ZANCO Journal of Pure and Applied Sciences

TYPE Original article
DOI http://dx.doi.org/10.21271/ZJPAS.37.3.13

OPEN ACCESS

*Corresponding author
Rojgar Qarani Ismael
rojgar.ismael@su.edu.krd

RECEIVED :09 /01 /2025
ACCEPTED :09/03/ 2025
PUBLISHED :30/ 06/ 2025

KEYWORDS:
Deep Learning, U-Net, FCN, Mask R-CNN, Sequential Integration

Sequential Hybrid Integration of U-Net and Fully Convolutional Networks with Mask R-CNN for Enhanced Building Boundary Segmentation from Satellite Imagery

Rojgar Qarani Ismael*,Haval Abduljabbar Sadeqa

Department of Geomatics (Surveying) Engineering, College of Engineering Salahaddin University-Erbil, Erbil, Kurdistan Region-Iraq.

ABSTRACT

In the recent years, building boundary segmentation obtained significant advancement through using deep learning. The present algorithms, such as Convolutional Neural Network (CNN) are unable to detect buildings in challenging urban areas like occlusions. This study investigates the integration of U-Net and Fully Convolutional Networks (FCN) with Mask R-CNN to improve building boundary segmentation using high-resolution satellite imagery. A sequential hybrid approach has been developed for combining semantic and instant segmentation. The integration between the U-Net with Mask R-CNN has been achieved by feeding the segmentation result from the U-Net as an input into the Mask R-CNN. A similar procedure was applied in the integration of the FCN with Mask R-CNN. The integration of U-Net with Mask R-CNN resulted in an improvement in the recall by 9.9% and an increase by 4.3 % in the F1-score, demonstrating its capability in segmenting boundary precision and fine-grained details. Similarly, FCN combined with Mask R-CNN has shown an enhancement of recall by 9.9% and precision by 7.6%, assuring its capability in the capture of global context. Further analysis through comparison between integration U-Net with Mask R-CNN with results from previous studies, demonstrates that the proposed integration scheme outperforms the existing results. The performance evaluation across RGB and panchromatic datasets highlights the flexibility of these integrations by proving their efficiency in different applications. Despite the minor challenges that appeared in boundary alignment, the results brought out the potential of such hybrid models for applications in urban planning, cadastral mapping, and disaster management.

1.Introduction

In recent years, the extraordinary advancement in deep learning has led to an unprecedented improvement in image segmentation, which can be regarded as a new landmark in the field of building boundary extraction. Building boundary segmentation from satellite images has been essentially important in remote sensing and urban planning. Applications of accurate segmentation of the building boundaries assisted in different tasks including urban mapping (Luo et al., 2021), cartography, infrastructure monitoring (Yu et al., 2024), disaster management (Bousias Alexakis and Armenakis, 2022), environmental planning (Yan et al., 2023), and 3D city modeling (Zhao et al., 2020). With the rapid development of satellite imaging, recently high-resolution images are frequently acquired for obtaining detailed information about urban landscapes. This provides an effective extraction of building boundaries from such images for improving urban analysis and planning activities with higher precision.

However, building boundary segmentation tasks from satellite imageries faces many challenges due to the diversity in building shapes, sizes, and orientations apart from occlusions, shadows, trees, and other complex features that contribute to complicating segmentation performance (Zhang et al., 2021). These complexities are reflected to be higher in providing an accurate result by traditional image processing techniques. Recent improvements regarding deep learning have been promising for improving segmentation accuracy, yet there are challenges in obtaining a model architecture for robustly leveraging optimum outcomes (Li et al., 2023).

The objectives of this research are integrating U-Net with Mask R-CNN in addition to the FCN with Mask R-CNN to gain better accuracy in building boundary segmentation from satellite images. By leveraging of their strengths, the aim is to create a robust hybrid segmentation model that is capable of addressing different challenges. Furthermore, analyzing the impact of RGB and panchromatic modalities on the segmentation process. The expected goals are enhancing the accuracy and efficiency of building boundary delineation; seeking to overcome issues related to occlusions at different scales and architectural styles.

The contribution of the research is as follows: 1. Demonstrating a novel hybrid integration strategy: A new strategy of integration will be proposed, which fuses the strengths of U-Net with Mask R-CNN as well as FCN with Mask R-CNN for improving building boundary segmentation accuracy. 2) Comprehensive evaluation: extensive experiments are conducted to show the performance of the integrated model on many data sets of satellite images, showing improvements over existing methods. 3) Addressing challenges in detecting complex building boundaries.

The structure of the research is as follows, the introduction is discussed in the section 1, and the literature is illustrated in the section 2. Section 3 is related to the methodology and section 4 is related to the experiment result. The discussion and conclusions are in sections 5 and 6 respectively.

2. Related Work

The segmentation of building boundaries from satellite images has undergone remarkable development, especially with the recent application of deep learning techniques. Traditional methods primarily relied on manual digitization or semi-automatic processes using edge detection and thresholding (Ahmadian et al., 2024, Xia et al., 2021), or region growing (Bai et al., 2024). It has always been labor-intensive and could hardly deal with complex urban environments and different conditions of the images (Das, 2024).

With the emergence of machine learning, more sophisticated methods for building extraction were introduced, such as Support Vector Machines (SVM) and Random Forests (RF). Such methods improve image processing but still need extensive feature engineering that is sensitive to the quality of the input features (Li et al., 2019, Thakur et al., 2019). In addition, their performance is normally not robust in terms of the quality and relevance of the extracted features (Reda and Kedzierski, 2020).

Deep learning, normally the state-of-the-art convolutional neural networks (CNN), has brought a revolution in segmentation of building

boundaries. They learn features hierarchically directly from raw image data to provide better accuracy, hence making them robust (Raghavan et al., 2022). Traditional and machine learning methods still show limitations with regard to high-resolution images, occlusions, shadows, and features related to urban environments (Wang et al., 2024).

U-Net, FCN, and Mask R-CNN are deep learning architectures that are widely adopted as the framework in photogrammetry, remote sensing, urban planning, and computer vision. The U-Net which proposed by (Ronneberger et al., 2015) was originally used for biomedical image segmentation but has since then been adopted for the segmentation of building boundaries due to its encoder-decoder structures, which is important to effectively capture the spatial contextual information. Recent studies have identified U-Net as being quite effective in extracting building boundaries from high-resolution satellite images. However, the obtained results still suffer from imperfection due to the presence of occlusion particularly in complex urban areas and high-performance computational resources is required. (Alsabhan et al., 2022; Sariturk and Seker, 2022; Ayala et al., 2021; Li et al., 2021a; Liu et al., 2020; Wagner et al., 2020; Li et al., 2019; Prathap and Afanasyev, 2018; Guillaume et al., 2017).

