ICIC Express Letters
Part B: Applications
Volume **16**, Number **3**, March **2025**

ICIC International ©2025 ISSN 2185-2766
pp. **309–316**

BUILDING SEGMENTATION IN DRONE PHOTOS ACROSS VARIED FLIGHT ALTITUDES USING MASK R-CNN

Kanokwan Khiewwan and Duangduen Asavasuthirakul*

Department of Computer Science and Information Technology
Faculty of Science
Naresuan University
99 Moo 9, Thapo Sub-District, Muang District, Phitsanulok 65000, Thailand
kanokwankh63@nu.ac.th; *Corresponding author: duangduenr@nu.ac.th

Received April 2024; accepted July 2024

Abstract. *This research proposed a precise altitude-adaptive building segmentation method which employs Mask R-CNN to accurately partition buildings in drone photos taken at various altitudes. We used transfer learning to train our model, starting with a pre-trained COCO model and fine-tuning it with our training dataset. Optimal performance was achieved in 100-200 epochs with ResNet50 and a learning rate of 0.005. Evaluation used precision, recall, and F1-Score metrics. Models trained and tested at the same flight altitude consistently achieved the highest accuracy while the model trained with a combination of images at different altitudes yielded F1-Scores exceeding 80% across all testing datasets. The visualization of the building segmentation results confirmed the superior performance in scenarios involving partial obstructions, adjacent buildings, and small-sized structures. This research provides an effective method for accurate building partitioning and offers valuable insights for future studies in the field.*
Keywords: Deep learning, ResNet, Unmanned aerial vehicle (UAV), Instance segmentation

1. **Introduction.** Unmanned aerial vehicle (UAV), commonly referred to as drones, have gained widespread adoption for visual data acquisition in remote or inaccessible areas. The utilization of drone imagery is motivated by its ease of deployment and relatively lower costs compared to satellite imagery [1,2].

In real-world drone photography, flights can occur at different altitudes, leading to variations in the images they capture [3]. For example, images taken at altitudes of 90, 120, and 150 meters exhibit distinct characteristics, such as variations in object size. At lower altitudes, more details can be detected with higher object resolutions [4,5].

Deep learning is employed in research for image segmentation, necessitating the teaching of computers to understand the features of target objects [6,7]. When using drone photos captured in different areas and at different altitudes, these variations can affect the objects or image features, potentially impacting model training and performance [8]. In image segmentation, when the goal is to separate individual instances of objects, such as individual buildings, a technique known as instance segmentation is usually employed [9,10]. One popular instance segmentation technique is Mask R-CNN, a deep learning model that combines object detection and instance segmentation. The output from this technique includes labeling objects with their class, bounding box, and pixel-level mask [11,12].

Previous research has indicated that training a Mask R-CNN model using images captured at the same altitude typically yields favorable results. Nathania et al. [13] proposed a deep learning-based approach for automatically detecting buildings in high-density cities using aerial imagery. Yu et al. [14] introduced a vehicle detection method for mid-to-high

DOI: 10.24507/icicelb.16.03.309

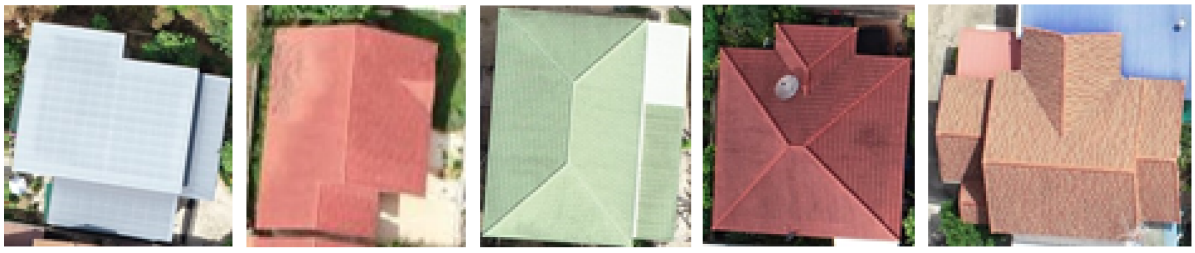
altitude UAV images based on visual attention and spatial-temporal information. They utilized a deep semantic segmentation network to segment road areas in the image, from which stability regions were extracted as candidate regions for vehicles. Ham et al. [15] presented a deep learning-based system for detecting unregistered buildings in UAV images captured at the specific altitude of 120 meters. This methodology involves training a deconvolution network using geospatial data and performing semantic segmentation to generate a building probability map. By comparing this map with existing GIS data, unregistered buildings can be identified. The proposed approach effectively combines deep learning and geospatial data analysis.

