

BUILDING SEGMENTATION IN AGRICULTURAL LAND USING HIGH RESOLUTION SATELLITE IMAGERY BASED ON DEEP LEARNING APPROACH

L. -Y. Liu ¹, C. -K. Wang ^{2*}

¹ Dept. of Geomatic, National Cheng Kung University, Tainan, Taiwan - david31009@gmail.com

² Dept. of Geomatic, National Cheng Kung University, Tainan, Taiwan - chikuei@ncku.edu.tw

KEY WORDS: Building Segmentation, Deep Learning, High Resolution Satellite Imagery.

ABSTRACT:

Understanding building area in agricultural land is important since arable land area in Taiwan is limited. One of the practical ways is manual digitization on high resolution satellite imagery, which can avoid field investigation and achieve satisfying results. However, such practice is tedious and labor intensive. Past researches have shown that deep learning methods are useful to segment buildings in different cities using satellite imagery. In this study, ENVINet5 model was trained and used to segment buildings from high resolution Pleiades pansharpened imagery. The training images (with the size of 2500 pixels × 2500 pixels) were randomly selected from 9 counties/cities to increase diversity since each county/city has different building patterns. The performance of ENVINet5 model reached 0.977, 0.814, 0.847, and 0.829 respectively on accuracy, precision, recall, and F1 score. Since evaluation by pixels can be difficult to show geometry of buildings, we evaluated the model by counting the number of inference building segments, which was post-processed from inference result of ENVINet5 trained model. Further analysis by counting the inference building segments is discussed in this study.

1. INTRODUCTION

Building segmentation in agricultural land is an important issue in Taiwan since arable land area can be influenced by building area. One of the practical ways to estimate building area in agricultural land is to digitize them manually on high resolution satellite imagery, which is rich in spatial information and helpful for building visual interpretation. However, manual digitization is time-consuming and labor intensive. Deep learning approaches have been applied to segment buildings automatically on satellite imagery in many studies (Boonpook et al., 2018; Maltezos et al., 2017; Vakalopoulou et al., 2015). Results on building segmentation can be different depending on various building patterns, and deep learning models (Zhang et al., 2020).

In this study, ENVINet5 (built in ENVI version 5.6, and ENVI Deep Learning version 1.1) is used to segment buildings in agricultural land in Taiwan from high resolution Pleiades imagery, which includes multispectral and panchromatic images. The resolution is 2 meters for colour and near-infrared bands (Figure 2a), and the resolution is 0.5 meter for panchromatic band (Figure 2b). ENVINet5 is an encoder-decoder fully convolutional network. Its architecture (Figure 1) is based on U-Net with some modifications on layers of convolution and the size of input and output. The input of ENVINet5 is patch with agricultural buildings. And the output is a probability map, where the pixel values range from 0 to 1. The architecture has the characteristic of U-Net, which can work on few training images and yield precise segmentations (Ronneberger et al., 2015).

ENVINet5 is special with four proprietary hyperparameters that can improve the performance of model. First, Class Weight introduces a biased selection of patches, so the model extracts patches that contain more feature pixels. Second, Patch Sampling Rate can control the density of sampling. Since feature pixels are often sparse comparing to background pixels, high density of sampling rate can generate more patches with more feature pixels.

Third, Loss Weight biases loss function to make more adjustment on identifying feature pixels. Fourth, Blur Distance helps the model to learn building borders by blurring the edges and decreasing the blur during training.

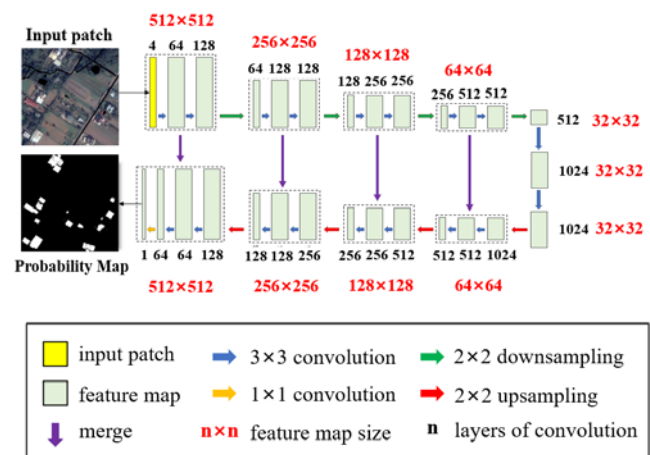


Figure 1. ENVINet5 architecture.