Fully Convolutional Networks (FCNs) proposed for pixel-wise segmentation tasks by (Long et al., 2015), have been successfully applied to building segmentation, demonstrating robust performance in capturing the overall structure of the buildings. Recent investigations have found FCNs to work efficiently for the segmentation of building boundaries from satellite imagery, but the algorithm faces challenges in extracted buildings, especially in occluded urban areas, and a large training dataset is required (Liu et al., 2019; Bittner et al., 2018, Ji et al., 2018, Mace et al., 2018, Mou and Zhu, 2018).

Mask R-CNN, developed by (He et al., 2017), extended the Faster R-CNN algorithm by adding a branch for predicting segmentation masks. These modifications, makes the model to perform object detection and instance segmentation, thus being particularly suitable for detecting individual buildings. Recent studies have revealed that Mask R-CNN detects the instance-level boundary with high accuracy and performs segmentation from high-resolution satellite images quite efficiently. Some studies which are focused on this area are presented in (NourEldeen and Wahed, 2024; Sakeena et al., 2023; Susetyo et al., 2023; Li et al., 2021b; Carvalho et al., 2020).

Despite the significant advancements in building boundary segmentation using U-Net, FCN, and Mask R-CNN, still standalone implementations face limitations in terms of accuracy and contextual understanding (Gashti et al., 2024, Zhang et al., 2020). Conventional models like U-Net and FCN, though effective for segmentation tasks, often fail to preserve fine boundary details in high-resolution satellite imagery (Aryal and Neupane, 2023, Lussange et al., 2023). The down-sampling process inherent in these architectures leads to a loss of spatial information, which is critical for delineating building edges (Neupane et al., 2021). This will cause inaccurately delineating building boundaries, especially in complex urban environments (Zorzi and Fraundorfer, 2019). While Mask R-CNN efficient in instance segmentation, struggles to identify smaller and large building features in complex urban environments (Wang et al., 2023; Gao et al., 2022) also computational complexity and high resources keeps challenging task (Dalal et al., 2020).

A hybrid model could potentially leverage the strengths of architectures and overcome the existing limitations. Recent studies demonstrated that combining the strengths of U-Net, FCN, and Mask R-CNN architectures lead to improve building boundary segmentation and extraction from satellite imagery. For example, (Anh et al., 2022) examined the integration of Mask R-CNN with U-Net for building footprint extraction. The combined models yield more stable and accurate results compared to single models, demonstrating the effectiveness of multi-model approaches in building segmentation tasks. Similarly, (Han et al., 2021) integrated features from SegNet and U-Net into Mask R-CNN to enhance building detection in remote sensing images. This fusion improves the model's ability

to accurately locate and segment building boundaries. (Zhang and Chi, 2020) introduced Mask-R-FCN, a network that combines Fully Convolutional Networks (FCNs) and Mask R-CNN to improve semantic segmentation, particularly for small objects in remote sensing images. The fusion of pixel-level and object-level information enhances segmentation performance.

From the literature, it can be noticed that the standalone algorithms still struggle to detect building footprints accurately. Therefore, in this research integration of U-Net and FCN with Mask R-CNN architectures from high-resolution RGB and panchromatic satellite imagery have been investigated to achieve better performance than the existing methods.

3. Methodology

The overall workflow adopted in this study consists of four main steps: study area and preprocessing, model architectures, model training, and integration strategy.

3.1 Study Area and Preprocessing

This study utilizes high-resolution stereo satellite imagery captured by the GE01 sensor on November 8th, 2020. The spatial resolution of the images is 50 cm, providing details which are necessary for precise building segmentation tasks. The study focuses on the urban area of Erbil, Kurdistan Region, Iraq, which is known for its dense and diverse building structures, making it ideal case for testing the effectiveness of integration of U-Net, or FCN, with Mask R-CNN model for building boundary delineation, Figure 1.

The stereo images are processed via photogrammetry to generate DSMs which have been used for generating panchromatic and multispectral ortho-rectified imagery. The produced ortho-rectified images had a spatial resolution of 0.5 m for panchromatic and 2.0 m for multispectral images. Thus, the resolution of the multispectral is four times lower than the panchromatic images.

To train CNN architectures, pan-sharpened (RGB) imagery has been used. Which is produced by merging high resolution panchromatic image 0.5 m spatial resolution with lower resolution (2.0 m) multispectral image to obtain a high-resolution RGB image using ENVI software. Thus, the datasets used in the researchconsist of two subsets of ortho-rectified high-resolution: panchromatic, and pan-sharpened (RGB) imagery. The panchromatic imagery is finer in spatial resolution, therefore providing detailed information on the spatial characteristics of buildings, while the multi-spectral imagery provides very important spectral information that helps in discriminating different land cover types. This process is achieved through pan-sharpening; the multi-spectral imagery from multispectral bands fused with high spatial resolution of panchromatic imagery (Pushparaj and Hegde, 2017).

(a) (b)

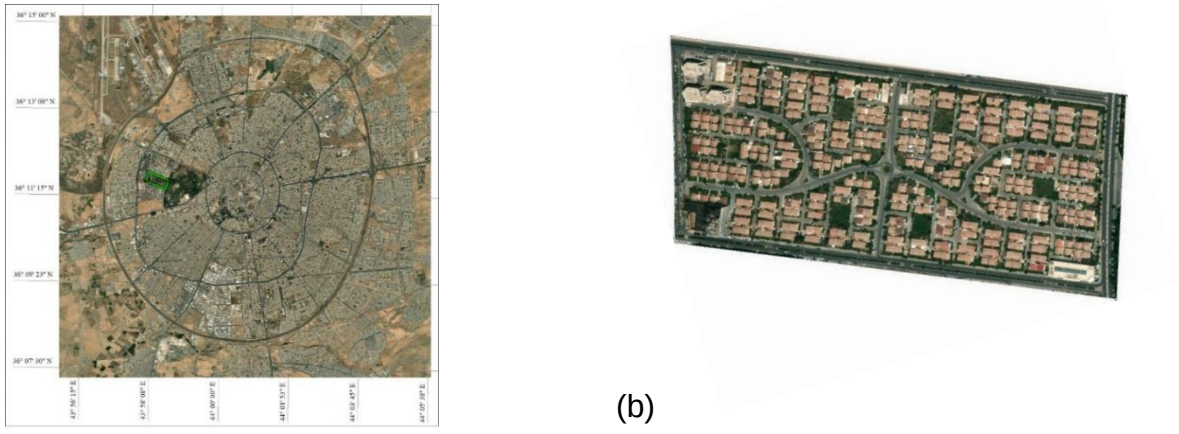


Figure 1: (a) Erbil city, showing the study area which is marked with a green rectangle (b) study area that has been used in the research.

The dataset has been pre-processed for the training the convolutional neural network models. This includes tasks such as resizing, annotation, and augmentation. The images were resized into the 256 x 256 pixels. Later, a subset of images is manually annotated to create the ground truth masks, which will demarcate the boundary of buildings. These annotated images will serve to train and validate the U-Net, FCN, and Mask-RCNN models. The Annotation was done both in the Roboflow framework and VGG Image Annotator VIA (Dutta and Zisserman, 2019). Data augmentation techniques which including geometric transformations and color adjustments are specified to artificially increase the variety and number of training data to avoid overfitting and enhancing generalization ability (Amarù et al., 2023, Shorten and Khoshgoftaar, 2019). The augmentation for the dataset has been performed through a Python library called "imgaug"(https://github.com/aleju/imgaug).

