.

For example, if the target is a house footprint, it is better not to limit the number of houses appearing in an image to five or ten houses. However, it is recommended to set the size as power of 2, for example, 128 or 256, to conform the limitation of particular AI models on the input size. Figure 6 represents the example image with the resolution of 10 meters and $10,980^2$ pixels split into pieces with the window size and the step size equal to 256 pixels so that windows do not overlap with each other.

The step size depends on the total number of training images. Considering that only the limited number of training images can be accessed, it is necessary to allow split images to overlap with each other aiming to ensure obtaining the sufficient number of training images. However, the overlapping rate may affect the performance and the efficiency of the training process. It should be noted that the optimal window size and step size may be different depending on a particular project, and therefore, they should be set indivdually.

3.3 Augmentation

Augmentation is also used to generate more training data. In theory, the target object in a remote sensing image should not change when the image is rotated or flipped, and therefore, it is possible to simultaneously rotate the image and to label it as a new one to generate the larger number of training images. Figure 7 represents an example corresponding to the way of generating a dataset six times larger than the training one (by rotating 90 degrees, 180 degrees, 270 degrees, flipping up and down, and flipping left and right).

3.4 Test Data Sampling

Test data sampling, also referred to as train test splitting, can be used to evaluate the final model performance accurately. It is rather difficult for humans to understand and inference the trained weights for input variables in an AI model. Therefore, the only way to evaluate the applicability of an AI model is to evaluate the performance of its final prediction result. Generally, it is a common approach to split the whole dataset into the training and testing dataset randomly. However, once split images are randomly selected as the training and testing data so

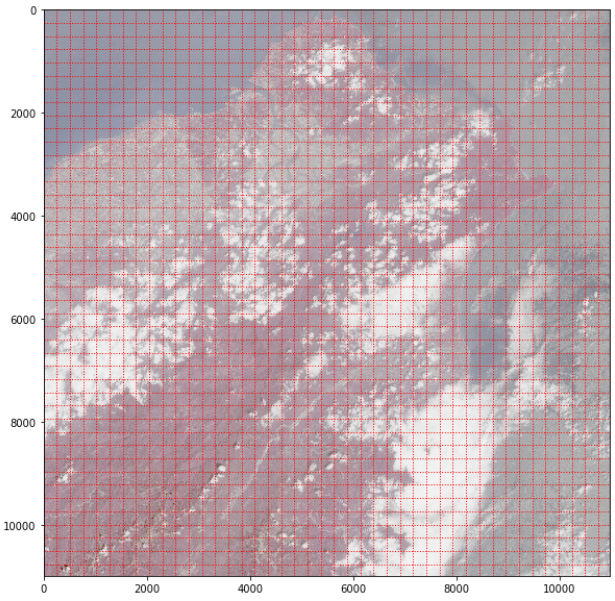


Figure 6. Splitting.

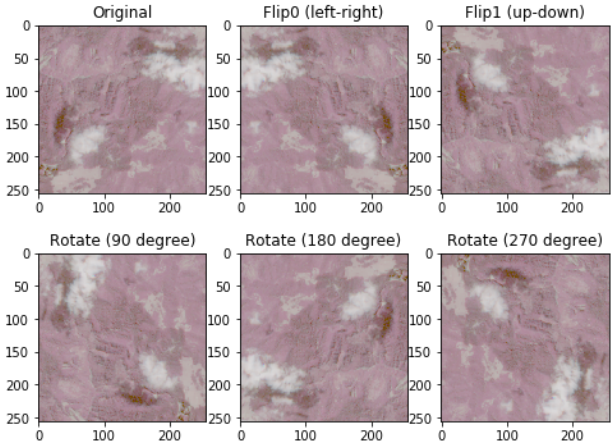


Figure 7. Augmentation.

that they overlap with each other, testing images may also be seen by the model, and accordingly, the evaluation result may be the overfitting result.

The appropriate way to perform test data sampling is to sample particular areas in a remote sensing image as the test data and ensure that these data will not be seen by the model. Moreover, the sampled test areas should include all kinds of characteristics in the image that may influence the final recognition output. For example, Figure 8 illustrates the sampled area in an image used for crop farmland recognition. Considering that the characteristics of spring onion farmlands are rather similar to crop farmlands on satellite images, the west part of the image is sampled as the test area, thereby assisting the model to learn that the split images in the area may not correspond to crop farmlands even if they have similar characteristics.

