DOI: 10.24996/ijs.2026.67.9.29

ISSN: 0067-2904

Deep Learning-Based Building Detection from HR Aerial Imagery: Application of Mask R-CNN in Baghdad

Aseel K. Hasheem*, Fanar M. Abed

1Department of Surveying Engineering, College of Engineering, University of Baghdad, Baghdad, Iraq..

Received: 1/2/2025 Accepted: 8/ 7/2025 Published: xx

Abstract

This paper suggests a practical application method that effectively detects buildings in HR aerial photos in Baghdad by creating a segmentation mask using the Mask R-CNN deep learning model. The paper used aerial 10 cm resolution images of Baghdad as the main case study, focusing on a particular urban region neighborhood with various land feature types, building sizes, and population density. Another public access aerial image with a resolution of 7.5 cm was used in New Zealand to assess the DL model's performance for different building architectures. The study found that Mask R-CNN has effectively extracted buildings from HR aerial images of Baghdad, outputting indices with reasonable accuracy. The method achieved a segmentation accuracy of 81.655% and a bounding box accuracy of 83.585% on test data from various areas within the city. Analysis revealed that most incorrect predictions occurred within Baghdad's Al-Sadr district. This was attributed to irregular building layouts, closely spaced or adjacent buildings, and high image brightness caused by sunlight. This issue could be mitigated by considering image acquisition time to minimize the effects of sun angle and its resulting high brightness. Effective processing parameters and image enhancement techniques mitigated occlusion issues and reduced false positives, enabling these methods to facilitate the successful detection of buildings in Al-Mansour and Al-Karrada districts that were not initially present in the ground truth data.

Keywords: Building Detection; Computer Vision; Deep Learning; HR Aerial Images, Mask RCNN; Segmentation.

الكشف عن المباني باستخدام التعلم العميق من الصور الجوية عالية الدقة: تنفيذ نموذج *Mask-RCNN* في بغداد

اسيل كريم هاشم* , فنار منصور عبد

قسم هندسة المساحة، كلية الهندسة، جامعة بغداد، بغداد، العراق

الخلاصة

يقترحُ هذا البحثُ طريقةً تكشفُ بشكلٍ فعّالٍ عن المباني في الصورِ الجويةِ عاليةِ الدقةِ عن طريق إنشاءِ ماسك تجزئةٍ باستخدامِ نموذجِ التعلُّمِ العَميق Mask R-CNN. استخدمَ البحثُ صورًا جويةً بدقةِ 10 سم لبغداد كمنطقة دراسة رئيسية، مع التركيزِ على منطقةِ حيٍّ حضريٍّ مُحدَّدٍ بأنواعٍ مُختلِفةٍ من معالمِ الأرض، وأحجامِ

*Email : aseel.kareem2208@coeng.uobaghdad.edu.iq

مبانٍ، وكثافةٍ سكانيةٍ. ولأغراضِ التقييم، تمَّ استخدامُ صورٍ جويةٍ أخرى مُتاحةٍ لجميع المستخدمين بدقةِ 7.5 سم فوق نيوزيلندا لتقييمِ أداءِ نموذجِ التعلُّمِ العَميق على هندسةٍ معماريةٍ مُختلِفةٍ للمباني. وجدتِ الدراسةُ أنَّ Mask R-CNN استخلصَ المباني بشكلٍ فعّالٍ من الصورِ الجويةِ عاليةِ الدقةِ لبغداد. حقَّقَ النموذجُ دقةَ تجزئةٍ بنسبةِ 81.655٪ ودقةَ مربعِ إحاطةٍ بنسبةِ 83.585٪ على بياناتِ اختبارٍ غيرِ مرئيةٍ من مناطقَ مُختلِفةٍ داخلَ المدينة. كشفَ التحليلُ أنَّ غالبيةَ التنبؤاتِ غيرِ الصحيحةِ حدثت داخلَ حيِّ مدينةِ الصدر في بغداد. يُعزى ذلك إلى التخطيطاتِ غيرِ المنتظمةِ للمباني، والمباني المُتقاربةِ أو المُتجاوِرة، والسطوعِ العالي للصورةِ الناتجِ عن ضوءِ الشمس. يُمكنُ للبحثِ المُستقبلي تخفيفَ هذه المشكلةِ من خلالِ مراعاةِ وقتِ التقاطِ الصورةِ لتقليلِ آثارِ زاويةِ الشمس وسطوعِها العالي الناتج. خفَّفتْ مُعلماتُ المعالجةِ الفعّالةِ، وتحسينُ الصورةِ، وتقنياتُ زيادةِ البياناتِ مشكلاتِ حجب المعالم للظواهر الارضية كالمباني وقلَّلتْ من التنبؤاتِ الخاطئة، ممَّا أدَّى إلى نتائجِ تجزئةٍ واعدةٍ في مناطقَ مُحدَّدةٍ. سهَّلتْ هذهِ الطرقُ الكشفَ الناجحَ عن المباني في منطقتي المنصور والكرادة اللتين لم تكونا مكتشفتين في بياناتِ الحقيقةِ الأرضيةِ في البداية.

1. Introduction

Significant progress in space technology, photography, and remote sensing methods has enabled the widespread availability of precise information. The comprehensive texture and semantic details offered by high-resolution (VHR) images are invaluable for building extraction and land cover classification [2]. Automatic detection and extraction of buildings from remote-sensing images are crucial for estimating population density, facilitating urban planning, and producing and updating topographic maps, among other applications [3]. While building extraction has garnered significant interest, it remains a challenging endeavor due to factors such as noise, occlusion, and the intricate nature of backgrounds in original remote sensing images [4]. Traditional methods for building extraction, such as edge detection, Object-Based Image Analysis (OBIA), template matching, etc., predominantly depend on manual and expert analyses of the statistical characteristics of building shapes and textures obtained from remote sensing imagery [5]'[2]. Nevertheless, these techniques are often hindered by their limited accuracy, which poses difficulties in addressing the requirements of extensive datasets and the need for intelligent automated updates [6].

In recent years, advancements in computer hardware, coupled with the availability of extensive high-quality aerial imagery datasets, have facilitated the application of approaches grounded in Deep Learning (DL) for building extraction [7]'[8]. A prominent model within the realm of DL is the convolutional neural network (CNN), which has demonstrated considerable success in various tasks, including image classification, object detection, image semantic segmentation, and Internet of Things (IoT) [9]'[10]'[11]'[12]. DL architectures, particularly Convolutional Neural Networks (CNNs), have demonstrated significant efficacy in various applications within the realm of Computer Vision (CV) [13]. Neural networks have been applied in risk assessment following numerous natural disasters, including earthquakes [14] and forest fires [15]. Notably, CNNs tailored for semantic segmentation, such as Mask R-CNN, can autonomously recognize and outline objects within images, including structures like buildings [16]. This advancement provides a viable alternative to conventional methods, enabling the scalable, precise, and efficient extraction of buildings from remote sensing data.

The objective of this study is to automatically detect and differentiate buildings from the background and the surrounding objects in HR aerial images using the Mask R-CNN DL model [17]. This process is typically categorized as a segmentation task, a well-established topic within CV. However, numerous challenges may arise in accurately segmenting buildings from HR images, especially in densely populated urban environments with informal housing, such as in cities. To evaluate the performance and generalizability of the suggested approach for practical use, it is tested under difficult urban conditions.

2. Review of Literature

In recent years, many scholars have focused on enhancing and improving building detection from aerial images. This is challenged by several limitations, including the use of a labeled dataset, buildings covered by vegetation, unclear building boundaries, and variations in their size and shape. Therefore, scholars have developed different methods involving CNN design, as well as preprocessing and postprocessing techniques. Without a pre-processing phase and dependent on a patch-based routine, the CNN model includes five convolutional layers, and Global Average Pooling (GAP) is replaced with a fully connected layer. Hereby, Alshehhi *et al.* (2017) [20] developed the CNN method to enhance the prediction of building and road segmentation. This needs further processing to enhance the accurate identification of urban feature boundaries.