3.2 Model Architecture

The integrations of deep learning architectures of U-Net, FCN, and Mask R-CNN have been proposed to implement building boundary segmentation individually. Resized satellite image patches to 256x256 for RGB (3-channel) and panchromatic (1-channel) images are served as input data. While the output is represented by binary masks where each pixel is classified as building or non-building, Figure 2. The proposed CNN architectures were implemented using Python v.3.7 and TensorFlow library. The model architectures were built from scratch without using pre-defined or fine-tuned models utilizing the integrated development environment (IDE) of PyCharm.

3.2.1 U-Net

The U-Net model includes 10 Conv2D layers in the contracting path (encoder). In this step, the spatial resolution was gradually reduced, leading to the loss of detail. An expansive path is made up of 4 Conv2D layers; it is the decoder. During the upsampling, the spatial resolution has been recovered to its original size in the output. In order to link encoder and decoder layers, skip connections were incorporated to ensure the preservation of spatial details.

3.2.2 FCN

The FCN-8s architecture, which is fully convolutional network was proposed to eliminate dense layers while maintains spatial dimensions throughout the network. The FCN-8s encoder evolves of 12 Conv2D layers and 1 Conv2D transpose layer. Decoder layers are composed of 4 Conv2D layers and 4 Conv2D transpose layers. Skip connections are utilized to improve the precision of buildings boundary through integrating detailed features from earlier layers.

3.2.3 Mask R-CNN

Mask R-CNN architecture with ResNet-101 feature extraction backbone has been proposed to be used for extracting detailed and rich representations from image patches. The network architecture has two major modules: The Region Proposal Network (RPN) and the instance segmentation module. The RPN selects the regions that have a high probability of containing buildings, While the instance segmentation module then generates precise binary masks for each detected building.

3.3 Model Training

The training sessions of U-Net, FCN, and Mask R-CNN architectures were conducted individually through a systematic process using split RGB and Panchromatic image training data (80 % of the dataset); this is to ensure the balance between the building and non-building classes and robust performance of the model. No pre-trained weights were used during the training initializations of U-Net and FCN models, while for Mask R-CNN a pre-trained weight that performed on ResNet-101 backbone was utilized. A batch sizes of 8 and 16 were implemented during the training of the U-Net and FCN models, while for Mask R-CNN the batch size 1 and threshold 0.3 was selected. The Adam optimizer with a learning rate of 0.001 has been employed during the training of proposed CNN architectures. A binary cross-entropy loss function was used for training of U-Net and FCN models, while the Mask R-CNN loss function is a composite of the loss that integrates three components: classification loss, bounding box regression loss, and mask loss. The total loss is the weighted sum of these three components, ensuring balanced optimization of classification, localization, and mask prediction during training.

Up to 1000 epochs, early stopping based on validation loss was used to prevent overfitting during training of U-Net and FCN model. The epoch number of the Mask R-CNN was fixed at 100. A sigmoid activation function was performed in last layer of the architectures during the trainings, this turned the raw outputs of the model into probability outputs, which are meaningful for pixel-wise classification among the building and no building classes. The performance was checked for the validation set after every epoch using evaluation metrics, which give an insight into the model's performance in delineating building boundaries.

After training, the model was tested on an independent dataset to check its generalization capability by comparing predictions with ground truth annotations. The training process was run using a hardware configuration of: Core i9 CPU, 64 GB RAM, 8.0 GB NVIDIA GeForce GTX1070 GPU. Totally, six models have been generated as shown in figure 2.

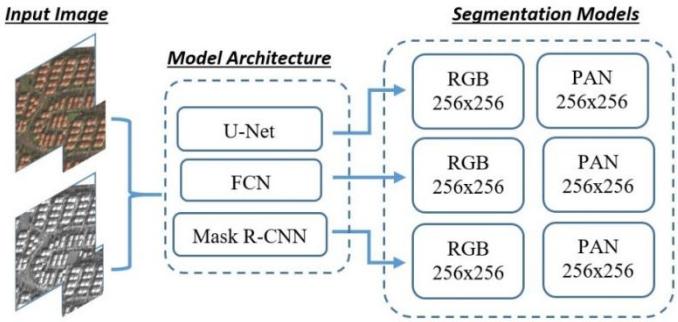


Figure 2: Model generations flow chart for building extraction

3.4 Integration Strategy

In order to enhance the building boundary segmentation, it is expected that by the integrate U-Net, FCN, and Mask R-CNN, will lead to better results. In this study, two sequential hybrid integration methodologies were developed to perform both semantic and instance segmentation by combining each U-Net and FCN with Mask R-CNN architectures. A semantic segmentation was produced using the U-Net and FCN model, taking the advantage of the encoder-decoder structure with skip connections. The models were independently trained on RGB and panchromatic datasets based on the hyperparameters that were mentioned in section 3.3.

The output of the U-Net and FCN was a pixel-wise segmentation mask for identifying the building regions in the input imagery. The final masks were then used as pre-processed inputs to the Mask R-CNN architecture and served as enhanced feature maps that provided precise boundary details to the region proposal network (RPN) in Mask R-CNN. Then, the RPN detects the region of interest (ROI) in the U-Net and FCN mask outputs, and by the use of a ResNet-101 backbone, it refined the segmentation further to provide instance-level masks for individual buildings. This also allowed Mask R-CNN to correct boundary imprecisions of U-Net and FCN, improving building edge detection accuracy. Figure 3 illustrates the methodology of the integration approach.

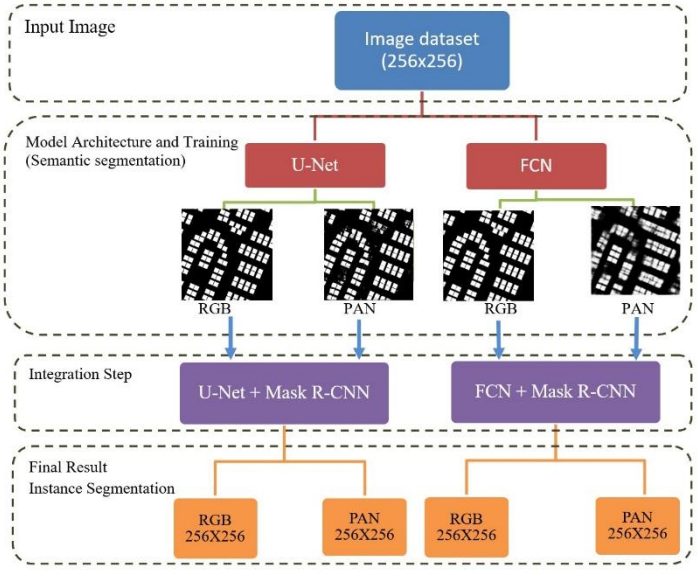


Figure 3: Hybrid integration methodologies

4.Experimental Results

In this section, the performance of standalone architectures and the hybrid integration approaches of U-Net and FCN with Mask R-CNN architectures for segmentation building boundaries from satellite images has been assessed using a comprehensive dataset. The used dataset includes two different subsets of image type: panchromatic and RGB described in sections 3.1. The study area covers a total of 405 buildings. The dataset consists of 201 images,

with a splitting into the training, validation, and test sets in a ratio of 80%, 10%, 10% respectively. This distribution will ensure robust training and an adequate amount of data for validation and testing using these models for generalization ability.

