APPLYING DEEP LEARNING TO AUTOMATICALLY DETECT BUILDINGS FROM TRUE ORTHOIMAGES

Hao Chan Hsieh [1*] Chia-Chang Hsu[2] Shih-Hong Chio[3]

[1*]Graduate Student, [2]Master, [3]Professor,

Department of Land Economics, National Chengchi University

No.64, Sec.2, ZhiNan Road,

Wenshan District, Taipei 11605, Taiwan

[1*] Email:: love0983906400@gmail.com ; [2] Email: a0960088227@gmail.com, [3]Email: chio0119@gmail.com

KEY WORDS: Building detection, Deep learning, Digital surface model, Digital height model

ABSTRACT **:** This study selected Soczi Island in Taipei City as the experimental area to investigate the application of deep learning in automatically detecting buildings in aerial true orthoimages. Different time period aerial color true orthoimages were used to conduct deep learning MS-FCN (Multi-Scale Fully Convolutional Network) model building detection. The study incorporated DSM and DHM data to investigate the benefits of elevation on the model. The results demonstrated that compared to using only true orthoimages, the inclusion of elevation information from DSM and DHM could enhance the building recognition capabilities of the model, achieving F1-scores of 87.16% and 87.65%, respectively. According to the findings, applying deep learning in conjunction with high-resolution true orthoimages to aid building recognition is feasible.

1.**INTRODUCTION**

In recent years, inspired by the structure of the human brain and aided by advancements in computer technology, deep learning has attracted increasing attention. Many studies in image recognition have chosen Convolutional Neural Networks (CNNs) as the foundational architecture, achieving impressive recognition results. Previous literature has also highlighted the effectiveness of applying CNNs to image recognition or classification tasks, particularly in grid-based data. (Marmanis et al., 2016; Zhu et al., 2017; Duan et al., 2018)

However, due to the fact that the final layer of a CNN uses fully connected layers to obtain classification prediction probabilities, these probabilities are one-dimensional and simultaneously lose spatial information from the input. To address this issue, Long et al. introduced the concept of a Fully Convolutional Network (FCN). FCN replaces one or more fully connected layers with convolutional layers, performing upsampling on the feature maps to obtain feature maps of the same resolution as the input image. Furthermore, FCN can accommodate input images of arbitrary sizes and perform pixel-wise detection by processing each individual pixel.

In summary, this study uses true orthoimages as a foundation, combining dense matching Digital Surface Models (DSMs) and Digital Height Models (DHMs) for building recognition using MS-FCN (Multi-Scale Fully Convolutional Network) deep learning model(Zeng and Zhu, 2018).

2.STUDY DATA AND METHODS

2.1 Study Data

The study area is Soczi Island in Taipei City, located at the confluence of the Keelung River and Tamsui River, covering an approximate area of 300 hectares. The study collected original aerial images, interior and exterior orientation parameters, and 1/1000 topographic maps for Soczi Island in the years 2007 and 2021. The relevant data about aerial images is tabulated as Table 1. The aerial images in 2007 were collected by airplane that were flown by flying 1706 meters above ground along 6 flight lines, comprising 4 east-west lines and 2 north-south lines with a forward and side overlap of 70% and 50% respectively, resulting in a ground sampling resolution (GSD) of 12.6 cm/pixel. The aerial images in 2021 were collected by airplane that were flown by flying height 1709 meters above ground along 6 flight lines, comprising 3 east-west lines and 3 north-south lines with a forward and side overlap of 80% and 30% respectively, resulting in an approximate GSD of 9.7 cm/pixel.

Table 1. Study data

Year	Camera	focal length	pixel size	image size	Total images
2007	DMC	120 mm	12μm	7680 × 13824	101
2021	UltraCam XP	100 mm	6μm	11310 × 17310	86

In both years, Pix4D Mapper was used to generate dense point clouds for producing Digital Surface Models (DSMs) and true orthoimages. These outputs maintained the same resolution as the original images. Figure 1 shows the DSMs of Soczi Island in 2007 and 2021 after manual editing to eliminate outliers. Since the elevation derived from the collected exterior orientation parameters was based on ellipsoidal height, subsequently, a geoid undulation correction was applied to convert them into orthometric height. This study employed the geoid undulation calculation platform provided by the Ministry of the Interior's Land Administration, Taiwan, to assess the geoid undulation. An average geoid undulation was approximately 20 meters.