However, it remains uncertain whether training the model with drone images taken at different altitudes can be effective. In real-world scenarios, UAV can vary in altitude during operation [16]. Therefore, this research proposes an instance segmentation method to accurately partition individual buildings in images captured at various altitudes. Our method employs Mask R-CNN, which is trained on the COCO dataset and augmented with custom data, thereby enhancing building feature description and segmentation accuracy.

2. Materials and Method.

2.1. Study area and data preparation.

The dataset used to experiment was obtained from drone-captured photos at three different altitudes: 90, 120, and 150 meters, in a village near Naresuan University, Phitsanulok Province, Thailand. The original resolution of each photo is 5472 × 3648 pixels in RGB format, showing the rooftops from a straight-up view. The images were resized to 1000 × 667 pixels to enable efficient handling by GPUs and reduce training time. The model was trained using a dataset containing five roof types: gable, flat, hip, pyramid, and complex, as illustrated in Figure 1. We followed the approaches proposed by Zhang et al. [17] and Wang et al. [18] to develop a model suitable for various datasets.



(a) (b) (c) (d) (e)

FIGURE 1. The different roof types in the training dataset: (a) Flat; (b) gable; (c) hip; (d) pyramid; (e) complex: the combination of roof types in (a)-(d)

The training datasets used in the experiment were categorized into four groups based on flight altitudes, as detailed in Table 1. Group 1 was collected from an altitude of 90 meters, Group 2 from 120 meters altitude, Group 3 from 150 meters, and Group 4 was a combination of data from all three altitudes. The testing dataset was divided into three groups, with each containing images captured at 90, 120, or 150 meters altitude, as detailed in Table 2.

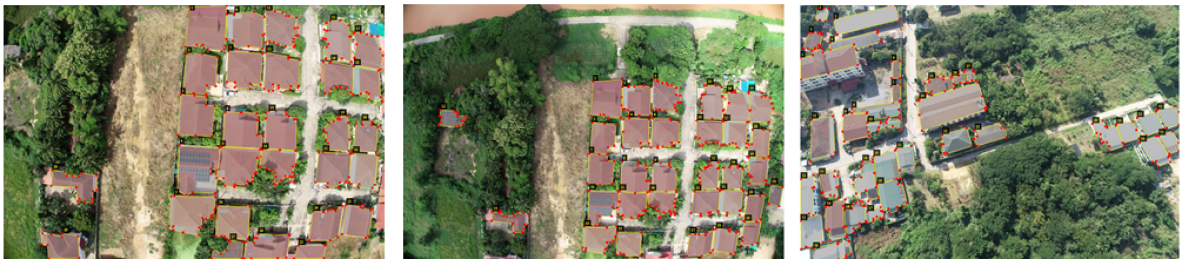
Building rooftop photos were annotated using the free online tool, VGG Image Annotator (VIA), to create an annotation file in JSON format, as detailed in [19]. One building rooftop was labeled as one object/building. Even though multiple buildings are constructed to the main house with different roof colors, they are treated as the same building/object. Townhouses with a common roof with no distinction between separate

TABLE 1. The details of the training datasets used in the experiment

Datasets	90-m altitude		120-m altitude		150-m altitude		Combined altitudes	
	No. of images	No. of buildings	No. of images	No. of buildings	No. of images	No. of buildings	No. of images	No. of buildings
Training	7	228	7	308	7	239	9	337
Validate	3	91	3	98	3	52	3	80