2. METHOD

2.1 Data Pre-processing

In this study, nearest-neighbor diffusion-based (NNDiffuse) pan sharpening technique was applied to fuse multispectral and panchromatic image since the algorithm can preserve sharp spatial features from panchromatic images and spectral information from multispectral images (Sun et al., 2013). Pansharpened image with high resolution (0.5 meters) and multispectral bands (R, G, B, NIR) was obtained after fusion. Next, Non-agricultural mask (Figure 2c), provided by Taiwan Agricultural Research Institute, was overlaid with pansharpened

* Corresponding author

image to extract agricultural land area. Image masking technique can restrict analysis to a subset region instead of using whole image scene (Kastens et al., 2005). The pansharpened image with only agricultural land is shown in Figure 2d. Then, the pixel value of non-agricultural mask in pansharpened image was set as the maximum value for the byte type of the image in case that agricultural land area has similar pixel value to non-agricultural mask.

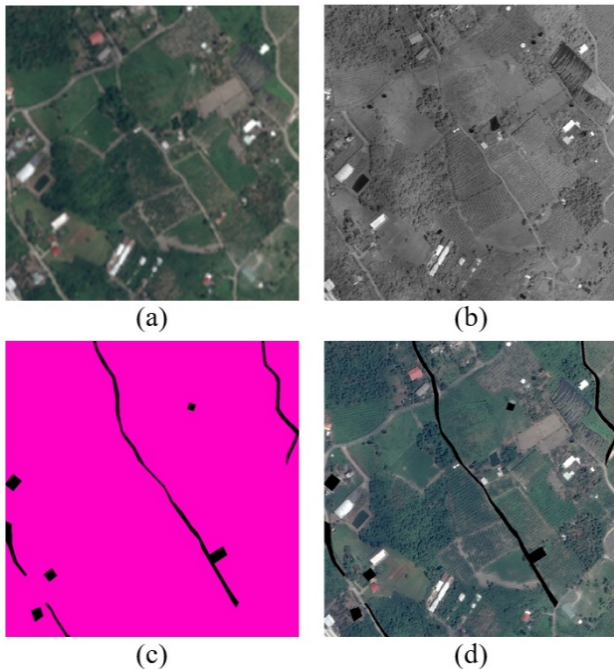


Figure 2. Data pre-processing. (a) multispectral image; (b) panchromatic image; (c) non-agricultural mask shown as black area; (d) pansharpened image with agricultural area.

2.2 Manual digitization

Since building patterns in agricultural land are complicated, manual digitization results can be unstable depending on

difference of human interpretation. Therefore, consistent digitization is necessary for our model to correctly identify and segment the buildings. In this study, each building pattern was digitized in a building polygon as follows. Firstly, shadow is excluded from digitization (cf. Figure 3a and 3f). Next, some buildings cut by the non-agricultural mask (Figure 3b) were kept several pixels inside the building for the digitization. Because the pixel value of the non-agricultural mask was set as the maximum value for the byte type of the image, the model can be confused if non-agricultural mask is digitized in the building polygons (Figure 3g). Then, adjacent buildings with no space in between (Figure 3c) or buildings in close proximity (Figure 3d) were digitized as one single large building polygon since it was challenging to digitize each building separately. Finally, when the buildings were occluded by vegetation (Figure 3e), the occlusion part is kept out for digitization (Figure 3j).

2.3 Model Training

The study area covered 9 counties/cities in Taiwan, from Miaoli to Taitung (Figure 4). Since each county/city has different building patterns, 500 sub-images (with the size of 2500 pixels \times 2500 pixels) were randomly selected (Figure 4) to increase diversity in the training data. The buildings in each sub-image were manually digitized. The 500 sub-images were randomly divided into training sets and validation sets with the proportion of 8:2.

ENVINet5 is trained using patch-based convolutional neural network. A patch is a certain region in sub-image (Figure 5a), and a batch is number of patches being trained during iterations (Figure 5b). In this study, the patch size was set as 512 \times 512 pixels, and the batch size was set as 64. During every epoch, entire 10,000 patches are trained batch by batch. The parameters are learned and updated every iteration. ENVINet5 model was trained on a workstation with NVIDIA GeForce RTX 2080 Ti GPU.

