Current Applied Science and Technology e-ISSN: 2586-9396

Curr. Appl. Sci. Technol. ….., Vol... (No...), e0265709

Research article

Integrating RGB and DSM Data for Enhanced Building Segmentation in UAV Images

Kanokwan Khiewwan[1], Duangduen Asavasuthirakul[2] and Sutasinee Chimlek[1]*

[1]Department of Computer Science and Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand
[2]Drone AI Solutions Co., Ltd., Kamphaeng Phet, Thailand

Received: 17 December 2024, Revised: 15 May 2025, Accepted: 7 July 2025, Published: 1 December 2025

Abstract

Accurate building segmentation in unmanned aerial vehicle (UAV) orthophotos remains a significant challenge due to the visual similarity between buildings and non-target elements such as trees, roads, and background clutter. This study proposes an enhanced segmentation method—referred to as RGB-DSM-IMP (M3)—which integrates RGB imagery, Digital Surface Model (DSM) data, and a novel background removal preprocessing step. The Mask Region-Based Convolutional Neural Network (Mask R-CNN) framework was employed to evaluate three segmentation strategies: a baseline model using only RGB imagery, a second model combining RGB imagery with DSM data, and the proposed model that incorporates both data types along with preprocessing. All models were trained and tested on drone-acquired images representing a variety of building types and environmental conditions. Performance was evaluated using precision, recall, F1-score, average precision (AP), mean intersection over union (mIoU), and mean average precision (mAP). The enhanced model achieved the highest results across all metrics, with an average F1-score of 0.74, mIoU of 0.74, and mAP of 0.63. These findings highlight the benefit of integrating elevation data to enhance spatial differentiation and demonstrate the effectiveness of background removal in reducing misclassifications caused by visually similar objects. In addition, the method maintained a practical inference time per image, supporting its real-world applicability. Overall, the study demonstrates that combining height-based information with strategic preprocessing significantly improves the accuracy and robustness of building segmentation in complex aerial imagery.

Keywords: background removal; building segmentation; Digital Surface Model (DSM); Mask R-CNN; UAV Imagery

1. Introduction

The use of UAVs, or drones, has become essential for capturing high-resolution aerial images across various applications, including land use analysis (Yang et al., 2018; Al-

*Corresponding author: E-mail: sutasineec@nu.ac.th
https://doi.org/10.55003/cast.2025.265709
Copyright © 2024 by King Mongkut's Institute of Technology Ladkrabang, Thailand. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Najjar et al., 2019; Rizk et al., 2022), disaster monitoring (Chen, et al., 2021; Chueprasert et al., 2025) and infrastructure planning (Yi et al., 2019). When processed into orthophotos, UAV-captured images provide precise georeferenced data that is critical for accurate building segmentation in both urban (Yi et al., 2019; Kamarulzaman et al., 2023) and rural landscapes (Amo-Boateng, et al., 2022; Wang et al., 2022). However, effective segmentation remains challenging due to the visual complexity of scenes where elements such as trees, roads, and open spaces can closely resemble building structures (Boonpook et al., 2020; Li, et al., 2021). These similar background elements interfere with segmentation algorithms, making accurate building isolation more difficult and highlighting the necessity for advanced methodologies.

To address these challenges, recent research has increasingly focused on integrating Digital Surface Model (DSM) data with RGB imagery, utilizing DSM's elevation data to provide spatial context. A Digital Surface Model (DSM) is a type of 3D elevation data that represents the Earth's surface including all objects on it, such as buildings, vegetation, and infrastructure. Unlike a Digital Terrain Model (DTM), which depicts only the bare ground surface, a DSM captures the top surfaces of features, making it especially valuable for urban analysis (Jiang et al., 2008; Chea et al., 2019). In this study, DSM data was generated through photogrammetry techniques applied to UAV imagery, and it provides crucial height information for distinguishing buildings from visually similar objects like trees or flat surfaces such as roads (Li et al., 2021). By adding elevation context to the RGB images, the DSM significantly enhances the segmentation model's ability to identify and isolate building structures with greater precision (Boonpook et al., 2020). This elevation data is invaluable for distinguishing objects with similar spectral characteristics, such as separating tall trees from buildings or roads in urban and natural landscapes (Jiang et al., 2008). By incorporating height-based features, DSM enhances segmentation accuracy in complex environments, enabling better discrimination between target and non-target elements.

Mask R-CNN is a two-stage framework, for instance segmentation that extends Faster R-CNN by adding a segmentation branch. The framework consists of five key components. The first component is a backbone convolutional network (e.g., ResNet50) that extracts deep features from the input image. The second component, Region Proposal Network (RPN), is used to propose candidate regions where objects might be located. The third component is the RoIAlign layer, which accurately extracts fixed-size feature maps from the proposed regions. The next component is the classification and regression head, which assigns class labels to the objects and refines the bounding boxes. Finally, the segmentation head, implemented as a fully convolutional network (FCN), generates a binary mask for each detected object, enabling pixel-level segmentation.

During training, Mask R-CNN jointly optimizes three losses including classification, bounding box regression, and mask prediction. This multi-task structure makes the network particularly robust in environments with overlapping or visually similar objects, such as buildings and vegetation. The RoIAlign operation also improves mask precision by avoiding misalignment caused by quantization, which is crucial in UAV images where accurate edge detection is required. For these reasons, we adopted Mask R-CNN as the segmentation model in this study. Its ability to simultaneously detect, classify, and segment buildings at the pixel level makes it well-suited for the complexities of aerial imagery, especially when enhanced with elevation data from DSM and preprocessing techniques like background removal. The operational workflow of Mask R-CNN is illustrated in Figure 1 (Snyder et al., 2021).

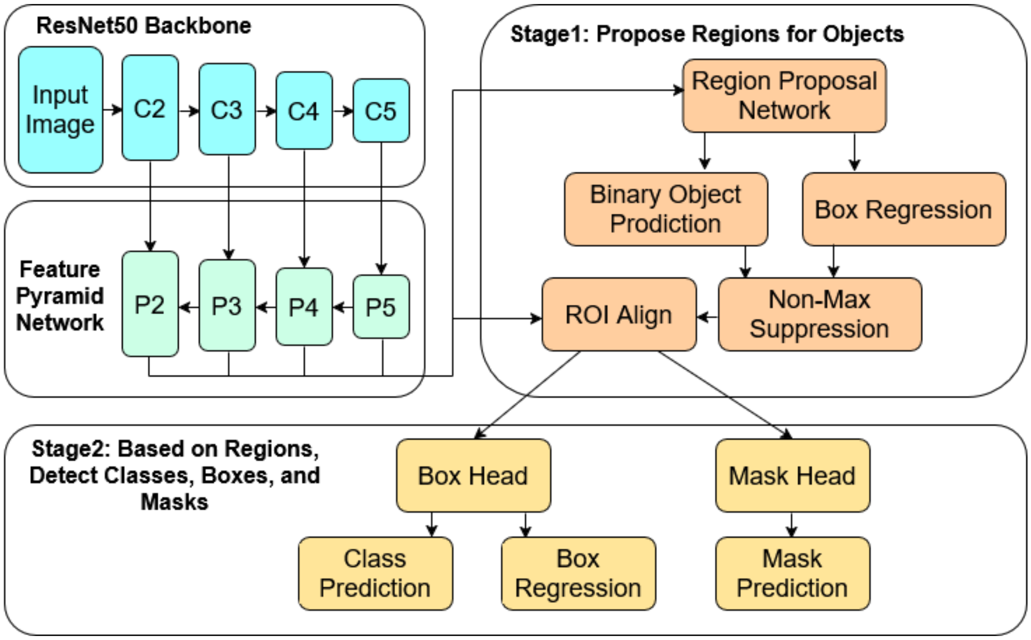


Figure 1. The working architecture of Mask R-CNN

Mask R-CNN can employ height and spatial information in combination with DSM data to enhance boundary precision and reduce misclassification. For instance, integrating DSM data with RGB imagery has demonstrated significant improvements in segmentation accuracy, such as using the Visible Band Difference Vegetation Index (VDVI) to differentiate buildings from trees (Boonpook et al., 2020). Despite these advancements, DSM has notable limitations that must be addressed. For example, Wang et al. (2018) highlighted that DSM struggles to accurately model building boundaries in the presence of non-target objects, while KC et al. (2021) reported suboptimal performance when DSM was applied to short vegetation, such as trees with heights ranging from 0.2 to 0.3 m.