Later, Xu *et al.* (2018) proposed a novel framework that utilizes guided filters after the DL network application to optimize the classification results of VHR images and remove noise caused by salt and pepper. Additionally, it successfully preserved the building's borders within the image. However, some buildings covered by trees have shapes that are difficult to identify, and some boundaries are challenging to classify due to their hazy and uneven nature [21]. Xia *et al.* (2021) introduced a semi-supervised deep learning approach based on the Edge Detection Learning (EDLED) network, which combines semi-supervised learning and edge detection neural networks. This demonstrates how this approach can be generalized and offers a fresh perspective on learning generalization, highlighting the importance of applying post-processing to enhance the results [22]. Tian *et al.* (2022) designed encoder-decoder networks. They later suggest integrating atrous convolution, deformable convolution, attention mechanisms, and pyramid pooling modules to enhance the efficacy of feature extraction within the encoding path [23].

Then, Qiu et al. (2023) introduced the Attentional Feature Learning Neural Network (AFL-NET) [18], which enabled the reduction of loss in detailed information during the downsampling process and ensured a rich acquisition of features. The multiscale feature fusion module, grounded in a self-attention mechanism, adeptly learns the interrelationships among various features, facilitating optimal utilization of building characteristics while mitigating the risk of under-segmentation outcomes. Additionally, the shape feature refinement module can intelligently adapt to the shape features of buildings, enhancing the smoothness of their edges and concurrently reducing the likelihood of segmentation, which can help to overcome the inaccurate distinction between buildings and complex backgrounds resulting from the high contrast within building layers, even while the contrast between buildings and non-building objects in the image is low [18].

In the same way, to extract small-sized buildings and buildings with undistinguished borders and irregular shapes from aerial photographs, Yan *et al.* (2023) employed DL model where the Multi-Scale context Fusion Module (MSFM) [2] is integral to main architecture, as it effectively captures and enhances a variety of global multi-scale feature where it adeptly utilizes complex representations of spatial dimensions present in both high-level and low-level features and considered Such functionality is particularly beneficial for delineating building boundaries that display discontinuities and for accurately depicting the characteristics of irregularly shaped structures. Hence, the main model significantly improves its ability to capture subtle details [2], enhancing the overall efficacy of extraction tasks [24]. Further, the U-Net [25] (The U-Shaped

Architecture Network) model introduced by Ronneberger, Fischer, and Brox (2015) is highly regarded for its efficacy in semantic segmentation across multiple fields. It proves incredibly beneficial for applications where training data is scarce, such as in the segmentation of biomedical images [26]. By combining its advantages with the benefits of residual learning, atrous spatial pyramid pooling, and focal loss, the deep residual U-Net (RU-NET) [4] was proposed, successfully reducing the number of parameters and mitigating network degradation. Despite its efficacy, the approach may not achieve sufficient border accuracy for urban objects and may fail to recognize buildings wrapped by trees, misidentify shadows from the concrete floor as buildings, and exhibit inadequate recognition of aligned buildings [4].

All previous approaches aimed to increase the accuracy of building extraction and improve the efficiency of DL models in CV, considering the developments in contemporary computer resources. However, in Iraq, particularly in Baghdad, there is an urgent need to extract buildings for various purposes, such as studying urban expansion or conducting statistical analyses. Therefore, in this study, a segmentation methodology was applied using the MASK-RCNN to extract buildings in Baghdad from HR aerial images and compare outcomes with public dataset results to study and investigate the outputs in such a challenging implementation.

3. Materials and Methods

The methodology presented in this work includes four main stages. The first stage involves study planning, which includes selecting the study case and assessing data availability about the research requirements. The second stage is preparing the data for the case study, which goes through several steps, namely assigning (labeling) the buildings in the aerial images. As in the Baghdad dataset, the brightness of the images was improved when needed. The next stage was cutting the aerial images to a fixed and appropriate size and dividing them into training, validation, and testing datasets. Data training was applied in the third stage, and Mask R-CNN was utilized after fine-tuning. At the same time, the last stage comprises analyzing the results obtained and evaluating the accuracy of the outcome achieved. The synoptic structure of the methodology workflow is illustrated in Figure 1.

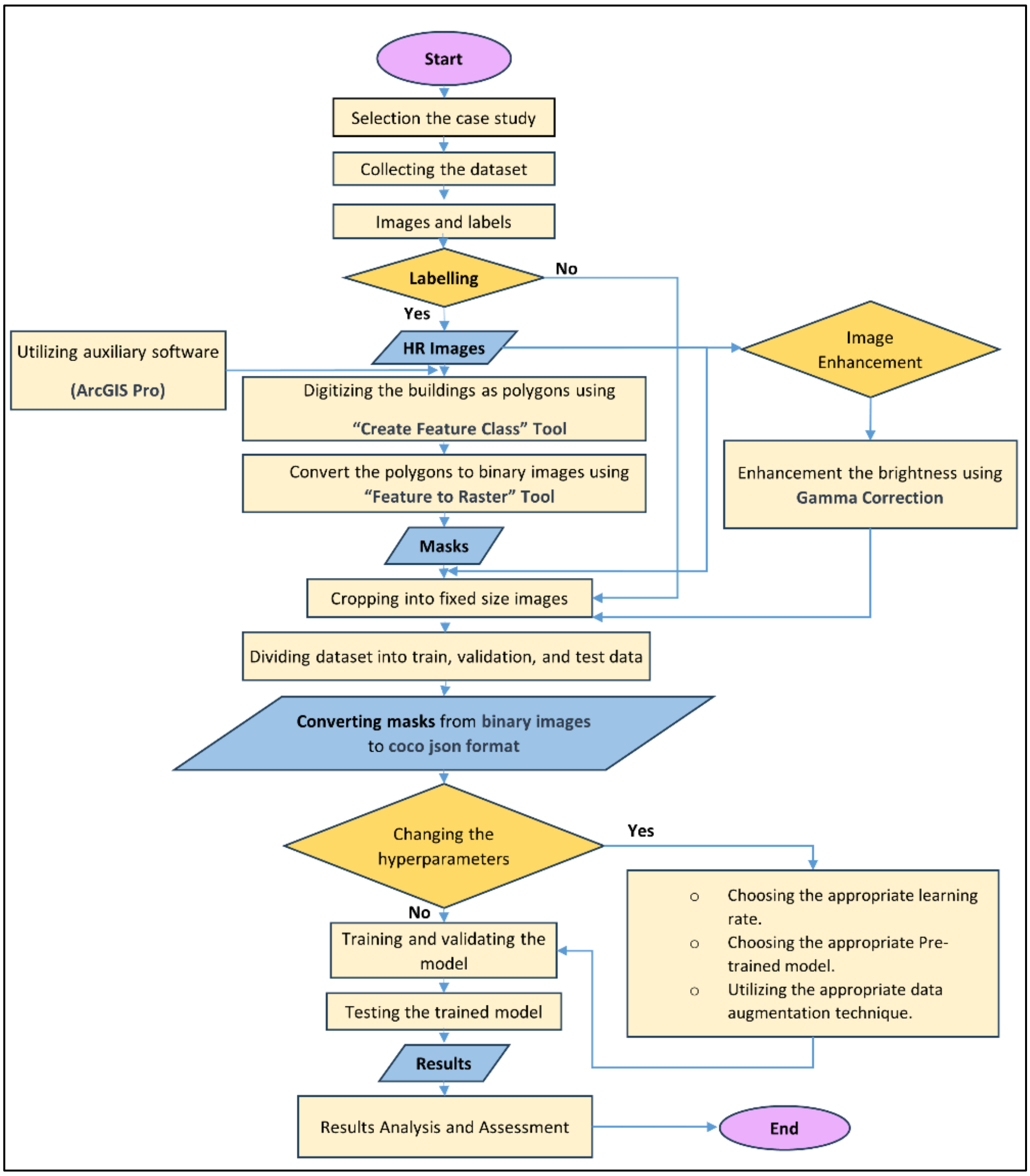


Figure 1: The Flow chart of the implemented methodology.