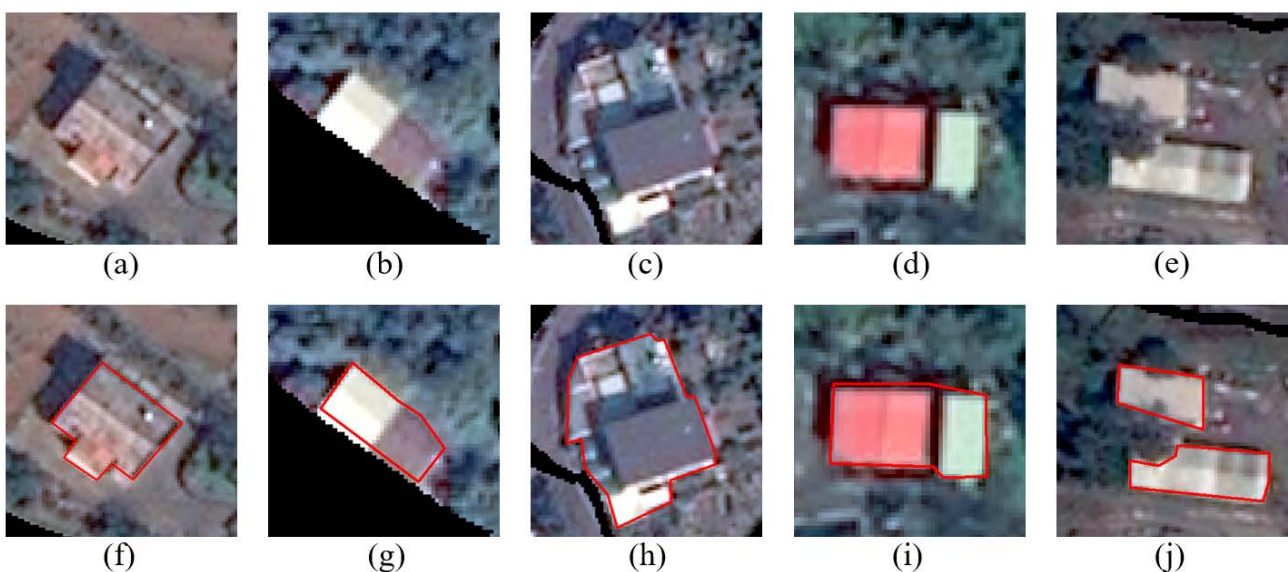


Figure 3. Examples of manual digitization results from different building patterns. (a) A building with shadow; (b) A building cut by non-agricultural mask; (c) Adjacent buildings; (d) Buildings in close proximity; (e) Buildings occluded by vegetation; (f)(g)(h)(i)(j) Manual digitization results of above building patterns.

