

# A Deep Learning Approach for Building Segmentation in Taiwan Agricultural Area Using High Resolution Satellite Imagery

Liang-Yi Liu<sup>1</sup> Chi-Kuei Wang<sup>2\*</sup> An-Te Huang<sup>3</sup>

## Abstract

Understanding buildings in agricultural area is important because the arable land in Taiwan is limited. One of the practical ways is manual digitization from high resolution satellite imagery, which can acquire satisfying results without field investigation. However, such practice is tedious and labour intensive. Given these reasons, past research devoted to deep learning approaches have shown that convolutional neural networks are useful for building segmentation using satellite imagery. In this study, ENVINet5 model was trained and utilized from high resolution Pléiades pansharpened imagery. The training images (with the size of 2500 pixels × 2500 pixels) were randomly selected from 9 counties/cities to increase diversity because each county/city has different building patterns. The performance of ENVINet5 model was evaluated based on pixels and polygons, respectively. The pixel-based evaluation showed that the trained model can find 84% of building pixels. The polygon-based evaluation was carried out through calculating the number of building segments and comparing them with the reference data using IoU (Intersection of Union). The results showed that 92% of building segments were found, and the IoU of most building segments range between 0.6 and 0.9. The trained model was validated on the testing images for the transferability test. Moreover, an image tiling and stitching technique was proposed to deal with large satellite imagery. Finally, we compared the time costs of labelling with and without the aid of deep learning approach. The results showed that the time costs decreased by 7.3% with the help of deep learning approach.

**Keywords:** Building Segmentation, Deep Learning, High Resolution Satellite Imagery

## 1. Introduction

Building segmentation from high resolution imagery has many applications such as urban planning and disaster management (Yang *et al.*, 2018a). Due to various building shapes, color, and complex background in the satellite images, building segmentation is one of the challenging tasks in the field of remote sensing. With the rapid development of sensor technology, high resolution satellite imagery has become more accessible and affordable (Yang *et al.*, 2018b). Therefore, performing building segmentation from high resolution satellite imagery is widely applied to many studies (Li *et al.*, 2019; Mou & Zhu, 2018; Huang *et al.*, 2017).

The existing approaches of building segmentation can be divided into two categories: pixel-based methods and object-based methods (Khosrav *et al.*, 2014). Pixel-based methods tend to perform better with images of coarse spatial resolution (Kaszta *et al.*, 2016). However, such methods have difficulties dealing with high resolution imagery because of the pixel heterogeneity, mixed pixels, and spectral similarity (Esetlili *et al.*, 2018). The performance using high resolution imagery by pixel-based methods often result

in obvious salt and pepper effect (Blaschke *et al.*, 2000). This problem can be solved using object-based method (Quynh Trang *et al.*, 2016). In object-based methods, images are first segmented into the clusters of homogeneous pixels (objects), then the objects are classified with their spatial properties (Wu *et al.*, 2017). Objects not only consider spectral features, but also include contextual and geometric features (Pu *et al.*, 2011). However, there are some problems when applying object-based methods. According to Radoux & Defourny (2008), classification results highly depend on the quality of segmentation, which relies on low-level handcrafted features (such as edges, corners, texture, shadow, and multispectral properties) in the remote sensing imagery. Accordingly, the representative ability is limited due to the lack of high-level features (Shrestha & Vanneschi, 2018), which restricts the performance of the object-based methods.

Over the past ten years, deep learning approaches have achieved state-of-the-art performance on image segmentation with convolutional neural networks (CNN) (Gu *et al.*, 2017; Khan *et al.*, 2020). CNN can effectively extract different levels of information including corners (low-level), object parts (mid-level) and the whole object (high-level) from remote sensing imagery through multiple convolutional layers

<sup>1</sup> Master, Department of Geomatics, National Cheng Kung University

<sup>2</sup> Professor, Department of Geomatics, National Cheng Kung University

<sup>3</sup> Master Student, School of Civil Engineering, Purdue University

\* Corresponding Author, E-mail: chikuei@ncku.edu.tw

Received Date: Jan. 14, 2022

Revised Date: Feb. 25, 2022

Accepted Date: Mar. 04, 2022

(Nogueira *et al.*, 2016), and the performance is closer to visual interpretation in object recognition (Zhang *et al.*, 2020). Besides, CNN is able to combine spatial and spectral information based on the input image without preprocessing (Alshehhi *et al.*, 2017). Nowadays, deep learning approaches have been applied to the information extraction of remote sensing such as buildings (Chen *et al.*, 2020). For example, SegNet was implemented to segment buildings in Boonpook *et al.* (2018) along riverbank using UAV images, and the overall accuracy reached more than 90%. Maltezos *et al.* (2017) extracted buildings with convolutional neural networks (CNN) based structure using orthoimages and additional height feature with dense image matching point clouds. The method outperformed linear kernel SVM and the RBF kernel SVM classifiers. A pretrained ImageNet network was transferred in Vakalopoulou *et al.* (2015) by integrating additional spectral information. The quantitative validation indicated high completeness and correctness rates.

In this study, a deep learning approach was applied to segment buildings in Taiwan agricultural area from high resolution Pléiades satellite imagery. Buildings in the agricultural area include farmhouses, factories, residential housings etc. Since the arable land is limited in Taiwan, monitoring buildings in the agricultural area can understand the situation of land use.

## 2. Method

### 2.1 Data Pre-processing

In this study, high resolution Pléiades imagery and the non-agricultural mask were utilized. Pléiades satellites provide multispectral (Figure 1(a)) and panchromatic imagery (Figure 1(b)), which are both stored in the 16-bit data type. The spatial resolution is 2 meters for the color and the near-infrared bands, and the spatial resolution is 0.5 meter for the panchromatic band. The data include 142 Pléiades satellite images, which were acquired between the years of 2016 and 2017. Non-agricultural mask was provided from Taiwan Agricultural Research Institute, and it was used to distinguish the agricultural area from the non-agricultural area. Before training EVNINet5 model, Pléiades satellite images were pre-processed as follows.

(1) Pan Sharpening: In order to obtain high resolution imagery with multispectral bands, nearest-neighbor diffusion-based (NNDiffuse) pan-sharpening algorithm was applied to fuse multispectral and panchromatic images. The algorithm was proposed by Sun *et al.* (2013), which can preserve the sharp spatial features from panchromatic images and the spectral information from multispectral images. NNDiffuse pansharpened image (Figure 1(c)) with

high resolution and multispectral bands was obtained after fusion.

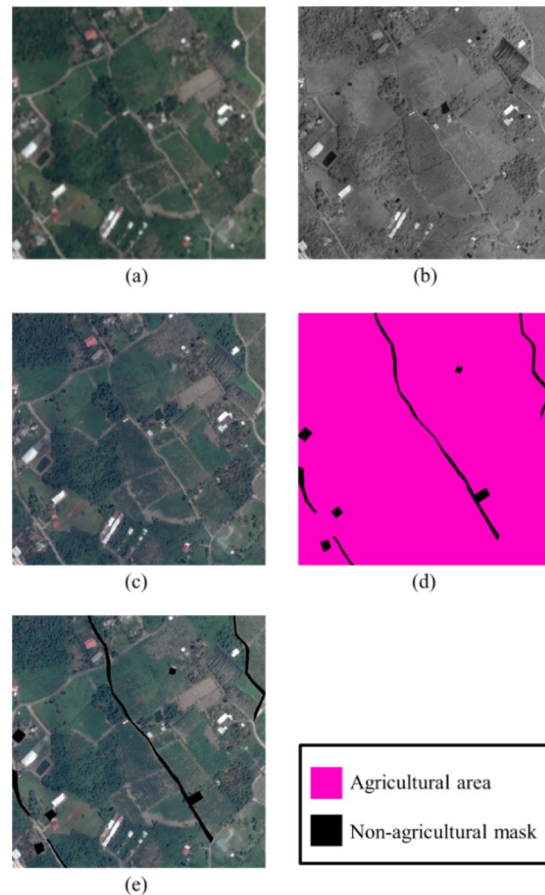


Figure 1 Data pre-processing (a) multispectral image; (b) panchromatic image; (c) NNDiffuse pansharpened image; (d) agricultural area (denoted as pink) and non-agricultural mask (denoted as black); (e) NNDiffuse pansharpened agricultural image

(2) Image Masking: Image masking is a useful technique that can restrict analysis to a subset region instead of using whole image scene (Kastens *et al.*, 2005). In this study, we focused on the agricultural area in Taiwan; therefore, non-agricultural area was masked out using non-agricultural mask (Figure 1(d)). After image masking, NNDiffuse pansharpened agricultural image is shown in Figure 1(e).