To address these limitations, background removal, a technique commonly used in image analysis, offers a promising preprocessing solution that has not been widely applied to UAV-based building segmentation. Chea et al. (2019) demonstrated that background removal improved classification in plant imagery by eliminating non-target elements, allowing models to focus on relevant features. Similarly, Ran et al (2019) introduced a method for segmenting rock images by removing irrelevant elements, such as grass and soil, to improve classification accuracy. Although effective in other domains, the application of background removal in conjunction with DSM data for UAV-based building segmentation remains largely unexplored. Adding this step could significantly improve building segmentation accuracy by focusing the model exclusively on buildings.

This research addresses existing gaps by introducing an image processing method that combines DSM and RGB data with background removal to enhance building segmentation accuracy. By eliminating irrelevant background elements, such as trees and roads, this approach streamlines segmentation, allowing Mask R-CNN to focus exclusively on buildings. The DSM-based preprocessing step adds a spatial layer that enhances Mask R-CNN’s ability to distinguish building features from similarly colored objects, reducing misclassification risks due to spectral overlap.

These enhancements have practical implications across various geospatial domains. The proposed method improves building segmentation accuracy in UAV-based mapping, supporting applications such as urban planning, environmental monitoring, and disaster response. By increasing segmentation robustness and precision, especially in complex and cluttered scenes, this approach advances current capabilities in remote sensing and geoinformatics.

The main contribution of this study is the development of an enhanced building segmentation method called RGB-DSM-IMP (M3), which combines RGB imagery, Digital Surface Model (DSM) data, and a tailored background removal preprocessing step. This integrated approach significantly improves segmentation accuracy and robustness in UAV imagery, especially in complex environments with vegetation and overlapping structures. The results demonstrate that M3 consistently outperforms traditional RGB-based and RGB-DSM models, confirming the value of multimodal data fusion and preprocessing in real-world aerial mapping tasks.

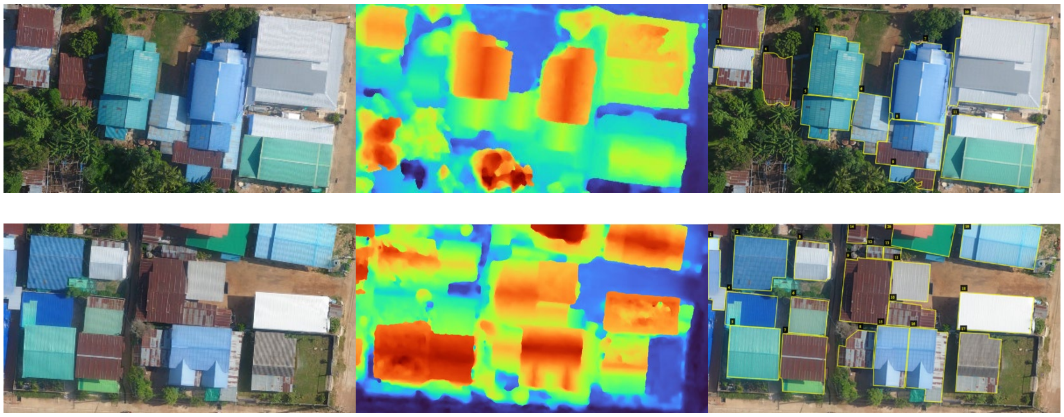
2. Materials and Methods

2.1 Dataset

The dataset used in this study was derived from UAV images captured at an altitude of 150 m in a village near Naresuan University, Phitsanulok, Thailand. A DJI Phantom 4 Pro drone with an FC6310R camera, positioned perpendicularly to the ground, was used for image acquisition. The raw images (5472 × 3648 pixels) were processed through photogrammetry to generate orthophotos and the corresponding Digital Surface Model (DSM) data. These orthophotos provided top-down views of rooftops and included images of various building types, such as detached houses, row houses, schools, and warehouses. To enhance GPU efficiency and reduce training time, the orthophotos and DSM data were divided into smaller patches measuring 1000 × 667 pixels. Although the original images contained multiple object types, this study focused exclusively on buildings. A total of 510 labeled building objects were extracted from 17 image patches. These were randomly split into a training set of 12 patches (approximately 70%) and a test set of 5 patches (approximately 30%). Each image contained one or more buildings, and annotations followed the COCO format. Experimental results showed that the number of training images had limited impact on model performance. In contrast, the number of annotated building objects was a more decisive factor. This finding emphasizes the significance of object-level diversity and quantity over the mere number of image samples in improving segmentation accuracy. Figure 2 presents examples from the dataset used in this study. Column 1 shows RGB orthophotos. Column 2 presents the corresponding Digital Surface Model (DSM) data generated through photogrammetric processing. Column 3 displays manually annotated ground truth segmentation masks in COCO format.

2.2 Experiment setup

The experiments and computational analyses for this study were conducted on a computer equipped with an Intel Core i7 processor, 16 GB of RAM, and an NVIDIA GeForce GTX 1060 graphics card. This configuration provided sufficient computational power to handle large datasets and enabled efficient training and testing of the deep learning models. The NVIDIA GeForce GTX 1060 GPU played a crucial role in accelerating neural network



RGB images DSM images Ground truth segmentation mask

Figure 2. Sample RGB, DSM, and ground truth segmentation images used for training and testing