3.1 Case Study Description

The main case study in this research lies in the central region of Iraq (Baghdad city), which was selected due to its residential, natural, urban, and industrial diversity and its remarkable urban expansion during this decade. This requires continuous updating of the data for multiple classification and segmentation processes, and therefore, it has become the focus of this study. Aerial HR orthophotos with 10cm resolution were obtained from previous studies [27] of three urban regions in Baghdad, which were selected based on differences in shape, size, and the area of buildings, afforestation, and surrounding spaces. Samples from the regions of AL-Mansour, AL-Karrada, and AL-Sadr neighborhoods were selected through Figure 2 due to the diversity in the residential and commercial buildings, where there are small buildings that are close to each other and have few spaces, as in the AL-Sadr district, in contrast to large buildings that were scattered from each other and surrounded by trees, as in the AL-Mansour district. However, AL-Karrada includes most types of this diversity, which is needed to train the DL model.

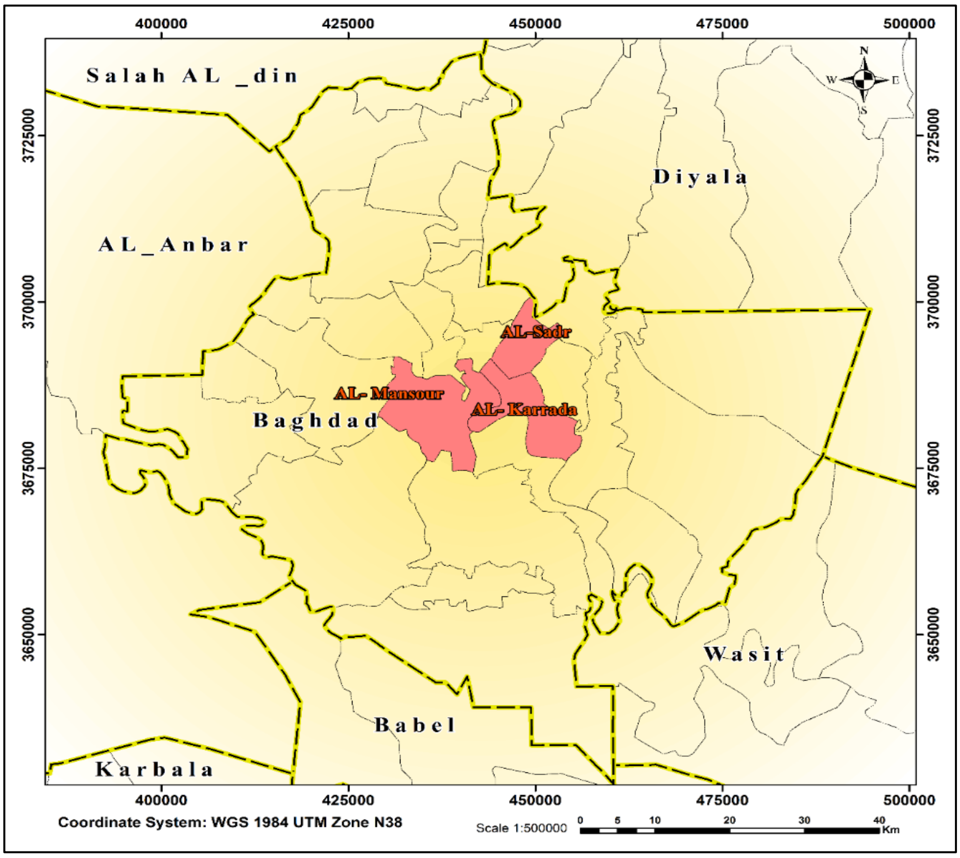


Figure 2: Baghdad case study showing DL application districts.

Furthermore, a pilot case study was selected for assessment purposes, focusing on Christchurch, a city in New Zealand. AIRS [28] (Aerial Images for Roof Segmentation) is a public dataset of aerial images with a resolution considered among the most accurate and available for scientific research. The aerial images have a resolution of 7.5 cm and include more than 220,000 buildings. This dataset covers approximately 457 km², provided by the LINZA data service [29]. The images were taken between 2015 and 2016 and manually labeled in 2016 as binary [30]. This dataset is for the city of Christchurch, one of New Zealand's largest cities. The pixel size is 10,000×10,000 pixels, comprising over a thousand images. Refer to Table 1 for a description of the dataset.

3.2 Dataset Preparation

The data preparation phase encompasses all operations performed on the data prior to running the training process of the DL model. This includes the labeling process for the aerial images, which was conducted solely for the main data of the study. The labeling of the Baghdad dataset was implemented following the ArcGIS Pro software using the (create feature) tool. Later, resulting polygons were converted to binary images and exported the same size as the corresponding aerial images using the (feature to raster) tool, which was 20000×20000. The next phase was applied to the mathematical operations by cropping the aerial orthophotos and masks into a fixed size of 512×512 pixels. In implementation, samples of images were selected for each dataset (Table 1).

Table 1: Dataset descriptions.

Dataset	Train images	Val images	Test images	Tile size	Total	Resolution
Baghdad	949	204	204	512*512 pixel	1357	10 cm
AIRS	1252	269	269	512*512 pixel	1790	7.5 cm

3.3 The DL Proposed Model

Mask R-CNN was applied at this stage, one of the important and distinctive DL models proposed by [17]. It is one of the most renowned systems in object detection and identification, recognized for its accuracy, efficiency, and ability to perform this task effectively. It is considered an advanced model for Faster R-CNN [11] by incorporating the feature of predicting the object mask, in addition to the Faster R-CNN outputs, which include classification and bounding boxes. The general structure of the model proposed in this research consists of the following basic formations:

1. The backbone is the Convolutional Neural Network (CNN) model with a Feature Pyramid Network FPN, the basic structure for extracting feature maps from images, as shown in Figure 3. In this research, the backbone RESNET101-FPN was used.
2. Region Proposal Network (RPN) generates suggestions for areas likely to contain objects to be detected, depending on the backbone outputs, which are the feature maps.
3. The RPN outputs enter the third stage, which includes the two paths. The first path predicts the coordinates of the specified square for the object in each proposed area outside of RPN, while the other path predicts the mask for the object within that square, as the two paths work in parallel.

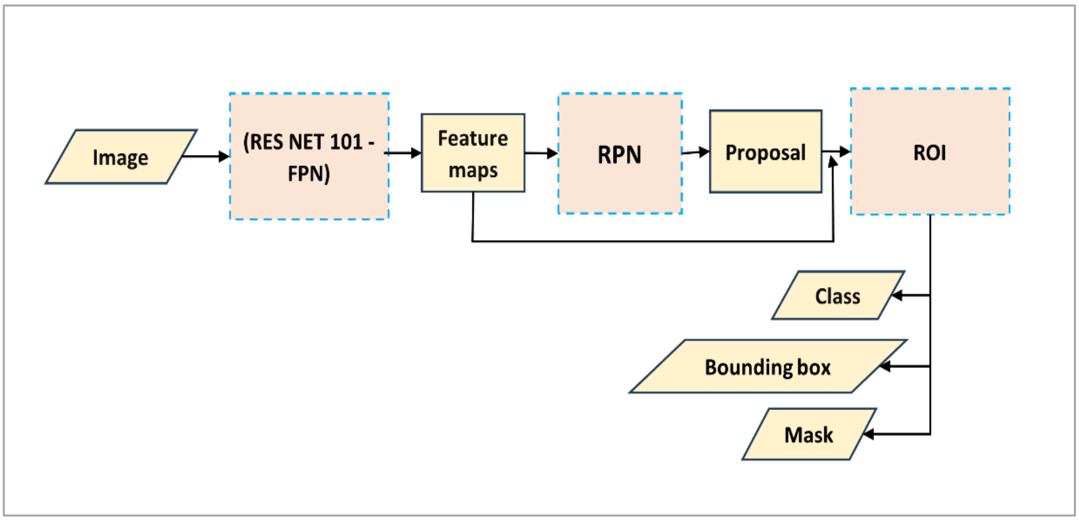


Figure 3: Mask R-CNN Architecture.