3.5 Model Building

In the AI field, a convolutional neural network (CNN) is one of the most popular and basic algorithms that has been proposed by Lecun et al., (1998) to deal with the image data. Compared with conventional pixel-based prediction approaches, CNN can

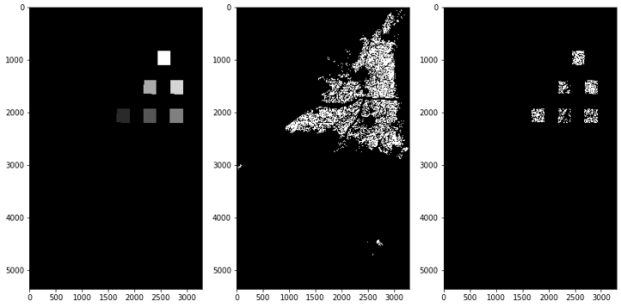


Figure 8. Sampling.

be used to incorporate neighbor pixels while performing the final prediction. Figure 9 provided in Lecun et al., (1998) represents the structure of a CNN model. The left side of the figure illustrates the input of the model, which is used to extract the detail textures of the original image. When later convolution layers are applied, the wider characteristics can be calculated. Then, the final prediction will consider all neighbor pixels and obtain a considerably better performance. There are two major CNN-based segmentation models defined in the AI field: U-Net and mask R-CNN that are introduced in the following para-graph.

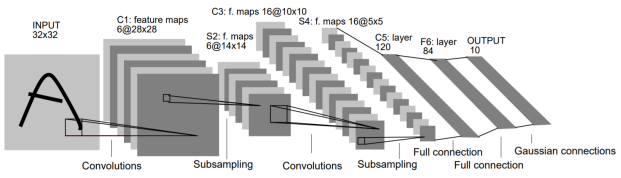


Figure 9. Convolutional neural network (CNN).

3.5.1 U-Net The architecture of the U-Net model, as illustrated in Figure 10, is proposed by Ronneberger et al., (2015). It is based on the fully convolutional network (FCN), which is one of the most basic segmentation algorithms in the AI field. The network includes the extracting and expanding paths, thereby forming the U-shaped architecture. The extracting path is used to derive and concentrate the important features of an image into representative image vectors with smaller dimensions. In its turn, the expanding path is utilized to perform predictions on each pixel based on the representative image vectors. Practically, U-Net is usually employed for irregular and continuous shapes, such as clouds and landslide.

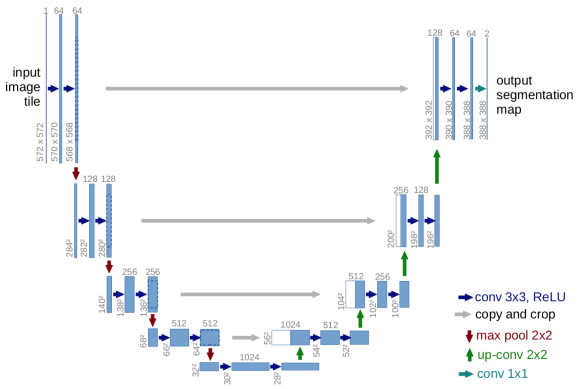


Figure 10. U-Net structure.

3.5.2 Mask R-CNN The architecture of the mask R-CNN model is proposed by He et al., (2008). Unlike U-Net, mask R-CNN first executes faster R-CNN object detection and obtains a bounding box for each object. Then, FCN segmentation is performed within each bounding box. Faster R-CNN object detection outputs a bounding-box proposal, and for each box, it decides on the possibility whether the box contains the object. This step is followed by FCN segmentation applied to a bounding box. Practically, mask R-CNN is deemed suitable for regular and artificial shapes, such as house footprint.

3.6 Prediction and Merging

Once the AI model is trained, it can be used to inference on the unseen data. The problem associated with the inference step in the field of remote sensing is that the results of predicting on split images cannot be combined directly into an original-sized image. Otherwise, the edge of a split image may have defects as it would lack neighbor pixels for the model to inference. Figure 11 represents the cloud detection result (right) and the raw image (left). The thin black line denotes the defect appears a result of an attempt to directly combine the outcomes of predicting on the split images.

As a result, the appropriate way to merge the prediction results is to first split them according to the original window size and with a smaller step size. This allows ensuring that every pixel is overlapped. Then, the edge of the prediction result should be cut by 2 to 3 pixels. Finally, we calculate the mean value of the prediction result for each pixel in an image as its final prediction result.