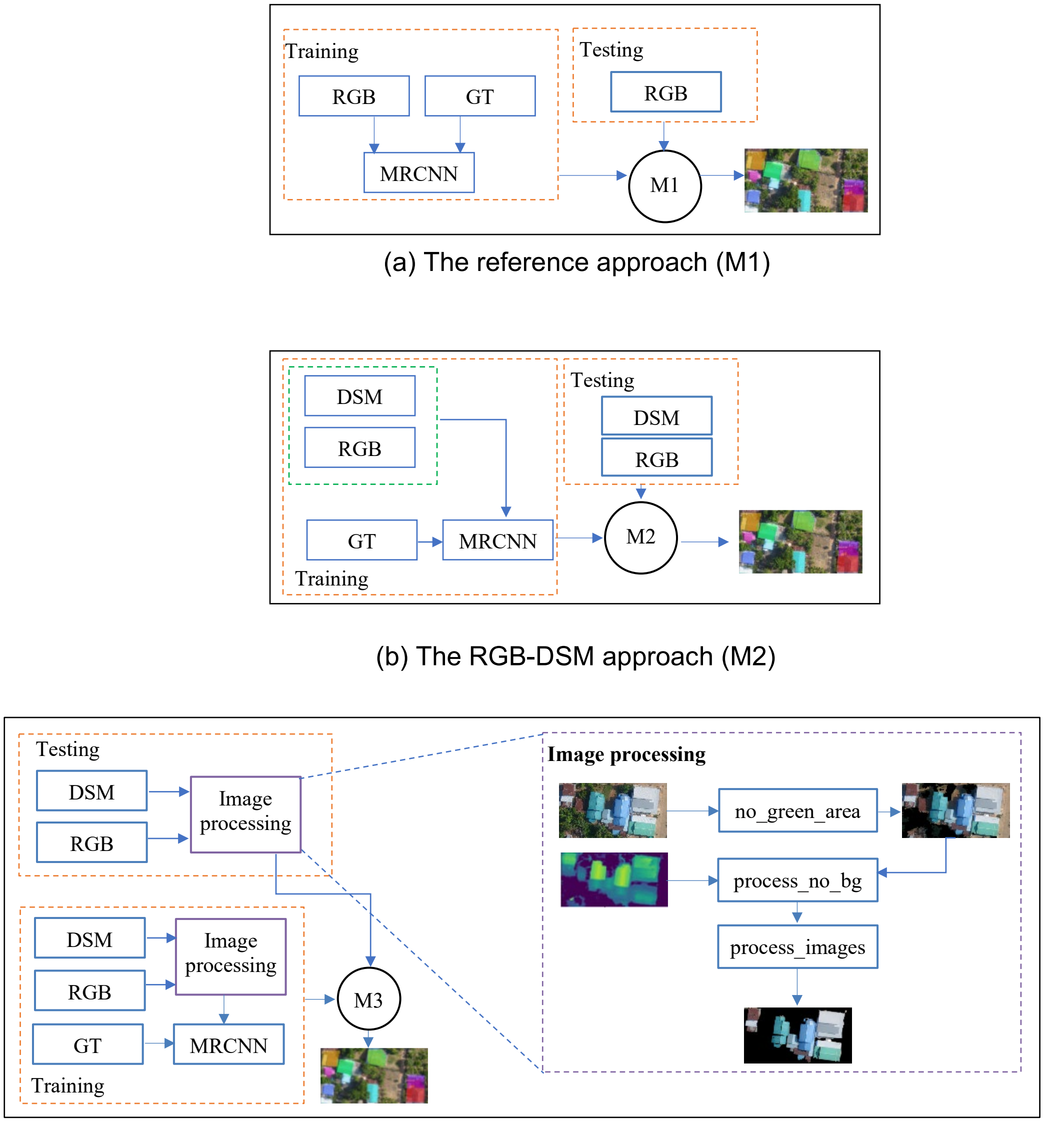
operations and image processing tasks, significantly reducing computation time and enhancing the overall experimentation efficiency. The building segmentation program was implemented in Python using TensorFlow 1.5.0 and Keras 2.1.2 libraries. Three segmentation approaches were developed and implemented in this study, as illustrated in Figure 3.

Figure 3 illustrates the following approaches in detail:

a) M1 – reference approach:
The input of this approach is RGB images and ground truth annotations. The process involves directly feeding these inputs into the Mask R-CNN (MRCNN) model without any additional data or preprocessing steps. The output is the predicted segmentation masks of building areas.

(b) M2 – RGB-DSM approach:
This approach uses RGB images, DSM data, and ground truth annotations as input. The process includes combining RGB imagery with DSM data to provide the model with spatial and elevation context. These enriched inputs are then used to train the MRCNN model. The resulting output is a set of building segmentation masks with improved accuracy, especially around object boundaries.

(c) M3 – RGB-DSM-IMP approach:
The input of this approach includes RGB images, DSM data, and ground truth. A preprocessing step is applied to remove irrelevant background elements such as trees and roads from the RGB images. The DSM data is also refined through image enhancement techniques. These processed inputs are then fed into the MRCNN model. The output is a more precise segmentation result with enhanced focus on building structures and reduced influence from non-building features.



(a) The reference approach (M1)

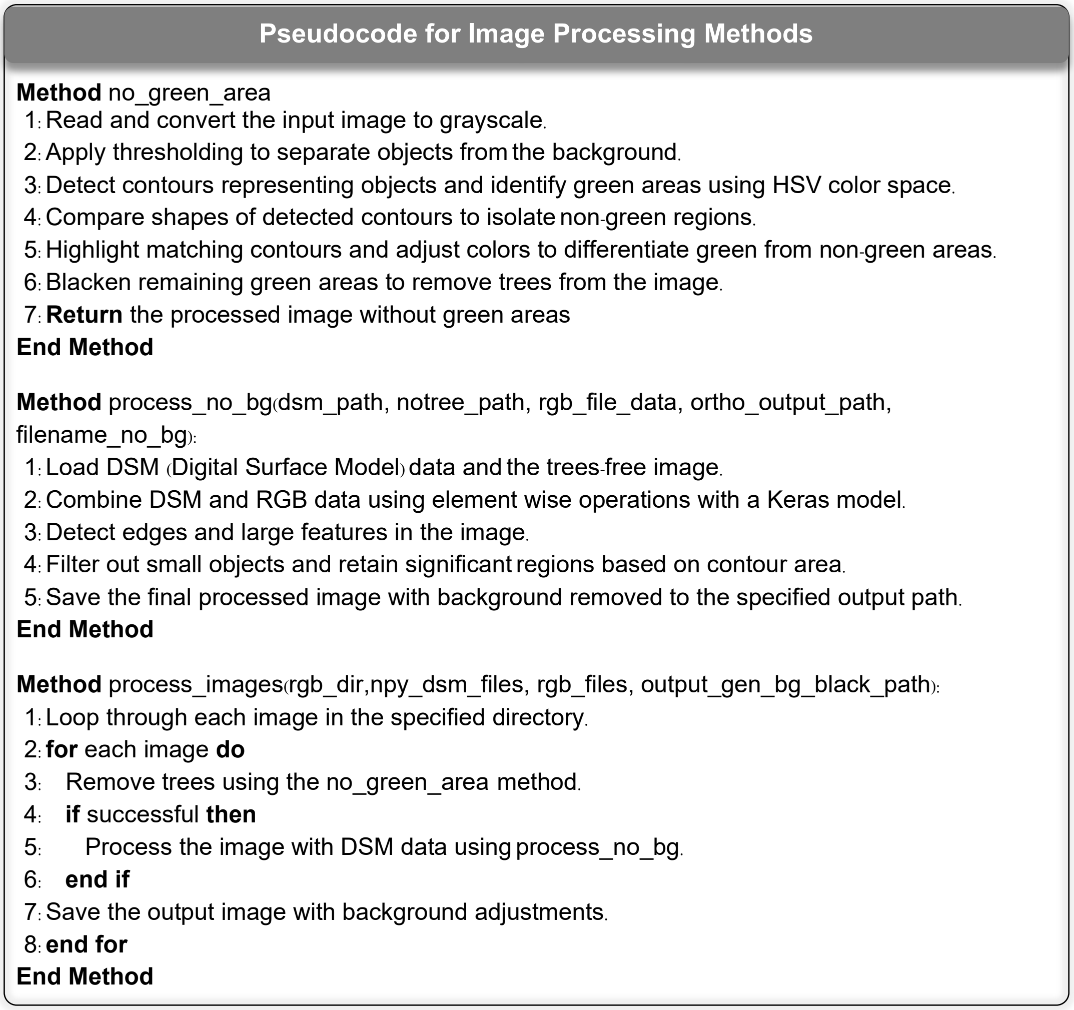
(b) The RGB-DSM approach (M2)