### 2.2 Study Area

The study area is the agricultural area of Taiwan. EVNINet5 was trained using 500 sub-images from Miaoli to Taitung cities/counties clockwise, and it was tested using 10 sub-images from Taichung to Pingtung cities/counties counter clockwise (Figure 2). The size of the sub-image is 2500 pixels  $\times$  2500 pixels, and each of them was randomly collected. Collecting sub-images from different cities was to increase the diversity since

each county/city has different building patterns. The training images were randomly divided into the training sets and the validation sets with the proportion of 8:2. The testing images were used to test the transferability of the trained model.

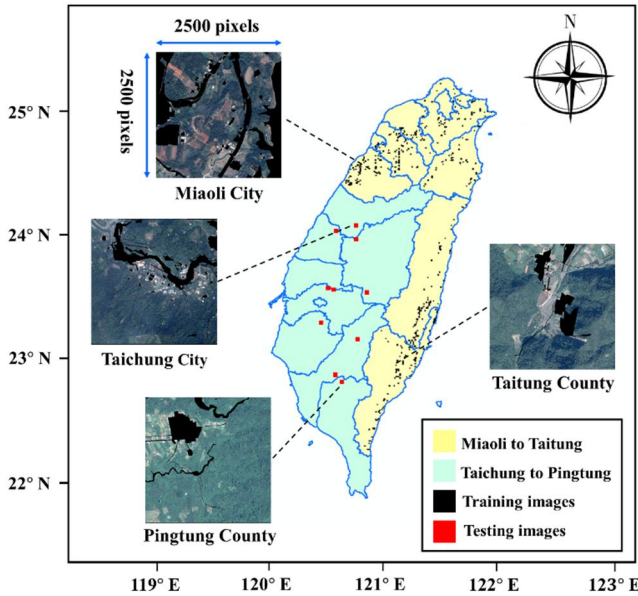


Figure 2 Study area

### 2.3 Data Labelling

Due to the high degree of variety of building patterns in the agricultural area, labelling results can be unstable depending on different labellers. Therefore, consistent labels as reference are necessary for ENVINet5 to correctly identify and segment the buildings. In this study, each building pattern was labelled as a building polygon as follows. Firstly, shadow is excluded from the label (Figure 3(a)). Second, labels of the buildings, which is in contact with the non-agricultural area, were kept several pixels away from the non-agricultural area (Figure 3(b)). Third, adjacent buildings with no space in between (Figure 3(c)) were labelled as one single building polygon since it was challenging to label each building separately. Fourth, buildings occluded by vegetation were kept out for the label (Figure 3(d)).

### 2.4 Architecture of ENVINet5

The model we used in this study is a U-Net (Ronneberger *et al.*, 2015) based architecture called ENVINet5. It is a Deep Learning module (v1.1) built in a commercial software ENVI (v5.6). ENVINet5 has the characteristics of U-Net. For example, it can be trained in an end-to-end fashion from few training images and yield precise segmentations. Moreover, it includes concatenate operation by combining high-level

semantic information and low-level detailed features. Since U-Net is one of the effective architectures in object segmentation (Soni *et al.*, 2020), many studies also improved its architectures to perform building segmentation (He *et al.*, 2020; Guo *et al.*, 2020; Yi *et al.*, 2019; Xu *et al.*, 2018). In ENVINet5, four proprietary hyperparameters were introduced. Class Weight brings in a biased selection of patches, so the model can extract patches with more feature pixels. Next, Patch Sampling Rate can control the density of sampling. Because the feature pixels are often sparse comparing to the background pixels, high density of sampling rate can generate more patches with more feature pixels. Then, Loss Weight biases loss function to make more adjustment on identifying the feature pixels. Finally, Blur Distance helps the model to learn the building borders by blurring the edges and decreasing the blur during training (ENVI Development Team, 2020).

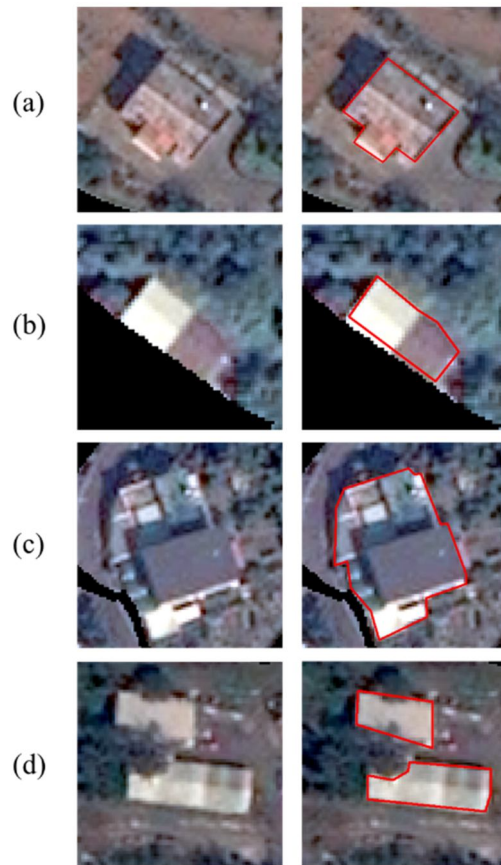


Figure 3 Examples of different building patterns (first column) and their labelling results (second column) (a) a building with shadow; (b) a building in contact with non-agricultural area; (c) adjacent buildings; (d) buildings occluded by vegetation

ENVINet5 was trained using the patch-based convolutional neural network. The input of ENVINet5 is a patch with the agricultural buildings. And the



output is the probability map, where the pixel values range from 0 to 1 in the form of floating-point numbers. The brighter pixels denote higher probability, and the darker pixels denote lower probability. The architecture of ENVINet5 is shown in Figure 4. The network went through four times of downsampling and upsampling. It also merges high resolution features with low resolution features (the purple arrow shown in Figure 4).

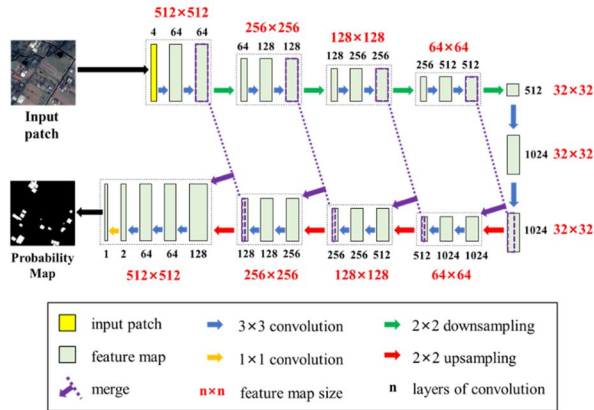


Figure 4 The architecture of ENVINet5

A patch is a certain region in the sub-image (Figure 5(a)), and a batch is a number of patches being trained for every iteration (Figure 5(b)). The parameters were updated in every iteration. In this study, the patch size was set as  $512 \times 512$  pixels, and the batch size was set as 64.