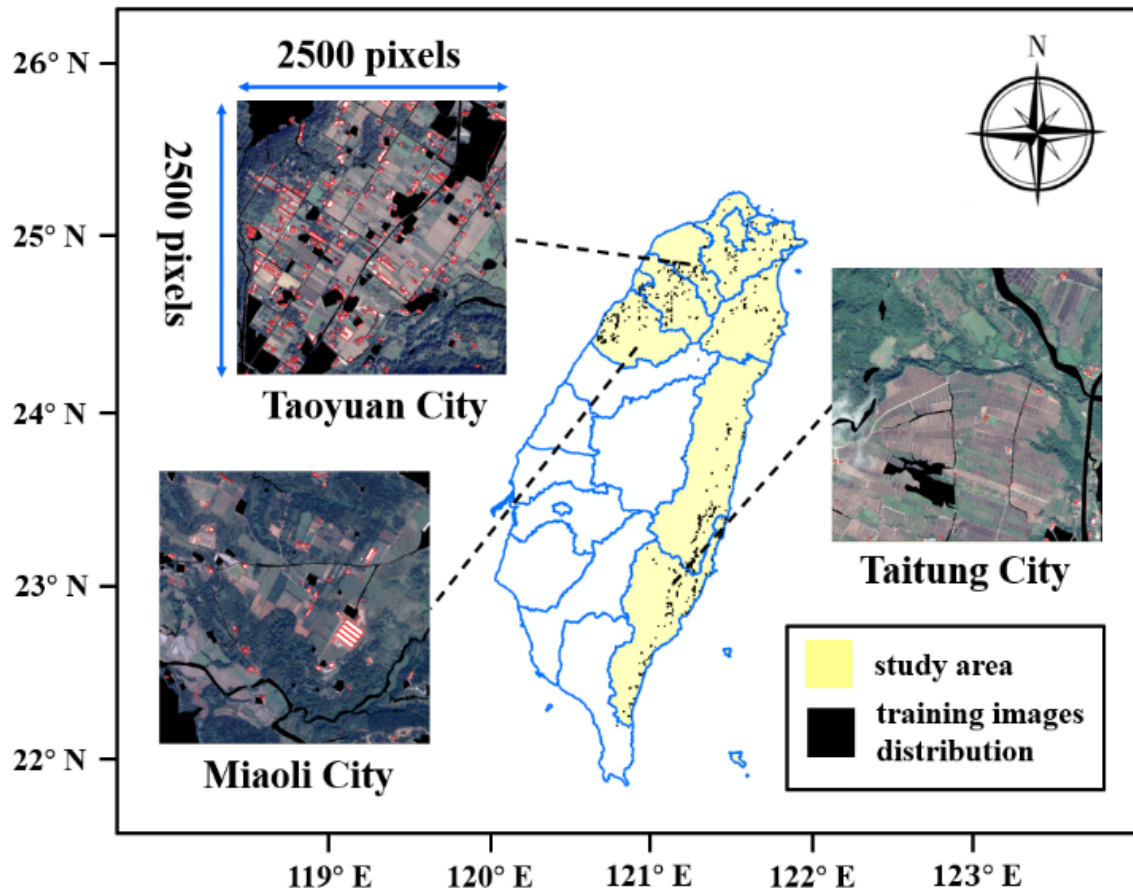


Figure 4. Study area and locations of training images.

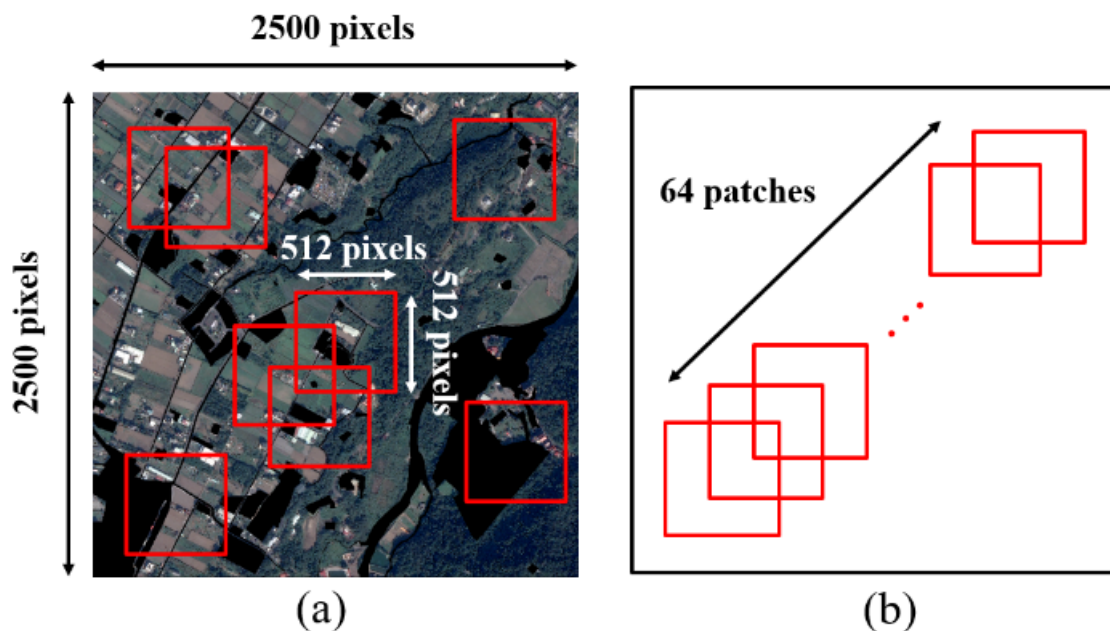


Figure 5. Illustration of patch and batch: (a) sub-image with several patches (in red frame); (b) a batch of 64 patches.

2.4 Model Evaluation

In this study, confusion matrix and assessment indices are used to evaluate the model. The assessment indices include accuracy, precision, recall, and F1 score. The model is evaluated by pixels

within each image using 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).

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

Table 1. Confusion matrix.

In assessment indices, accuracy is overall correctness including building pixels and background 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 et al., 2018). The equations of assessment indices are listed below:

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

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

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

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

2.5 Data post-processing

The inference result is a probability map from the output of a trained ENVINet5 model. Each pixel in probability map indicates the probability of being a building pixel. Higher probability is likely to be a building pixel. The values of pixels range from 0 to 1 in the form of floating-point numbers (Figure 6b). In this study, four post-processing steps are carried out as follows.

Step 1: Set the threshold for the probability map. The probability map is converted to a binary map with the probability threshold of 0.6 (Figure 6c). If a pixel value is more than or equal to 0.6 in probability, it is considered as a building pixel. Otherwise, it is a non-building pixel.

Step 2: Vectorize the binary map. The binary map is vectorized to obtain building segments with outlines. (cf. Figure 6c and 6d)

Step 3: Remove the overlapping polygons. Overlapping polygons were generated from the vectorization of non-building pixels as shown in figure 6d.