(c) The RGB-DSM-IMP approach (M3)

Figure 3. The three building segmentation approaches used in this study

2.2.1 RGB-DSM approach

The integration of Digital Surface Model (DSM) data with RGB imagery provides a valuable enhancement in training MRCNN for effective image segmentation. DSM data supplies elevation information that distinguishes objects based on their height, allowing MRCNN to differentiate between objects that may share similar RGB characteristics but vary in elevation, such as trees and buildings. Integrating DSM data as an additional input channel enhances the network's spatial context, facilitating the distinction of regions that require separate segmentation, despite similar color characteristics. For instance, regions with a green hue in the range of 30–90 in RGB data, which may represent trees or certain structures, can be further classified based on their height profiles in the DSM data. This additional layer of information enhances MRCNN’s ability to produce precise segmentation masks, particularly in complex environments, resulting in a model that is more robust to variations in color and texture. Consequently, the combined RGB-DSM approach improves model accuracy and reduces misclassification errors by reinforcing object differentiation through spatial and height-based cues. The pseudocode of the image processing methods is illustrated in Figure 4.



Pseudocode for Image Processing Methods

```
Method no_green_area
Read and convert the input image to grayscale.
Apply thresholding to separate objects from the background.
Detect contours representing objects and identify green areas using HSV color space.
Compare shapes of detected contours to isolate non-green regions.
Highlight matching contours and adjust colors to differentiate green from non-green areas.
Blacken remaining green areas to remove trees from the image.
Return the processed image without green areas
End Method

Method process_no_bg(dsm_path, notree_path, rgb_file_data, ortho_output_path,
filename_no_bg):
Load DSM (Digital Surface Model) data and the trees-free image.
Combine DSM and RGB data using element wise operations with a Keras model.
Detect edges and large features in the image.
Filter out small objects and retain significant regions based on contour area.
Save the final processed image with background removed to the specified output path.
End Method

Method process_images(rgb_dir,npy_dsm_files, rgb_files, output_gen_bg_black_path):
Loop through each image in the specified directory.
for each image do
  Remove trees using the no_green_area method.
  if successful then
    Process the image with DSM data using process_no_bg.
  end if
Save the output image with background adjustments.
end for
End Method
```

Figure 4. Pseudocode for image processing methods

2.2.2 RGB-DSM-IMP approach

The RGB-DSM-IMP approach addresses the building segmentation problem by removing non-target elements, such as trees, grass and other vegetation. This method also integrates Digital Surface Model (DSM) data with RGB images to provide additional spatial and elevation context. The objective is to enhance segmentation accuracy by reducing misclassification and guiding the model's attention toward relevant structures. The overall workflow of the RGB-DSM-IMP approach consists of three main steps, which are described in detail below.

Step 1: Green area removal using the no_green_area method

In this step, green areas such as trees and grass—which often share similar color and texture characteristics with rooftops in RGB images—are identified and removed to minimize confusion during segmentation. The process begins with the conversion of the RGB image into the HSV (Hue, Saturation, Value) color space, which facilitates the isolation of green tones commonly associated with vegetation. A thresholding technique is then applied to detect pixels within the green hue range. Once these areas are identified, they are darkened or masked, effectively removing vegetation and retaining only non-green features such as buildings and roads. The result is a cleaner RGB image with vegetation suppressed, making it easier for the segmentation model to learn from more relevant input data. This process is illustrated in Figure 5, which displays examples of the original RGB images, their corresponding DSM data, and the combined RGB+DSM outputs after green area removal.

Figure 5 shows sample outputs from the preprocessing stage in the RGB-DSM-IMP approach. The left column presents the original RGB images containing visible vegetation. The middle column shows the corresponding Digital Surface Model (DSM) data, which captures elevation features. The right column illustrates the RGB+DSM combined input after applying the no_green_area method. In this version, green areas (e.g., trees and grass) have been detected and suppressed to improve focus on building structures during segmentation.

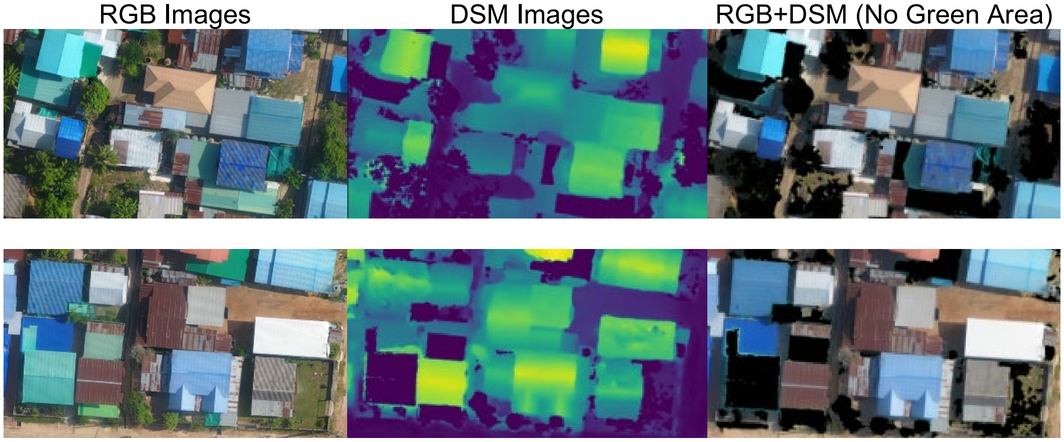


Figure 5. Green area removal and RGB+DSM data integration

Step 2: Fusion of DSM and RGB data using the process_no_bg method
In the second step, after green areas have been removed, the cleaned RGB image is fused with the corresponding Digital Surface Model (DSM) data. This integration enhances the model's ability to distinguish buildings from other objects such as roads and flat surfaces by incorporating both spatial texture and elevation information. The fusion process begins with the alignment of the DSM data to the RGB image using a decomposition-based operation. Edge detection techniques are then applied to emphasize structural features, such as building boundaries. Subsequently, small and irrelevant objects are filtered out based on their size, retaining only significant structures that contribute meaningfully to the segmentation task. The result of this step is a refined input image that combines visual and height cues. The integration of DSM data with RGB imagery provides richer contextual information, improving the accuracy of building segmentation by clearly separating target structures from non-target elements.

Figure 6 illustrates the input and preprocessed images used in the RGB-DSM-IMP approach. The left column shows the original RGB images captured by UAV. The middle column presents the corresponding DSM images, which convey elevation data. The right column displays the fused RGB+DSM images generated using element-wise multiplication after background elements were removed. This combination enhances both spatial and elevation features, supporting more precise building segmentation.