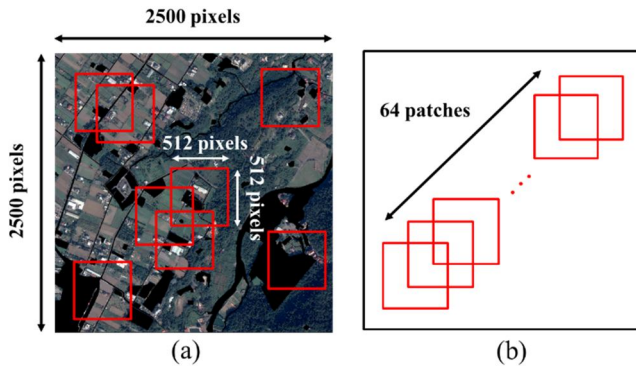


Figure 5 Illustration of a patch and a batch (a) a sub-image with several patches (in red frame); (b) a batch of 64 patches

## 2.5 Data Post-processing

To obtain building segments from the probability map, four post-processing steps are carried out in the following steps. The example shown in Figure 6(a) and 6(b) is adjacent buildings and its probability map from ENVINet5.

(1) Thresholding: The probability threshold was set as 0.6. If a pixel value is greater than or equal to 0.6,

it is considered as a building pixel. Otherwise, it is a non-building pixel. The reason for setting the probability threshold as 0.6 were discussed in section 3.2. The probability map was then converted to the binary map (Figure 6(c)).

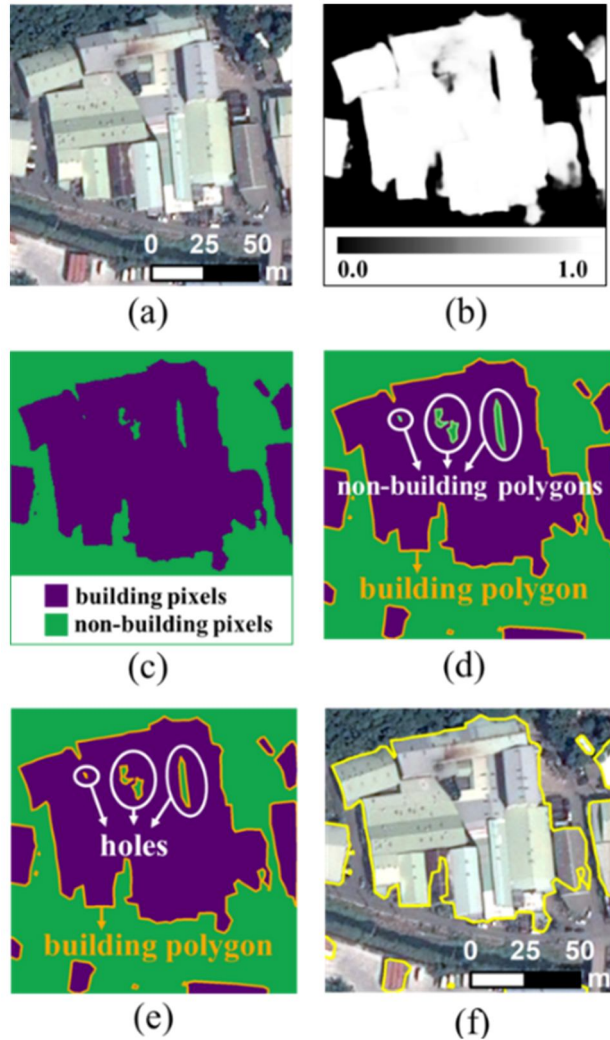


Figure 6 Data post-processing (a) adjacent buildings shown in the pan-sharpened agricultural image; (b) probability map from ENVINet5; (c) binary map with building and non-building pixels; (d) vectorization result from binary map, where building and non-building polygons were due to building and non-building pixels; (e) building polygon with holes; (f) building segment after filling holes in (e)

(2) Vectorization: Vectorization was to acquire vector data from raster data. The process was completed using the “Raster to Polygon” tool in ArcMap (ArcGIS Development Team, 2019). The output produces polygons with smoothed outlines using its proprietary algorithm. After vectorization, building pixels and non-building pixels were vectorized to building and non-building polygons,

respectively (Figure 6(d)). The process also assigned different codes (referred to as gridcode in ArcMap) to building polygons and non-building polygons. Building polygons were coded as 1, and non-building polygons were coded as 0.

- (3) Keep Building Polygons: The building polygons, where the code of 1, were kept using the “Make Feature Layer” tool in ArcMap.
- (4) Filling Holes of Building Polygon: Building polygon was left with several holes after removing the non-building polygons (Figure 6(e)). The appearance of the holes is mainly caused by shadow of the adjacent buildings; however, they are still part of the building. The holes were filled up using the “Eliminate Polygon Parts” tool in ArcMap. The threshold condition was the area of holes. If the area of each hole is less than 25 percent of the building polygon, the hole will be filled up. The building segments (Figure 6(f)) were obtained after filling the holes. The building segments are polygons in the shapefile format.

## 2.6 Evaluation

### 2.6.1 Pixel-based Evaluation

Confusion matrix and assessment indices are used to evaluate ENVINet5 model in the study. The assessment indices include accuracy, precision, recall, and F1 score. The model is evaluated with each sub-image using the validation sets. In confusion matrix (Table 1), correctly predicted building and non-building pixels are defined as true positive (TP) and true negative (TN); incorrectly predicted building and non-building pixels are defined as false negative (FN) and false positive (FP).

Table 1 Confusion matrix

	Reference	
Prediction	building pixel	non-building pixel
building pixel	True Positive (TP)	False Positive (FP)
non-building pixel	False Negative (FN)	True Negative (TN)

In assessment indices, accuracy is overall correctness including building pixels and non-building pixels. Precision is the ratio of correctly predicted building pixels within all positive prediction. Recall shows the proportion of reference building pixels being predicted. F1 score is a harmonic combination between precision and recall, which keeps the correctness of precision and completeness of recall value (Prathap & Afanasyev, 2018). The equations of assessment indices

are listed below:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \dots\dots\dots(1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \dots\dots\dots(2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \dots\dots\dots(3)$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \dots\dots\dots(4)$$

### 2.6.2 Polygon-based Evaluation

In order to evaluate the quality of segmentation and calculate the number of the building segments, polygon-based evaluation was carried out. In Table 2, six different cases were considered. Case1 is omission building, which showed the model failed to segment the building. Case 2 is commission building, where the model mistakenly segmented non-building pixels as building. Case 3 is one-to-one correspondence, where the segmented building corresponded to one reference building. Case 4 is many-to-one correspondence, which means several buildings were segmented, and all of which corresponded to one reference building. Case 5 is one-to-many correspondence, which means a building was segmented, and it corresponded to several reference building. Case 6 is many-to-many correspondence, where several buildings were segmented, and all of which corresponded to several reference buildings. For all the cases in this study, “many” indicates greater than or equal to 2.