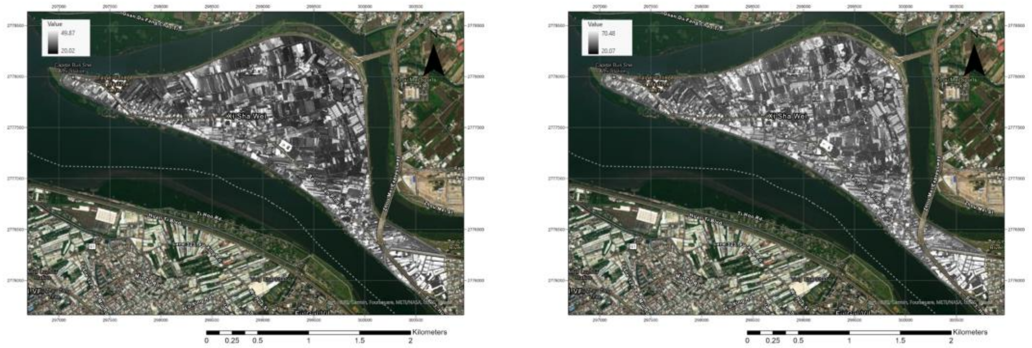


Figure 1. DSMs of Soczi Island in 2007 (left) and 2021 (right)

Because there were no significant changes in elevation within the Soczi Island region, this study adopted the Digital Elevation Model (DEM) derived from the refined 2021 DSM of the Soczi Island area as its foundation. By subtracting DEM from DSMs in 2007 and 2021, Digital Height Models (DHMs) for 2007 and 2021could be obtained. Figure 2 shows the DHMs of Soczi Island in 2007 and 2021. By integrating the true orthoimages, DSMs, and DHMs, the study aimed to explore detection accuracy using MS-FC deep learning model.

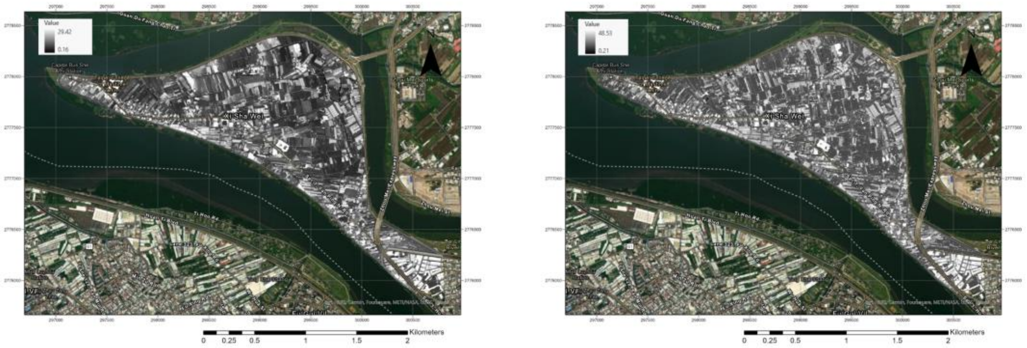


Figure 2. DHMs of Soczi Island in 2007 (left) and 2021 (right)

In this study, the image generated from the building vector layer of the 1:1000 topographic map was used as a reference to generate the label data for the model test. For this study, the true orthoimages, true orthopimages merged with DSMs and true orthoimages merged with DHMs for the Soczi Island area in 2007 were used as training data. The test data consisted of images of the same type from the year 2021.

2.2 **Study Methods**

This study employs true orthoimages, true orthoimages merged with DSMs and true orthoimages merged with DHMs as input data to investigate the building recognition in the Soczi Island area of Taipei City. The study flowchart is shown in Figure 3.