TABLE 2. The details of the testing datasets used in the experiment

Datasets	Flight altitudes	No. of images	No. of buildings
Test#1	90-m	3	84
Test#2	120-m	3	124
Test#3	150-m	3	119



(a) 90-m flight altitude (b) 120-m flight altitude (c) 150-m flight altitude

FIGURE 2. The example images with building annotations at different flight altitudes

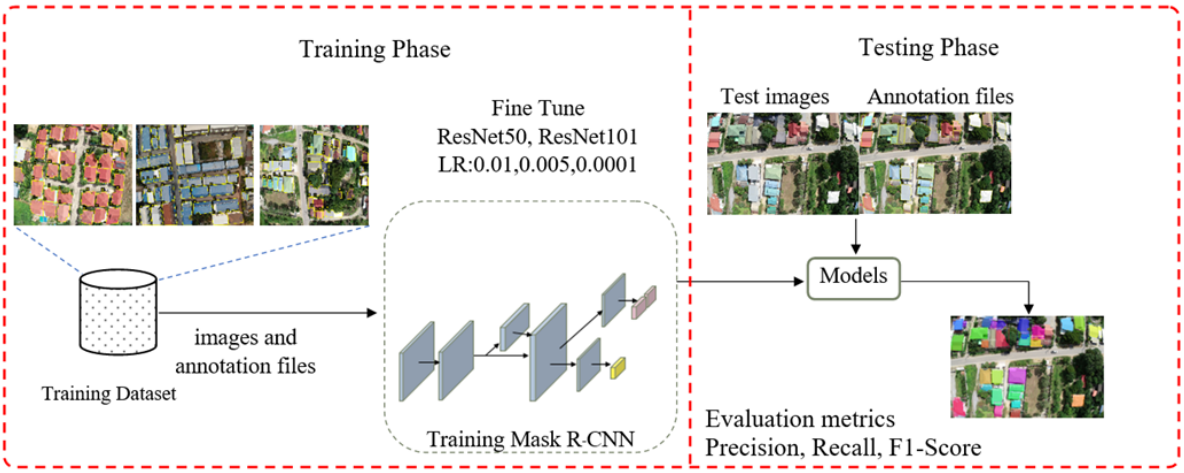


FIGURE 3. An overview of our proposed methodology for building segmentation

units are labeled as one building. The example images and annotations used in the experiment at different flight altitudes are shown in Figure 2.

2.2. **Methodology.** The software and hardware environment configuration included a computer equipped with a Core i7 CPU with a GTX1060 GPU card and 20GB of RAM and Windows 10 as the operating system. The building segmentation program was written in Python using TensorFlow 1.5.0 and Keras 2.1.2 libraries. The methodology used in the study is shown in Figure 3.

As shown in Figure 3, there are two phases of operation in the method, namely the training phase and the testing phase.

2.2.1. *Training and testing phase.* In the training phase, we used Mask R-CNN, which leveraged transfer learning from the COCO dataset. We started with a pre-trained model from the COCO dataset, and then we further trained it using our training dataset referred to Section 2.1. We also compared its performance using two different backbones: ResNet50 and ResNet101 for extracting features, along with three different learning rates (LR): 0.01, 0.005, and 0.0001. The model was trained for 200 epochs on a single GPU with a batch size of 8. As a result, we obtained models for each training dataset group and each learning rate. Comprehensive details of the parameter settings, as cited in [12], are shown in Table 3.

Table 3. Summarization of hyperparameters

Hyperparameters	Detail
Batch size	8
Epoch	200
Learning rate	(0.01, 0.005, 0.0001)
Backbone	(ResNet50, ResNet101)
Weight decay	0.0001
GPU_count	1
Images_per_GPU	1
Momentum	0.9

In the testing phase, we used three datasets of captured images taken at altitudes of 90, 120, and 150 meters, as shown in Table 2. These datasets were tested with the best models that we had identified and trained by each training group and each learning rate and then the model's performance was evaluated.

