

The impact of using multi-source remote sensing images on building segmentation with U-Net model

PHAM Trung Dung^{1*}, PHAM Ngoc Hung²

¹ Hanoi University of Mining and Geology, Hanoi, Vietnam

² Phenikaa University, Hanoi, Vietnam

* Corresponding email: phamtrungdung@humg.edu.vn

Abstract: *The building extraction from remote sensing (RS) images has been a significant area of research in the photogrammetric and remote sensing communities, especially with the development of deep learning for over a decade. With the availability of multi-source data from RS images, accurately identifying buildings with different spatial image resolutions has become a challenging task. In this study, we assessed how the unalignment of image resolution between the training and testing datasets affects the ability to extract buildings. Image resolution plays a crucial role in the performance of building extraction. Our experiments found that as the image resolution decreased from 10 cm to 50 cm, the efficiency of building segmentation reduced from 0.759 to 0.585 according to the IoU metric. Besides, the ability and accuracy of building segmentation significantly decreased when the difference in image resolution between the training and testing datasets increased. In the case study, we use the model trained on a 10 cm resolution dataset to predict for 50 cm resolution data, the IoU drops significantly to 0.299. This research offers important insights into building segmentation tasks using multi-source data from satellite, airborne, and UAV images.*

Keywords: *Semantic segmentation; Building extraction; U-Net model; multi-source remote sensing images; Image resolution*

1. Introduction

The building is the centre of urban areas and one of the key elements of digital mapping. As urban areas develop, it has become increasingly important to create and update the location and information of buildings for urban planning, land use, and population management. With the advancement in remote sensing, computer vision, and deep learning technologies, the automatic extraction of buildings from RS images has become an increasingly popular topic (Lv, Peng et al. 2020), (Zhang, He et al. 2022).

Remote sensing technologies have been continuously developed, providing high-resolution (HR) or even very high-resolution (VHR) aerial photogrammetry. VHR stands for RS data with a spatial resolution of 5 cm to 4 m (Li, Huang et al. 2022). The resolution indicates a small object can be represented in the image. In the meter-level of image resolutions (e.g., satellite images with about 2 m resolution like QuickBird and WorldView-2), buildings appear to have uniform reflectance with little variation. Conversely, in higher image resolutions, such as 0.5 m of spaceborne images (Tejeswari, Sharma et al. 2022), 0.1 m airborne images (Li, Xin et al. 2021), and centimeter-level resolution of aerial photogrammetry from unmanned aerial vehicles (UAV) (Stiller, Stark et al. 2019), buildings exhibit diverse reflectance with relatively high variation. This is because nearly all structures' installations, such as antennas, water tanks, and chimneys, can be visible in the image (Lee, Lee et al. 2008). Additionally, the presence of surroundings like vegetation and shadows results in high local contrast. As a result, the VHR images open the door to extracting roof-building information on a large scale in the field of surveying and mapping (Wu 2022).

Computer vision has experienced significant advancements over the past decade, largely due to the rapid development of deep learning networks. One important task in this field is semantic segmentation, which is the classification of images at the pixel level. This task is typically performed using a convolutional neural network (CNN) (LeCun, Bottou et al. 1998). Other DL models for the semantic segmentation task include SegNet (Badrinarayanan, Handa et al. 2015), Fully Convolutional Networks (FCNs) (Long, Shelhamer et al. 2015), and U-net (Ronneberger, Fischer et al. 2015). Building extraction from aerial images using deep learning models has attracted increasing attention in recent years (Raghavan, Verma et al. 2022), (Yu, Ji et al. 2021), (Mo, Seong et al. 2021). Automatic mapping of buildings has remarkable success directly from UAV and airborne imagery (Boonpook, Tan et al. 2021), (Zheng, Ai et al. 2020), (Pilinja

Subrahmanya, Haridas Aithal et al. 2021). However, the extraction of buildings from multiple sources of RS images is still a challenge.

In deep learning, the multi-source RS images mean that the image resolution of the training set and testing dataset are not aligned. Additionally, discussions regarding the varying spatial resolutions in building segmentation using multi-source RS images are not adequate. In agricultural applications, the issue of spatial resolution's impact on plant species classification is discussed in the research by Roth, Roberts et al. (2015). Their findings highlight how different levels of spatial resolution can affect the classification of plant species. Similar to this study direction, Liu, Yu et al. (2020) identified the suitable resolution of images for the classification of vegetation types using UAV imagery. Besides, the resolution of an image significantly affects the performance of semantic segmentation algorithms can be found in (Guo, Wu et al. 2019). In that study, the super-resolution technique is proposed to use multi-source images in a wide range of resolutions. The proposed method improved approximately 19% in Jaccard and Kappa metrics when the image resolution is unaligned. In addition, Guo, Shi et al. (2023) evaluated the super-resolution technique applied for low image resolution in deep learning models. Apart from the super-resolution technique, image transform-based, data augmentation-based (Raghavan, Verma et al. 2022), and transfer learning-based methods (P, Soni et al. 2022) can be applied to overcome this issue.