4.1 Performance Metrics

For the quantitative assessment purposes, the performance of the segmentation models have been evaluated utilizing metrics shown below:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

$$IoU = \frac{A \cap B}{A \cup B} \quad (4)$$

$$Improvement\ accuracy = \frac{New\ value - Original\ value}{Original\ value} \times 100 \quad (5)$$

Where |A∩B| is the area of overlap between the predicted segmentation (A) and the ground truth (B), and |A∪B| is the area of their union, TP is the number of true positives and FP is the number of false positives. FN is the number of false negatives.

4.2 Results and Evaluation

The obtained results from applying the deep learning are illustrated in Figures 4 through 6. It shows the results of building segmentations performed by U-Net, FCN, and Mask R-CNN individually. All the buildings are identified with both algorithms and using the RGB and Pan data. The results of hybrid integration models are shown in figure 7.

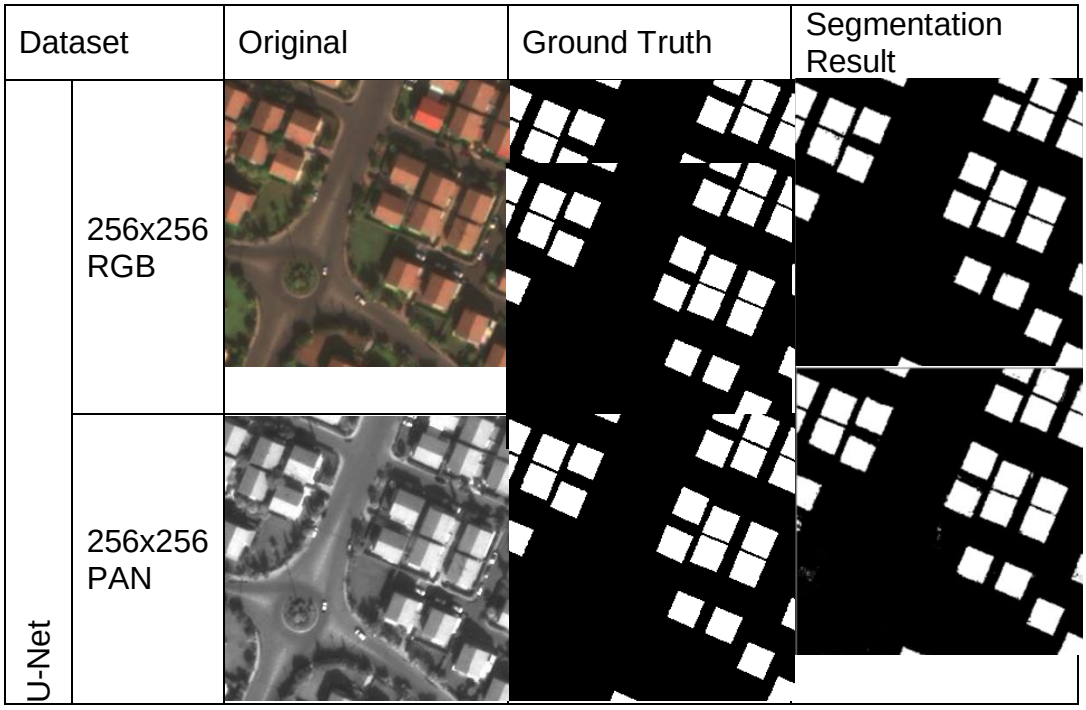


Figure 4: U-Net segmentation results

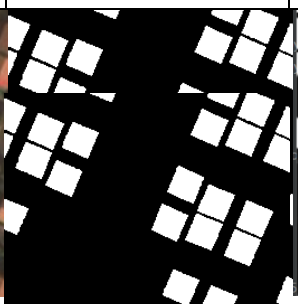
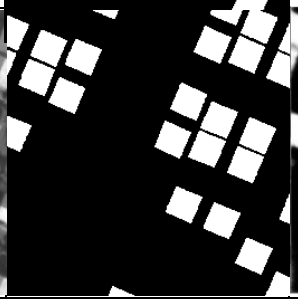
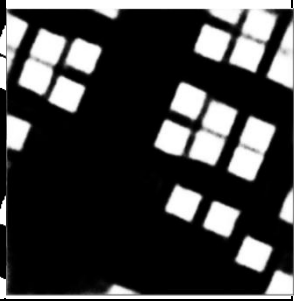
Dataset		Original	Ground Truth	Segmentation Result
FCN	256x256 RGB			
	256x256 PAN			

Figure 5: FCN segmentation results

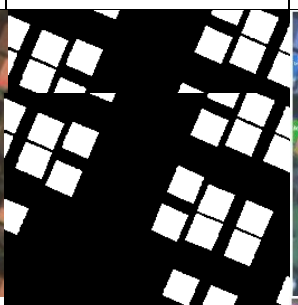
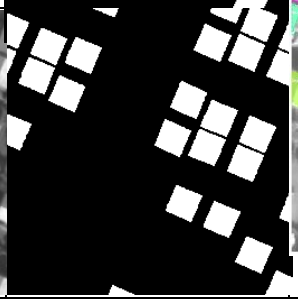
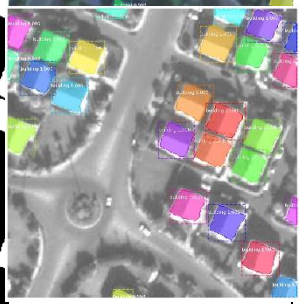
Dataset		Original	Ground Truth	Segmentation Result
Mask-RCNN	256x256 RGB			
	256x256 PAN			