2.2.2. *Evaluation of model.* The model's performance was evaluated by counting the number of correctly identified buildings, considering the overlap of pixels in the segmented area compared to its corresponding ground truth. This evaluation employs the Intersection over Union (IoU) metric, defined in Formula (1), with a threshold exceeding 50%. Subsequently, precision, recall, and the F1-Score were computed. The specific formula for each accuracy evaluation metric is as follows [20]:

$$\text{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A is the segmented area, and B is the ground truth.

$$\text{Precision (P)} = \frac{\text{The number of objects correctly segmented}}{\text{The total number of ground truth}} \tag{2}$$

$$\text{Recall (R)} = \frac{\text{The number of objects correctly segmented}}{\text{The total number of objects segmented}} \tag{3}$$

$$\text{F1-Score (F1)} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

3. **Results and Discussion.** We examined the loss values of the Mask R-CNN models across four experimental groups, as depicted in Figure 4. It was observed that the loss value of each model reached a stable state after approximately 100 epochs, for both the ResNet50 and ResNet101 architectures. To prevent overfitting, we followed the approach described in [21] and limited training to 200 epochs, in line with the guidance from [12]. The experiment was performed on three LR – 0.01, 0.005, and 0.0001. We discovered that the model trained with the LR of 0.005 produced the highest F1-Score.

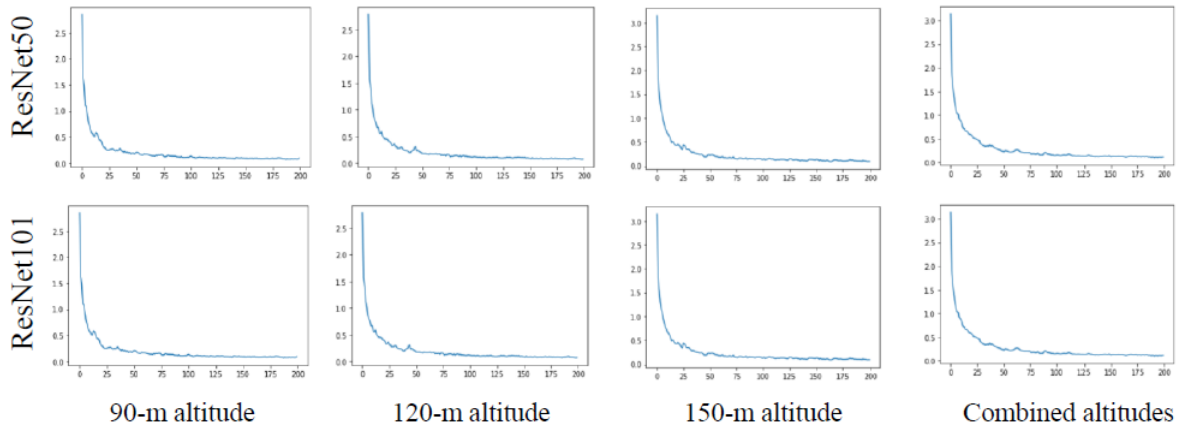


FIGURE 4. The training loss from ResNet50 and ResNet101

3.1. **Building segmentation performance.** Table 4 illustrates how well the model performed with different training groups and learning rates using the testing datasets. Specifically, Table 4 shows the performance of the models trained with ResNet50 and ResNet101 using a learning rate of 0.005.