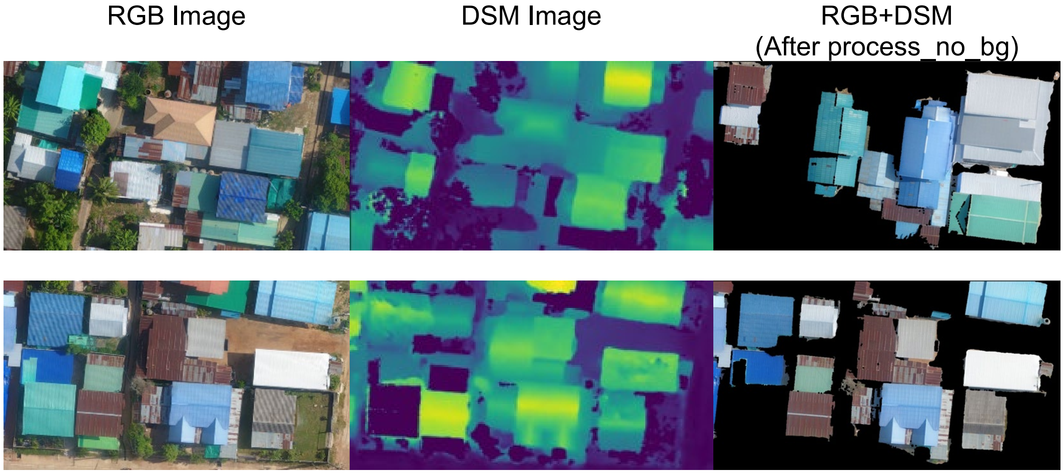


Figure 6. DSM-enhanced RGB input after background removal

Step 3: Automated batch processing using the process_images method
The final step involves automating the preprocessing workflow using the process_images method to efficiently handle large datasets. This method ensures consistency in data preparation and significantly reduces the time required to process multiple images. The automation process begins by iterating through all images stored in the dataset directory. For each image, the no_green_area method is first applied to eliminate vegetation and other non-target green areas. The resulting cleaned image is then passed to the process_no_bg method, which fuses the image with its corresponding DSM data and performs further refinement, including object filtering and edge enhancement.

All processed images are saved to a specified output directory. The outcome of this automated procedure is a complete set of preprocessed images with vegetation removed and DSM information integrated. These images are then ready to be used as input for training or evaluating the building segmentation model based on the Mask R-CNN framework.

2.2.3 Computational time

All experiments were conducted on a computer equipped with an Intel Core i7 processor, 16 GB of RAM, and an NVIDIA GeForce GTX 1060 GPU. Each model was trained for 120 epochs under identical hardware and software conditions. The training time per epoch varied depending on the complexity of the input data and preprocessing involved in each approach.

The M1 model (Reference Approach), which utilized only RGB images, required approximately 199 s per epoch, resulting in a total training time of approximately 6.6 h. The M2 model (RGB-DSM Approach), which combined RGB images with DSM data, took 247 s per epoch, yielding a total training time of about 8.2 h. The M3 model (RGB-DSM-IMP Approach), which incorporated both DSM data and background removal preprocessing, required the most time, averaging 295 s per epoch, or approximately 9.8 h in total.

In terms of inference time, the average time to segment a single test image was measured under the same hardware conditions. The M1 model completed segmentation in approximately 60 s per image, while the M2 model required about 80 s. The M3 model, with additional preprocessing, took around 100 s per image. These times reflect the full segmentation pipeline, including image loading, preprocessing, and prediction.

2.3 Performance measurement

This study evaluates model performance using metrics such as Precision, Recall, F1-Score, Intersection over Union (IoU) and mean average precision (mAP). IoU measures the overlap between predicted and ground truth segmentation masks. A match is considered valid if the Intersection over Union (IoU) between the predicted mask and the ground truth mask is greater than or equal to 0.5 (IoU ≥ 0.5).

To calculate these metrics, the following terms are defined:

1) True Positives (TP): Predicted masks that correctly match ground truth masks (IoU ≥ 0.5).

2) False Positives (FP): Predicted masks with no matching ground truth (IoU < 0.5 or unmatched)

3) False Negatives (FN): Ground truth masks with no matching predictions (IoU < 0.5 or unmatched)

The equations used are:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{1}$$

The model’s performance is evaluated using the following calculated metrics.

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

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

The equation for the F1-score, which combines Precision and Recall for a balanced measure of accuracy, is computed as:

$$\text{F1-Score} = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \tag{4}$$

In addition to standard evaluation metrics, this study reports the average precision (AP) at an Intersection over Union (IoU) threshold of 0.5, which indicates the model's ability to detect and segment building objects with at least 50% overlap with the ground truth. Given that the task involves a single class (building), the mean average precision (mAP) is equivalent to the average AP across all test images, and is calculated as:

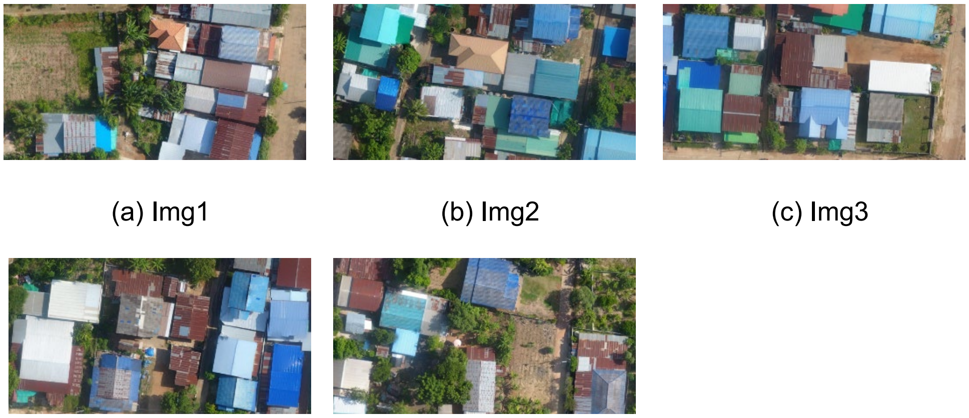
$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{5}$$

where N denotes the total number of images, and AP_i refers to the AP computed for the *i*-th image.

3. Results and Discussion

3.1 Evaluation of test images

The performance of the three segmentation approaches—M1, M2, and M3—was evaluated using five test images (Img1 to Img5), which represent varying degrees of complexity in building segmentation. These images, shown in Figure 7, were selected to include a mix of simple and dense urban environments with overlapping structures and background noise. Figure 7 shows that the test images (Img1–Img5) illustrated varying complexity in building segmentation tasks. These examples were used to evaluate the performance of the three proposed models.



(a) Img1

(b) Img2

(c) Img3

(d) Img4

(e) Img5

Figure 7. The test images illustrating varying complexity in building segmentation tasks

3.2 Quantitative comparison of model performance

Evaluation metrics used in this study included precision, recall, F1-score, and Average Precision (AP), as shown in Table 1, along with Mean Intersection over Union (Mean IoU) in Table 2. These metrics provide a comprehensive basis for comparing the performance of the three segmentation models—M1 (Reference Approach), M2 (RGB-DSM Approach), and M3 (RGB-DSM-IMP Approach)—across five test images with varying levels of complexity.

Table 1. Precision (P), Recall (R), F1-score (F1), and Average Precision (AP) for models M1–M3 across five test images

	M1				**M2**				**M3**			
Image	**P**	**R**	**F1**	**AP**	**P**	**R**	**F1**	**AP**	**P**	**R**	**F1**	**AP**
Img1	0.83	0.79	0.81	0.74	1.00	0.53	0.69	0.47	0.94	0.79	**0.86**	0.74
Img2	0.65	0.68	0.67	0.63	0.80	0.55	0.65	0.49	0.87	0.59	**0.70**	0.54
Img3	0.69	0.55	0.61	0.50	0.83	0.50	0.63	0.44	0.85	0.55	**0.67**	0.50
Img4	0.67	0.31	0.42	0.26	0.83	0.77	**0.80**	0.71	0.69	0.69	0.69	0.63
Img5	0.67	0.31	0.42	0.26	0.77	0.77	0.77	0.71	0.83	0.77	**0.80**	0.72
Average	0.70	0.53	0.59	0.48	0.85	0.62	0.71	0.56	0.84	0.68	**0.74**	0.63
			mAP=	**0.48**			**mAP=**	**0.56**			**mAP=**	**0.63**

Table 2. Mean Intersection over Union (IoU) scores for each model on the same test set

Image	**M1**	**M2**	**M3**
Img1	0.77	0.73	0.73
Img2	0.62	0.68	0.73
Img3	0.70	0.73	0.74
Img4	0.74	0.74	0.75
Img5	0.65	0.65	0.73
Average	0.70	0.71	**0.74**

Model M1, which utilizes only RGB input, demonstrates moderate precision (0.70) but relatively low recall (0.53). This results in the lowest average F1-score (0.59) and mean Average Precision (mAP = 0.48) among the three models. These results indicate that M1 is more prone to false negatives, especially in complex scenes where visual similarities between rooftops and background elements such as trees or roads often lead to segmentation errors.