Step 4: Fill up the polygon holes. Inference building segment was left with several holes after removing the overlapping polygons (Figure 6e). The appearance of the polygon holes is noise mainly caused by shadow or occlusions, but they are still part of the building. The polygon holes were filled up using threshold of area. If the area of each hole is 25 percent less than the whole inference building segment, the hole will be filled up.

At last, the inference building segment is obtained after four post-processing steps as shown in figure 6f.

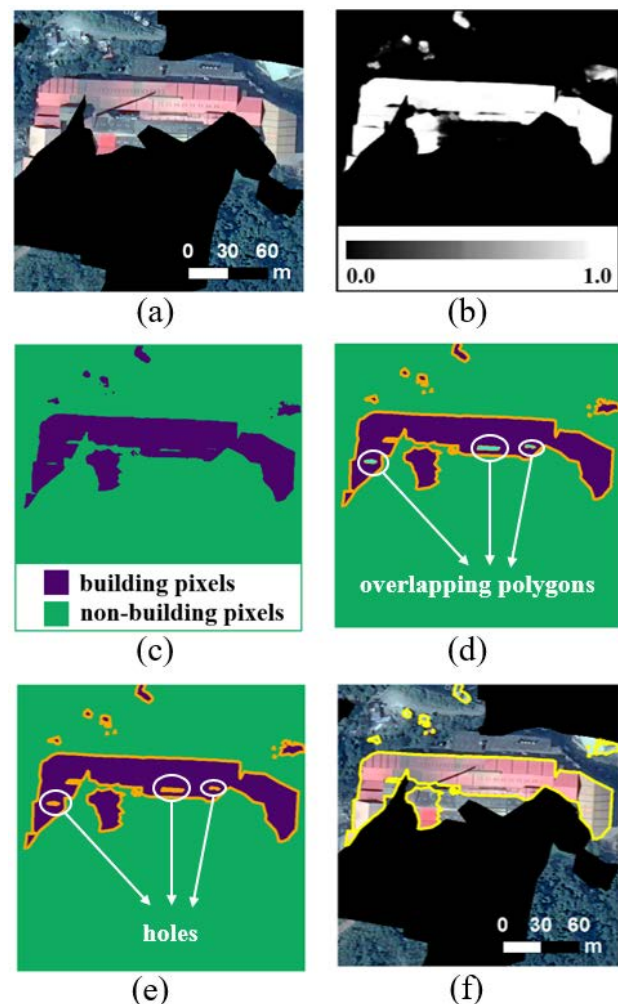


Figure 6. An example of adjacent buildings in post-processing. (a) Buildings shown in pansharpened image. (b) Probability map from ENVINet5. Brighter pixels denote higher probability. (c) Binary map with building and non-building pixels after applying the probability threshold of 0.6. (d) Vectorization result from binary map, where overlapping polygons were due to non-building pixels. (e) Inference building segment with holes after removing overlapping polygons. (f) Inference building segment overlaid on the pansharpened image after filling holes in (e).

2.6 Evaluation of inference building segments

In this section, inference building segments are analysed by counting their numbers since evaluation by pixels can be difficult to display the geometry of each building. In addition, a large number of TNs case inflated the accuracy of the inference result. In order to conduct the accuracy assessment based on counting the numbers of inference building segments, the following cases were considered:

Case 1. Omission building. Omission building is defined as the building our model miss to predict and find out (Figure 9a).

Case 2. Commission building. Commission building is defined as the building our model overpredict (Figure 9b).

Case 3. One-to-one correspondence. One inference building segment overlap with one reference building polygon (Figure 9c).

Case 4. Many-to-one correspondence. Many inference building segments overlap with one reference building polygon (Figure 9d).

Case 5. One-to-many correspondence. One inference building segment overlap with many reference building polygons (Figure 9e).

Case 6. Many-to-many correspondence. Many inference building segments overlap with many reference building polygons (Figure 9f).

For easy understanding, case 1 and case 2 are the non-overlapping situations, and case3, 4, 5, 6 are the overlapping situations in this study.