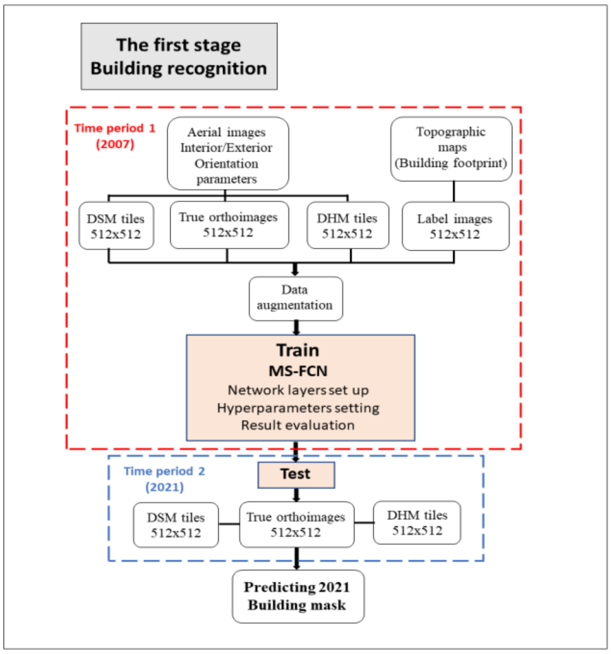


Figure 3. Building Recognition Study Flowchart

2.2.1 Data Preprocessing: Data preprocessing includes label data generation, elevation data normalization, and data augmentation. Each step is described as follows:

(1) Label data generation

Label data was generated by overlaying the vector data of permanent houses, temporary structures, and open structures from the 1/1000 topographic map of Soczi Island onto true orthoimage. Three types of building vector data were converted from vector data to image raster data, called building image, by using ArcGIS Pro software with same GSD of true orthoimage. The building image consists of value 0 and 255, where 0 represents non-building pixel and 255 represents building pixels. Due to the generation of true orthoimage by Pix4D Mapper software with only the interior and exterior orientation parameters as initial values without using any control points, the systematic shift occurs. To ensure proper alignment between the true orthoimage and the building image, this study manually added control points based on visually apparent and reliable building corners for image registration. Finally, the processed building image was cropped to a size of 512 × 512 pixels, called building label image, corresponding to the same dimension of their corresponding true orthoimage, called building label image data, as depicted in Figure 4.

(a) True orthoimage (b) Building label image

Figure 4. Building Label Data

(2) Elevation data normalization

Elevation data normalization includes the normalization of the DSMs and DHMs. The elevation values were normalized to a range from 0 to 255, as seen in Figures 5 and 6. The processed DSM and DHM images were also cropped into a size of 512 × 512 pixels, called DSM and DHM label images.

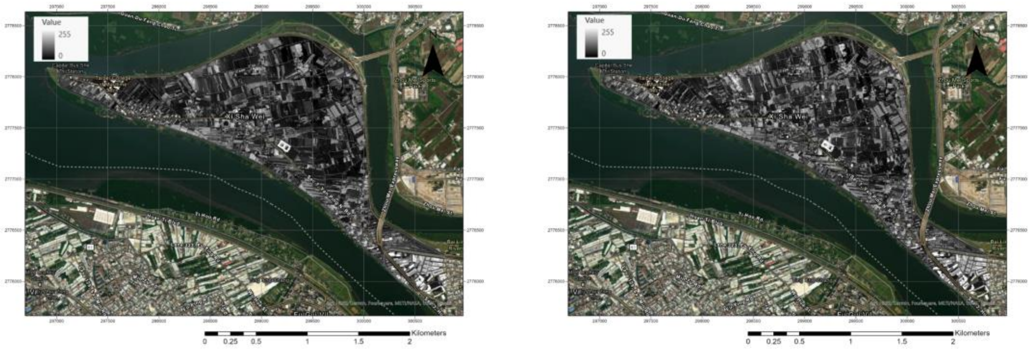


Figure 5. Normalized DSM Image Data in 2007(left) and 2021(right)

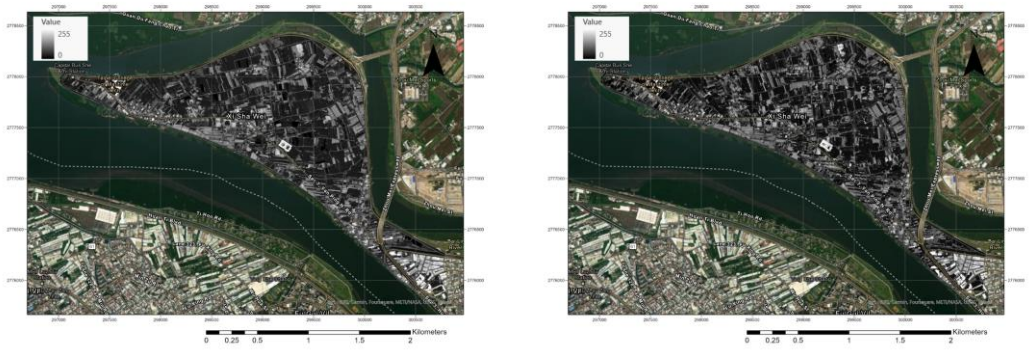


Figure 6. Normalized DHM Image Data in 2007(left) and 2021(right)