Figure 6: Mask R-CNN segmentation results

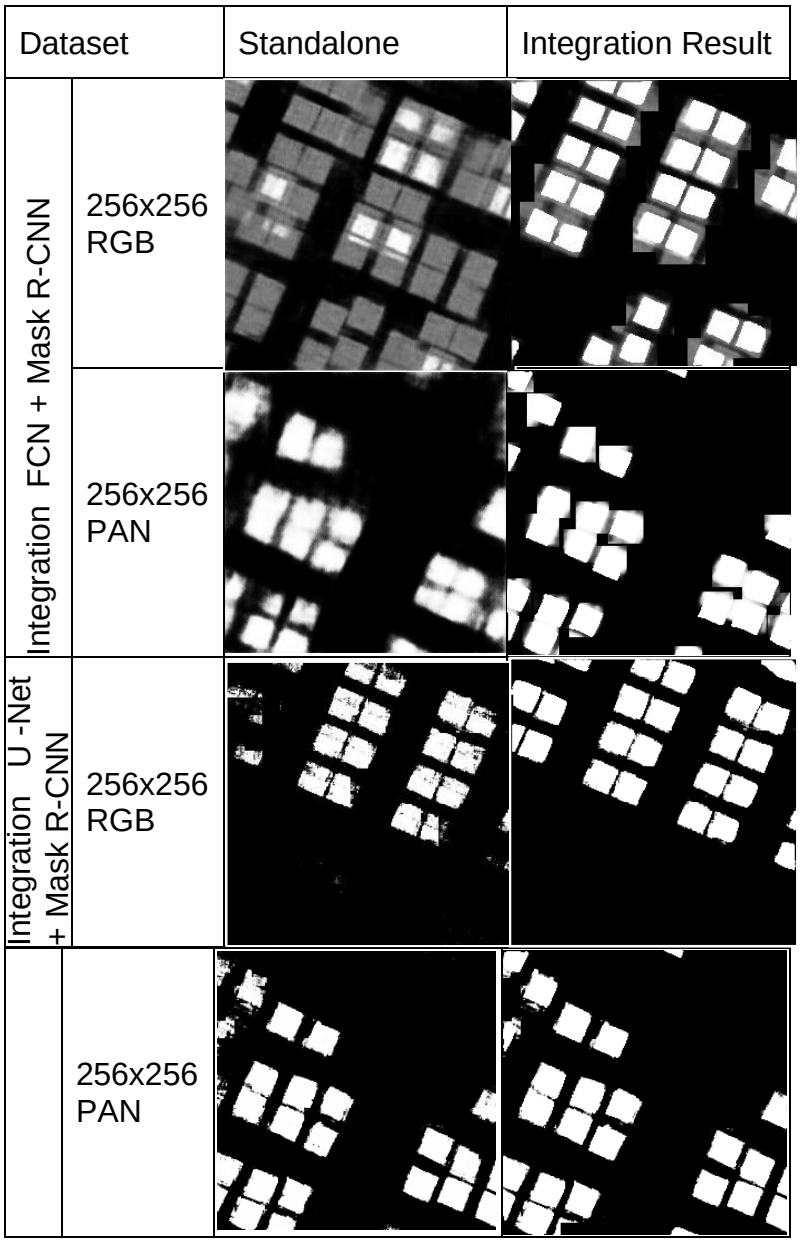


Figure 7: Integration results

4.2.1 Quantitative Evaluation

For the quantitative assessment of the generated models with U-Net, FCN, and Mask R-CNN, the four metrics were used: recall, precision, F1-Score, and IoU. These were comprehensively analyzed for both RGB and panchromatic datasets. Table 4.2 represents the independent results of U-Net, FCN, and Mask R-CNN architectures. It can be noticed that U-Net achieved better accuracy from the RGB dataset; considering Recall, F1-Score and IoU of 0.902, 0.950 and 0.982 respectively, as shown in figure 8. The Recall, F1-Score and IoU metrics emphasizes the performance of the FCN model in panchromatic dataset through 0.923, 0.950 and 0.975, respectively, figure 9. Similarly, for Mask R-CNN, a metrics of Recall, Precision, and F1-Score of 0.936, 0.927, and 0.931 respectively; demonstrates that the model obtained better performance in panchromatic dataset.

Table 4.2 Evaluation metrics of architectures

Metrics	U-Net		FCN		Mask R-CNN	
	PAN	RGB	PAN	RGB	PAN	RGB
Recall	0.892	0.902	0.923	0.860	0.936	0.893
Precision	0.986	0.983	0.897	0.952	0.927	0.831
F1-Score	0.940	0.950	0.950	0.920	0.931	0.860
IoU	0.981	0.982	0.975	0.973		

Regarding the integration results of U-Net with Mask R-CNN, it can be noticed from Table 4.3 and Figure 8 that the hybrid model provides significant improvements compared to the standalone model. The improvement accuracy has been computed utilizing equation 5. The results for U-Net with Mask R-CNN have shown an increase in recall of 9.9% for panchromatic and 8.1% for RGB datasets, proving its capability to detect more true positives. It achieved F1-Score improvement of 4.3% and 2.6% for PAN and RGB, respectively, reflecting balanced performance between recall and precision. Decreased IoU by -0.4% in PAN and -1.1%, in RGB indicating struggles in achievement precise overlaps of predicted and ground truth segmentations.

Table 4.3 Evaluation metrics of Integration U-Net + Mask RCNN architectures

Metrics	Standalone U-Net		Integration U-Net+ Mask R-CNN		Improvement accuracy (%)	
	PAN	RGB	PAN	RGB	PAN	RGB
Recall	0.892	0.902	0.980	0.975	**9.9**	8.1
Precision	0.986	0.983	0.980	0.976	-0.6	-0.7
F1- Score	0.940	0.950	0.980	0.975	**4.3**	2.6
IoU	0.981	0.982	0.977	0.971	-0.4	-1.1