In this study, we analyzed the impact of image resolutions on the accuracy of building performance using the U-net model. We conducted both qualitative and quantitative evaluations to assess the influence of different resolutions between the training and testing datasets on building segmentation performance. The results of our experiments offer a comprehensive understanding of how a hyperparameter trained on a specific image resolution can predict buildings on images with different resolutions.

The rest of this paper is organized as follows: Section 2 outlines the structure of the U-Net model and the metrics used to evaluate building segmentation. Section 3 presents the experiment, including computer configurations, software environment, and a description and pre-processing of the dataset. The results and discussion are analyzed in Section 4. Finally, the conclusion and future work are summarized in Section 5.

2. Methodology

In this paper, we aim to evaluate the efficiency of building extraction using the U-net model. We will assess its performance using various metrics, including accuracy, precision, recall, F1-score, Jaccard index (IoU), and dice-score. The following sections will outline the structure of the U-net model and the evaluation metrics.

2.1 U-net model

The U-net, developed by Ronneberger, Fischer et al. (2015), is a popular model for image semantic segmentation due to its high accuracy and efficiency. This network is a fully convolutional neural network (FCNN) that employs both a contracting and an expanding path during the convolution process. The U-net has a symmetrical structure, forming a U-shape with encoder-decoder parts (see Fig. 1).

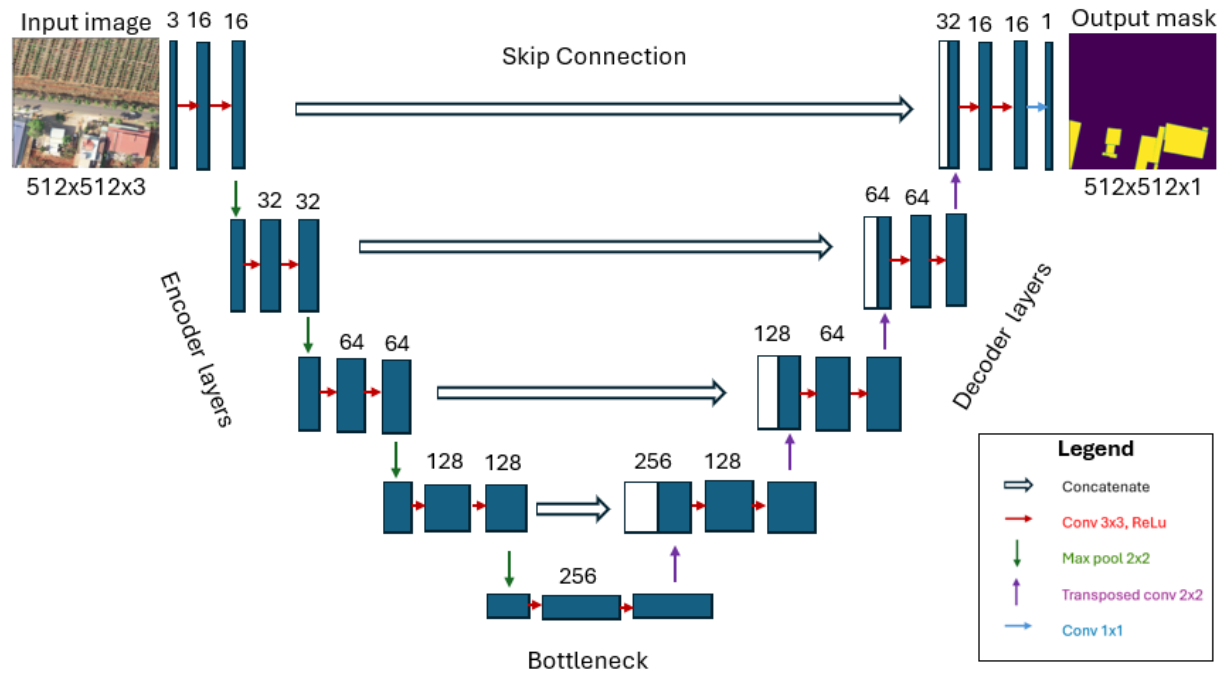


Fig. 1. U-net model with an RGB input image of size 512×512×3 and output mask of size 512×512×1.