(3) Data augmentation

To mitigate the issue of insufficient training samples leading to subpar model performance, this study employs data augmentation techniques to generate new training samples from existing data, aiming to prevent overfitting and enhance the quality of model training. Initially, the generated true orthoimage of 36467 × 52958 pixels and building image in 2017 was cropped into 512 × 512 pixels image segments. Three common spatial transformation methods involving geometric aspects of the image are chosen for data augmentation: horizontal flipping, vertical flipping, and transposition, as shown in Figure 7. Through spatial transformations based on image geometry, not only is it possible to increase the number of training samples, but it also allows the segmentation targets to appear at different positions within the images to enhance the model's learning capabilities. When using geometric transformations for data augmentation, it's crucial to apply the same transformations to both the label data and the corresponding elevation information images. This ensures that the alignment between the input data and the labels is maintained throughout the augmentation process.

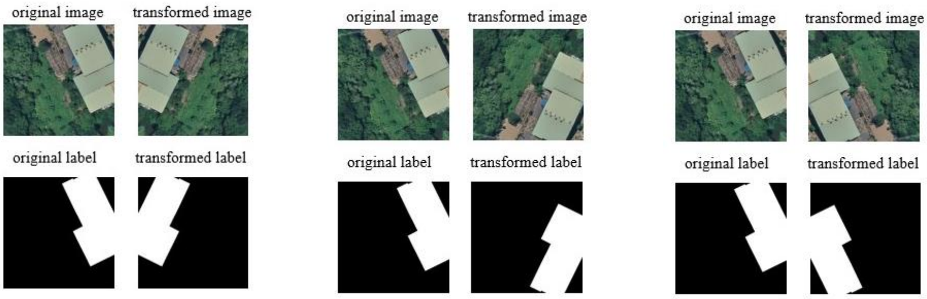


Figure 7. Left: Horizontal flipping, Middle: Vertical flipping, Right: Transposition

2.2.2 Deep Learning Network Model: This study refers to the model proposed by Zeng and Zhu (2018)for human, vehicle, and background segmentation, adopting the Multi-Scale Fully Convolutional Network (MS-FCN) to aid in building recognition, as shown in Figure 8.

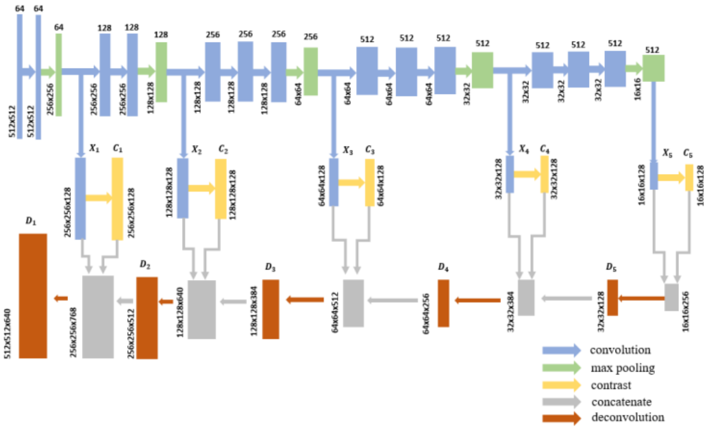


Figure 8. MS-FCN Building Recognition Model Architecture(Zeng and Zhu, 2018)

The model architecture offers several advantages: (He et al., 2017；Zeng and Zhu, 2018)

1.Direct FCN design: The model is designed based on the design principles of Fully Convolutional Networks (FCN), enabling the direct application of convolutional neural networks to pixel-level segmentation tasks.

2.Multi-Scale fusion: The utilization of multi-scale fusion enhances segmentation accuracy and effectively handles objects of varying sizes.

3.Feature utilization: The architecture maintains feature maps of different resolutions for input data at the same resolution. This approach fully utilizes image information from each layer of features and promotes efficient sharing of feature extraction, making it adaptable for various image segmentation tasks.