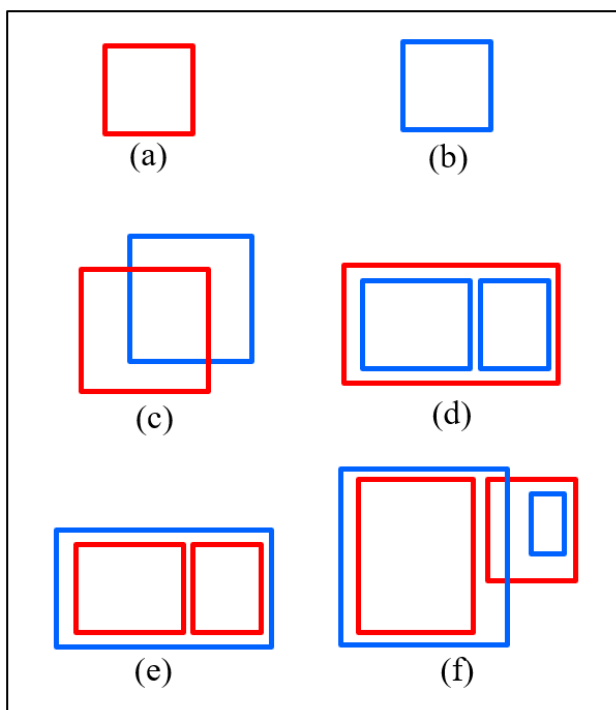


Figure 9. Six different cases for the non-overlapping and the overlapping situations. Inference building segments are denoted as blue and reference building polygons are denoted as red. (a) Omission building; (b) Commission building; (c) One-to-one correspondence; (d) Many-to-one correspondence; (e) One-to-many correspondence; (f) Many-to-many correspondence.

The analysis starts with the calculating the number of omission buildings and commission buildings by overlaying inference building segments and reference building polygons. Then, omission error rate (OER) and commission error rate (CER) were calculated using the formula below. Note that the overlapping cases are considered as correct predictions as long as reference building polygons and inference building segments are overlapped.

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

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

Through calculating OER and CER, it is obvious to understand whether our model can localize and segment all buildings on high resolution satellite images.

Next, IoU is utilized for the evaluation of the overlapping cases (case 3 to case 6) because not every of them are segmented properly. IoU, intersection over union, is the most common metric to compare similarity between two arbitrary shapes (Rezatofghi et al., 2019). Values of IoU range from 0 to 1 in floating point numbers. If the value is closed to 1, it represents that two shapes have higher similarity. The formula of IoU is shown below.

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

In this study, IoU is calculated for reference building polygons and inference building segments. Since not every overlapping case was one-to-one correspondence, several IoUs were calculated. Then, the largest IoU was kept to ensure one inference building segment overlap with one reference building polygon. Take many-to-one correspondence as example, several IoUs will be calculated since many inference buildings segments overlap with one reference building. Then, inference building segment with the largest IoU was kept, and the rest of the overlapping inference building segments were removed. Similar process was carried out in the case of one-to-many correspondence. The difference is reference building polygons with the largest IoU was kept, and the rest of the overlapping reference building polygons were removed. As for the case of many-to-many correspondence, both processes were executed. Overlapping inference building with the largest IoU was selected first, followed by overlapping reference building polygons with the largest IoU. The above analysis with six different cases of how OER and CER are calculated and how the selection of largest IoU works are shown in the pseudo code below.

*RB stands for reference building polygon
*IB stands for inference building segment

Input: shapefiles of RBs and IBs

Output: calculation of OER, CER, and a list of IoU

```

1. begin
2. for each sub-image in validation set do
3.   Overlay two shapefiles
4.   Count number of RBs
5.   Count omission buildings and commission buildings
6.   Find overlapping cases of two shapefiles
7.   for each overlapping case do
8.     if IB > 1 and RB = 1 then
9.       Compute IoU for every IB
10.      Keep one IB with the largest IoU
11.     end if
12.     if IB = 1 and RB > 1 then
13.       Compute IoU for every RB
14.       Keep one RB with the largest IoU
15.     end if
16.     if IB > 1 and RB > 1 then
17.       for every RB do
18.         Compute IoU for every IB
19.         Keep IB with the largest IoU
20.       end for
21.       for every IB do
22.         Compute IoU for every RB
23.         Keep RB with the largest IoU
24.       end for
25.     end if
26.   end for
27. end for
28. Calculate OER and CER
29. List IoU
30. end

```

3. RESULTS AND DISCUSSIONS

It took 26 hours to train ENVINet5. To avoid overfitting problem, the training process was stopped at 100 epochs since training loss and testing loss tend to diverge. The training process is shown in figure 7. Performance of the model is evaluated by pixels with validation sets. The values of assessment indices are 0.977, 0.814, 0.847, and 0.829 respectively for accuracy, precision, recall, and F1 score (Figure 8).