TABLE 4. Building segmentation accuracy on the different testing datasets (values in %)

Testing datasets	Feature extraction network	Building segmentation accuracy											
		90-m altitude			120-m altitude			150-m altitude			Combined altitudes		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
90-m	ResNet50	89.31	92.69	**90.97**	91.83	82.47	86.90	87.02	68.47	76.64	86.96	86.65	**86.80**
	ResNet101	84.72	84.13	84.42	95.23	84.37	89.47	67.84	55.78	61.22	89.31	77.20	82.81
120-m	ResNet50	82.09	96.08	88.54	89.84	96.58	**93.09**	75.02	75.69	75.35	90.49	87.11	**88.77**
	ResNet101	81.26	94.19	87.25	88.48	90.27	89.37	74.39	68.71	71.44	84.31	85.40	84.85
150-m	ResNet50	57.86	80.61	67.37	61.35	75.52	67.70	78.78	70.22	74.25	85.36	75.93	**80.37**
	ResNet101	63.70	73.52	68.26	66.26	65.96	66.11	81.38	72.94	**76.93**	77.95	74.51	76.17

The ResNet50 models consistently outperformed the ResNet101 models across most testing datasets. Although the ResNet101 model that was trained with the 150-meter altitude dataset achieved the highest F1-Score, this difference was not statistically significant compared to the performance of the ResNet50 model. The superior performance of ResNet50 can be attributed to its use of fewer layers in the feature extraction networks. This design choice offers advantages, such as reduced hardware requirements and shorter training times, enabling more efficient exploration of hyperparameters [22]. Furthermore, increasing the number of layers in the network, ResNet101, may result in the loss of critical data for the segmentation process, aligning with findings reported in [23].

It was also observed that models trained and tested at the same flight altitude consistently achieved the highest segmentation accuracy while the results of the model using combined altitudes demonstrated an improvement in overall segmentation performance across different altitudes.

3.2. **Visual analysis of building segmentation.** The visualization of the building segmentation results from all four groups was conducted using the ResNet50 model with a learning rate of 0.005 to verify the building segmentation performance. The outcomes are presented in Figure 5. While the highest F1-Score varied across different testing datasets, the combined altitudes model consistently demonstrated superior performance in most scenarios, particularly when dealing with buildings partially obstructed by trees (Figure 5(a)), adjacent buildings with limited gaps (Figure 5(b)), and small-sized buildings (Figure 5(c)). This visual analysis consistently highlighted the precision and accuracy of the combined altitudes model in achieving excellent building segmentation outcomes.