3.4 Experiment

The DL methodology was implemented using Anaconda Navigator [31], which served as a tool for managing environments and packages. This approach enabled the systematic organization of the methodology and the practical configuration of the work environment. Additionally, Jupyter Notebook [31] provides an interactive interface, enabling the smooth writing and execution of code, as well as integrating scripts and visualizations to analyze the results comprehensively. Detectron2 [32] represents the advanced iteration of the library developed by Facebook Artificial Intelligence Research, offering cutting-edge algorithms for detection and segmentation tasks. This library is designed to facilitate a variety of CV research initiatives as well as production applications within Facebook. Therefore, it was the optimal library for our research. The proposed approach was implemented in a computing environment defined by the hardware specifications detailed in Table 2.

Table 2: PC Specifications used to process the data.

CPU	13th Gen Intel®Core (TM)i7-13650HX
RAM	16 GB
Storage	500GB SSD
GPU	NVIDIA GeForce RTX 4060

The training process was conducted over 6,000 iterations. During the first thousand iterations, a warm-up phase was used with an introductory learning rate of 0.001 and a warm-up factor of 0.001. This step ensures the stability of the training and reduces the risk of overfitting that the model may face during training (Figure 4).

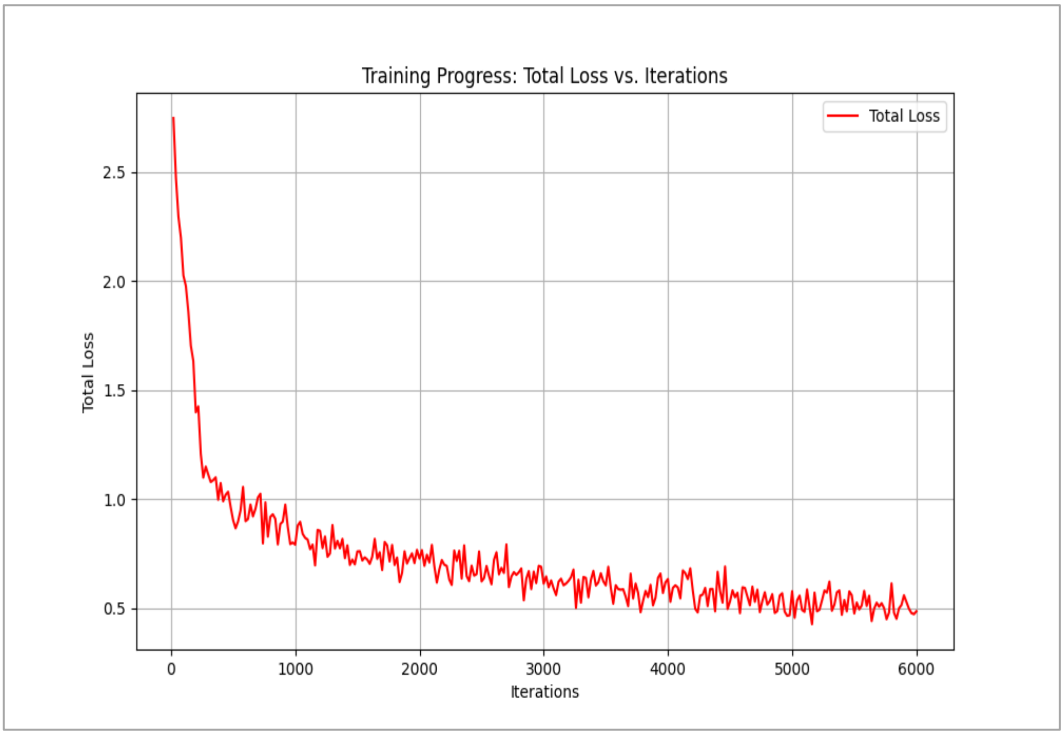


Figure 4: Total loss of Mask R-CNN Model.

The momentum value used was 0.9, and the weight decay was 0.0001. These values were chosen over their counterparts because they improved the model's average precision. At the intersection over union equal to 0.5, the model reached 4.255% for the bounding boxes and 4.841% for the segmentation if the learning rate was reduced from 0.01 to 0.001. A pre-trained model on the COCO dataset was utilized to accelerate the training process and leverage the previous weights.

The data was divided into training, validation, and test sets at a ratio of 70%, 15%, and 15%, respectively. The training data contains approximately 6190 buildings, while the validation data contains 1205. The validation process was carried out after 50 iterations during training, using a validation dataset that contains 1,294 buildings.

3.5 Evaluation Metrics

Coco metrics were used to evaluate the performance of the proposed model. AP was calculated using a set of thresholds of the intersection over union (IOU), ranging from 0.5 to 0.95 in increments of 0.05. Additional metrics, such as AP@0.5 and AP@0.75, were also reported.

Similarly, the average recall (AR) was calculated for the same set of IOU thresholds. The AP and AR were calculated for large, medium, and small-sized buildings, through which the model's performance is evaluated in discovering buildings of different sizes (Figure 5).

```
Average Precision  (AP) @[ IoU=0.50:0.95 | area=   all | maxDets=100 ]
Average Precision  (AP) @[ IoU=0.50      | area=   all | maxDets=100 ]
Average Precision  (AP) @[ IoU=0.75      | area=   all | maxDets=100 ]
Average Precision  (AP) @[ IoU=0.50:0.95 | area= small | maxDets=100 ]
Average Precision  (AP) @[ IoU=0.50:0.95 | area=medium | maxDets=100 ]
Average Precision  (AP) @[ IoU=0.50:0.95 | area= large | maxDets=100 ]
Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets=  1 ]
Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets= 10 ]
Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets=100 ]
Average Recall     (AR) @[ IoU=0.50:0.95 | area= small | maxDets=100 ]
Average Recall     (AR) @[ IoU=0.50:0.95 | area=medium | maxDets=100 ]
Average Recall     (AR) @[ IoU=0.50:0.95 | area= large | maxDets=100 ]
```

Figure 5: Coco metrics used to assess the detection process.

4. Results and Discussion

Following data testing, the proposed model was trained according to the criteria mentioned in the experiment section, using the learning ratio and the pre-trained model of the Baghdad dataset (Table 3).

Table 3: Building extraction results of Mask R-CNN of the Baghdad dataset.

	Metrics	ResNet50-FPN	ResNet50-FPN +DAT	ResNet101-FPN+DAT
Bounding box	AP 0.5:0.95	54.788	55.580	57.724
	AP 0.5	81.001	81.445	83.585
	AP 0.75	56.374	58.123	59.933
	AR 0.5:0.95	68.0	70.3	72.1
Segmentation	AP 0.5:0.95	53.810	54.550	55.982
	AP 0.5	79.782	80.470	81.655
	AP 0.75	56.732	58.596	60.258
	AR 0.5:0.95	66.4	68.6	69.3

The results demonstrate the improvement achieved by the model when utilizing the ResNet101-FPN backbone and the Data Augmentation Technique (DAT), including rotation by 15 degrees and horizontal flipping. The AP for the bounding boxes increased from 54.788 to 57.724, calculated for different values of the IOU. Specifically, at an IOU of 0.5, the AP increased from 81.001 to 83.585, while at an IOU of 0.75, it increased from 56.374 to 59.933. As for the AR, it improved significantly by about 4.1%.