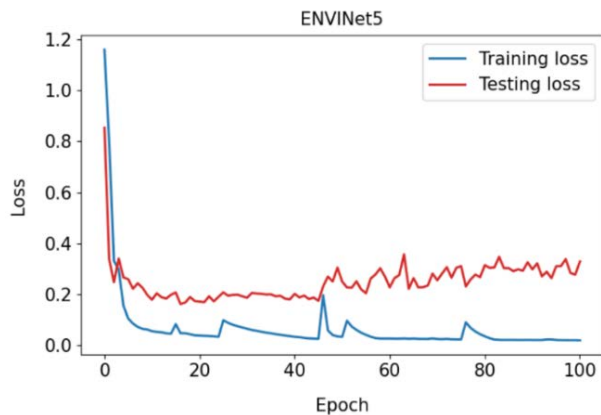


Figure 7. Training process of ENVINet5.

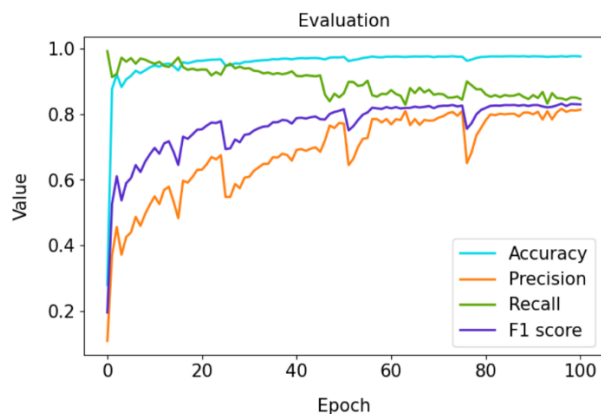


Figure 8. Evaluation of ENVINet5 using validation sets with accuracy, precision, recall, and F1 score indices.

The result of OER and CER is shown in table 2. Among 100 sub-images in validation sets, 5727 reference building polygons are digitized manually. Our model misses to predict 484 buildings and overpredicts 5044 buildings. OER and CER are 8.45% and 88.07% respectively.

	numbers of buildings	error rate
reference buildings	5727	NA
omission buildings	484	8.45 %
commission buildings	5044	88.07%

Table 2. Statistics of omission and commission buildings with the calculation of OER and CER.

OER and CER varied with the different settings of probability threshold from inference results. We intentionally kept OER low with the expense of high CER value because our goal is not to miss any buildings on the high resolution satellite image.

After applying selection of the largest IoU within all the overlapping cases, 4770 buildings are kept as one-to-one correspondence. The IoUs of them are plotted as histogram shown in figure 10. In visual interpretation, we considered IoU more than or equal to 0.6 as acceptable predictions. For further understanding of similarity that different IoUs display, figure 11 illustrates examples of various IoUs with inference and reference results.

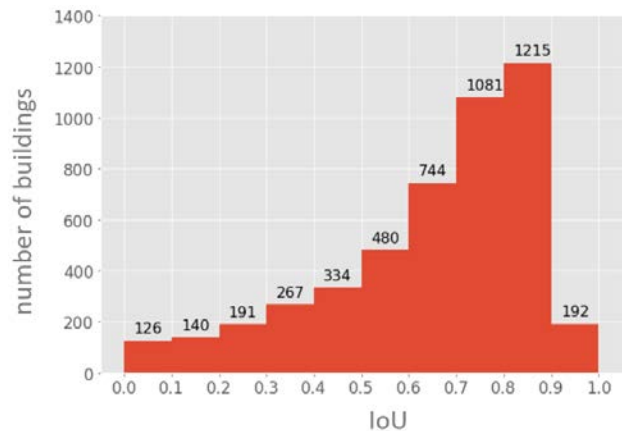


Figure 10. Histogram for IoU, where number of buildings were reported for each IoU bin.

CONCLUSION

In this study, we demonstrated the feasibility of deep learning approach to segment buildings automatically in agricultural land using high resolution Pleiades pansharpened imagery. In order to cover various building patterns in different districts, ENVINet5 was trained on random 500 sub-images (with the size of 2500 pixels \times 2500 pixels) from 9 cities/counties around Taiwan. To maintain the quality of manual digitization results, each building pattern is digitized in a consistent building polygon. The performance of our model was evaluated by pixels within sub-images using validation sets. The result reached 0.977, 0.814, 0.847, and 0.829 respectively on accuracy, precision, recall, and F1 score. However, evaluation by pixels can be difficult to show the geometry of the buildings. Therefore, we evaluate our model by counting the number of the inference building segments, which are post-processed from the inference result of ENVINet5. Post-processing includes four steps, which are 1. setting the threshold for probability map, 2. vectorizing the binary map, 3. removing the overlapping polygons, and 4. filling up the polygon holes. Next, the inference building segments were analysed with OER and CER to check the reliability of our model making correct predictions. Finally, the overlapping cases were analysed using IoU. To ensure every inference building segment correspond to single reference building polygon, the largest IoU was kept. The majority of building segments have IoUs over 0.6, which are seen as acceptable for visual interpretation.