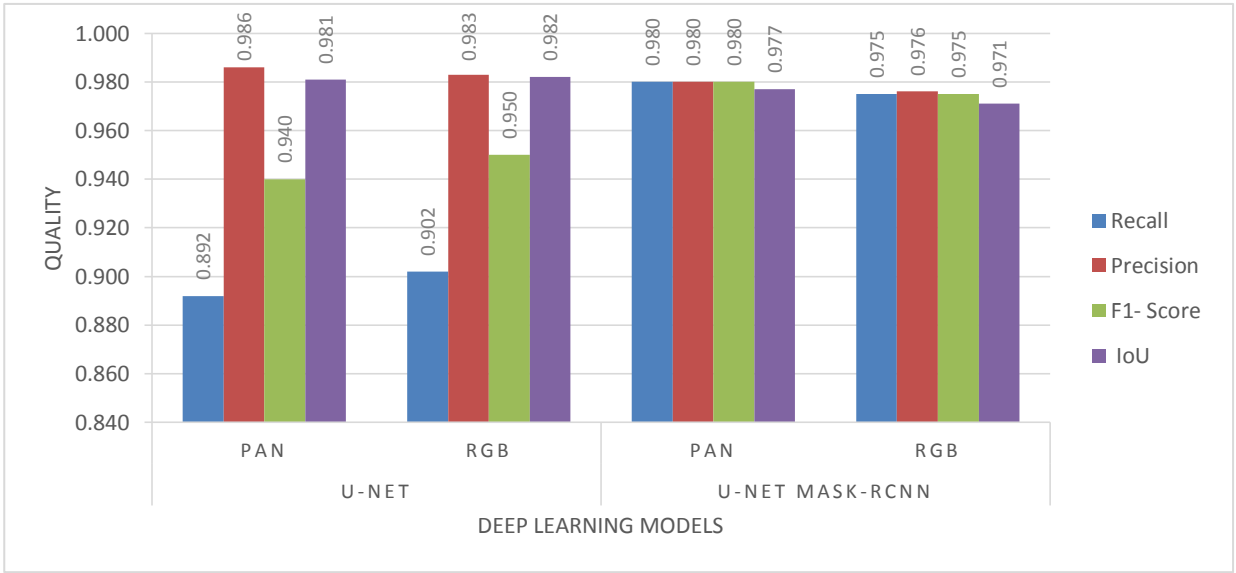


Figure 8: Standalone and integration of U-Net and Mask R-CNN

In contrast to the integration of FCN with Mask R-CNN, Figure 9 shows a remarkable enhancement of the hybrid model specifically in F1-Score compared to the standalone. Table 4.4 demonstrated a remarkable result of the integration approach in recall on RGB datasets, improving as much as 9.9%, while for panchromatic datasets, the improvement is 4.1%. Besides, precision for PAN datasets has a great increment of 7.6%, which shows a decrease in false positives. Despite these strengths, its IoU was substantially weakened: -2.1% PAN and -4% RGB, which are indications of coarse boundary predictions and reducing alignment accuracy.

Table 4.4 Evaluation metrics of Integration FCN+Mask R-CNN architectures

Metrics	FCN		FCN + Mask R-CNN		Improvement accuracy (%)	
	PAN	RGB	PAN	RGB	PAN	RGB
Recall	0.923	0.860	0.961	0.945	4.1	**9.9**
Precision	0.897	0.952	0.965	0.955	7.6	0.3
F1-Score	0.950	0.920	0.962	0.947	1.3	**2.9**
IoU	0.975	0.973	0.955	0.934	-2.1	-4.0

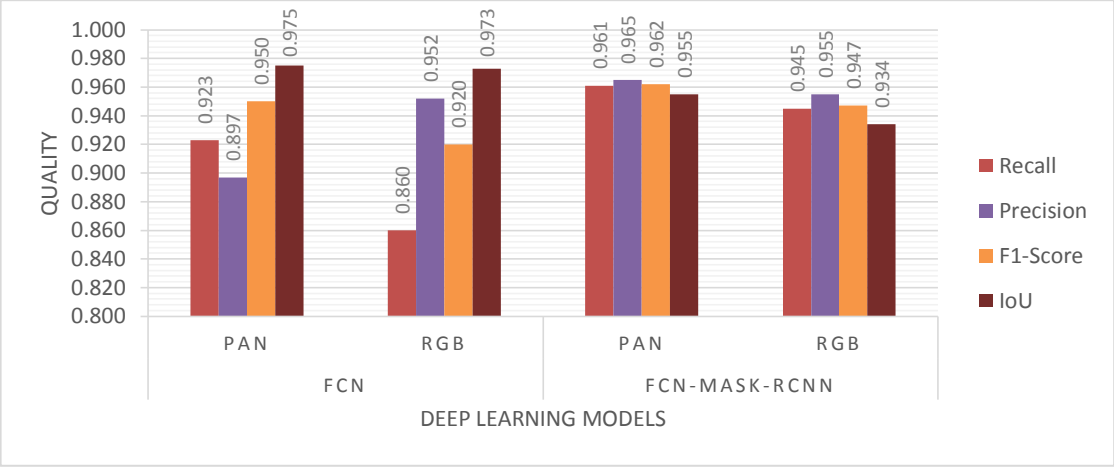


Figure 9: Standalone and integration of FCN and Mask-RCNN

5.Discussion

This work investigates the integration of U-Net and FCN with Mask R-CNN for constructing boundary segmentation from high-resolution satellite imagery. The results have shown that each integration leverages the strengths of the individual models in order to handle challenges inherent in the delineation of complex building boundaries, such as occlusions, varying building scales, and diverse architectural styles.

Integration of U-Net with Mask R-CNN resulted in notable improvements of 9.9% in recall and 4.3% in the F1-score within the panchromatic datasets. This reflected the strength of the U-Net's encoder-decoder architecture in capturing fine-grained spatial details that have been further refined by region proposal and instance segmentation in Mask R-CNN. Minor reductions in IoU by -0.4% for PAN and -1.1% for RGB; argued that predictions with preserved boundary precision might be improved if aligning to ground truth better, especially overlapping and irregular structures.

The FCN and Mask R-CNN integration showed outstanding performance in recall for RGB datasets with an improvement of 9.9%, this is indicative that FCN has captured the global context. Most importantly, precision has improved by 7.6% for PAN datasets, showing that this network is performing well in reducing false positives. However, IoU showed a significant drop of -2.1% for PAN and -4% for RGB, indicating the challenge of maintaining boundary alignment. This might be improved by

incorporating boundary-aware loss functions or sophisticated post-processing methods. Results further emphasize the relative performances through bar charts for the two integration methods, Figures 8 and 9. The integration of U-Net with Mask R-CNN outperformed in preserving boundary precision and handling high-resolution data, while FCN with Mask R-CNN combination excelled in global context awareness. The performance of integration U-Net with Mask R-CNN model has been further analyzed by comparing to the integration results obtained by Alsabhan et al., 2022, and Li et al., 2021for the same integration. Figure 10 reveals the precision, F1-score, and IoU of 0.98, 0.98, and 0.977 respectively of the proposed integration scheme, which outperforms 0.88 and 0.93 precision, 0.84 and 0.85 F1-score-score, and 0.74 IoU for both Alsabhan et al., 2022, and Li et al., 202, respectively.

These findings, in fact, reflect the complementary nature of the two hybrid approaches and suggest that even better segmentation results could be obtained by fusing their outputs. Each integration has shown a different strength and weakness. Comparative characteristics between both proposed integration methods have been shown in Table 5.1. Future work should therefore investigate ensemble methods or multi-stage architectures that leverage the advantages of both integrations for further improvement in the accuracy and robustness of segmentation.