By leveraging these features, the MS-FCN model proves to be effective in the task of building recognition, aligning with the objectives of this study. Then, Focal Loss function introduced by Lin et al. (2017), along with Precision, Recall, F1 Score, and Intersection over Union (IoU) metrics derived from confusion matrix calculations are employed to evaluate the discrepancies between predicted and ground truth. Focal Loss function adjusts the weights of training samples to focus more on challenging-to-classify samples during the training process and can mitigate class imbalance issues. F1 Score is computed using Precision and Recall and ranges between 0 and 1, with higher values indicating better detection accuracy by the model. IoU measures the similarity between two sets, and a higher IoU indicates a higher similarity between the model's detection results and the ground truth. These metrics collectively provide a comprehensive assessment of the model's prediction quality.

3.STUDY DATA

The tests utilized image data produced from true orthoimages, DSM, and DHM of Soczi Island in 2007 with a GSD of 12.6 cm. The generated true orthoimage size was 36467 × 52958 pixels, and it was cropped into 512 × 512 pixels segments as input for the deep learning model.

For Test 1, true orthoimage was used as input images for the model. After data augmentation, the training dataset was split into an 8:2 ratio, with 4928 images (512 × 512 pixels) for training and 1232 images for validation. In Test 2, true orthoimage was merged with DSM image to create 4-band image data. In Test 3, true orthoimage was merged with DHM image to produce 4-band image data. The test data consisted of relevant image data from Soczi Island in 2021. The generated true orthoimage size was 11310 × 17310 pixels, and it was cropped into image patch of 512 × 512 pixels as input for the deep learning model for test. The test dataset was 1920 images (512 × 512 pixels).

3.1 Deep learning model parameter setting

In this study, the building recognition study utilized the MS-FCN architecture. The convolutional process involved utilizing VGG-16 as the backbone of the MS-FCN network, combining the strengths of both architectures to extract building features effectively. After the convolutional layers, max-pooling layer was used to retain essential features within localized regions. A contrast layer was incorporated to enhance feature extraction, particularly in complex backgrounds. Additionally, the reverse convolution technique was used to restore the image dimensions. The final classification layer utilized the sigmoid function, commonly used in binary classification tasks.

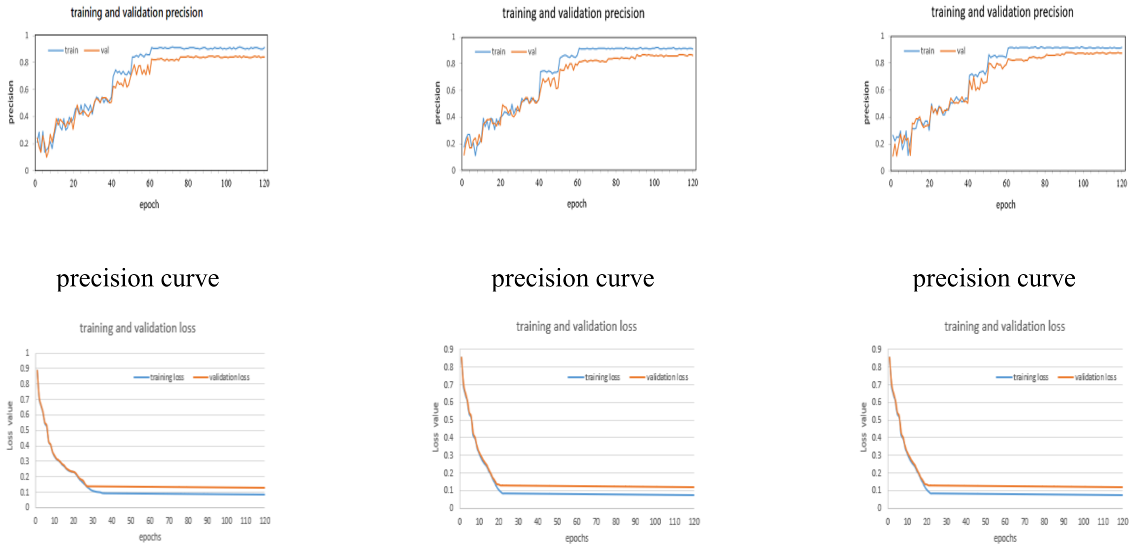
During the training of the deep learning model, the input image information can be influenced by the background, resulting in the inability to effectively segment building information. To address class imbalance issues, Focal Loss is commonly used (Lin et al., 2017). Throughout the training process, a learning rate of 0.0001 was set for the deep learning model. The optimization method chosen was the widely used Adam optimizer. Employing mini-batch training has several advantages, such as reducing training time, alleviating GPU memory stress, mitigating overfitting, enhancing generalization, and consequently improving accuracy on the test dataset (Keskar et al., 2016; Master and Luschi, 2018). Therefore, this study set the batch size to 2 for the deep learning model. The training was conducted over 120 epochs, utilizing pre-trained network weights from VGG-16 as the initial weights.