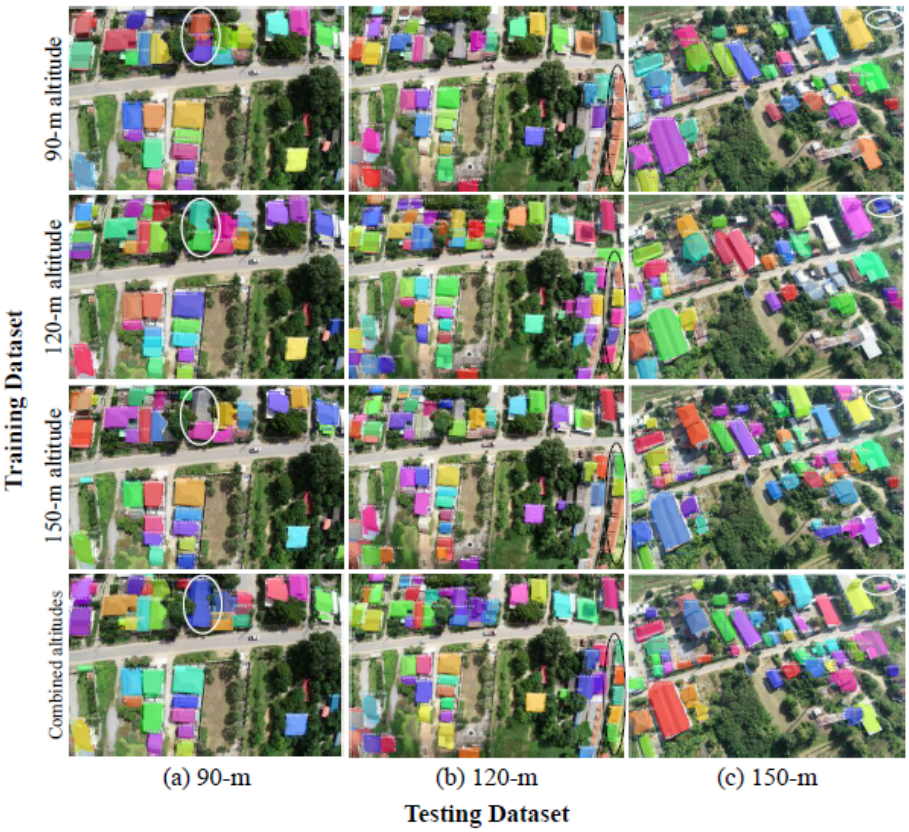


FIGURE 5. Comparison of building segmentation results using the four Mask R-CNN models with the three testing datasets

As shown in Table 4 and Figure 5, the models trained and tested on images taken at the same altitude consistently achieved the highest segmentation accuracy. The combined altitudes model, which used images from all altitudes, also consistently achieved F1-Scores exceeding 80% across the three testing datasets. Furthermore, the segmentation results of the Figure 5 images further validated the superiority of the combined altitudes model, particularly in scenarios involving partial building obstructions, adjacent buildings with limited gaps, and small-sized buildings.

4. **Conclusion.** This research proposed a precise altitude-adaptive building segmentation method which employs Mask R-CNN to accurately partition buildings at various altitudes. We used transfer learning to train our model, starting with a pre-trained COCO model and fine-tuning it with our training dataset. Optimal performance was achieved in 200 epochs with ResNet50 and a learning rate of 0.005.

In future work, our objective is to include a wider variety of building types. In this study, our focus was on the training dataset consisting of images from a housing estate. The next task is to assess whether the model still performs well when introduced to locations with more diverse building types. Additionally, we will explore different instance segmentation models such as U-Net and YOLOv8, and compare their performance to Mask R-CNN. Our purpose is to enhance pixel-level segmentation accuracy, which is crucial for applications such as automated area measurement. This can be achieved by improving the alignment of the segmented building boundaries with the ground truth.

Acknowledgement. We appreciate Mapedia Co., Ltd. for providing the drone photographs utilized in this research. We also thank Mr. Roy I. Morien of the Naresuan University Graduate School for his editing of the grammar, syntax and general English expression in this manuscript.

REFERENCES

[1] H. A. H. Al-Najjar, B. Kalantar, B. Pradhan, V. Saeidi, A. A. Halin, N. Ueda and S. Mansor, Land cover classification from fused DSM and UAV images using convolutional neural networks, *Remote Sensing*, vol.11, no.12, 2019.

[2] K. Ali and B. A. Johnson, Land-use and land-cover classification in semi-arid areas from medium-resolution remote-sensing imagery: A deep learning approach, *Sensors*, vol.22, 2022.

[3] N. Anders, M. Smith, J. Suomalainen et al., Impact of flight altitude and cover orientation on digital surface model (DSM) accuracy for flood damage assessment in Murcia (Spain) using a fixed-wing UAV, *Earth Sci. Inform.*, vol.13, pp.391-404, https://doi.org/10.1007/s12145-019-00427-7, 2020.

[4] E. Seifert, S. Seifert, H. Vogt, D. Drew, A. J. van, A. Kunneke and T. Seifert, Influence of drone altitude, image overlap, and optical sensor resolution on multi-view reconstruction of forest images, *Remote Sensing*, vol.11, no.10, https://doi.org/10.3390/rs11101252, 2019.

[5] E. Ahmed, R. Mostafa, G. Essam, M. Rasha and T. Ahmed, The influence of flight height and overlap on UAV imagery over featureless surfaces and constructing formulas predicting the geometrical accuracy, *NRIAG Journal of Astronomy and Geophysics*, vol.11, no.1, pp.210-223, DOI: 10.1080/20909977.2022.2057148, 2023.

[6] I. H. Sarker, Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions, *SN Comput. Sci.*, vol.2, 420, https://doi.org/10.1007/s42979-021-00815-1, 2021.