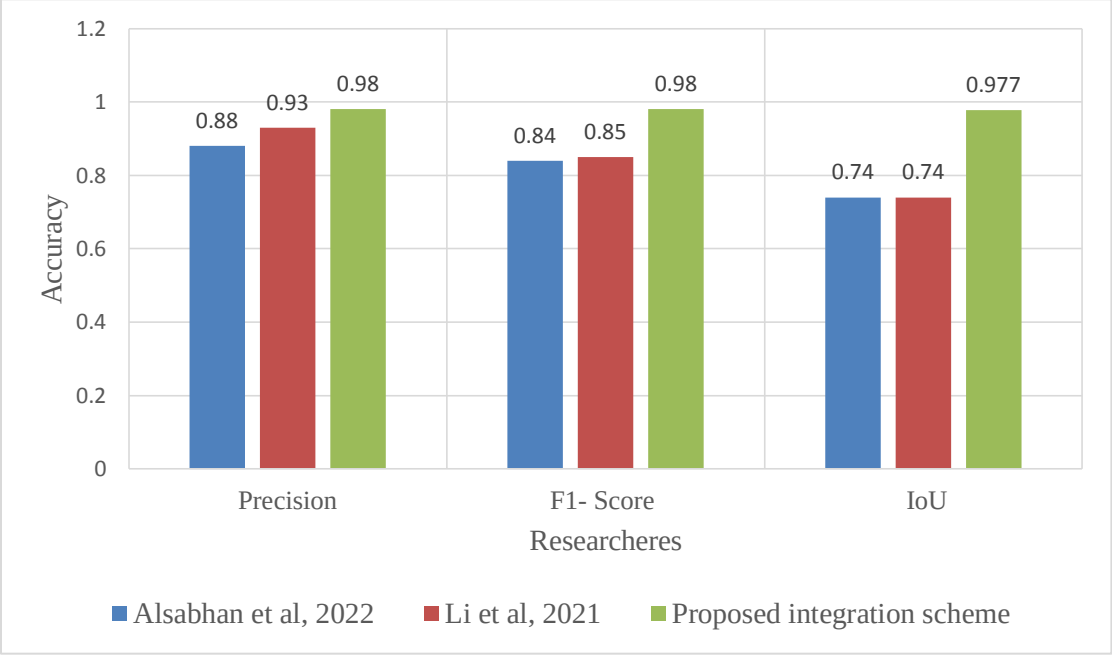


Figure 10: Comparison between U-Net with Mask R-CNN model with other researchers

Table 5.1 Comparative characteristics

Aspect	U-Net + Mask R-CNN	FCN + Mask R-CNN
Strengths	Boundary preservation, better precision	Superior recall, good global context
Weaknesses	Slightly lower IoU	Coarse boundaries, lower precision
Best Dataset	Panchromatic (PAN)	RGB
Applications	Detailed mapping, cadastral tasks	Urban planning, large-scale detection

6.Conclusion

This study investigated the use of sequential hybrid deep models, specifically U-Net and FCN integration with Mask R-CNN, for improving the segmentation of building boundaries using high-resolution satellite imagery. The proposed integration technique enhances satellite image segmentation by leveraging the powers of both instant and semantic segmentation techniques. The results have shown that U-Net with Mask R-CNN combination can yield a promising performance regarding boundary precision, while giving balanced segmentation performance on high-resolution satellite images of the panchromatic features. Furthermore, the results also show a remarkable improvement in the recall and F1-Score metrics. On the other hand, FCN with Mask R-CNN integration proved to yield much better results on recall, especially for RGB datasets, reflecting an ability of global spatial context encoding, despite some limitations in boundary alignment. A comparison of RGB and panchromatic datasets in both integrations underlined the adaptability of those integrations to changing data modalities. The U-Net integration fits best in applications where fine boundary delineation was necessary, while FCN integration was considered better when detecting larger building areas with complex layout. These findings reveal the potential of hybrid models for handling the building segmentation challenges, especially considering diverse urban and rural scenarios. Moreover, the paper contributes to developing a framework that is scalable and robust for the accurate segmentation of buildings which requires a correct topology and accurate building shape models.

Acknowledgment: Authors would like to thank Salahaddin University-Erbil for their support throughout the research.

Financial support: No financial support.

Potential conflicts of interest: Authors declare no conflicts of interest relevant to this article.

References

Ahmadian, N., Sedaghat, A., Mohammadi, N. & Aghdami-Nia, M. 2024. Deep-Learning-Based Edge Detection for Improving Building Footprint Extraction from Satellite Images. *Environmental Sciences Proceedings,* 29**,** 61.

Alsabhan, W., Alotaiby, T. & Dudin, B. 2022. Detecting Buildings and Nonbuildings from Satellite Images Using U-Net. *Computational Intelligence and Neuroscience,* 2022**,** 4831223.

Amarù, S., Marelli, D., Ciocca, G. & Schettini, R. 2023. DALib: A Curated Repository of Libraries for Data Augmentation in Computer Vision. *Journal of Imaging,* 9**,** 232.

Anh, H. T., Tuan, T. A., Long, H. P., Ha, L. H. & Thang, T. N. Multi Deep Learning Model for Building Footprint Extraction from High Resolution Remote Sensing Image. 2022 Singapore. Springer Nature Singapore, 246-252.

Aryal, J. & Neupane, B. 2023. Multi-Scale Feature Map Aggregation and Supervised Domain Adaptation of Fully Convolutional Networks for Urban Building Footprint Extraction. *Remote Sensing,* 15**,** 488.

Ayala, C., Sesma, R., Aranda, C. & Galar, M. 2021. A deep learning approach to an enhanced building footprint and road detection in high-resolution satellite imagery. *Remote Sensing,* 13**,** 3135.

Bai, J., Jia, C., Yu, S., Sun, L., Zhang, L., Chang, Z. & Hou, A. Building Extraction from High-Resolution Remote Sensing Images Using Improved HRNet Method. IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium, 2024. IEEE, 7982-7985.

Bittner, K., Adam, F., Cui, S., Körner, M. & Reinartz, P. 2018. Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 11**,** 2615-2629.

Bousias Alexakis, E. & Armenakis, C. 2022. Improving CNN-Based Building Semantic Segmentation Using Object Boundaries. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLIII-B3-2022**,** 41-48.

Carvalho, O. L. F. D., De Carvalho Junior, O. A., Albuquerque, A. O. D., Bem, P. P. D., Silva, C. R., Ferreira, P. H. G., Moura, R. D. S. D., Gomes, R. a. T., Guimaraes, R. F. & Borges, D. L. 2020. Instance segmentation for large, multi-channel remote sensing imagery using mask-RCNN and a mosaicking approach. *Remote Sensing,* 13**,** 39.

Dalal, A.-A., Shao, Y., Alalimi, A. & Abdu, A. 2020. Mask R-CNN for geospatial object detection. *International Journal of Information Technology and Computer Science (IJITCS),* 12**,** 63-72.

Das, S. Automated Building Segmentation in Areal Images Using Boundary Edge Detection. 2024 Singapore. Springer Nature Singapore, 237-250.

Dutta, A. & Zisserman, A. The VIA annotation software for images, audio and video. Proceedings of the 27th ACM international conference on multimedia, 2019. 2276-2279.

Gao, J., Zhang, B., Wu, Y. & Guo, C. Building Extraction from High Resolution Remote Sensing Images Based on Improved Mask R-CNN. 2022 4th International Conference on Robotics and Computer Vision (ICRCV), 2022. IEEE, 1-6.

Gashti, E. H., Delavar, M. R., Guan, H. & Li, J. 2024. Semantic Segmentation Uncertainty Assessment of

Different U-net Architectures for Extracting Building Footprints. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* 10**,** 141-148.

Guillaume, C., Aramburu, C. & Bougdal-Lambert, I. 2017. Satellite image segmentation for building detection using U-Net. *Computer Science*.

Han, Q., Yin, Q., Zheng, X. & Chen, Z. 2021. Remote sensing image building detection method based on Mask R-CNN. *Complex & Intelligent Systems***,** 1-9.