During training, the model's performance was monitored by computing the loss value on the validation dataset. If there was no reduction in the validation dataset's loss over ten consecutive epochs, training was early-stopped to prevent unnecessary computations. This approach aims to optimize the training process and improve the model's effectiveness.

3.2 The analysis of the experimental results

In terms of training times, Test 1 required approximately 3 hours, while Test 2 and 3 took around 6.5 hours each. Figure 9 shows the precision and loss curves for model training and validation for Test1, 2, and 3. Figure 10 illustrates the recognition results the entire labeled image along with the outcomes from Test 1 to 3. Additionally, Figure 11 provides a detailed view of some exemplary results. The validation and testing accuracy results are summarized in Table 2. From the test outcomes, several points can be made:

A. The precision and recall of the three test results were all above 83%, with F1-scores exceeding 85%, and IoU values surpassing 80%. This indicates that utilizing deep learning for intelligent building recognition yields promising predictive capabilities.



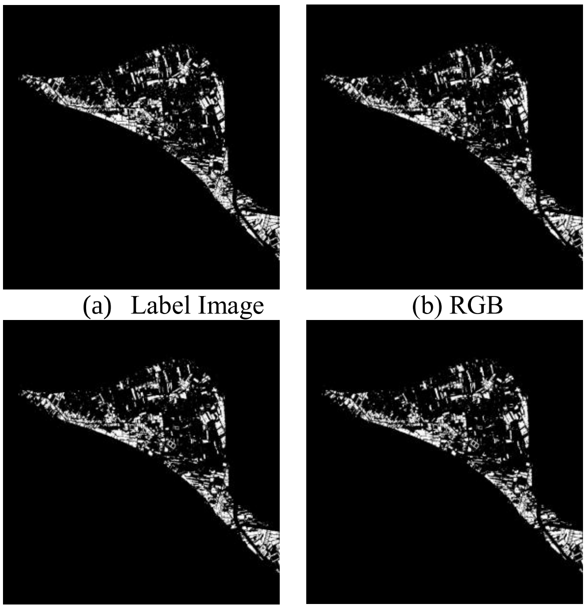
precision curve precision curve precision curve

loss curve
(a) Test1

loss curve
(b) Test2

loss curve
(c) Test3

Figure 9. Precision curve and loss curve of building recognition tests



(a) Label Image (b) RGB

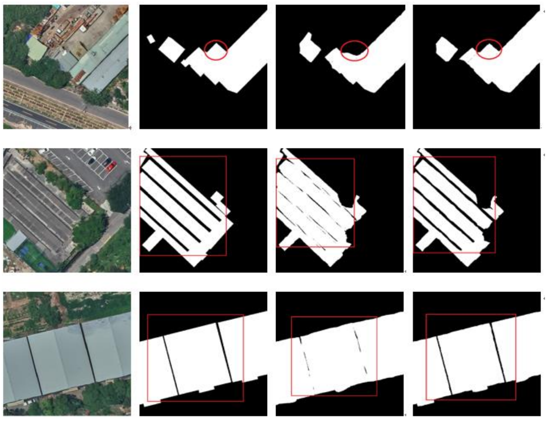
(c) RGB+DSM (d) RGB+DHM

Figure 10. Building Recognition Results

B. The results indicate that the deep learning model benefited from the elevation information provided by DSM and DHM. Both tests (Test 2: RGB+DSM and Test 3 : RGB+DHM) outperformed Test 1 (RGB only) in terms of accuracy metrics. In Test 2, precision, recall, and IoU was increased by 1.71%, 1.66%, and 1.08% respectively compared to Test 1. In Test 3, precision, recall, and IoU increased by 2.46%, 1.87%, and 1.19% respectively compared to Test 1.