The encoder layers on the left side are responsible for downsampling and extracting the image feature. These layers are to reduce the spatial resolution of feature maps while increasing their depth. The main purpose of encoder layers is to capture abstract representations of the input. The encoder path contains five layers of convolution blocks. Each layer includes two 3x3 convolution operations followed by a rectified linear unit (ReLU) activation function. Then, they are downsampled by using a 2x2 max pooling operation with a stride of 2.

On the other hand, the decoder layers on the right side are to decode the data and locate the features. The spatial resolution of the input images is maintained in this operation. The decoder path functions similarly to the encoder path, where up-convolution is used to double the width and height of the image.

The upsampled image in the encoder path is combined with the corresponding feature map from the decoder path through a skip connection. This skip connection helps to maintain spatial information and accurately locate features.

2.2 Evaluation metrics

The performance of machine learning algorithms can be assessed using metrics in two main ways. First, the model is optimized through a loss function. Second, its performance is validated and evaluated. In this paper, the performance of the U-Net model for building extraction tasks is measured using six general metrics: accuracy, precision, recall, F1-score, Jaccard index (IoU), and Dice coefficient. The first four metrics are derived from the confusion matrix, which includes true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). In the context of image segmentation, TP refers to the number of pixels that the model correctly predicts. FP represents the number of pixels that the model incorrectly classifies. TN indicates the number of background (negative) pixels that the model correctly identifies, and FN is the number of background pixels that the model incorrectly classifies.

Accuracy is a popular metric in semantic segmentation, defined as the number of correctly segmented pixels divided by the total number of pixels, as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision is computed as the number of correctly segmented pixels divided by the total number of pixels predicted by the model:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

By contrast, recall measures the ratio of true positives to the sum of true positives and false negatives:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-score is a trade-off between precision and recall and is calculated by:

$$F1 = 2 \left(\frac{Precision \times Recall}{Precision + Recall} \right). \tag{4}$$

Apart from the four above general metrics, Jaccard and Dice scores are preferred to use as metrics in the semantic segmentation task. These scores belong in the category of intersection over union (IoU) metrics. The IoU is defined by the ratio of intersection and union between the predicted (P) and actual segmentation areas (A) as:

$$IoU = \frac{|P \cap A|}{|P \cup A|} = \frac{|P \cap A|}{|P| + |A| - |P \cap A|}. \tag{5}$$

Like the Jaccard/IoU score, the Dice score is another reliable metric in semantic segmentation. This score is to measure similarity between two samples, and it can be computed by:

$$Dice = \frac{2|P \cap A|}{|P| + |A|}. \tag{6}$$

3. Experiments

3.1 Experimental design

The experiments were carried out using the PyTorch 1.11.0 deep learning framework. Google Colaboratory was utilized to access the CUDA parallel computing platform and graphical processing units (GPUs). The GPU significantly reduced training time by allowing separate computations to run in parallel. The specific GPU configuration used was the Tesla T4, a professional graphics card by NVIDIA with 16 GB GDDR6 memory and a 256-bit memory interface. Additionally, it features 320 tensor cores that enhance the speed of deep learning applications. The batch size was set to 8, and the total number of epochs was determined to be 100. We used cross-entropy as the loss function and employed stochastic gradient descent (SGD) as the optimizer to train the model parameters. The initial learning rate for SGD was established at 0.01.

3.2 Dataset description

The study area covers several provinces in three main regions: North, Middle, and South of Vietnam. The buildings in the area vary in shape and size, mostly being rectangular, some L-shaped, or square. They can be grouped into four main categories: modern residences, factories, urban areas, and rural areas (see Fig. 2). The buildings in urban and rural areas are typically narrow and long, ranging from 50 to 200 m², with 1 to 5 stories, and are surrounded by houses, roads, and trees. In contrast, buildings in modern and industrial areas are usually larger, ranging from 150 to 1000 m², and are surrounded by open space. Different types of buildings used for the dataset aim to increase complexity and diversity during the training of the model.



Fig. 2. Four types of buildings are used for training the model

The images in the dataset were captured by a UAV, providing a spatial resolution of 10 cm per pixel (each pixel corresponds to a 10x10 cm² area on the ground). Due to limited GPU memory, the original images must be divided into smaller segments for the experiment. The UAV images were divided into patches of size 512 × 512 pixels, which are referred to as input images or tiles. The entire research area contains approximately 40,000 buildings, with 10,000 corresponding tiles.

To assess the impact of image resolution on predicted performance, the original images were downsampled at given ratios of 2, 3, 4, and 5, resulting in corresponding image resolutions of 20, 30, 40, and 50 cm for each experimental case. These images with downsampled, were organized into four datasets.