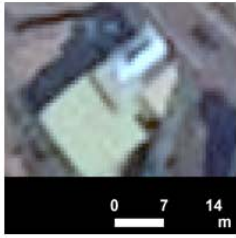
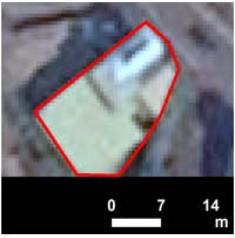
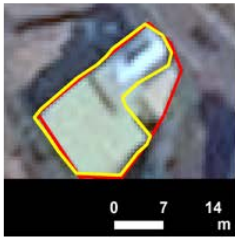
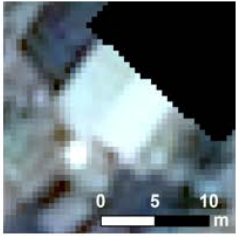
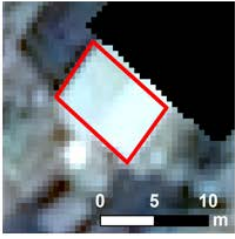
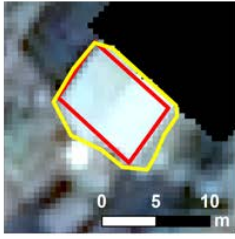





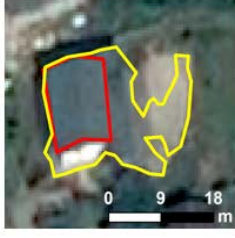



Values of IoU	Pansharpened images	Reference building polygons	Reference building polygons and inference building segments
IoU: 0.96			
IoU: 0.77			
IoU: 0.68			
IoU: 0.48			
IoU: 0.37			
IoU: 0.15			

Figure 11. Examples of inference results with various IoUs. Reference building polygons are denoted as red, and inference building segments are denoted as yellow. The first column is values of IoU; the second column is the pansharpened images; the third column is the reference building polygons digitized manually; the last column is the overlay of the reference building polygons and the inference building segments

REFERENCES

- Boonpook, W., Tan, Y., Ye, Y., Torteeka, P., Torsri, K., Dong, S., 2018. A Deep learning approach on building detection from unmanned aerial vehicle-based images in riverbank monitoring. *Sensors* 2018, 18(11), 3921. <https://doi.org/10.3390/s18113921>
- ENVI Development Team, 2020. Exelis Visual Information Solutions software, Version 5.6. L3HARRIS Geospatial Corporation. www.l3harrisgeospatial.com (May 2020)
- Kastens, J. H., Kastens, T. L., Kastens, D. L. A., Price, K. P., Martinko, E. A., Lee, R. Y., 2005. Image masking for crop yield forecasting using AVHRR NDVI time series imagery. *Remote Sensing of Environment*, 341-356. <https://doi.org/10.1016/j.rse.2005.09.010>
- Maltezos, E., Doulamis, N., Doulamis, A., 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). <https://doi.org/10.1117/1.JRS.11.042620>
- Prathap, G., Afanasyev, I., 2018. Deep learning approach for building detection in satellite multispectral imagery. *2018 International Conference on Intelligent Systems (IS)*, 461-465. <https://doi.org/10.1109/IS.2018.8710471>
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Retrieved from <http://arxiv.org/abs/1505.04597>
- Rezatofghi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese S., 2019. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. Retrieved from <https://arxiv.org/abs/1902.09630>
- Sun, W., Chen, B., Messinger, D., 2013. Nearest-neighbor diffusion-based pan-sharpening algorithm for spectral images. *Optical Engineering*, 53(1). <https://doi.org/10.1117/1.OE.53.1.013107>
- Vakalopoulou, M., Karantzas, K., Komodakis, N., Paragios, N., 2015. Building detection in very high resolution multispectral data with deep learning features. *2015 IEEE international Geoscience and Remote Sensing Symposium (IGARSS)*, 1873-1876. <https://doi.org/10.1109/IGARSS.2015.7326158>
- Zhang, L., Wu, J., Fan, Y., Gao, H., Shao, Y., 2020. An efficient building extraction method from high spatial resolution remote sensing images based on improved Mask R-CNN. *Sensors* 20, no. 5: 1465. <https://doi.org/10.3390/s20051465>