4. RESULTS

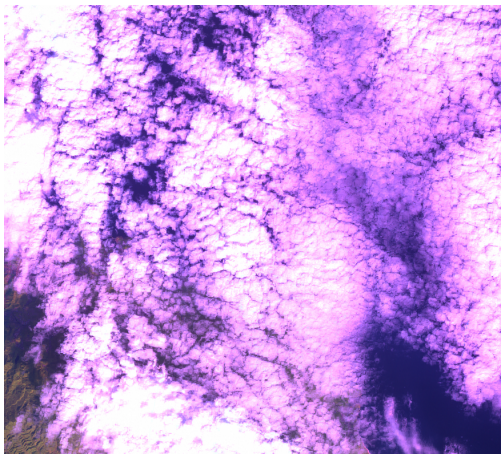
In this research, three cases are considered to evaluate the proposed general segmentation process: landslide segmentation task, crop farmland segmentation task, and cloud segmentation task. In this section, the ways employed to perform test data sampling and to build models in these three tasks are explained. Moreover, the evaluation results are provided.

4.1 Landslide Segmentation Task

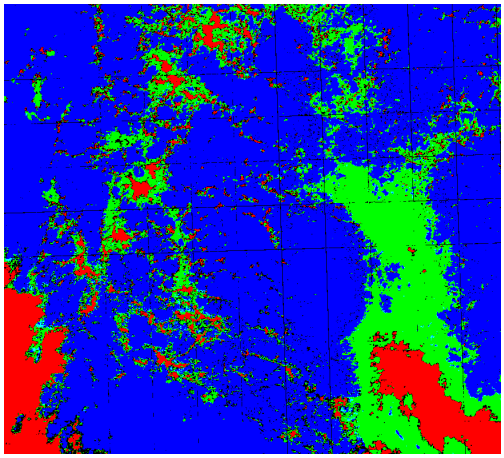
In the landslide segmentation task, 28 UAV images are split into 20 training images and 8 testing ones, as demonstrated in Figure 12. The blue bounding boxes represent the training images, while the red boxes correspond to the testing images.

The U-net model with 5×2 layers is constructed in this task. The model architecture consists of five down-sampling layers and five up-sampling ones. The input and output image shapes corresponding to each layer are 512, 128, 32, 16, 8, 16, 32, 128, and 512.

In this task, IOU of 72% is reached on the testing images. Figure 13 shows the segmentation results for the two out of eight testing images. The black areas are labeled as the landslide ones, while the red polygons denote the model predicted landslide areas. The main error in this case is that the model misrecognizes the river as the landslide, as they have rather similar color.



(a) Raw image.



(b) Prediction.

Figure 11. Merged prediction error.

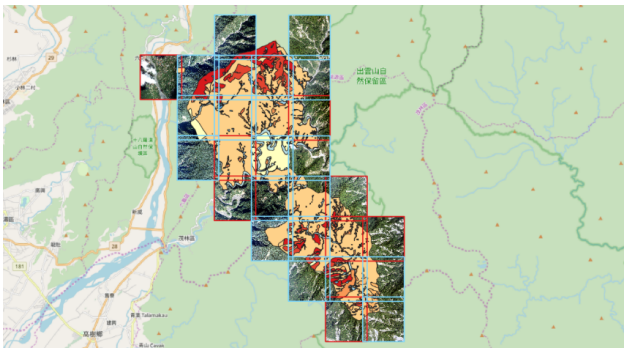
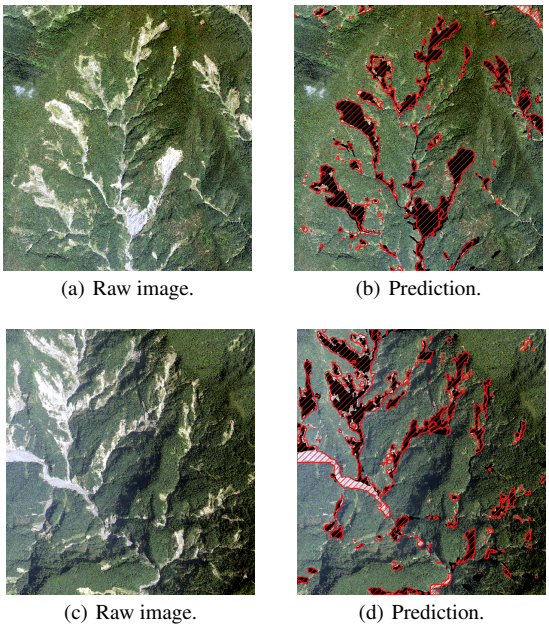


Figure 12. Landslide test data sampling.