Each dataset was further divided into subsets for the training dataset, validation dataset, and testing dataset according to an 8:1:1 proportion.

3.3 Dataset pre-processing

The pre-processing data for training process of building extraction consists of labelling and augmentation tasks. First, the training set was manually labelled into two classes which are buildings and non-buildings. Then, the buildings and non-buildings were converted to binary masks with values of 1 and 0, respectively (see Fig. 3). Next, the augmentation technique was utilized during the training process to increase the number of input images. In this study, the augmentation technique included random rotation, horizontal flips, resize, random brightness contrast, blur, and sharpen.

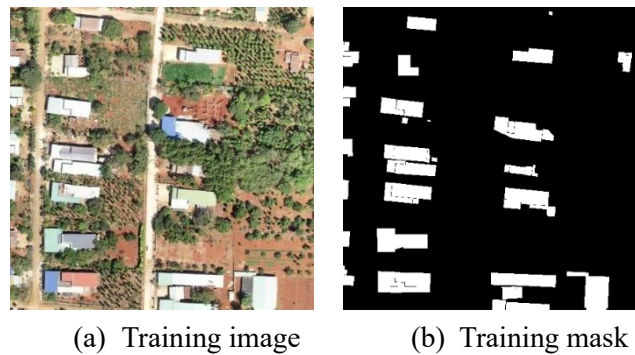


Fig. 3. The training image (a) and the corresponding mask

4. Results and discussions

To demonstrate the effects of varying image resolutions between the training and testing datasets, both qualitative and quantitative evaluations of semantic segmentation performance are presented. This impact is demonstrated by comparing the efficiency of building extraction measured by six metrics, consisting of accuracy, precision, recall, F1-Score, IoU, and Dice, in two scenarios. First, buildings are extracted from the UAV image using hyperparameters trained on a UAV image with the same image resolution. Second, buildings are extracted from the UAV image with different resolutions as the training set. The training sets in five experiments were carried out in the same location with corresponding image resolutions of 10, 20, 30, 40, and 50 cm, respectively.

4.1 Qualitative evaluation

Fig. 4 provides a qualitative assessment of the performance of building extraction at different image resolutions. Fig. 4 (1) and (2) represent the RGB image captured from a UAV and the ground truth mask, while (3) and (4) show the building extraction performance by the U-net model in two scenarios. In the first scenario, both the training and testing sets have similar image resolutions. In the second scenario, the training set has an image resolution of 10 cm, while the testing set has image resolutions of 20, 30, 40, and 50 cm in Fig. 4 (a), (b), (c), and (d), respectively. We compared the building extraction performance between the first scenario (3) and the second scenario (4), with variations in image resolution at four levels.

The difference in image resolution between the training set and the testing set has a significant impact on the efficiency of building extraction. As the difference increases, the efficiency of building extraction decreases. Specifically, Fig. 4 (a) depicts the building extraction for the testing set with a 20 cm image resolution, where buildings are predicted using hyperparameters trained on image resolutions of 20 cm and 10 cm in (3) and (4), respectively. Intuitively, the results show a small difference in resolution between the training and testing sets, leading to almost all buildings predicted in (3) and (4) being relatively comparable (See Fig. 4 (a)).

In contrast, Fig. 4 (d) shows a significant difference in building extraction when using two different sets of hyperparameters trained on image resolutions at 50 cm and 10 cm in (3) and (4), respectively. In this case, the efficiency of building extraction is dramatically reduced, with several buildings in UAV images that cannot be predicted by the model (See Fig. 4 (d)).

4.2 Quantitative evaluation

The quantitative results generated by the differing image resolutions in the training and testing datasets are shown in Tab. 1 and Tab. 2. Tab. 1 shows the comparison in efficiency of building segmentation measured by six metrics when the image resolutions of the training set and testing set are the same. The efficiency of building extraction, measured by relevant metrics, reduces with lower image resolution.

Except for accuracy, five metrics, including precision, recall, F1-Score, IoU, and dice, reduced from 0.896, 0.896, 0.834, 0.759, and 0.906 to 0.853, 0.652, 0.731, 0.585, and 0.863 with the decrease of image resolution from 10 to 50 cm, respectively. The reason for this phenomenon is that the lower image resolution often contains more noise and blurring, making it difficult to extract features from the images. In other words, the difference in image resolutions between the training set and the testing set leads to a difference in the distribution of these datasets. The findings align with those of the study in (Guo, Shi et al. 2023), which identified a significant impact of spatial image resolution on segmentation results. The IoU decreases from approximately 0.70 to 0.45 as the image resolution decreases from 0.3 m to 2.4 m.

