Combining dual attention mechanism and efficient feature aggregation for road and vehicle segmentation from UAV imagery

Trung Dung Nguyen, Trung Kien Pham, Chi Kien Ha, Long Ho Le, Thanh Quyen Ngo, Hoanh Nguyen
Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

Article Info

Article history:

Received May 15, 2023
Revised Oct 2, 2023
Accepted Oct 24, 2023

Keywords:

Deep learning
Dual attention
Feature aggregation
Road and vehicle segmentation
Semantic segmentation

ABSTRACT

Unmanned aerial vehicles (UAVs) have gained significant popularity in recent years due to their ability to capture high-resolution aerial imagery for various applications, including traffic monitoring, urban planning, and disaster management. Accurate road and vehicle segmentation from UAV imagery plays a crucial role in these applications. In this paper, we propose a novel approach combining dual attention mechanisms and efficient multi-layer feature aggregation to enhance the performance of road and vehicle segmentation from UAV imagery. Our approach integrates a spatial attention mechanism and a channel-wise attention mechanism to enable the model to selectively focus on relevant features for segmentation tasks. In conjunction with these attention mechanisms, we introduce an efficient multi-layer feature aggregation method that synthesizes and integrates multi-scale features at different levels of the network, resulting in a more robust and informative feature representation. Our proposed method is evaluated on the UAVid semantic segmentation dataset, showcasing its exceptional performance in comparison to renowned approaches such as U-Net, DeepLabv^{3+}, and SegNet. The experimental results affirm that our approach surpasses these state-of-the-art methods in terms of segmentation accuracy.

Corresponding Author:

Hoanh Nguyen
Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City
12 Nguyen Van Bao, Ho Chi Minh City, Vietnam
Email: nguyenhoanh@iuh.edu.vn

1. INTRODUCTION

Road and vehicle segmentation from unmanned aerial vehicle (UAV) imagery represents a critical cog in the machinery of several applications, encompassing traffic monitoring, autonomous navigation, urban planning, and infrastructure management [1]–[4]. These segmentations facilitate enhanced comprehension and modeling of traffic trends, heightened situational awareness, and more strategic decision-making in the aforementioned domains. Given the rapid progression in UAV technology and the surge in the availability of high-resolution aerial visuals, the need for powerful, efficient segmentation methods capable of tackling intricate urban landscapes and yielding accurate results has become increasingly paramount [5]–[7].

Image segmentation tasks have witnessed remarkable achievements thanks to the advancements in deep learning-based approaches, specifically convolutional neural networks (CNNs). Several architectures, such as fully convolutional neural network (FCN) [8], U-Net [9], DeconvNet [10], and SegNet [11], have been proposed and exhibited impressive performance across diverse semantic segmentation tasks [12]–[17]. Based on the success of these architectures, various methods have been proposed to improve performance of road and vehicle segmentation tasks. Zhang *et al.* [18] presents a deep learning approach for road extraction using a modified U-Net architecture with residual connections. The method significantly improves road segmentation performance in remote sensing imagery, as demonstrated by experimental results on multiple datasets. Wan *et al.* [19], introduce a novel deep learning structure that incorporates a dual attention mechanism to improve road extraction performance. The experimental results demonstrate that the proposed DA-RoadNet model achieves superior segmentation accuracy in high-resolution satellite imagery compared to existing methods. Ren *et al.* [20] introduce a novel method that combines a capsule-based U-Net structure with dual attention mechanisms. The method enhances the model's ability to capture both local and global contextual information, resulting in improved road extraction performance in remote sensing imagery. Kestur *et al.* [21] propose the undecimated (UFCN) model that utilizes a FCN for road extraction in high-resolution RGB imagery obtained from UAVs. The UFCN model efficiently processes large-scale aerial images, demonstrating improved road segmentation performance compared to traditional methods. Varia *et al.* [22] present a deep CNNs-based approach called DeepExt for road extraction using high-resolution RGB imagery from UAVs. The proposed convolutional neural network model demonstrates improved road segmentation performance by leveraging the features of aerial images, outperforming traditional approaches in road extraction tasks. Qian *et al.* [23], propose a deep CNNs-based structure called DLT-Net that simultaneously detects drivable areas, lane lines, and traffic objects in a single framework. The model demonstrates improved efficiency and accuracy compared to separate detection methods, showcasing its potential for real-time applications in intelligent transportation systems. Lo Bianco *et al.* [24] present a method that combines the semantic segmentation of road objects and lanes within a single CNN framework. This unified approach results in improved accuracy and efficiency compared to separate segmentation methods, offering potential benefits for autonomous vehicle navigation and intelligent transportation systems. Teichmann *et al.* [25] propose a method called MultiNet, which enables real-time simultaneous semantic reasoning for autonomous driving applications. The model efficiently processes input data within a single deep learning framework, simultaneously detecting objects, estimating drivable areas, and segmenting lanes, thereby improving overall performance, and reducing computational overhead.