C. In Figure 11, the first and second rows compare the detection results of building roof colors with similar colors of grounds or roads within the red boxes. It's evident that integrating RGB with DHM imagery (see the fourth columns) effectively enhances building detection due to the rich elevation information provided by DHM. In the third row within the red box, buildings with smaller spacing are accurately recognized due to the inclusion of DHM's elevation information (also see the fourth columns).

D. From Table 2, the results demonstrate that Test 3 (RGB+DHM) outperforms Test 2 (RGB+DSM) in terms of precision, recall, F1-score, and IoU, with improvements of 0.75%, 0.21%, 0.49%, and 0.11% respectively. This shows that DHM provides more accurate building elevation information compared to DSM.



(a) Original Image (b) Label (c)RGB (d)RGB+DHM

Figure 11. Results Showing Using Only RGB Images and Using RGB+DHM Images

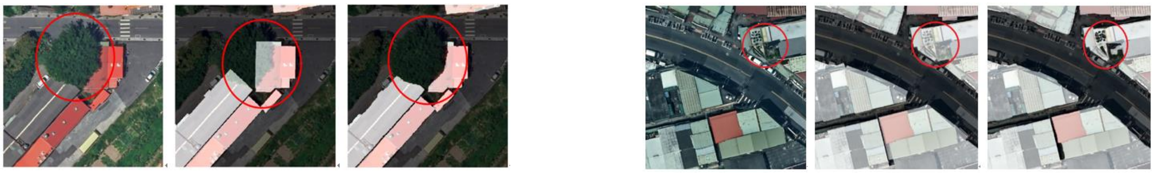
Table 2. Deep Learning Building Recognition Verification/ Test Accuracy Results

Verification/ Test Data	RGB	RGB+DSM	RGB+DHM
precision	84.67%/83.46%	86.23%/85.17%	87.08%/85.92%
recall	88.62%/87.58%	90.36%/89.24%	90.63%/89.45%
F1-score	86.60%/85.57%	88.25%/87.16%	88.82%/87.65%
IoU	81.34%/80.27%	82.65%/81.35%	82.84%/81.46%

Based on the analysis of building recognition from Figure 10, it was observed that the deep learning model still encountered situations of omission or misclassification due to factors like occlusion, shadows, or issues with the quality of the true orthoimages during the learning process. For the purpose of visualization, adjustments were made to the labeled images and recognition outcomes. These encountered situations are further explained below.

(1) Tree occlusion or Vegetation on the rooftop

Deep learning is image-based and relies on learning features from images to automatically segment objects. However, in reality, many building areas are obstructed by objects such as trees, which poses challenges to accurate recognition. In this study, when identifying buildings, only visible building structures could be detected. In cases where obstructions by trees occur, within the red circle shown in the left of Figure 12, the model cannot accurately recognize the complete structure of the building. Due to the presence of other plants on the rooftops of buildings, the model is unable to recognize the complete structures of the buildings, as shown in the right of Figure 12.



(a)image (b) label (c)predict

(a)image (b) label (c)predict

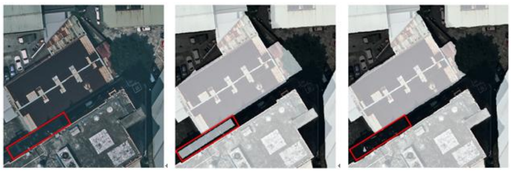
Figure 12. Diagram of buildings obstructed by trees (left) and vegetation on the rooftop (right).

(2) Shadow over buildings

Due to buildings being obscured by shadows, the model is unable to discern the structure of the buildings, as shown in Figure 13.

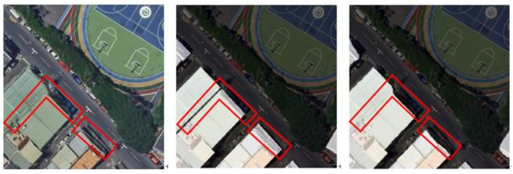
(3) Quality of true orthoimages

The true orthoimages utilized in this study were generated using the Pix4D Mapper software through automated aerial triangulation and adjustment. This process offers advantages such as speed and minimal manual intervention. However, certain areas in the images still exhibit distortion and deformation issues, as depicted in Figure 14.