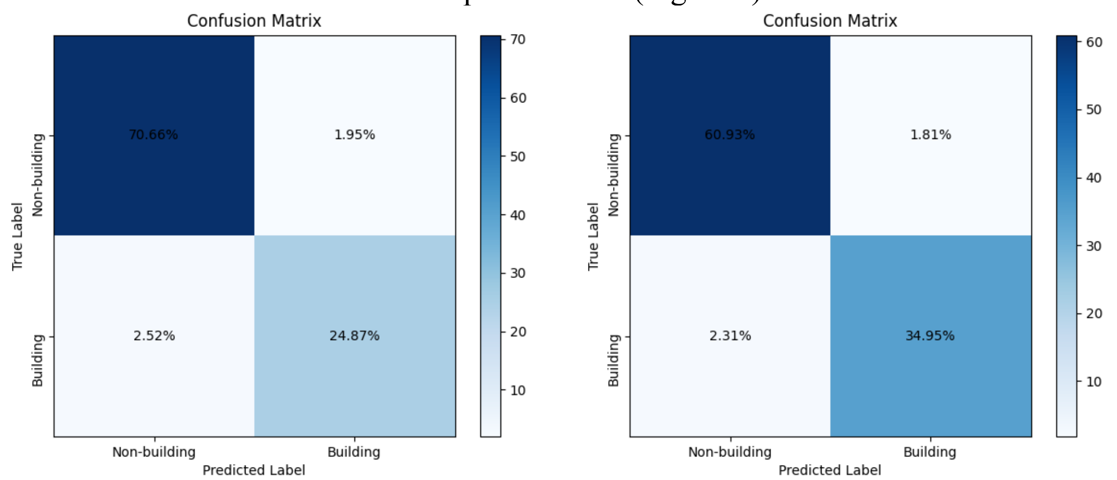
As for the segmentation results, the increase in the AP was acquired from 53.810 to 55.982 at IOU=0.5:0.95, while at 0.75, the increase was from 79.782 to 81. 655. Thus, for IOU= 0.75,

AP increased from 56.732 to 60.258. As for the AR at 0.5:0.95, the increase rate was 3%. This indicates that the model does well in finding the correct buildings.
For a comparison of the results for both study cases, refer to Table 4 for further details. The approach proved highly effective on the New Zealand data, as evidenced by the metric values, which showed a significant increase. This suggests that the model performs better at detecting buildings in Christchurch than in Baghdad. The model's ability to detect the largest number of correct objects was good and high in bounding boxes, and the segmentation results were perceptually accurate at 7.2% and 7.3%.

Table 4: Building Extraction Results Using Mask R-CNN on Both Datasets

	Metrics	ARIS Building Dataset	Baghdad Dataset
Bounding box	AP 0.5:0.95	73.882	57.724
	AP 0.5	92.576	83.585
	AP 0.75	81.490	59.933
	AR 0.5:0.95	79.2	72.1
Segmentation	AP 0.5:0.95	71.387	55.982
	AP 0.5	91.960	81.655
	AP 0.75	79.554	60.258
	AR 0.5:0.95	76.6	69.3

Despite the ARIS dataset yielding higher precision in most evaluation metrics compared to the Baghdad dataset, this difference was interpreted quantitatively without applying statistical testing, such as a T-test. In addition to evaluating the proposed model using Coco metrics, the confusion matrix was used to assess performance (Figure 6).

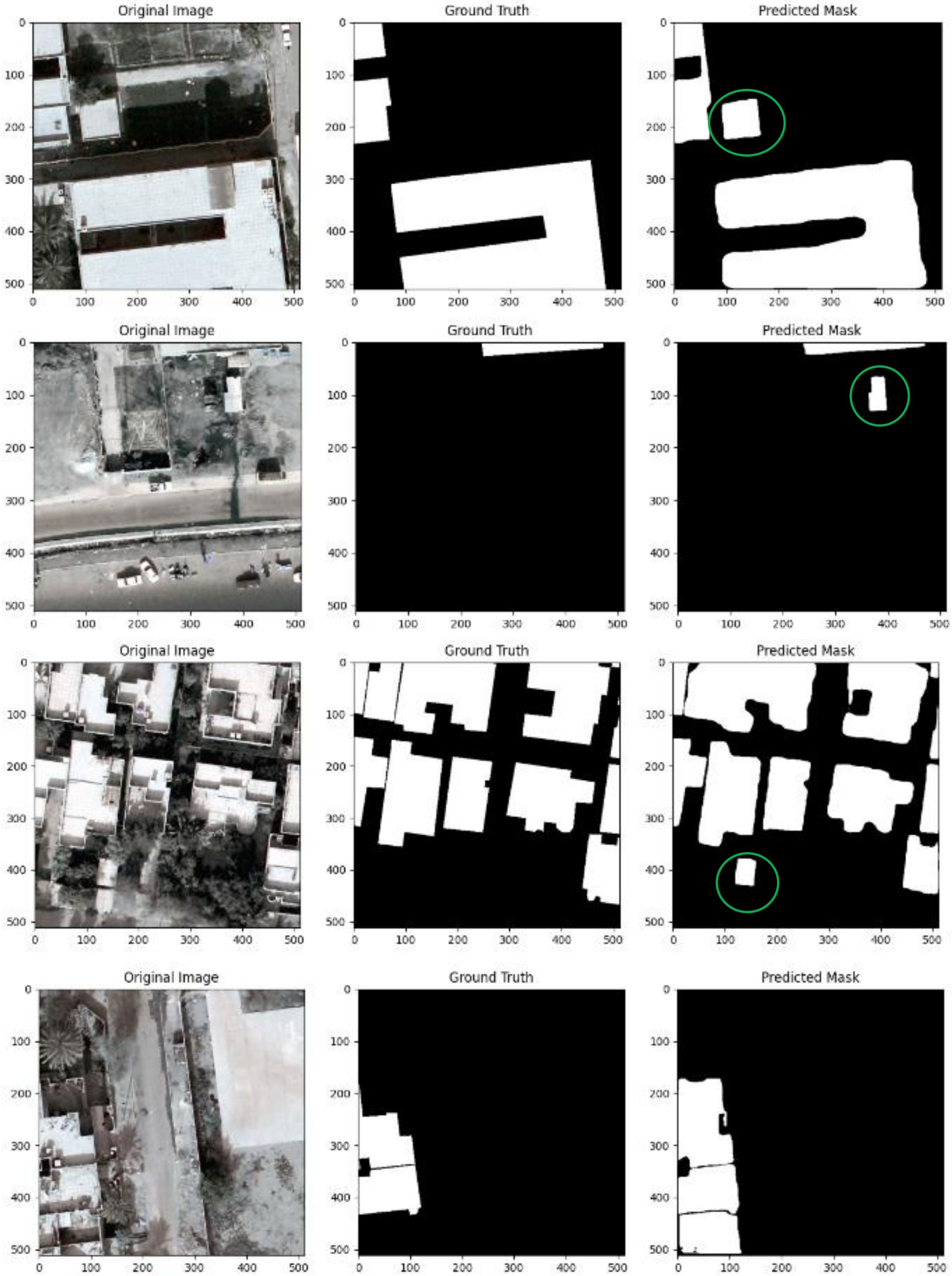


(a) **(b)**

Figure 6: Confusion matrix results. a.Confusion matrix of Baghdad dataset. b. Confusion matrix of the AIRS dataset.