He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. Proceedings of the IEEE international conference on computer vision, 2017. 2961-2969.

Ji, S., Wei, S. & Lu, M. 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing,* 57**,** 574-586.

Li, C., Fu, L., Zhu, Q., Zhu, J., Fang, Z., Xie, Y., Guo, Y. & Gong, Y. 2021a. Attention Enhanced U-Net for Building Extraction from Farmland Based on Google and WorldView-2 Remote Sensing Images. *Remote Sensing,* 13**,** 4411.

Li, L., Zhang, T., Oehmcke, S., Gieseke, F. & Igel, C. 2023. BuildSeg: a general framework for the segmentation of buildings. *arXiv preprint arXiv:2301.06190*.

Li, W., He, C., Fang, J., Zheng, J., Fu, H. & Yu, L. 2019. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sensing,* 11**,** 403.

Li, Y., Xu, W., Chen, H., Jiang, J. & Li, X. 2021b. A novel framework based on mask R-CNN and histogram thresholding for scalable segmentation of new and old rural buildings. *Remote Sensing,* 13**,** 1070.

Liu, P., Liu, X., Liu, M., Shi, Q., Yang, J., Xu, X. & Zhang, Y. 2019. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. *Remote Sensing,* 11**,** 830.

Liu, Z., Chen, B. & Zhang, A. Building segmentation from satellite imagery using U-Net with ResNet encoder. 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 25-27 Dec. 2020 2020. 1967-1971.

Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 3431-3440.

Luo, L., Li, P. & Yan, X. 2021. Deep Learning-Based Building Extraction from Remote Sensing Images: A Comprehensive Review. *Energies,* 14**,** 7982.

Lussange, J., Yu, M., Tarabalka, Y. & Lafarge, F. 2023. 3D detection of roof sections from a single satellite image and application to LOD2-building reconstruction. *arXiv preprint arXiv:2307.05409*.

Mace, E., Manville, K., Barbu-Mcinnis, M., Laielli, M., Klaric, M. & Dooley, S. 2018. Overhead detection: Beyond 8-bits and rgb. *arXiv preprint arXiv:1808.02443*.

Mou, L. & 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*.

Neupane, B., Horanont, T. & Aryal, J. 2021. Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. *Remote Sensing,* 13**,** 808.

Noureldeen, A. & Wahed, M. E. 2024. Enhanced building footprint extraction from satellite imagery using Mask R-CNN and PointRend. *Bulletin of Electrical Engineering and Informatics,* 13**,** 3601-3608.

Prathap, G. & Afanasyev, I. Deep learning approach for building detection in satellite multispectral imagery. 2018 international conference on intelligent systems (IS), 2018. IEEE, 461-465.

Pushparaj, J. & Hegde, A. V. 2017. Evaluation of pan-sharpening methods for spatial and spectral quality. *Applied Geomatics,* 9**,** 1-12.

Raghavan, R., Verma, D. C., Pandey, D., Anand, R., Pandey, B. K. & Singh, H. 2022. Optimized building extraction from high-resolution satellite imagery using deep learning. *Multimedia Tools and Applications,* 81**,** 42309-42323.

Reda, K. & Kedzierski, M. 2020. Detection, Classification and Boundary Regularization of Buildings in Satellite Imagery Using Faster Edge Region Convolutional Neural Networks. *Remote Sensing,* 12**,** 2240.

Ronneberger, O., Fischer, P. & Brox, T. 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, 2015. Springer, 234-241.

Sakeena, M., Stumpe, E., Despotovic, M., Koch, D. & Zeppelzauer, M. 2023. On the Robustness and Generalization Ability of Building Footprint Extraction on the Example of SegNet and Mask R-CNN. *Remote Sensing,* 15**,** 2135.

Sariturk, B. & Seker, D. Z. 2022. A Residual-Inception U-Net (RIU-Net) Approach and Comparisons with U-Shaped CNN and Transformer Models for Building Segmentation from High-Resolution Satellite Images. *Sensors,* 22**,** 7624.

Shorten, C. & Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of big data,* 6**,** 1-48.

Susetyo, D. B., Harintaka, H. & Rizaldy, A. The application of mask R-CNN for building extraction. The 9th International Seminar on Aerospace Science and Technology, 2023 Bogor, Indonesia. AIP Publishing.

Thakur, V., Doja, M., Ahmad, T. & Rawat, R. 2019. Cadastral boundary extraction and image classification using OBIA and machine learning for National Land Records Modernization Programme in India. *J. Remote Sens. GIS,* 8.

Wagner, F. H., Dalagnol, R., Tarabalka, Y., Segantine, T. Y., Thomé, R. & Hirye, M. C. 2020. U-net-id, an instance segmentation model for building extraction from satellite images—case study in the joanópolis city, brazil. *Remote Sensing,* 12**,** 1544.

Wang, W., Shi, Y., Zhang, J., Hu, L., Li, S., He, D. & Liu, F. 2023. Traditional village building extraction based on

improved Mask R-CNN: a case study of Beijing, China. *Remote Sensing,* 15**,** 2616.

Wang, X., Tian, M., Zhang, Z., He, K., Wang, S., Liu, Y. & Dong, Y. 2024. SDSNet: Building Extraction in High-Resolution Remote Sensing Images Using a Deep Convolutional Network with Cross-Layer Feature Information Interaction Filtering. *Remote Sensing,* 16**,** 169.

Xia, L., Zhang, X., Zhang, J., Yang, H. & Chen, T. 2021. Building Extraction from Very-High-Resolution Remote Sensing Images Using Semi-Supervised Semantic Edge Detection. *Remote Sensing,* 13**,** 2187.

Yan, G., Jing, H., Li, H., Guo, H. & He, S. 2023. Enhancing Building Segmentation in Remote Sensing Images: Advanced Multi-Scale Boundary Refinement with MBR-HRNet. *Remote Sensing,* 15**,** 3766.

Yu, Y., Wang, C., Kou, R., Wang, H., Yang, B., Xu, J. & Fu, Q. 2024. Enhancing Building Segmentation With Shadow-Aware Edge Perception. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 18**,** 1-12.

Zhang, L., Dong, R., Yuan, S., Li, W., Zheng, J. & Fu, H. 2021. Making Low-Resolution Satellite Images Reborn: A Deep Learning Approach for Super-Resolution Building Extraction. *Remote Sensing,* 13**,** 2872.

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**,** 1465.

Zhang, Y. & Chi, M. 2020. Mask-R-FCN: A deep fusion network for semantic segmentation. *IEEE Access,* 8**,** 155753-155765.

Zhao, K., Kamran, M. & Sohn, G. 2020. Boundary Regularized Building Footprint Extraction from Satellite Images Using Deep Neural Networks. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.,* V-2-2020**,** 617-624.

Zorzi, S. & Fraundorfer, F. Regularization of building boundaries in satellite images using adversarial and regularized losses. IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, 2019. IEEE, 5140-5143.