We overlaid the building segments with the reference buildings to count the omission and commission buildings. Then, the omission error rate (OER) and the commission error rate (CER) were computed using the formulas below.

$$\text{OER} = \frac{\text{number of omission buildings}}{\text{number of reference buildings}} \times 100\% \dots\dots\dots(5)$$

$$\text{CER} = \frac{\text{number of commission buildings}}{\text{number of reference buildings}} \times 100\% \dots\dots\dots(6)$$

Next, the quality of the building segmentation was evaluated for Case 3 to Case 6. We use IoU (intersection over union) to check the similarity between each building segment and reference building since IoU is the most common metric to compare similarity between two arbitrary shapes (Rezatofighi *et al.*, 2019). The Value of IoU range from 0 to 1. If the value is closed to 1, the building segment has higher similarity with the reference building. The formula of IoU is shown below.

$$\text{IoU} = \frac{\text{Area of intersection}}{\text{Area of Union}} \dots\dots\dots(7)$$

Table 2 Discussion of six different cases in polygon-based evaluation. The reference building and segmented building are denoted as red and blue polygons, respectively

Case No.	Condition	Description	Illustration
1	Omission building	The model failed to segment the building.	
2	Commission building	The model mistakenly segment non-building pixels as building.	
3	One-to-one correspondence	The segmented building corresponded to one reference building.	
4	Many-to-one correspondence	Several buildings were segmented, all of which corresponded to one reference building.	
5	One-to-many correspondence	A building was segmented, which corresponded to several reference buildings	
6	Many-to-many correspondence	Several buildings were segmented, all of which corresponded to several reference buildings.	

Since the IoU evaluation requires one-to-one correspondence, a filtering process is adopted for case 4 to case 6 in order to identify representative correspondence between a building segment and a reference building (Figure 7).

Table 3 shows examples of different IoUs. The first column is the values of IoU, and the second column is the pansharpened images. The last column shows the building segments (yellow) overly with the reference buildings (red).

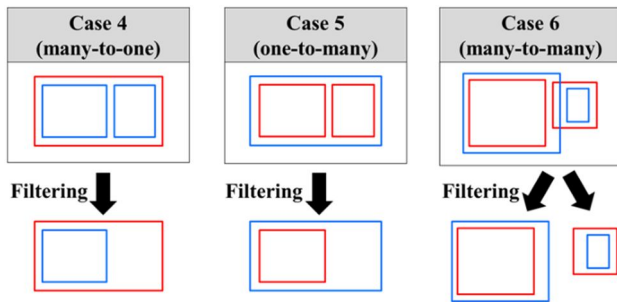


Figure 7 Filtering results for case 4, case 5, and case 6. The reference building and segmented building are denoted as red and blue polygons, respectively

In the real world result, the many-to-many condition (case 6) can be far more complicated than the schematic drawing shown in Figure 7. An example case of three reference buildings and three segmented buildings is shown in Figure 8 to elaborate the workflow of the filtering process adopted in this study. Reference building (denoted as red) and building segment (denoted as blue) are abbreviated to R and S, respectively. The reference buildings R1, R2, and R3 overlaps with S1 and S2. Five IoUs (IoU1 to IoU5) are calculated for each pair of overlapping polygons.

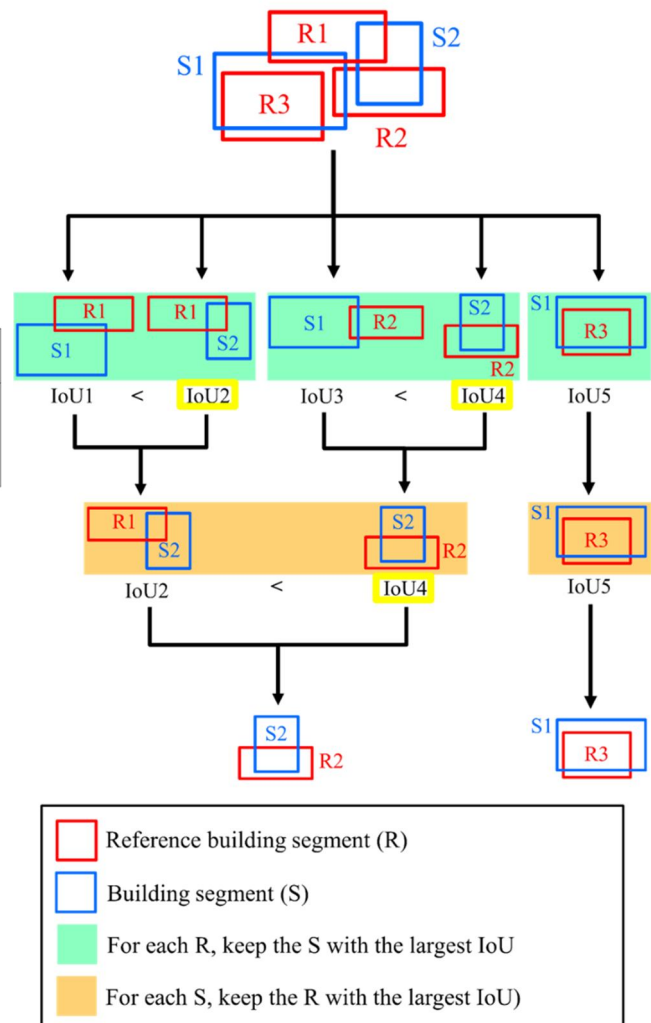



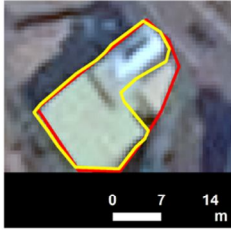






Figure 8 Illustration of the filtering process

- Reference building segment (R)
- Building segment (S)
- For each R, keep the S with the largest IoU
- For each S, keep the R with the largest IoU)

Table 3 Examples of various IoUs for different buildings

IoU	Pansharpened images	Examples of different IoUs
0.96		
0.77		
0.48		
0.15		

Firstly, the filtering process started with the reference buildings (the green block in Figure 8). Building segments that are overlapped with the same reference building were compared using IoU. The segmented building with the largest IoU is kept. For R1, S2 is kept because IoU2 is larger than IoU1. For R2, S2 is kept because IoU3 is larger than IoU4. For R3, S1 is kept because there is only one building segment. Next, the filtering process move to the building segments (the orange block in Figure 8). Reference buildings that are overlapped with the same building segment are compared using IoU. The reference building with the largest IoU is kept. For S1, R3 is kept because there is only one reference building. For S2, R2 is kept because IoU4 is larger than IoU2. After the filtering process, every building segment corresponds to one reference building. The results of the filtering process in Figure 8 are R2 and S2; R3 and S1.

## 2.7 Transferability

The transferability of the trained model was validated on 10 random testing images (with the size of 2500 pixels  $\times$  2500 pixels). The number of the testing images in each city/county is listed in Table 4. Both pixel-based and polygon-based evaluation were also carried out in the transferability test.

Table 4 The number of the testing images in different cities/ counties

City/County	Testing Image
Taichung City	1
Nantou County	1
Changhua County	2
Chiayi County	2
Tainan City	1
Kaohsiung City	2
Pingtung County	1

## 3. Results and Discussion

### 3.1 Training Process

It took 26 hours to train ENVINet5. To avoid overfitting problem, the training process was stopped at 100 epochs since the training loss and the testing loss tend to diverge. The training process is shown in Figure 9.

During the training process, 10,000 patches extracted from the training images were input to ENVINet5, and all the patches were trained batch by batch within 100 epochs. The patches were bias selected and generated with the control of Class Weight and Patch Sampling Rate in ENVINet5. Furthermore, the minimum and the maximum value were set for Class Weight and Blur Distance. The values indicate the degree of bias selection on patches, and a decaying gradient from the edge of the features. The maximum value is applied when the training begins. This value gradually decreases to the minimum value when the training ends. The settings of the hyperparameters (including ENVINet5 proprietary hyperparameters) are list in Table 5. The training of ENVINet5 model was implemented on a workstation with NVIDIA GeForce RTX 2080 Ti GPU.