In Figure 6(a), the model predictions for previously unseen test data from Baghdad are presented, with 70.66% correct predictions for background as background and 27.87% correct predictions for buildings as buildings. This yields approximately 95.5% correct predictions, while the others are incorrect predictions. Similarly, in Figure 6(b), the model predictions from the New Zealand test data yielded 60.93% correct predictions for background and 34.95% correct predictions for buildings, totaling more than 95.8% accuracy. This is clear evidence of the success of the Mask R-CNN methodology and its efficiency.

Qualitatively, the most distinctive feature of the model was its ability to detect buildings present in the aerial image but not in the ground truth data, as illustrated in Figure 7 by the green circle. This ability was only observed in the AL-Karrada and AL-Mansour regions, whereas in AL-Sadr City, most of the model's predictive errors occurred in this region. As for the buildings partially covered by trees in the Mansour and Karrada areas, the model accurately detected the buildings, and their shapes were somewhat regular, as illustrated in Figure 8 by the green circles.



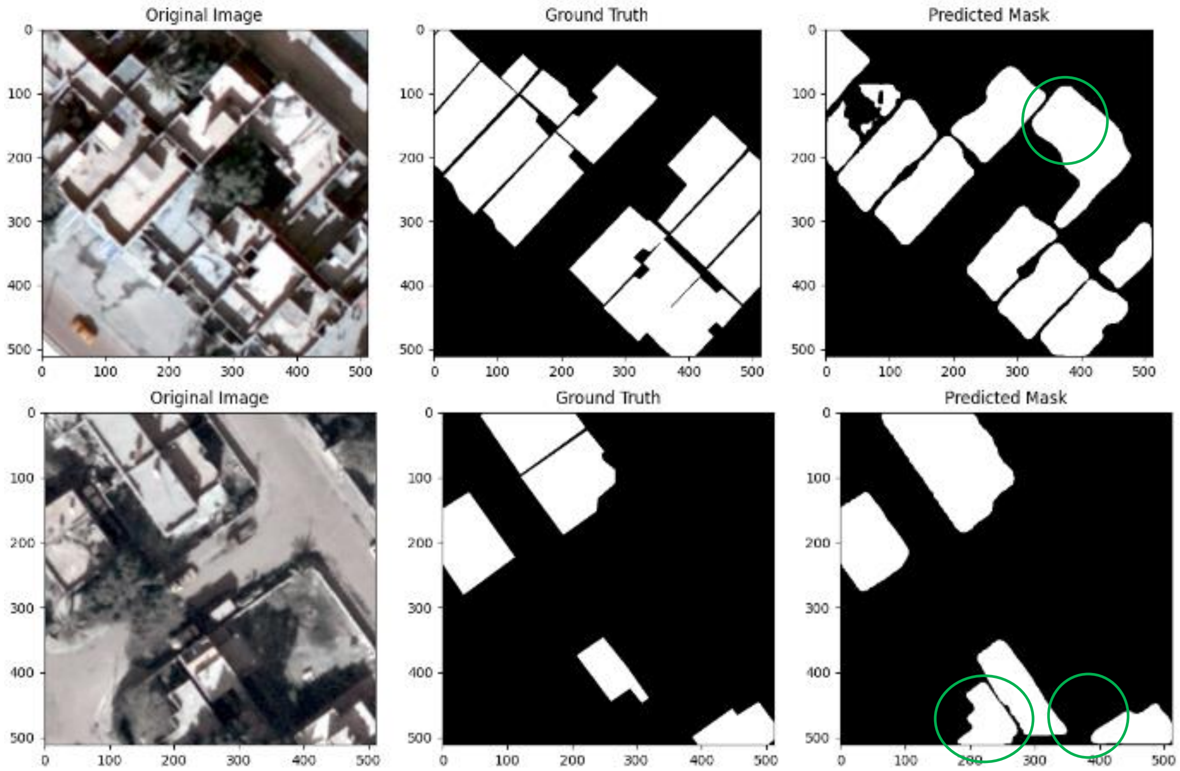


Figure :7 Samples showing the Model's efficiency in extracting buildings using the Baghdad dataset.

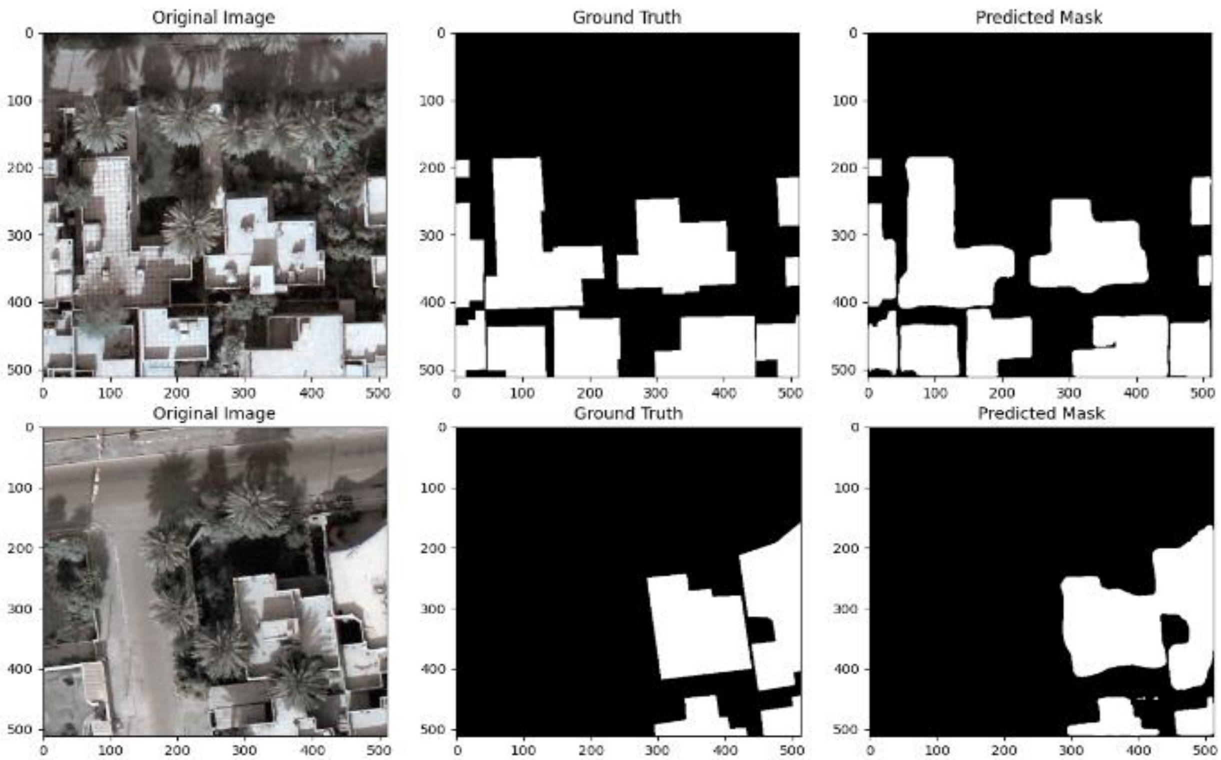


Figure 8: Tree occlusion samples in AL-Mansour and AL-Karrada districts.

In some test images, the model detected parts of the buildings under the shadows of their highest points, resulting from the sun's incidence angle (Figure 9, green circles). However, it was unable to detect buildings correctly in other areas (Figure 9, red circle). One of the weaknesses that should be considered is the model's inability, in addition to the above, to accurately detect many buildings in the AL-Sadr district region (Figure 10).