Tab. 1. Evaluation of building extraction with various image resolutions when the training and testing sets are similar to the image resolution

Image resolution		Metrics for evaluations					
Training set	Testing set	Accuracy	Precision	Recall	F1-Score	IoU	Dice
10 cm	10 cm	0.941	0.893	0.834	0.857	0.759	0.906
20 cm	20 cm	0.951	0.828	0.821	0.813	0.700	0.891
30 cm	30 cm	0.961	0.815	0.789	0.786	0.667	0.882
40 cm	40 cm	0.988	0.859	0.691	0.761	0.617	0.877
50 cm	50 cm	0.990	0.853	0.652	0.731	0.585	0.863

Tab. 2. Evaluation of building extraction with various image resolutions when the training and testing sets are different in image resolution

Image resolution		Metrics for evaluations					
Training set	Testing set	Accuracy	Precision	Recall	F1-Score	IoU	Dice
	10 cm	0.941	0.893	0.834	0.857	0.759	0.906
	20 cm	0.953	0.827	0.760	0.814	0.703	0.892
10 cm	30 cm	0.962	0.818	0.542	0.656	0.508	0.817
	40 cm	0.967	0.845	0.460	0.559	0.412	0.771
	50 cm	0.961	0.687	0.336	0.408	0.299	0.693

Tab. 2. compares the efficiency of building extraction on testing sets at different image resolutions using the hyperparameters trained on the 10 cm image resolution. These results show that the efficiency of building extraction decreases significantly when the difference in image resolution between the training and testing sets is large. The precision, recall, F1-Score, IoU, and dice metrics decrease from 0.893, 0.834, 0.857, 0.759, and 0.906 to 0.687, 0.336, 0.408, 0.299, and 0.693, respectively, as the corresponding image resolution decreases from 10 to 50 cm. Conversely, the accuracy metrics slightly increase, but these values are close to 1.

Tab. 3. The reduction in accuracy and efficiency of building extraction due to the difference in image resolution between the Training and Testing sets

Image resolution of testing sets	Metrics for evaluation					
	Accuracy	Precision	Recall	F1-score	IoU	Dice
20 cm	100%	100%	93%	100%	100%	100%
30 cm	100%	100%	69%	83%	76%	93%
40 cm	98%	98%	67%	73%	67%	88%
50 cm	97%	81%	52%	56%	51%	80%