Model M2, which integrates Digital Surface Model (DSM) data with RGB imagery, achieves the highest precision (0.85) and recall (0.62) values. It improves the average F1-score to 0.71 and increases mAP to 0.56. This suggests that the incorporation of height information from DSM helps distinguish buildings from background clutter. However, the model still exhibits inconsistencies in performance across different test images, implying sensitivity to DSM quality or scene-specific characteristics.

Model M3, which incorporates DSM data and a background removal preprocessing step, shows the most consistent and balanced performance. It achieves the highest average F1-score of 0.74 and mAP of 0.63, with strong precision (0.84) and recall (0.68). These improvements demonstrate the effectiveness of preprocessing in eliminating irrelevant features and enabling the model to focus more accurately on the target structures.

Overall, the quantitative results support the conclusion that while DSM data enhances segmentation accuracy (as seen in M2), the inclusion of background removal preprocessing (in M3) further improves both detection accuracy and consistency across varied environments.

3.3 Backbone comparison

In addition to evaluating the proposed segmentation approaches (M1–M3), this study also examined the impact of different backbone architectures within the Mask R-CNN framework. The architectures employed as feature extractors were ResNet50, ResNet101, and ResNeXt. Although ResNeXt achieved slightly higher F1-scores in some scenarios, it required substantially more computational resources, including longer training times and higher memory consumption. In contrast, ResNet50 demonstrated a more favorable balance between segmentation accuracy and computational efficiency. The choice of ResNet50 is further supported by our previous study on multi-altitude UAV imagery (Khiewwan & Asavasuthirakul, 2025), where extensive hyperparameter tuning and architectural comparisons revealed that ResNet50 consistently offered the best trade-off between performance and resource demands. That study showed that ResNet50 maintained stable segmentation accuracy across flight altitudes of 90 m, 120 m, and 150 m, with high F1-scores and faster convergence compared to deeper networks such as ResNet101. Although other convolutional neural network (CNN) backbones—such as DenseNet, EfficientNet, and MobileNet—were considered, this study focused on comparing deeper residual networks and ResNeXt variants to evaluate model depth and complexity.

3.4 Visual comparison of segmentation results

To further evaluate model performance beyond quantitative metrics, segmentation results for five representative test images (Img1–Img5) were visualized and compared across the three models, as shown in Figure 8. These images were selected to reflect a range of complexities in building segmentation scenarios, from relatively simple layouts to dense, vegetation-rich urban areas.

Figure 8 presents a visual comparison of the segmentation results produced by the three models across five representative test images. Each row corresponds to one test image, showcasing the original image followed by the output of Models M1, M2, and M3, respectively.

Model M1, which utilizes only RGB data, performs reasonably well in simpler environments (e.g., first and second rows), but it struggles in more complex scenes with background clutter, such as vegetation and roads. The segmentation masks frequently extend into non-building areas, reflecting the model’s limited ability to differentiate between

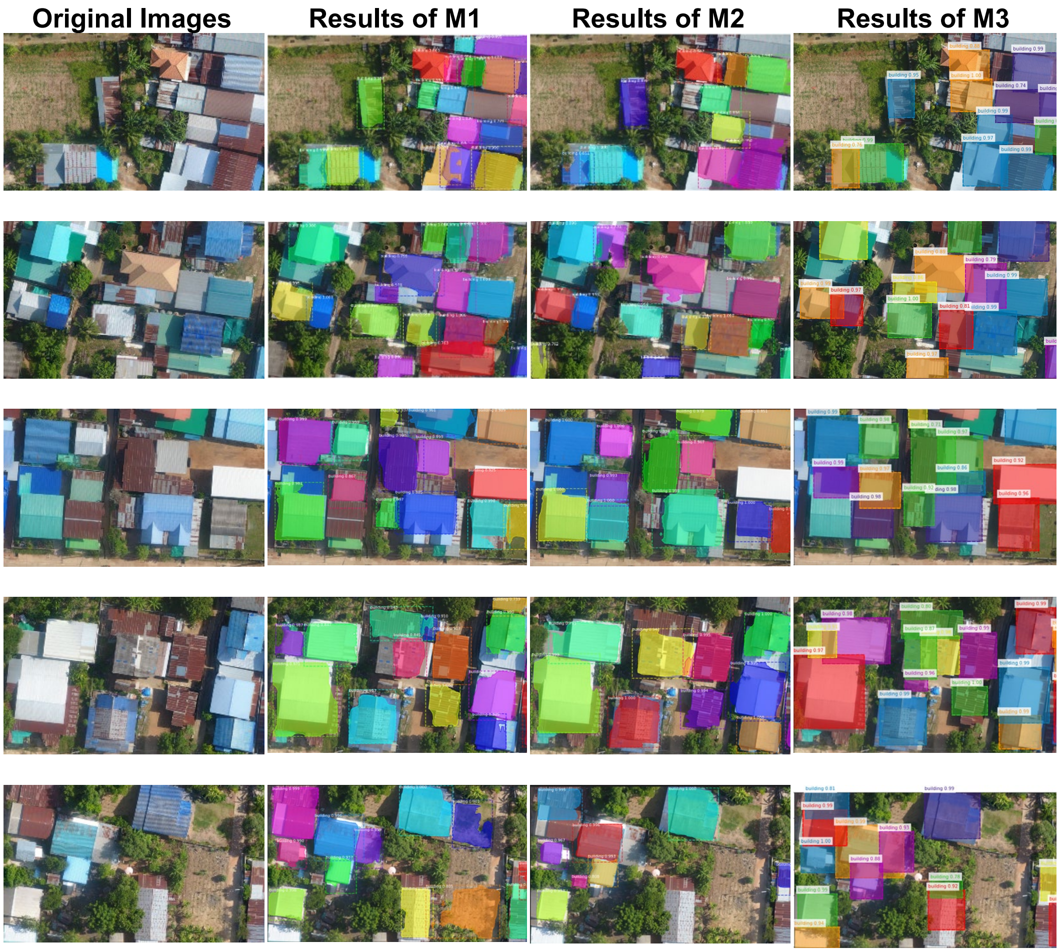


Figure 8. Segmentation results of Models M1, M2, and M3

similar textures and colors. Although these misclassifications are visually apparent, Model M1 still achieves relatively high F1-scores and IoU values in Tables 1 and 2. This can be explained by the object-level evaluation criterion, where predicted masks are considered correct if the Intersection over Union (IoU) is greater than or equal to 0.5. In such cases, masks that overlap significantly with the ground truth—despite partially including non-target regions—are counted as true positives. This highlights a common limitation of quantitative metrics in semantic segmentation, where numeric scores may not fully reflect the quality of boundary delineation or fine-scale visual accuracy (Zhang & Liu, 2020; Müller et al., 2022).