Table 5 Hyperparameters of the model

Patch size	512 $\times$ 512
Batch size	64
The number of patches	10,000
Epoch	100
Class Weight	min:2; max:4
Patch Sampling Weight	15
Loss Weight	1.5
Blur Distance	min:0; max:15



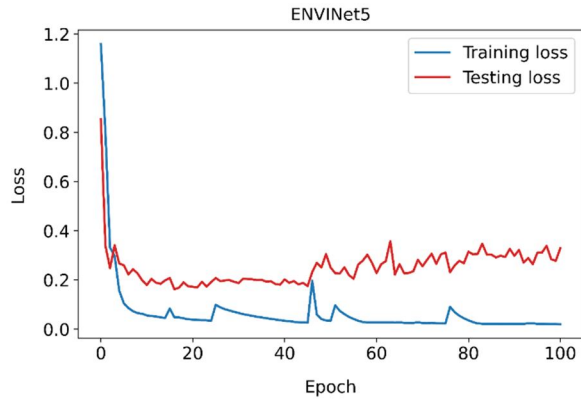


Figure 9 Training process of ENVINet5

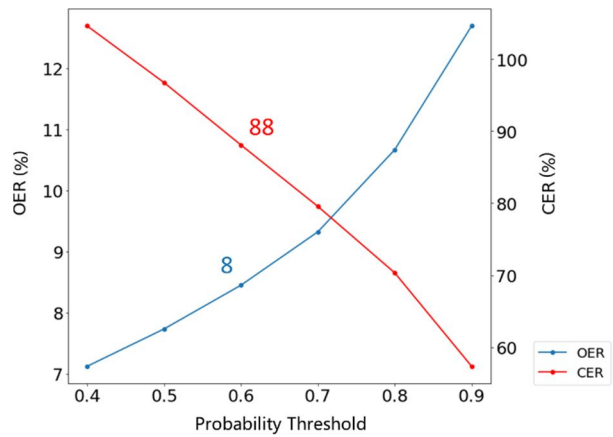


Figure 11 The values of OER and CER with different probability thresholds

### 3.2 Threshold of Probability Map

The results of pixel-based and polygon-based evaluation were influenced by the threshold of probability map. Therefore, the probability threshold from 0.4 to 0.9 were tested in this study. In pixel-based evaluation, the value of precision increased and the value of recall decreased when the probability threshold is greater (Figure 10).

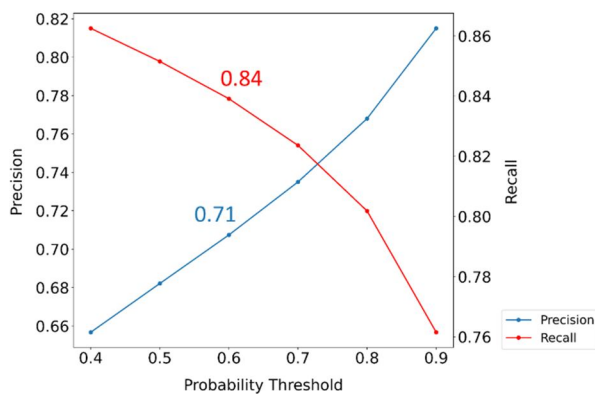


Figure 10 The values of precision and recall with different probability thresholds

In polygon-based evaluation, the larger the probability threshold, the higher the value of OER. On the contrary, the value of CER becomes lower when the probability threshold is smaller (Figure 11). The value of precision and OER were intentionally kept low with the expense of high recall and CER value because the ideal goal is not to miss any buildings from the high resolution satellite imagery. The trade-off of probability threshold is around 0.7 considering finding the most and correct building pixels and building segments. For the conservative estimate, the probability threshold was set as 0.6. The optimal probability threshold of 0.6 was also applied to the validation sets and testing data.

### 3.3 Accuracy Assessment

The pixel-based evaluation was carried out with the validation sets. The number of TP, FP, FN, and TN cases are shown in Table 6.

Table 6 The calculation of TP, FP, FN, and TN cases in confusion matrix using the validation sets

Prediction \ Reference	building pixel	non-building pixel
building pixel	3556048 (TP)	1470905 (FP)
non-building pixel	682058 (FN)	619290989 (TN)

The values of assessment indices are 0.99, 0.71, 0.84, and 0.77 respectively for accuracy, precision, recall, and F1 score. The results show that ENVINet5 can find 84% of building pixels according to the value of recall. Next, the polygon-based evaluation was also carried out with the validation sets. Omission buildings, commission buildings, OER, and CER were computed. The results are shown in Table 7.

Table 7 Statistics of omission and commission buildings with the computation of OER and CER using the validation sets

Prediction \ Reference	Number of building	Error rate
Reference building	5727	NA
Omission building	484	8%
Commission building	5044	88%

Among 100 sub-images in the validation sets, 5727 reference buildings were labelled manually. The values of omission buildings and the commission are 484 and 5044. OER and CER are 8% and 88% respectively. In other words, ENVINet5 can find 92% of building segments according to OER. After applying the filtering process over many-to-one, one-to-many, and many-to-many case, 4770 buildings were kept as one-to-one correspondence. The IoU for every



correspondence of building segment and reference building was plotted into a histogram shown in Figure 12. Most of the buildings have IoU gathering from 0.6 to 0.9.

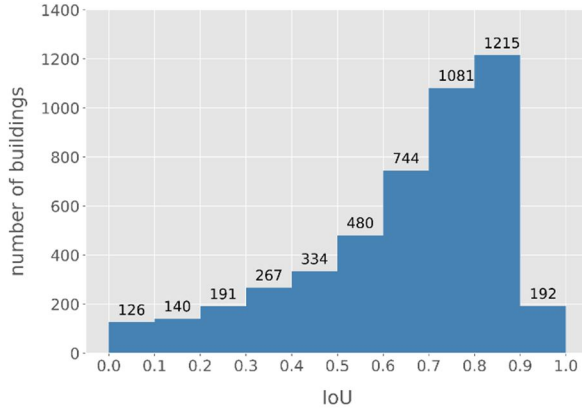


Figure 12 Histogram for IoU using the validation sets, where number of buildings were reported for each IoU bin

### 3.4 Transferability

The trained model was validated on 10 random testing images for the transferability test. Both the pixel-based and the polygon-based evaluation were carried out. For the pixel-based evaluation, the number of TP, FP, FN, and TN cases are shown in Table 8. The values of assessment indices are 0.99, 0.77, 0.80, and 0.78 respectively for accuracy, precision, recall, and F1 score.

Table 8 The calculation of TP, FP, FN, and TN cases in confusion matrix using the testing images

Reference \ Prediction	building pixel		non-building pixel	
	building pixel	non-building pixel	building pixel	non-building pixel
building pixel	537392 (TP)	182442 (FP)		
non-building pixel	124752 (FN)	61655414 (TN)		

For the polygon-based evaluation, the numbers of omission building and commission building with the calculation of OER and CER are shown in Table 9. 808 buildings were labelled manually among 10 random testing images. The values of omission buildings and the commission are 105 and 498. OER and CER are 12.99% and 61.63% respectively.

Table 9 Statistics of omission and commission buildings with the computation of OER and CER using the testing images

	Number of building	Error rate
Reference building	808	NA
Omission building	105	13%
Commission building	498	62%

After applying filtering process over many-to-one, one-to-many, and many-to-many case among 10 testing images, 656 buildings were kept as one-to-one correspondence. The IoU for each correspondence of building segment and reference building was plotted into a histogram shown in Figure 13. Most of the buildings have IoU gathering from 0.6 to 0.9. The transferability test shows the stable performance of ENVINet5.