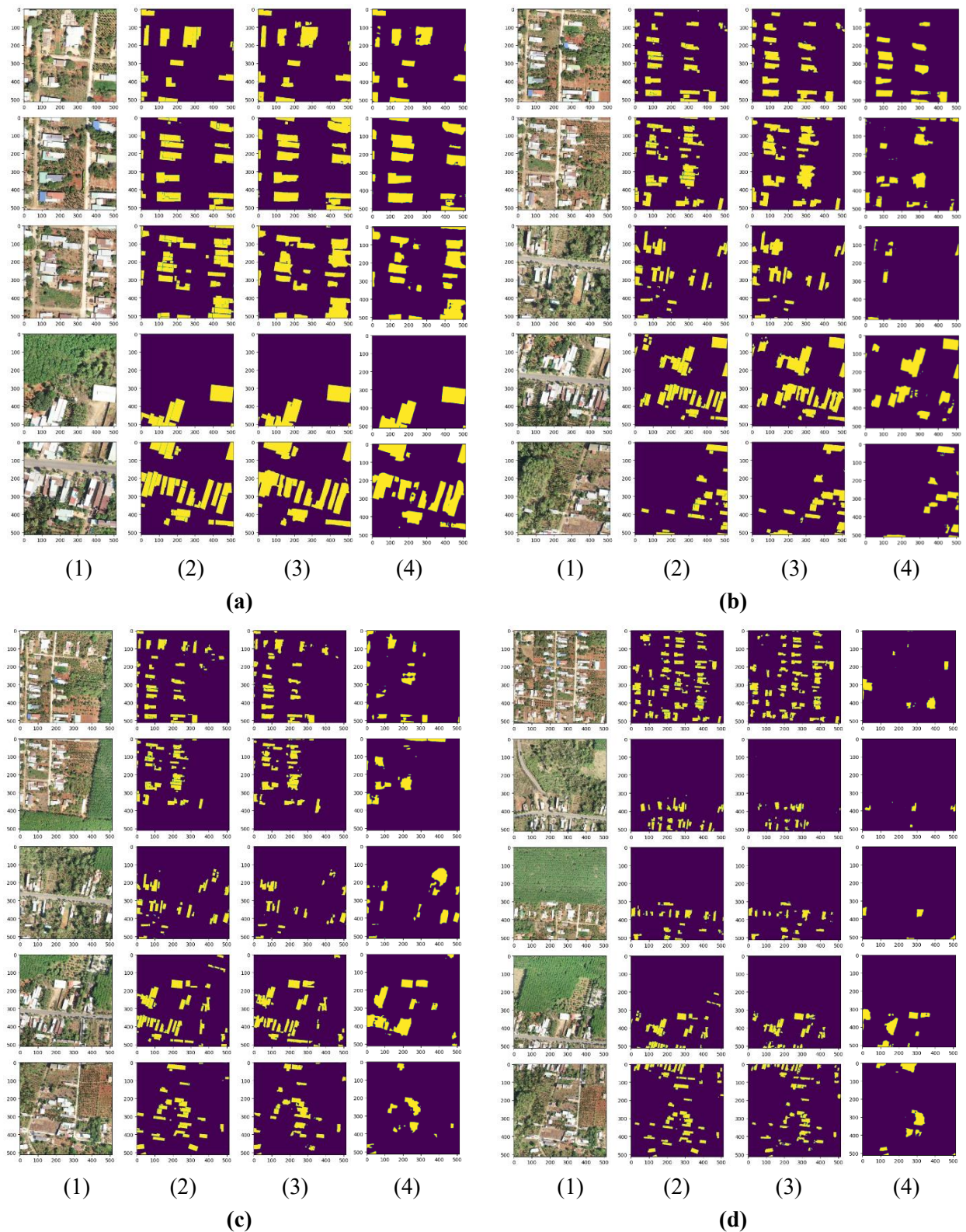


Fig. 4. Qualitative evaluation of building extraction. (1): ortho (RGB) UAV image; (2): ground truth masks for building footprints; (3) and (4) corresponding prediction by U-net