[7] L. Alzubaidi, J. Zhang, A. J. Humaidi et al., Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions, *J. Big Data*, vol.8, no.53, https://doi.org/10.1186/s40537-021-00444-8, 2021.

[8] H. Rizk, Y. Nishimur, H. Yamaguchi and T. Higashino, Drone-based water level detection in flood disasters, *Int. J. Environ. Res. Public Health*, vol.19, no.1, DOI: 10.3390/ijerph19010237, 2022.

[9] A. M. Hafiz and G. M. Bhat, A survey on instance segmentation: State of the art, *International Journal of Multimedia Information Retrieval*, vol.9, pp.171-189, 2020.

[10] K. He, G. Gkioxari, P. Dollár and R. Girshick, Mask R-CNN, *IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp.2980-2988, DOI: 10.1109/ICCV.2017.322, 2017.

[11] C. S. Ramesh and V. V. Kumar, A review on instance segmentation using mask R-CNN, *Proc. of the International Conference on Systems, Energy & Environment (ICSEE 2021)*, http://dx.doi.org/10.2139/ssrn.3794272, 2021.

[12] T. Wu, Y. Hu, L. Peng and R. Chen, Improved anchor-free instance segmentation for building extraction from high-resolution remote sensing images, *Remote Sensing*, vol.12, no.18, pp.2-15, 2020.

[13] D. A. D. A. Nathania, C. Surya, A. A. S. Gunawan and E. Irwansyah, Comparison of deep learning models for dense building segmentation in the city using aerial imagery data, *ICIC Express Letters, Part B: Applications*, vol.13, no.10, pp.1009-1018, 2022.

[14] R. Yu, H. Li, Y. Jiang, B. Zhang and Y. Wang, Vehicle detection for mid-to-high altitude UAV images based on visual attention and spatial-temporal information, *Sensors*, vol.22, no.6, pp.1-18, https://doi.org/10.3390/s22062354, 2022.

[15] S. Ham, Y. Oh, K. Choi and I. Lee, Semantic segmentation and unregistered building detection from UAV images using a deconvolutional network, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol.XLII-2, pp.419-424, https://doi.org/10.5194/isprs-archives-XLII-2-419-2018, 2018.

[16] F. X. Viana, G. M. Araujo, M. F. Pinto, J. Colares and D. B. Haddad, Aerial image instance segmentation through synthetic data using deep learning, *Journal of the Brazilian Society on Computational Intelligence (SBIC),* vol.18, no.1, pp.35-46, 2020.

[17] X. Zhang, Y. Fu, A. Zang, L. Sigal and G. Agam, Learning classifiers from synthetic data using a multichannel autoencoder, *arXiv Preprint*, arXiv: 1503.03163, 2015.

[18] Y. Wang, S. Li, F. Teng, Y. Lin, M. Wang and H. Cai, Improved Mask R-CNN for rural building roof type recognition from UAV high-resolution images: A case study in Hunan province, China, *Remote Sensing*, vol.14, no.2, pp.1-25, 2022.

[19] S. M. Noe, T. T. Zin, P. Tin and I. Kobayashi, Automatic detection and tracking of mounting behavior in cattle using a deep learning-based instance segmentation model, *International Journal of Innovative Computing, Information and Control*, vol.18, no.1, pp.211-220, 2022.

[20] K. Zhou, Y. Chen, I. Smal and R. Lindenbergh, Building segmentation from airborne VHR images using Mask R-CNN, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol.XLII-2/W13, pp.155-161, 2019.

[21] T. Suwannaphong, S. Chavana, S. Tongsom, D. Palasuwan, T. H. Chalidabhongse and N. Anantrasirichai, Parasitic egg detection and classification in low-cost microscopic images using transfer learning, *arXiv Preprint*, abs/2107.00968, 2021.

[22] K. K. Bressem, L. C. Adams, C. Erxleben, B. Hamm, S. M. Niehues and J. L. Vahldiek, Comparing different deep learning architectures for classification of chest radiographs, *Scientific Reports*, vol.10, 2020.

[23] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp.770-778, 2016.