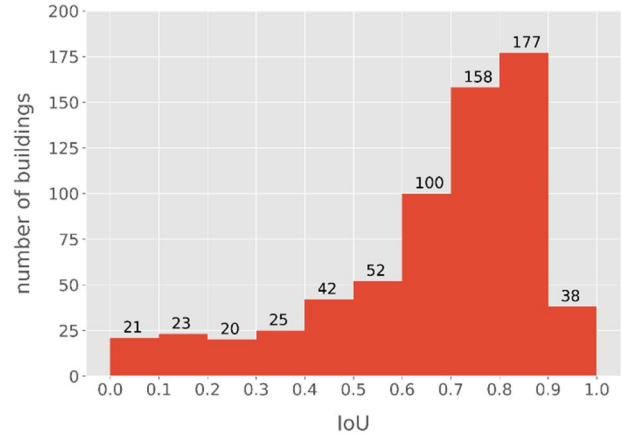


Figure 13 Histogram for IoU using the testing images, where number of buildings were reported for each IoU bin

## 3.5 Processing Considerations for Large Satellite Images

The size of remote sensing imagery is usually very large, which is difficult to be segmented directly using most deep learning networks and their associated hardware (Huang *et al.*, 2018). As a result, an image tiling and stitching technique was proposed as follows to deal with large satellite imagery.

### 3.5.1 Image Tiling

A large scale pansharpened agricultural image was tiled into several sub-images. The size of each sub-image is 2500 pixels  $\times$  2500 pixels. The illustration of image tiling is shown in Figure 14.

The tiling starts from the northwest of the image, moving from left to right and up to down. In Figure 14(a), the red frame shows the movement of the tiling to the next sub-image horizontally and vertically. The stride was set as 2400 pixels in order to create overlap, which is used to preserve the objects along the sub-image boundaries and prevent any miss (Ünel *et al.*, 2019). Because most of the height and width of the satellite image are not divisible by the size of sub-image, the rightmost and the downmost sub-images (denoted as dotted grey frame) are not the full-size of 2500 pixels  $\times$  2500 pixels. Therefore, the rightmost and

the downmost sub-images were moved left and up respectively to fit the boundary of the pansharpened agricultural image (denoted as green frame in Figure 14(b) and 14(c), respectively). For the sub-images that contain only non-agricultural pixels (denoted as blue frame in Figure 14(d)), they are discarded to save the processing time for inference.

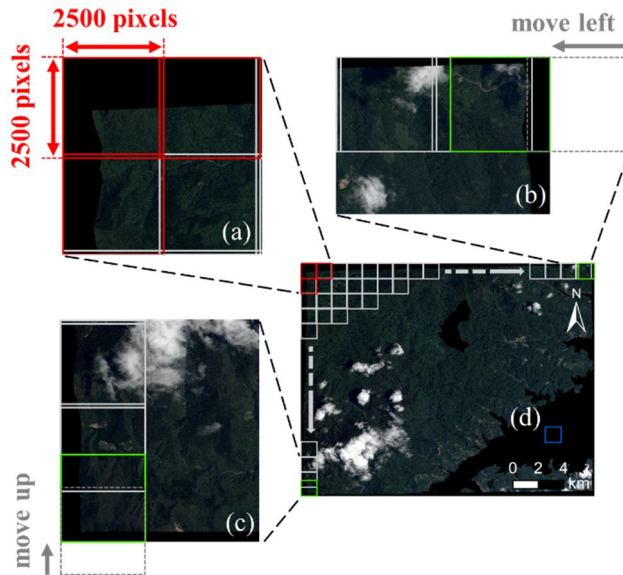


Figure 14 Illustration of image tiling using the pansharpened agricultural image from Yilan (a) movement of tiling with the stride of 2400 pixels horizontally and vertically; (b) rightmost sub-image moving left to fit the boundary of the pansharpened agricultural image; (c) downmost sub-image moving up to fit the boundary of the pansharpened agricultural image; (d) sub-image contains only non-agriculture pixels (denoted as black colour)

### 3.5.2 Image Stitching

Image stitching is the process that combine images with overlap and form a large image with high resolution (Wang & Yang, 2020). The way of stitching in this study was to compute the mean value of the overlapping pixels within the sub-images. This is to prevent any miss of buildings along the boundaries of the sub-images. Since the values of the pixel are probability, it is convenient and efficient to calculate. The stitching process was conducted using the “Mosaic To New Raster” tool in ArcMap. In Figure 15(a), one of the pansharpened agricultural images (with the size of 44688 pixels  $\times$  44836 pixels) across Nantou and Hualien County is shown as an example. The image was tiled into several sub-images and the inference result of each sub-image was stitched together to create a large probability map (Figure 15(b)).

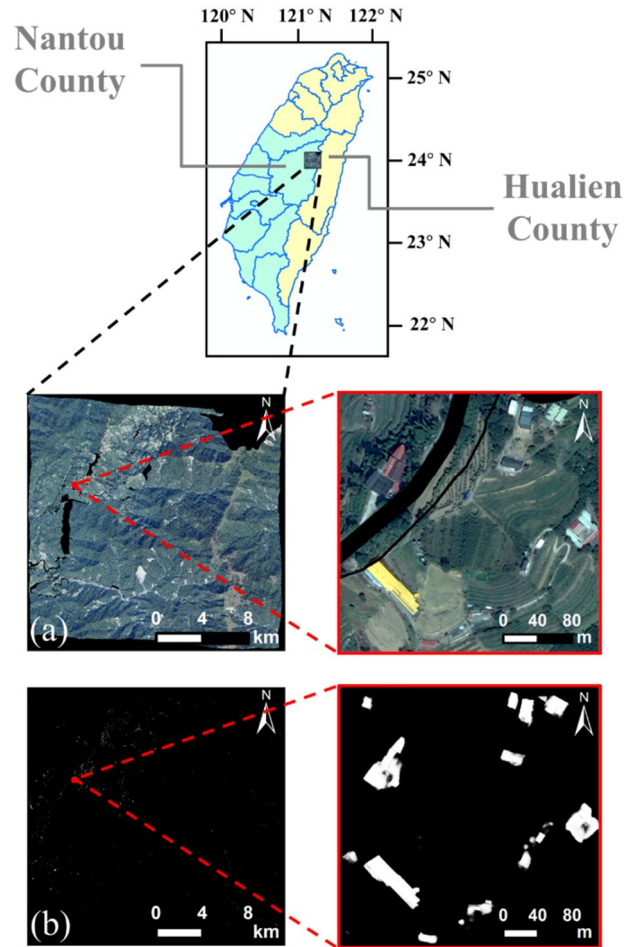


Figure 15 Results of image stitching using the pansharpened agricultural image across Nantou and Hualien County (a) a pansharpened agricultural image across Nantou and Hualien County; (b) a large probability map stitched by the inference results of each sub-image tiled from (a)

The advantage of calculating the mean value of the overlapping pixels is to produce finer building borders. Due to the different extent of the sub-images, the inference result of the overlaps can be different. Since the overlaps between sub-images have been inferred twice or four times by the trained model, averaging the inference results can balance different probabilities and yield smoother borders. The large probability map was post-processed using the proposed methods in section 2.6 to obtain building segments. Building borders with and without calculating the mean value of overlapping pixels were shown in Figure 16.