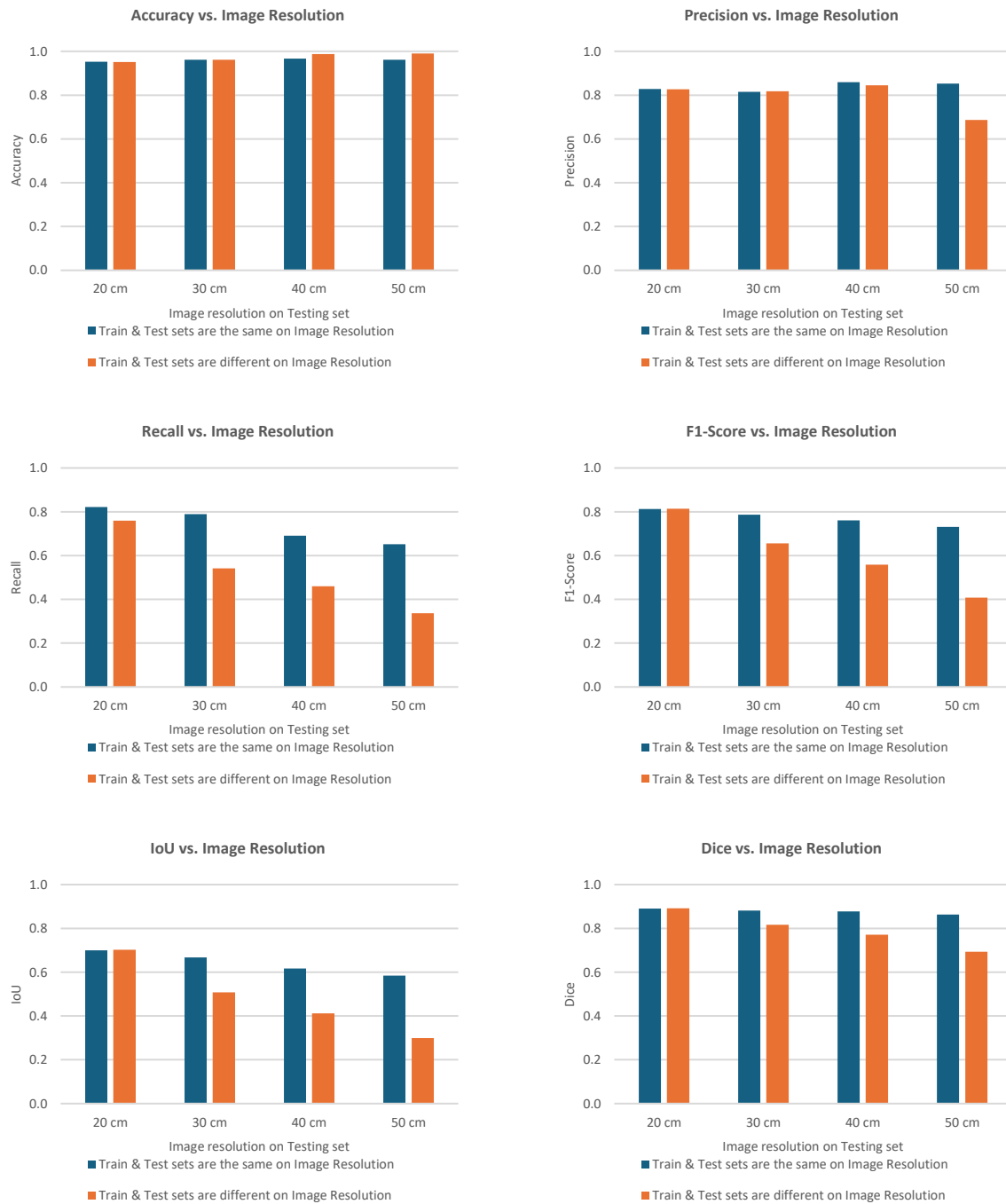


Fig. 5. The comparison of the accuracy and efficiency of building extraction with varying image resolution of the testing set

Fig. 5 compares the efficiency of building extraction using six metrics in two categories: (i) using the same image resolution for training and testing sets, and (ii) using different image resolutions for training and testing sets. It is evident that the segmentation efficiency of buildings significantly reduces with larger differences in the image resolution of the testing set. According to the quantitative assessment reported in Tab. 3, the image resolution of the testing set is smaller than twofold that of the training set (20 cm), and the efficiency of building segmentation is relatively unchanged. Conversely, when the image resolution of the testing set is five times smaller than that of the training set (50 cm), the efficiency decreases to 97% for accuracy, 81% for precision, 52% for recall, 56% for F1-score, 51% for IoU, and 80% for dice.

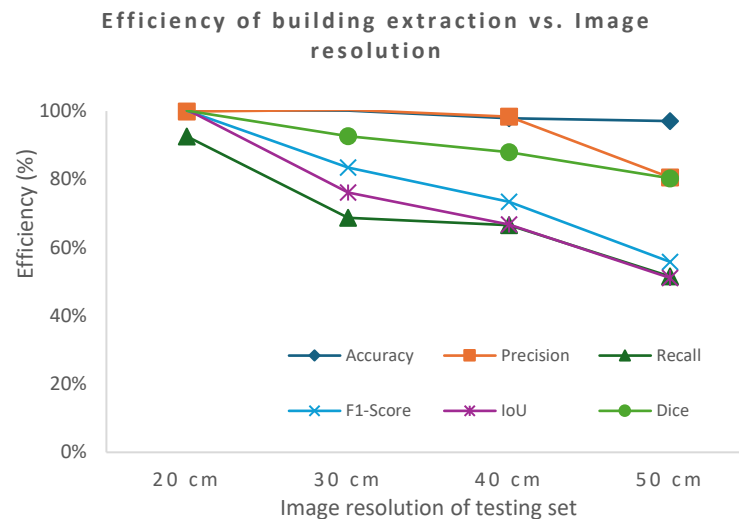


Fig. 6. The tendency towards efficiency in building segmentation

In Fig. 6 we can see how different image resolutions affect the accuracy of building segmentation. When predicting for the testing set with 30 cm image resolution, the recall, F1-score, and IoU values are 69%, 83%, and 76% respectively. However, these values drop to 52%, 56%, and 51% when predicting for the testing set with 50 cm image resolution. This indicates a significant decrease in efficiency for the lower resolution.

5. Conclusions

In our study, we examined how image resolution impacts the performance of building segmentation using the U-net model. We discovered that training the DL model with higher image resolutions can improve the building segmentation performance. Furthermore, we conducted both qualitative and quantitative assessments to determine how the difference in image resolutions between the training and testing sets affects building segmentation performance. Our experiments revealed that a significant difference in image resolution reduces the efficiency of building segmentation performance. However, we can use existing hyperparameters to predict buildings in images with comparable resolutions, thereby reducing the time required to establish a new training dataset and the training process. Suitable solutions for improving efficiency in the segmentation of buildings using multi-source remote sensing should be investigated in future work.

Acknowledgements

This research was funded by the Ministry of Education and Training of Vietnam under Grant number B2024-MDA-09.

Conflicts of Interest

The authors declare no conflict of interest.