Model M2, which incorporates DSM data, shows notable improvement in separating buildings from the background. The inclusion of elevation information helps the model identify structures that are not easily distinguished using RGB data alone. However, the results still exhibit some degree of misclassification, particularly in densely vegetated areas (e.g., fourth and fifth rows), suggesting that DSM data alone may not be sufficient in all cases.

Model M3, which combines DSM data with a preprocessing step for background removal, provides the most accurate and visually consistent results. The segmentation masks align more precisely with building boundaries, and the model demonstrates better generalization across various scene complexities. The results indicate that removing background elements such as trees and roads prior to segmentation significantly enhances the model's ability to focus on the target structures, thereby reducing false positives and improving overall segmentation quality.

Overall, the visual outputs support the quantitative findings presented in Tables 1 and 2, confirming that the M3 model achieves superior performance in terms of both accuracy and consistency across different environments.

3.5 Impact of background removal on misclassification

Segmentation errors in building extraction often result from visual similarities between target (e.g., rooftops) and non-target (e.g., trees or roads) elements. These challenges are particularly evident in models that lack preprocessing techniques to isolate buildings from the surrounding context. Figure 9 illustrates a case of misclassification by Model M1, which uses only RGB input for building segmentation. The highlighted area on the right (yellow ellipse) shows trees and vegetated regions that were incorrectly labeled as building structures. The model's inability to distinguish between rooftops and spectrally similar background elements results in a high rate of false positives. This example emphasizes the limitation of relying solely on RGB data without additional spatial or preprocessing enhancements. Although visually apparent, such errors do not significantly impact the F1-score (Table 1) because evaluation is based on an IoU threshold of ≥ 0.5, which accepts partial overlaps as correct (Zhang & Liu, 2020; Müller et al., 2022). Consequently, segmentation inaccuracies may be underrepresented in quantitative metrics. Similar to findings in scene classification research (Mungklachaiya & Salaiwarakul, 2024), visual similarity among objects—such as trees and buildings—can lead to misclassification, which may not be fully captured by quantitative metrics like IoU and F1-score.

By contrast, Figure 10 demonstrates the effectiveness of background removal in Model M3, where non-target elements such as trees and roads are excluded from the input through preprocessing. This step helps the model to focus exclusively on relevant structures—primarily buildings—resulting in improved segmentation accuracy. The output clearly shows that M3 minimizes confusion from background elements compared to M1 and M2.

Visual evidence from Figures 9 and 10 reinforces the quantitative advantages of M3, as also seen in Figure 8. In complex scenes, such as those with overlapping or ambiguous features, M1 and M2 often misclassify roads, trees, or low-contrast surfaces as rooftops due to their similar spectral profiles (Kampffmeyer et al., 2016). For instance, in Figure 9, the area marked by the yellow ellipse highlights such confusion, where the model fails to distinguish between trees and buildings.

The proposed model, M3, addresses these limitations by incorporating a background removal preprocessing step that eliminates irrelevant regions prior to model inference. As shown in Figure 10, this step enables better focus on buildings, reduces noise in the input, and improves both recall and precision (Dawn, 2024). The increase in segmentation accuracy achieved by M3, reflected in its higher F1-score and Mean IoU (Table 2), supports the claim that simplifying input features improves deep model performance (Dombrowski et al., 2022).

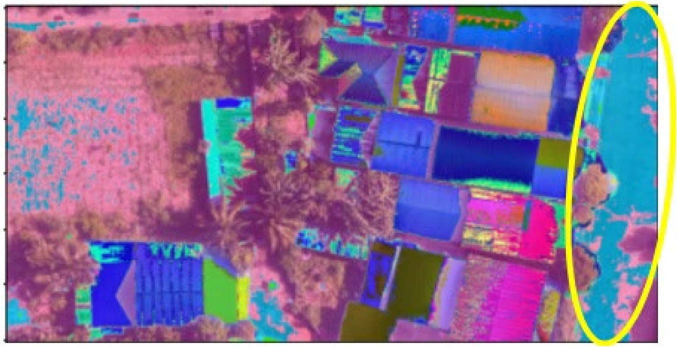


Figure 9. Example of misclassification in model M1

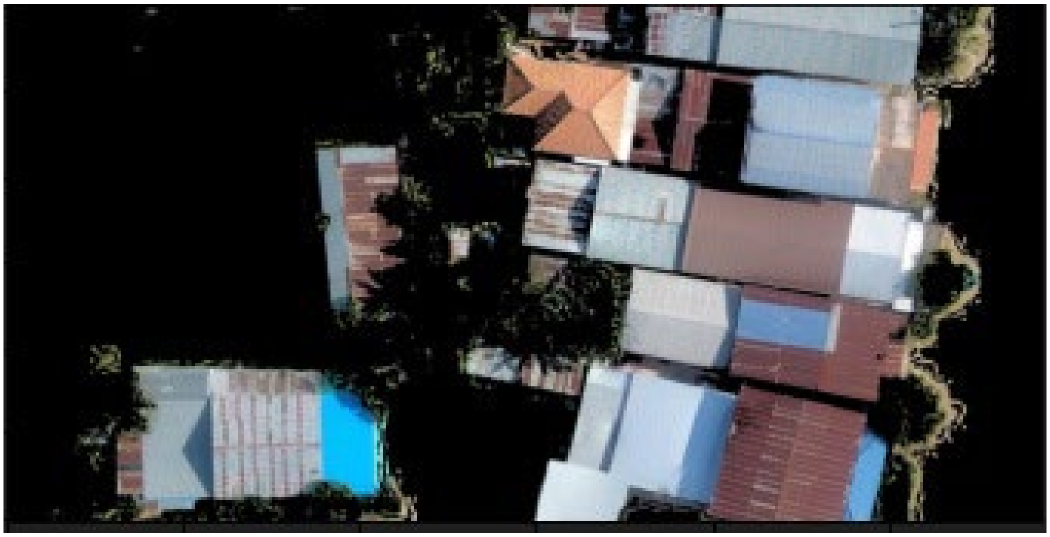


Figure 10. The effect of background removal in Model M3

This enhancement is consistent with established findings in image processing, which emphasize the importance of preprocessing for feature clarity and model robustness (Hu & Yu, 2023). Integrating DSM data with targeted preprocessing, as demonstrated in M3, proves particularly effective in complex urban environments where non-target elements overlap spatially and spectrally with buildings (Ma et al., 2024).

4. Conclusions

This study presents an enhanced building segmentation approach for unmanned aerial vehicle (UAV) imagery by integrating RGB data, Digital Surface Model (DSM) data, and a targeted background removal preprocessing technique, collectively referred to as the RGB-DSM-IMP (M3) approach. Through extensive quantitative and qualitative evaluations, the M3 model demonstrated superior performance compared to baseline methods—M1 (Red-Green-Blue only) and M2 (Red-Green-Blue with Digital Surface Model)—achieving the highest average F1-score (0.74), mean intersection over union (0.74), and mean average precision (0.63).

The results confirm that incorporating elevation information from Digital Surface Model data enhances spatial differentiation, while background removal significantly reduces misclassifications caused by vegetation and roads. These improvements enable the Mask R-CNN model to focus more accurately on building structures. This dual enhancement leads to greater segmentation accuracy and consistency across varied scene complexities, supporting its applicability in real-world use cases such as urban planning, infrastructure monitoring, and post-disaster assessment.