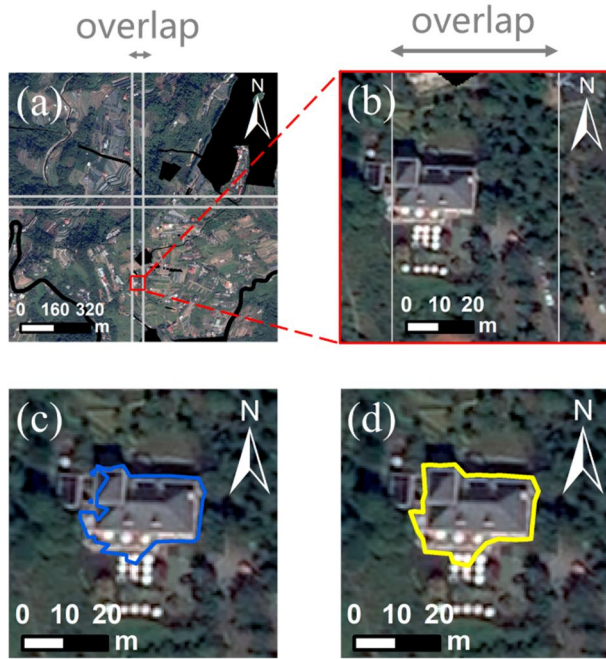


Figure 16 Building borders with and without calculating the mean value of overlapping pixels in the overlap of the sub-images (a) overlap generated from four sub-images; (b) a building within the overlap; (c) Borders of building segments calculating the mean value of the overlapping pixels; (d) Borders of building segments without calculating the mean value of the overlapping pixels

### 3.6 Time Cost Comparison

Automation of ENVINet5 improved the efficiency of building segmentation using high resolution satellite imagery. However, manual labelling is still irreplaceable due to the potential OER and CER. Therefore, a hybrid way of manual labelling with the aid of the deep learning results was carried out. The results of ENVINet5 were utilized as an additional layer that helps to locate the buildings, which is helpful for the manual labelling process. The hybrid method was compared with manual labelling based on the time cost. In this study, 75 random sub-images (with the size of 2500 pixels × 2500 pixels) were assigned equally to 3 professional data labellers. Each sub-image was labelled twice by the same data labeller. The first time was to label manually, and the second time was to label using the hybrid way. In case that the same sub-image was labelled twice in a short time, data labellers were required to label all the assigned images manually first before labelling using the hybrid method. The time spent on labelling using both methods were counted. The result is shown in Table 10.

For all the 75 sub-images, manual labelling took 1622 minutes and the hybrid way took 1504 minutes to complete. The result shows that manual labelling consume 7.3% more time than the hybrid method. It's

expected that the time cost for the manual labelling will increase obviously because fatigue can be triggered when dealing with large satellite images. And, the efficiency for labelling with the aid of deep learning results will highly improve.

Table 10 Time cost for assigned sub-images counted on each data labeller using the manual and the hybrid method

(time unit: minute)

Data labeller \ Sub-image	1		2		3	
	manual	hybrid	manual	hybrid	manual	hybrid
1 - 25	598	568	-	-	-	-
26 - 50	-	-	438	399	-	-
51 - 75	-	-	-	-	586	537

## 4. Conclusions

In this study, the feasibility of building segmentation from high resolution Pléiades pansharpened imagery in agricultural area based on deep learning approach was demonstrated. ENVINet5 was trained on random 500 sub-images (with the size of 2500 pixels × 2500 pixels) from 9 cities/ counties around Taiwan due to various building patterns. Each building pattern was labelled manually in a consistent building polygon as the reference data. In the training process, four proprietary hyperparameters were introduced to yield fine building borders. Four proprietary hyperparameters include Class Weight, Patch Sampling Rate, Loss Weight, and Blur Distance. The inference result from ENVINet5 is a probability map, which was post-processed to obtain building segments. Post-processing includes four steps: (1) thresholding, (2) vectorization, (3) keep building polygons, and (4) filling holes of building polygon.

To validate the trained model, both pixel-based and polygon-based evaluation were carried out. For the pixel-based evaluation, the trained model was validated by pixels within each sub-image in the validation sets. The results reached 0.99, 0.71, 0.84, and 0.77 respectively for accuracy, precision, recall, and F1 score. According to the value of recall, ENVINet5 can find 84% of building pixels. For the polygon-based evaluation, the number and the quality of the building segments were analysed. Six different cases were considered, which are omission building, commission building, one-to-one, many-to-one, one-to-many, and many-to-many (correspondence). OER, and CER were computed to show the buildings that are missed and overpredicted from ENVINet5. The result of OER and CER are 8% and 88%. According to the value of OER, ENVINet5 can find 92% of building segments. Next, the segmentation quality of each building segment was

evaluated by comparing with the reference data using IoU. Since not every case is one-to-one correspondence, the filtering process is needed. After filtering process, 4770 buildings were kept as one-to-one correspondence within validation sets, and most of the building segments have IoUs between 0.6 and 0.9.

For the transferability test of the model, 10 random testing images were collected from Taichung to Pingtung cities/counties counter clockwise around Taiwan. The values of assessment indices, OER, CER, and the statistics of IoUs for the building segments have shown the stability and performance of ENVINet5. Moreover, an image tiling and stitching technique was proposed to deal with the large satellite image across Nantou and Hualien County. Finally, the time cost of manual labelling and the hybrid way were compared by labelling 75 random sub-images twice. The time was counted by three professional data labellers. The results showed that the hybrid method cost 7.3% less time than manual labelling. And it is expected that more time can be saved using the hybrid way when dealing with large satellite images because the level of fatigue will increase.

## References

- Alshehhi, R., Marpu, P.R., Woon, W.L., and Dalla Mura, M., 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks, *ISPRS Journal of Photogrammetry and Remote Sensing*, 130: 139-149.
- ArcGIS Development Team, 2019. Powerful mapping and analytics software, Version 10.7.1., Esri Corporation, esri.com.
- Blaschke, T., Lang, S., Lorup, E., Strobl, J., and Zeil, P., 2000. Object-oriented image processing in an integrated GIS/remote sensing environment and perspectives for environmental applications, In: *Environmental Information for Planning, Politics, and the Public*, 2, Cremers, A. and Greve, K. (eds), Metropolis-Verlag, Marburg, Germany, pp. 555-570.
- Boonpook, W., Tan, Y., Ye, Y., Torteeka, P., Torsri, K., and Dong, S., 2018. A deep learning approach on building detection from unmanned aerial vehicle-based images in riverbank monitoring, *Sensors*, 18(11): 3921.
- Chen, Y., Tang, L., Yang, X., Bilal, M., and Li, Q., 2020. Object-based multi-modal convolution neural networks for building extraction using panchromatic and multispectral imagery, *Neurocomputing*, 386: 136-146.
- ENVI Development Team, 2020. Exelis visual information solutions software, Version 5.6., L3HARRIS Geospatial Corporation, www.l3harrisgeospatial.com
- Esetlili, T.M., Balcik, F.B., Sanli, F.B., Ustuner, M., Kalkan, K., Goksel, C., Gazioğlu, C., and Kurucu, Y., 2018. Comparison of object and pixel-based classifications for mapping crops using rapideye imagery: A case study of menemen plain, Turkey, *International Journal of Environment and Geoinformatics*, 5(2): 231-243.
- Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., and Chen, T., 2017. Recent advances in convolutional neural networks, *Pattern Recognition*, 77: 354-377.
- Guo, M., Liu, H., Xu, Y., and Huang, Y., 2020. Building extraction based on U-Net with an attention block and multiple losses, *Remote Sensing*, 12(9): 1400.
- He, N., Fang, L., and Plaza, A., 2020. Hybrid first and second order attention Unet for building segmentation in remote sensing images, *Science China Information Sciences*, 63(4): 140305.
- Huang, B., Reichman, D., Collins, L., Bradbury, K., and Malof, J.M., 2018. Tiling and stitching segmentation output for remote sensing: Basic challenges and recommendations, *arXiv preprint arXiv:1805.12219*.
- Huang, X., Yuan, W., Li, J., and Zhang, L., 2017. A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(2): 654-668.
- Kastens, J.H., Kastens, T.L., Kastens, D.L.A., Price, K.P., Martinko, E.A., and Lee, R.Y., 2005. Image masking for crop yield forecasting using AVHRR NDVI time series imagery, *Remote Sensing of Environment*, 99(3): 341-356.
- Kaszta, Z., Van de Kerchove, R., Ramoelo, A., Cho, M.A., Madonsela, S., Mathieu, R., and Wolff, E., 2016. Seasonal separation of African savanna components using worldview-2 imagery: A comparison of pixel- and object-based approaches and selected classification algorithms, *Remote Sensing*, 8(9): 763.
- Khan, A., Sohail, A., Zahoor, U., and Qureshi, A.S., 2020. A survey of the recent architectures of deep convolutional neural networks, *Artificial Intelligence Review*, 53(8): 5455-5516.
- Khosravi, I., Momeni, M., and Rahnemounfar, M., 2014. Performance evaluation of object-based and pixel-based building detection algorithms from very high spatial resolution imagery, *Photogrammetric Engineering and Remote Sensing*, 80: 519-528.