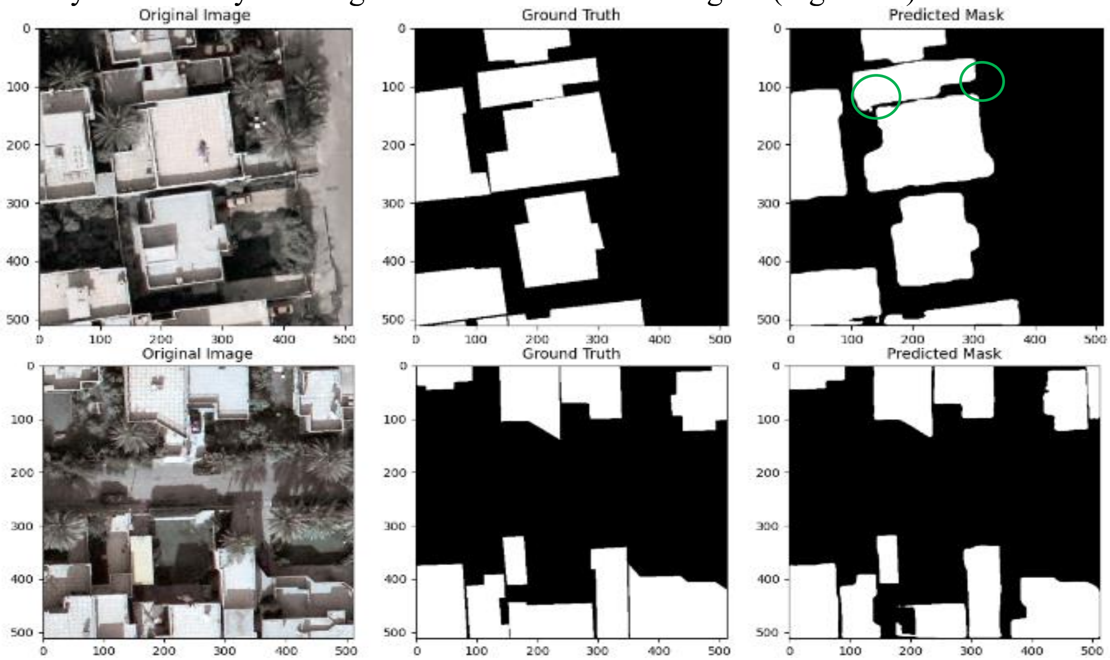


Figure 9: Samples of shadow effects in AL-Mansour and AL-Karrada districts.

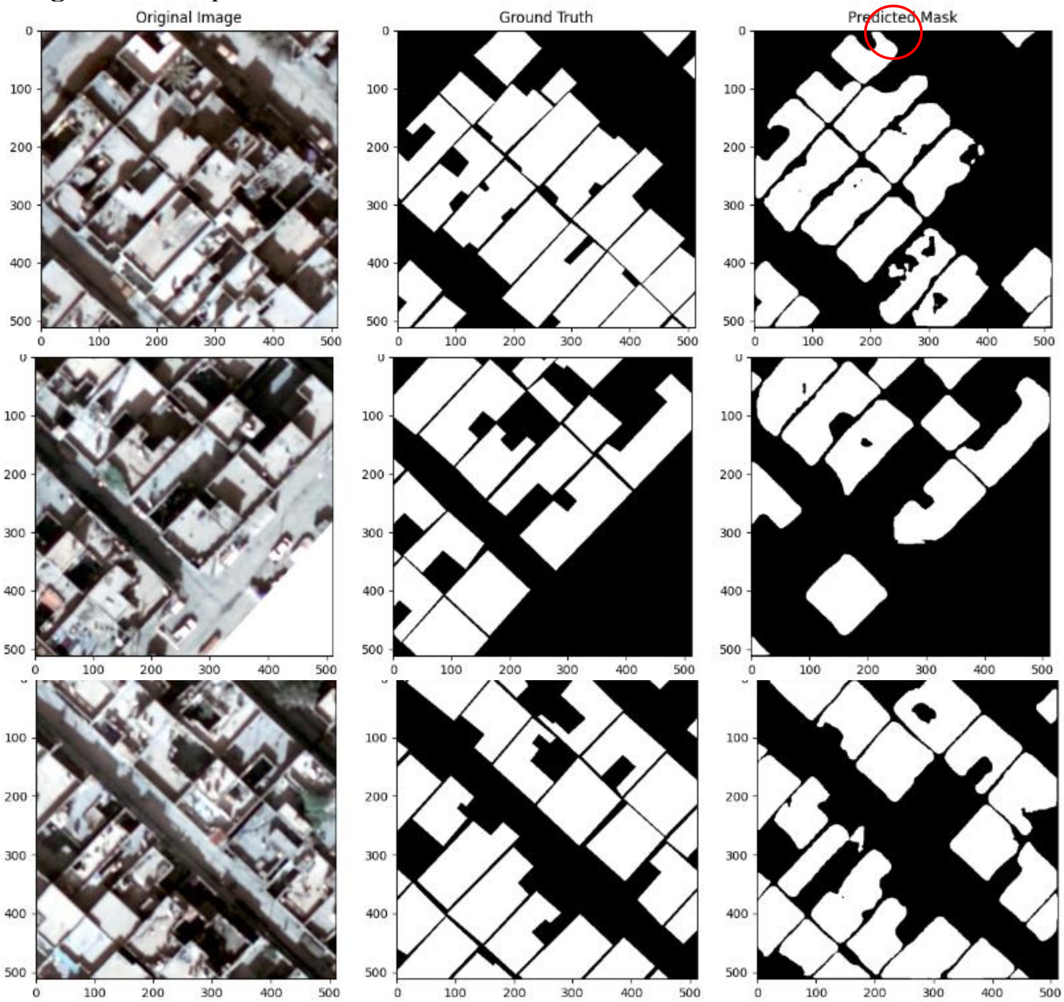


Figure 10: Representative samples of the model's inability to predict buildings in the AL-Sadr district.

The model classified the buildings as background instead of classifying them as buildings, which may be due to multiple reasons. The first is due to the strong adhesion between the buildings and the shadow resulting from this adhesion, which may give the model the illusion that it is a background. The second may be the quality of the aerial images taken, despite the enhancement that was used. The third is the time of image capture, which resulted in the intense brightness of the buildings' surfaces due to the sun's rays, making them nearly as bright as the nearby roads. Despite the advantages of automatic building extraction using the proposed DL model, several challenges remain, including errors in detecting attached or irregularly shaped buildings. In such cases, manual solutions can serve as a complementary approach. Manual or post-processing corrections yield high-accuracy outcomes, although they are a time-consuming process [20], particularly in complex urban scenes or with low-quality imagery [33]. These methods require extra time and effort [34].

5. Conclusion

This paper presented a practical developed segmentation approach to detect buildings in urban residential regions in HR aerial images of Baghdad city using the Mask R-CNN DL model. Despite the enhanced brightness and data augmentation technique, the model achieved 83.585 in AP@0.5 for the bounding boxes and 81.655 in AP@0.5 for the segmentation results. However, the model on the New Zealand dataset achieved an AR superiority ranging from 7.1% and 7.3% for detection and segmentation, respectively. The AP@0.5 achieved an excellence rate of 9% and 10% for detection and segmentation, respectively. This superiority resulted from the high resolution of the aerial images of the New Zealand dataset. Furthermore, the regular shapes of the buildings in this dataset, their spacing, and the materials used in the building roofs all contributed to improved outcomes. This has contributed to distinguishing buildings from their surroundings in the aerial image, compared to the results obtained from extracting buildings in the AL-Sadr region of Baghdad.