4.2 Crop Farmland Segmentation Task

In the crop farmland segmentation task, four satellite images are registered during different vegetation periods but in the same cultivation period. The, they are combined into a single image with 16 bands. The six areas in the red bounding boxes, as represented in Figure 14, are sampled as the training data, and the remaining area of an image is considered as the testing data.

The U-net model with 5×2 layers is implemented in this task. The model architecture comprises five down-sampling layers and five up-sampling ones. The input and output image shapes of each layer are 64, 48, 36, 24, 12, 24, 36, 48, and 64.



(a) Raw image. (b) Prediction.

(c) Raw image. (d) Prediction.

Figure 13. Landslide segmentation prediction results.

Figure 14. Crop farmland test data sampling.

In this task, IOU of 83% is achieved in the testing area. Figure 15 outlines the crop farmland segmentation result in the testing area. The green areas are labeled as crop farmlands, while the red polygons correspond to the model predicted crop farmlands.

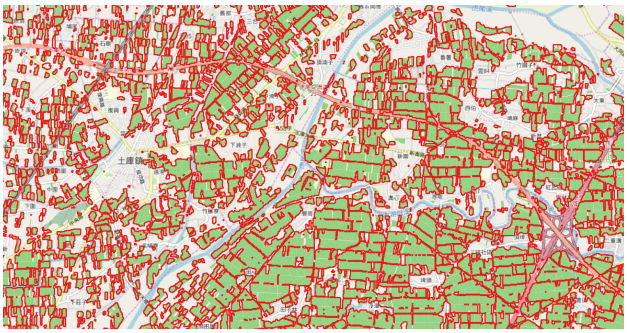


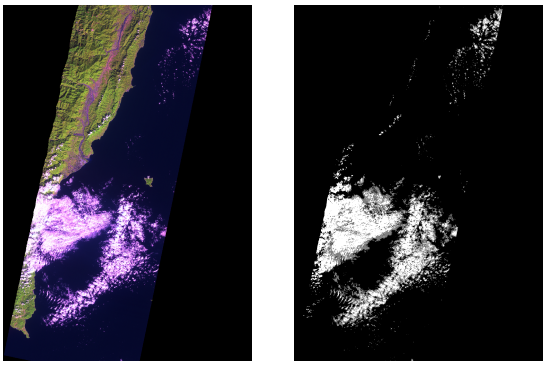
Figure 15. Crop farmland segmentation prediction result.

4.3 Cloud Segmentation Task

In the cloud segmentation task, two satellite images are used. Both of them are split in half so that one of the halves is used as a training area, and the other as a testing one.

The U-net model with 5×2 layers is built in this task. The model architecture includes five down-sampling layers and five up-sampling layers. The input and output image shapes of each layer are 101, 50, 25, 12, 6, 12, 25, 50, and 101.

In this task, IOU of 86% is achieved in the testing area. Figure 16 represents the obtained result. The white part corresponds to the areas that are fully covered by clouds, while the gray part denotes the areas that are covered by cloud halfway.



(a) Raw image. (b) Prediction.

Figure 16. Cloud segmentation prediction result.

5. CONCLUSION

In the present research, we aimed to develop a general segmentation process of applying AI segmentation techniques to remote sensing images. The proposed process was intentioned to address several difficult issues, such as the irregular shape of remote sensing images, lack of images for training, and normalization issues. By applying this generalized process, we could achieve an acceptable performance even when it was applied to different kinds of images and to various tasks. This confirmed that the segmentation process could be generalized with respect to different kinds of images and various tasks in the remote sensing field.

REFERENCES

Diakogiannis, F. I., Waldner, F., Caccetta, P., Wu, C., 2019. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. 2980-2988.

Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145, 60 - 77. Deep Learning RS Data.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

Li, Y., Zhang, H., Xue, X., Jiang, Y., Shen, Q., 2018. Deep learning for remote sensing image classification: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(6), e1264.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B. A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152, 166 - 177.

Ronneberger, O., Fischer, P., Brox., T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI*, 9351(1), 234–241.

Stoian, A., ans Jordi Inglada, V. P., Poughon, V., Derksen., D., 2019. Land Cover Maps Production with High Resolution Satellite Image Time Series and Convolutional Neural Networks: Adaptations and Limits for Operational Systems. *Remote Sens*, 11, 1986.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8–36.

Revised March 2020