(a)image (b) label (c)predict

Figure 13. Illustration of shadow over buildings

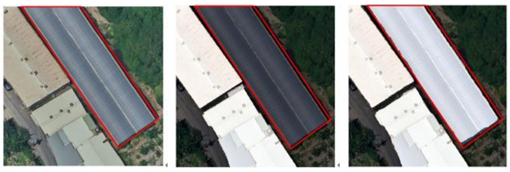


(a)image (b) label (c)predict

Figure 14. Quality issues with true orthoimages

(5) Buildings not labeled on the topographic map.

Unmarked buildings present in the used label data can still be recognized by the model, as depicted in Figure 15.



(a)image (b) label (c)predict

Figure 15. Unmarked building in label data

4 CONCLUSIONS AND RECOMMENDATIONS

4.1 Conclusions

This study used the MS-FCN deep learning model for building recognition, integrating high-resolution true orthoimage with DSM and DHM. The accuracy of the deep learning model for building recognition by true orthoimage incorporating DSM, the precision was 85.17%, recall was 89.24%, and IoU was 81.35%. Compared to using only true orthoimages, the precision was increased by 1.71%, recall by 1.66%, and IoU by 1.08%. Additionally, high-resolution true orthoimages were merged with DHM. The results indicated that incorporating DHM provided more accurate building elevation information than incorporating DSM. The precision was 85.92%, recall was 89.45%, and IoU was 81.46%. Compared to using DSM, the precision increased by 0.75%, recall by 0.21%, and IoU by 0.11%.

Furthermore, upon observing the test results, it is evident that when buildings are clearly visible, deep learning can recognize complete buildings. However, there are still areas prone to omission due to occlusions or shadows, as well as cases where the absence of building labels on the topographic map leads to false positives. Despite these challenges, deep learning is not solely reliant on proper parameter tuning for feature extraction, but rather has the capability to learn specific features. This aspect helps mitigate segmentation errors and avoids extracting unnecessary information.

4.2 **Recommendations**

The quality of true orthoimage is an important factor, the use of oblique photography alongside vertical photography to construct more comprehensive true orthoimages could be done in the future. The test area in this study is the Soczi Island area of Taipei City, Taiwan. Due to the diverse building colors and structures in various regions of Taiwan, along with the occurrence of unauthorized construction, the deep learning building detection model used in this study might not be directly applicable to other regions in Taiwan. For future studies in similar contexts, it is recommended to consider employing transfer learning. This approach involves utilizing pre-trained model weights to speed up the training process for new areas or similar domains, thereby enhancing efficiency. In our analysis of building recognition results, we observe that even with the integration of DHM elevation information with true orthoimages, the model still struggled to accurately identify buildings in shadow areas. For future study, it is recommended to explore adjustments in the weighting of the DHM input channel or consider preprocessing techniques to enhance the visibility of buildings within shadow regions. This could involve modifying the color tones or other factors related to the shadow to enable the model to more effectively discern buildings in these areas, ultimately improving the overall completeness of building recognition.

Acknowledgements

Acknowledgements of test data support for Urban Development Bureau Taipei City, Taiwan.

References

Duan, M., Li, K., Yang, C. and Li, K., 2018, “A hybrid deep learning CNN-ELM for age and gender classification” Neurocomputing, 275:448-461.

He, D., Yang, X., Liang, C., Zhou, Z., G.Ororbia, A., Kifer, D. and Giles, C.Lee., 2017, “Multi-scake FCN with Cascaded Instance Aware Segmentation for Arbitrary Oriented Word Spotting in The Wild” In Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 474-483.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. and Tang, P. T. P., 2016, “On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima” arXiv 2016, arXiv:1609.04836.

Lin, T.Y., Goyal, P., Girshick, R., He, K., and Dollár, P., 2017. Focal loss for dense object detection, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.2980-2988.

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M. and Stilla U., 2016, “SEMANTIC SEGMENTATION OF AERIAL IMAGES WITH AN ENSEMBLE OF CNNS. ISPRS Annals of the Photogrammetry” Remote Sensing and Spatial Information Sciences, III(3):473-480

Masters, D. and Luschi, C. (2018). Revisiting small batch training for deep neural networks. arXiv preprint arXiv:1804.07612.

Zeng, D. and Zhu, M., 2018, “Background Subtraction Using Multiscale Fully Convolutional Network” IEEE Access, 6:16010-16021.

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., 2017, “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources” IEEE Geoscience and Remote Sensing Magazine, 5(4):8-36.