- Li, W., He, C., Fang, J., Zheng, J., Fu, H., and Yu, L., 2019. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data, *Remote Sensing*, 11(4): 403.
- Maltezos, E., Doulamis, N., Doulamis, A., and Ioannidis, C., 2017. Deep convolutional neural networks for building extraction from orthoimages and dense image matching point clouds, *Journal of Applied Remote Sensing*, 11(4): 042620.
- Mou, L., and Zhu, X.X., 2018. RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images, arXiv preprint arXiv:1805.02091.
- Nogueira, K., Penatti, O.A.B., and dos Santos, J.A., 2016. Towards better exploiting convolutional neural networks for remote sensing scene classification, *Pattern Recognition*, 61: 539-556.
- Prathap, G., and Afanasyev, I., 2018. Deep learning approach for building detection in satellite multispectral imagery, *Proceedings of the International Conference on Intelligent Systems (IS)*, Funchal-Madeira, Portugal, pp.461-465.
- Pu, R., Landry, S., and Yu, Q., 2011. Object-based urban detailed land cover classification with high spatial resolution IKONOS imagery, *International Journal of Remote Sensing*, 32(12): 3285-3308.
- Radoux, J., and Defourny, P., 2008. Quality assessment of segmentation results devoted to object-based classification, In: *Object-Based Image Analysis*, Blaschke, T., Lang, S., and Hay, G.J. (eds), Springer, Berlin, Heidelberg, pp.257-271.
- Rezatofighi, H., Tsoi, N., Gwak, J.Y., Sadeghian, A., Reid, I., and Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp.658-666.
- Ronneberger, O., Fischer, P., and Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation, arXiv preprint arXiv:1505.04597
- Shrestha, S., and Vanneschi, L., 2018. Improved fully convolutional network with conditional random fields for building extraction, *Remote Sensing*, 10(7): 1135.
- Soni, A., Koner, R., and Villuri, V.G.K., 2020. M-UNet: Modified u-net segmentation framework with satellite imagery, *Proceedings of the Global AI Congress*, Kolkata, India, pp.47-59.
- Sun, W., Chen, B., and Messinger, D.W., 2013. Nearest-neighbor diffusion-based pan-sharpening algorithm for spectral images, *Optical Engineering*, 53(1): 013107.
- Quynh Trang, N.T., Toan, L.Q., Huyen Ai, T.T., Vu Giang, N., and Viet Hoa, P., 2016. Object-based vs. pixel-based classification of mangrove forest mapping in Vien An Dong Commune, Ngoc Hien District, Ca Mau Province using VNREDSat-1 images, *Advances in Remote Sensing*, 5: 284-295.
- Ünel, F.Ö., Özkalayci, B.O., and Çiğla, C., 2019. The power of tiling for small object detection, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, pp.582-591.
- Vakalopoulou, M., Karantzalos, K., Komodakis, N., and Paragios, N., 2015. Building detection in very high resolution multispectral data with deep learning features, *Proceedings of the IEEE international Geoscience and Remote Sensing Symposium (IGARSS)*, Milan, Italy, pp.1873-1876.
- Wang, Z., and Yang, Z., 2020. Review on image-stitching techniques, *Multimedia Systems*, 26: 413-430.
- Wu, Q., Zhong, R., Zhao, W., Fu, H., and Song, K., 2017. A comparison of pixel-based decision tree and object-based Support Vector Machine methods for land-cover classification based on aerial images and airborne lidar data, *International Journal of Remote Sensing*, 38(23): 7176-7195.
- Xu, Y., Wu, L., Xie, Z., and Chen, Z., 2018. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters, *Remote Sensing*, 10(1): 144.
- Yang, H., Wu, P., Yao, X., Wu, Y., Wang, B., and Xu, Y., 2018a. Building extraction in very high resolution imagery by dense-attention networks, *Remote Sensing*, 10(11): 1768.
- Yang, H.L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A., and Bhaduri, B., 2018b. Building extraction at scale using convolutional neural network: Mapping of the United States, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8): 2600-2614.
- Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W., and Zhao, T., 2019. Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network, *Remote Sensing*, 11(15): 1774.
- Zhang, L., Wu, J., Fan, Y., Gao, H., and Shao, Y., 2020. An efficient building extraction method from high spatial resolution remote sensing images based on improved mask R-CNN, *Sensors*, 20(5): 1465.

# 應用深度學習於高解析衛星影像臺灣農業區建物分塊

劉良逸<sup>1</sup> 王驥魁<sup>2\*</sup> 黃安德<sup>3</sup>

## 摘要

臺灣的可耕地面積有限，清查建物的面積有助於了解土地利用的狀況。為了瞭解建物在臺灣農業區所佔的總面積，現有的做法之一是透過高解析衛星影像進行人工辨識，此法可以掌握建物的邊界、改善現地調查的不便。然而，卻需要大量人力資源的投入。過去的研究顯示，深度學習的方法可以有效地在高解析衛星影像進行建物分塊。因此，本研究使用 ENVINet5 深度學習模型及 Pléiades 彩色融合影像進行訓練，針對臺灣的農業區進行建物分塊。因為各地區的建物型態皆不相同，所以本研究使用九個不同的縣市的影像進行訓練，每張訓練影像的尺寸為 2500 像素× 2500 像素。模型的評估是透過驗證集中的像素以及分塊後的建物多邊形進行計算。前者結果顯示，經訓練的模型可以找出 84%的建物像素；後者計算了建物多邊形的數量，並將其與參考建物以 IoU (Intersection of Union) 做比較。成果顯示，該模型可以在影像上偵測且分塊 92%的建物，其 IoU 集中於 0.6 到 0.9 之間。該模型也以測試集做可轉移性試驗。另外，本研究提出了影像切圖與拼接的方法以處理大範圍的衛星影像。最後，我們將 ENVINet5 的成果輔助人工辨識建物，可以節省 7.3% 的時間成本。

**關鍵詞：**建物分塊、深度學習、高解析衛星影像

<sup>1</sup> 國立成功大學測量及空間資訊學系 碩士

<sup>2</sup> 國立成功大學測量及空間資訊學系 教授

<sup>3</sup> 普渡大學土木工程學系 碩士生

\* 通訊作者, E-mail: chikuei@ncku.edu.tw

收到日期：民國 111 年 01 月 14 日

修改日期：民國 111 年 02 月 25 日

接受日期：民國 111 年 03 月 04 日