Literature – References

1. Badrinarayanan, V., A. Handa and R. J. a. p. a. Cipolla (2015). "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling."
2. Boonpook, W., Y. Tan and B. J. I. J. o. R. S. Xu (2021). "Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry." 42(1): 1-19.
3. Guo, Z., X. Shi, H. Zhang, D. Huang, X. Song, J. Yan, and R. J. a. p. a. Shibasaki (2023). "Enhancing building semantic segmentation accuracy with super resolution and deep learning: Investigating the impact of spatial resolution on various datasets."
4. Guo, Z., G. Wu, X. Song, W. Yuan, Q. Chen, H. Zhang, X. Shi, M. Xu, Y. Xu, and R. J. I. A. Shibasaki (2019). "Super-resolution integrated building semantic segmentation for multi-source remote sensing imagery." 7: 99381-99397.
5. LeCun, Y., L. Bottou, Y. Bengio and P. J. P. o. t. I. Haffner (1998). "Gradient-based learning applied to document recognition." 86(11): 2278-2324.
6. Lee, D. H., K. M. Lee, S. U. J. P. E. Lee, and R. Sensing (2008). "Fusion of lidar and imagery for reliable building extraction." 74(2): 215-225.

7. Li, J., X. Huang, L. Tu, T. Zhang, L. J. G. Wang and R. Sensing (2022). "A review of building detection from very high resolution optical remote sensing images." 59(1): 1199-1225.
8. Li, Z., Q. Xin, Y. Sun and M. J. R. S. Cao (2021). "A deep learning-based framework for automated extraction of building footprint polygons from very high-resolution aerial imagery." 13(18): 3630.
9. Liu, M., T. Yu, X. Gu, Z. Sun, J. Yang, Z. Zhang, X. Mi, W. Cao and J. J. R. S. Li (2020). "The impact of spatial resolution on the classification of vegetation types in highly fragmented planting areas based on unmanned aerial vehicle hyperspectral images." 12(1): 146.
10. Long, J., E. Shelhamer and T. Darrell (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition.
11. Lv, B., L. Peng, T. Wu and R. Chen (2020). Research on urban building extraction method based on deep learning convolutional neural network. IOP Conference Series: Earth and Environmental Science, IOP Publishing.
12. Mo, J. S., S. K. Seong, J. W. J. J. o. t. K. S. o. S. Choi, Geodesy, Photogrammetry and Cartography (2021). "Comparative evaluation of deep learning-based building extraction techniques using aerial images." 39(3): 157-165.
13. P, P. S., J. Soni and H. A. B (2022). Building extraction from remote sensing images using deep learning and transfer learning. IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium.
14. Pulinja Subrahmanya, P., B. Haridas Aithal and S. J. J. o. t. I. S. o. R. S. Mitra (2021). "Automatic extraction of buildings from uav-based imagery using artificial neural networks." 49(3): 681-687.
15. Raghavan, R., D. C. Verma, D. Pandey, R. Anand, B. K. Pandey and H. Singh (2022). "Optimized building extraction from high-resolution satellite imagery using deep learning." Multimedia Tools and Applications 81(29): 42309-42323.
16. Raghavan, R., D. C. Verma, D. Pandey, R. Anand, B. K. Pandey, H. J. M. T. Singh and Applications (2022). "Optimized building extraction from high-resolution satellite imagery using deep learning." 81(29): 42309-42323.
17. Ronneberger, O., P. Fischer and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer.
18. Roth, K. L., D. A. Roberts, P. E. Dennison, S. H. Peterson and M. J. R. s. o. e. Alonzo (2015). "The impact of spatial resolution on the classification of plant species and functional types within imaging spectrometer data." 171: 45-57.
19. Stiller, D., T. Stark, M. Wurm, S. Dech and H. Taubenböck (2019). Large-scale building extraction in very high-resolution aerial imagery using mask r-cnn. 2019 Joint Urban Remote Sensing Event (JURSE), IEEE.
20. Tejeswari, B., S. K. Sharma, M. Kumar, K. J. T. I. A. o. t. P. Gupta, Remote Sensing and S. I. Sciences (2022). "Building footprint extraction from space-borne imagery using deep neural networks." 43: 641-647.
21. Wu, Y. (2022). Deep learning based building extraction from high-resolution remote sensing images, University of Waterloo.
22. Yu, D., S. Ji, J. Liu, S. J. I. J. o. P. Wei and R. Sensing (2021). "Automatic 3d building reconstruction from multi-view aerial images with deep learning." 171: 155-170.
23. Zhang, P., H. He, Y. Wang, Y. Liu, H. Lin, L. Guo and W. J. I. A. Yang (2022). "3d urban buildings extraction based on airborne lidar and photogrammetric point cloud fusion according to u-net deep learning model segmentation." 10: 20889-20897.
24. Zheng, L., P. Ai and Y. Wu (2020). Building recognition of uav remote sensing images by deep learning. IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, IEEE.