Future research will aim to improve the model's generalization and adaptability by incorporating images captured under diverse environmental and lighting conditions. Further investigations will explore the model's robustness at different flight altitudes to ensure stable performance across varying image acquisition heights. In addition, future work will examine architectural modifications within the Mask R-CNN framework to assess whether such enhancements can further improve segmentation accuracy and feature discrimination.

These directions are intended to expand the model's capability to operate effectively under a wide range of real-world conditions, thereby promoting broader adoption of UAV-based geospatial analytics.

5. Acknowledgements

We extend our gratitude to Mapedia Co., Ltd. and Drone AI Solutions Co., Ltd. for providing the drone photographs used in this research.

6. Author's Contributions

Kanokwan Khiewwan designed research; implemented methodology; performed analysis; and wrote the paper. Duangduen Asavasuthirakul collected data; analyzed results; reviewed and edited the manuscript. Sutasinee Chimlek conceptualization and analysis; reviewed and edited the manuscript.

ORCID

Kanokwan Khiewwan https://orcid.org/0000-0003-3762-846X
Sutasinee Chimlek https://orcid.org/0000-0001-5371-8911

7. Conflicts of Interest

The authors declare that they have no conflicts of interest.

8. AI Declaration

The preparation of this manuscript by ChatGPT in order to improve language and readability was reviewed and edited as needed by Kanokwan Khiewwan, Duangduen Asavasuthirakul and Sutasinee Chimlek. The authors take fully responsible for the content of the publication.

References

Al-Najjar, H. A. H., Kalantar, B., Pradhan, B., Saeidi, V., Halin, A. A., Ueda, N., & Mansor, S. (2019). Land cover classification from fused DSM and UAV images using convolutional neural networks. *Remote Sensing, 11*(12), 1-18. https://doi.org/10.3390/rs11121461

Amo-Boateng, M., Sey, N., Amproche, A., & Domfeh, M. (2022). Instance segmentation scheme for roofs in rural areas based on Mask R-CNN. *Egyptian Journal of Remote Sensing and Space Science, 25*, 569-577.

Boonpook, W., Tan, Y., & Xu, B. (2020). Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry. *International Journal of Remote Sensing, 42*(1), 1-19. https://doi.org/10.1080/01431161.2020.1788742

Chea, C., Saengprachatanarug, K., Posom, J., Wongphati, M., & Taira, E. (2019). Sugarcane canopy detection using high spatial resolution UAS images and digital surface model. *Engineering and Applied Science Research, 46*(4), 312-317.

Chen, J., Wang, G., Luo, L., Gong, W., & Cheng, Z. (2021). Building area estimation in drone aerial images based on Mask R-CNN. *IEEE Geoscience and Remote Sensing Letters, 18*(5), 891-894. https://doi.org/10.1109/LGRS.2020.2988326

Chueprasert, T., Udomchaiporn, A., & Intagosum, S. (2025). Comparative analysis of deep learning models for building extraction from high-resolution satellite imagery. *Current Applied Science and Technology, 25*(1), Article e0260846. https://doi.org/10.55003/cast.2024.260846

Dawn, K. (2024, November 20). *Enhancing image segmentation using U2-Net: An approach to efficient background removal*. https://learnopencv.com/u2-net-image-segmentation/

Dombrowski, M., Reynaud, H., Baugh, M., & Kainz, B. (2022). *Foreground-background separation through concept distillation from generative image foundation models*. https://arxiv.org/pdf/2212.14306

Hu, Z., & Yu, T. (2023). *Dynamic spectrum mixer for visual recognition*. https://arxiv.org/pdf/2309.06721

Jiang, N., Zhang, J., Li, H., & Lin, X. (2008). Object-oriented building extraction by DSM and very high-resolution orthoimages. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 37*, 441-446.

Kamarulzaman, A. M. M., Jaafar, W. S. W, M., Saad, S. N. M., & Mohan, M. (2023). UAV implementations in urban planning and related sectors of rapidly developing nations: A review and future perspectives for Malaysia. *Remote Sensing, 15*(11), Article 2845. https://doi.org/10.3390/rs15112845

Kampffmeyer, M., Salberg, A. B., & Jenssen, R. (2016). Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1-9). IEEE. https://doi.org/10.1109/CVPRW.2016.105

KC, K., Yin, Z., Li, D., & Wu, Z. (2021). Impacts of background removal on convolutional neural networks for plant disease classification in-situ. *Agriculture, 11*, Article 827. https://doi.org/10.3390/agriculture11090827

Khiewwan, K., & Asavasuthirakul, D. (2025, March). Building segmentation in drone photos across varied flight altitudes using Mask R-CNN. *ICIC Express Letters Part B: Applications*, *16*(3), 309-316. https://doi.org/10.24507/icicelb.16.03.309

Li, J., Cai, X., & Qi, J. (2021). AMFNet: an attention-based multi-level feature fusion network for ground objects extraction from mining area's UAV-based RGB images and digital surface model. *Journal of Applied Remote Sensing, 15*(3), Article 036506. https://doi.org/10.1117/1.JRS.15.036506

Ma, X., Zhang, X., Pun, M.-O., & Huang, B. (2024). *MANet: Fine-tuning segment anything model for multimodal remote sensing semantic segmentation*. https://arxiv.org/pdf/2410.11160

Müller, D., Soto-Rey, I., & Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes, 15*, Article 210. https://doi.org/10.1186/s13104-022-06096-y

Mungklachaiya, S., & Salaiwarakul, A. (2024). Exploring deep learning features and bag-of-visual-words for scene classification. *ICIC Express Letters. Part B: Applications*, *15*(10), 1081-1088. https://doi.org/10.24507/icicelb.15.10.1081

Ran, X., Xue, L., Zhang, Y., Liu, Z., Sang, X., & He, J. (2019). Rock classification from field image patches analyzed using a deep convolutional neural network. *Mathematics, 7*, Article 755. https://doi.org/10.3390/math7080755

Rizk, H., Nishimur, Y., Yamaguchi, H., & Higashino, T. (2022). Drone-based water level detection in flood disasters. *International Journal of Environmental Research and Public Health, 19*(1), Article 237. https://doi.org/10.3390/ijerph19010237

Snyder, B., Kama, S., & ElGalaind, K. (2021, December 13). Distributed Mask R-CNN training with Amazon SageMakerCV. https://aws.amazon.com/blogs/machine-learning/distributed-mask-rcnn-training-with-amazon-sagemakercv

Wang, Q., Yan, L., Sun, Y., Cui, X., Mortimer, H., & Li, Y. (2018). True orthophoto generation using line segment matches. *Photogrammetric Record,* 33, 113-130.

Wang, Y., Li, S., Teng, F., Lin, Y., Wang, M., & Cai, H. (2022). Improved Mask R-CNN for rural building roof type recognition from UAV high-resolution images: A case study in Hunan Province, China. *Remote Sensing, 14*(2), Article 265. https://doi.org/10.3390/rs14020265

Yang, C., Rottensteiner, F., & Heipke, C. (2018). Classification of land cover and land use based on convolutional neural networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 4*(3), 251-258. https://doi.org/10.5194/isprs-annals-IV-3-251-2018

Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W., & Zhao, T. (2019). Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network. *Remote Sensing, 11*(15), Article 1774. https://doi.org/10.3390/rs11151774

Zhang, Y., & Liu, Y. (2020). Image segmentation evaluation: A survey of methods. *Artificial Intelligence Review, 53*(8), 5637-5674.