In the Baghdad dataset, the model successfully extracted buildings from images that were not initially present in the real data, indicating that it had learned effectively during the training phase. One key point to note is that the model accurately identified the parts of buildings covered by trees in this dataset, following the successful selection of processing settings. As for the parts of buildings covered by the shadows of their higher parts, it could extract these in many cases in some areas of Baghdad. However, the model was unable to recognize these parts of the buildings accurately. The reflectivity of the sunlight on the building surfaces and the buildings' adhesion were a real challenge for the model because they were difficult to distinguish, even for the human eye.

These results state not only the feasibility of using Mask R-CNN for urban building extraction in complex environments like Baghdad but also highlight the current limitations of deep learning models in handling occlusions, low contrast, and heterogeneous urban textures. Future work may explore the integration of attention mechanisms or hybrid models to improve robustness. Therefore, developing an approach that can distinguish buildings under this type of challenge may be achieved by utilizing other aerial image enhancement techniques that yield better results. Moreover, post-processing operations such as angle detection and regularizing the building boundaries can be performed to enhance the segmentation results.

References

[1] U. Bhattacharya and S. K. Parui, "An improved backpropagation neural network for detection of road-like features in satellite imagery," *Int. J. Remote Sens.*, vol. 18, no. 16, pp. 3379–3394, 1997, doi: 10.1080/014311697216937.

[2] G. Yan, H. Jing, H. Li, H. Guo, and S. He, "Enhancing Building Segmentation in Remote Sensing

Images: Advanced Multi-Scale Boundary Refinement with MBR-HRNet," *Remote Sens.*, vol. 15, no. 15, 2023, doi: 10.3390/rs15153766.

[3] S. Ji, S. Wei, and M. Lu, "Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, 2019, doi: 10.1109/TGRS.2018.2858817.

[4] H. Wang and F. Miao, "Building extraction from remote sensing images using deep residual U-Net," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 71–85, 2022, doi: 10.1080/22797254.2021.2018944.

[5] A. I. Hussein and M. J. Kelain, "Efficient Method to Segment Objects from Images Based on Enhanced Connected Component Labeling algorithm," *Iraqi J. Sci.*, vol. 59, no. 1B, pp. 420–425, Mar. 2018, doi: 10.24996/ijs.2018.59.1B.22.

[6] S. Saito and Y. Aoki, "Building and road detection from large aerial imagery," *Image Process. Mach. Vis. Appl. VIII*, vol. 9405, no. March, p. 94050K, 2015, doi: 10.1117/12.2083273.

[7] M. Yousif and B. Al-Khateeb, "Quantum Deep Learning: A Review," *Iraqi J. Sci.*, vol. 65, no. 8, pp. 4588–4605, Aug. 2024, doi: 10.24996/IJS.2024.65.8.37.

[8] I. Kadhim and F. M. Abed, "A Critical Review of Remote Sensing Approaches and Deep Learning Techniques in Archaeology," *Sensors*, vol. 23, no. 6, pp. 1–17, Mar. 2023, doi: 10.3390/s23062918.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[10] B. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2012.

[11] R. Girshick, "Fast R-CNN," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 1440–1448, 2015, doi: 10.1109/ICCV.2015.169.

[12] B. A. kadheem Hammood and A. T. Sadiq, "Deep Learning Approaches for IoT Intrusion Detection Systems," *Iraqi J. Sci.*, vol. 65, no. 11, pp. 6631–6646, Nov. 2024, doi: 10.24996/IJS.2024.65.11.36.

[13] S. Chen, Y. Ogawa, C. Zhao, and Y. Sekimoto, "Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, no. April 2022, pp. 129–152, 2023, doi: 10.1016/j.isprsjprs.2022.11.006.

[14] F. Chen and B. Yu, "Earthquake-Induced Building Damage Mapping Based on Multi-Task Deep Learning Framework," *IEEE Access*, vol. 7, pp. 181396–181404, 2019, doi: 10.1109/ACCESS.2019.2958983.

[15] P. Zhang, A. Nascetti, Y. Ban, and M. Gong, "An implicit radar convolutional burn index for burnt area mapping with Sentinel-1 C-band SAR data," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 50–62, Dec. 2019, doi: 10.1016/J.ISPRSJPRS.2019.09.013.

[16] T. Hou and J. Li, "Application of mask R-CNN for building detection in UAV remote sensing images," *Heliyon*, vol. 10, no. 19, p. e38141, 2024, doi: 10.1016/j.heliyon.2024.e38141.

[17] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 2980–2988, Dec. 2017, doi: 10.1109/ICCV.2017.322.

[18] Y. Qiu *et al.*, "AFL-Net: Attentional Feature Learning Network for Building Extraction from Remote Sensing Images," *Remote Sens.*, vol. 15, no. 1, 2023, doi: 10.3390/rs15010095.

[19] S. Wang, X. Hou, and X. Zhao, "Automatic Building Extraction from High-Resolution Aerial Imagery via Fully Convolutional Encoder-Decoder Network with Non-Local Block," *IEEE Access*, vol. 8, pp. 7313–7322, 2020, doi: 10.1109/ACCESS.2020.2964043.

[20] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, no. April, pp. 139–149, 2017, doi: 10.1016/j.isprsjprs.2017.05.002.

[21] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters," *Remote Sens.*, vol. 10, no. 1, p. 144, 2018, doi: 10.3390/rs10010144.

[22] L. Xia, X. Zhang, J. Zhang, H. Yang, and T. Chen, "Building extraction from very-high-resolution remote sensing images using semi-supervised semantic edge detection," *Remote Sens.*, vol. 13, no. 11, 2021, doi: 10.3390/rs13112187.

[23] Q. Tian, Y. Zhao, Y. Li, J. Chen, X. Chen, and K. Qin, "Multiscale Building Extraction with

Refined Attention Pyramid Networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3075436.

[24] Y. Li, L. Xu, J. Rao, L. Guo, Z. Yan, and S. Jin, "A Y-Net deep learning method for road segmentation using high-resolution visible remote sensing images," *Remote Sens. Lett.*, vol. 10, no. 4, pp. 381–390, 2019, doi: 10.1080/2150704X.2018.1557791.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," vol. 9351, no. Cvd, pp. 12–20, 2015, doi: 10.1007/978-3-319-24574-4.

[26] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, 2018, doi: 10.1109/LGRS.2018.2802944.

[27] A. Jebur, F. M. Abed, and M. U. Mohammed, "3D City Modelling by Photogrammetric Techniques," *Res. Gate*, no. July 2019, 2017, doi: 10.13140/RG.2.2.11494.06722.

[28] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, no. July, pp. 42–55, 2019, doi: 10.1016/j.isprsjprs.2018.11.011.

[29] LINZ Data Service." Accessed: Jan. 24, 2025. [Online]. Available: https://data.linz.govt.nz/

[30] "aerial-imagery-for-roof-segmentation." Accessed: Dec. 29, 2024. [Online]. Available: https://www.kaggle.com/datasets/atilol/aerialimageryforroofsegmentation

[31] Anaconda Navigator | Anaconda.org." Accessed: Jan. 24, 2025. [Online]. Available: https://anaconda.org/anaconda/anaconda-navigator

[32] GitHub - facebookresearch/detectron2: Detectron2 is a platform for object detection, segmentation and other visual recognition tasks." Accessed: Dec. 29, 2024. [Online]. Available: https://github.com/facebookresearch/detectron2

[33] X. Huang, W. Yuan, J. Li, and L. Zhang, "A New Building Extraction Postprocessing Framework for High-Spatial-Resolution Remote-Sensing Imagery," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. 2, pp. 654–668, Feb. 2017, doi: 10.1109/JSTARS.2016.2587324.

[34] E. Li, S. Xu, W. Meng, and X. Zhang, "Building Extraction from Remotely Sensed Images by Integrating Saliency Cue," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. 3, pp. 906–919, Mar. 2017, doi: 10.1109/JSTARS.2016.2603184.