While the aforementioned methods have indeed demonstrated noteworthy success in road and vehicle segmentation tasks, they often encounter challenges when applied to high-resolution UAV imagery. This type of imagery presents a diverse array of scales, orientations, and appearances of roads and vehicles that can confound conventional segmentation techniques. Firstly, the scale of objects in UAV imagery can vary dramatically. These wide-ranging scales can make it difficult for conventional segmentation methods to accurately differentiate between and identify roads and vehicles. Secondly, the orientation of roads and vehicles in UAV images can also pose a challenge. Roads can stretch in multiple directions, not just vertically or horizontally, and vehicles can be found in an assortment of orientations, depending on their direction of movement. This wide array of possible orientations can confuse conventional segmentation algorithms, leading to less accurate results. Moreover, roads and vehicles in UAV images can have a multitude of appearances due to differences in design, color, and lighting conditions. Appearance can also be influenced by factors such as the time of day or weather conditions, which can alter the visibility of the roads and vehicles. Lastly, complex urban scenes often come with their own set of challenges. Occlusions, such as one vehicle blocking the view of another, can make it difficult to accurately identify and segment each individual vehicle. Similarly, shadows can change the apparent shape and color of roads and vehicles, making them harder to segment correctly. Additionally, objects that look similar to roads or vehicles, such as rooftops or riverbanks, can confuse the segmentation task, leading to less accurate results. In this paper, we propose a novel approach that addresses these challenges and improves segmentation performance. Our method builds upon the strengths of the U-Net architectures while incorporating several novel components designed to handle the unique challenges associated with UAV imagery. We designed an efficient multi-layer feature aggregation strategy that integrates both deep and shallow features in the decoder branch. The feature aggregation strategy incorporates both spatial and channel self-attention mechanisms, allowing for adaptive feature refinement at each spatial location while effectively utilizing spatial and channel information during aggregation. The primary contributions of our work can be summarized as follows:

- We introduce a dual attention approach that effectively captures both contextual information at local and global scales, allowing the model to better distinguish roads, vehicles, and other objects in the scene.
- Our method employs an efficient multi-layer feature aggregation strategy that integrates multi-scale features and enhances the model's ability to segment objects of varying sizes, shapes, and appearances.
- We conduct extensive experiments using a publicly accessible dataset comprising high-resolution UAV images, demonstrating that our proposed method significantly outperforms renowned approaches such as U-Net, DeepLabv^{3+}, and SegNet in terms of segmentation accuracy and efficiency.
- We provide a thorough analysis of the results, highlighting the strengths and limitations of our method and suggesting avenues for future work.

The paper is structured as follows: section 2 elaborates on the proposed method, providing a detailed explanation. Section 3 analyzes the experimental results and compares them with other approaches. Finally, section 4 concludes the paper, highlighting future research avenues.

2. METHOD

2.1. The overall architecture

The overall pipeline of our model is illustrated in Figure 1. We employ an encoder-decoder architecture based on the U-Net [9] architecture for generating the output segmentation maps of roads and vehicles. U-Net has emerged as a widely adopted deep learning architecture for image segmentation tasks. U-Net consists of an encoder (contracting path) with successive layers of convolutional, ReLU activation, and max-pooling operations and a decoder (expanding path) with a combination of up-convolutional layers and skip connections from the corresponding layers in the encoder, giving it a symmetric U-shaped structure. The encoder captures contextual information, while the decoder combines the contextual information with spatial information to accurately segment the input image. We replaced the standard encoder part of U-Net with the RestNet-50 [26] architecture. This modification provides an improved feature extraction capability due to the deep and powerful ResNet-50 architecture. The decoder part remains the same as in the original U-Net, with up-convolutional layers and skip connections from the ResNet-50 encoder.

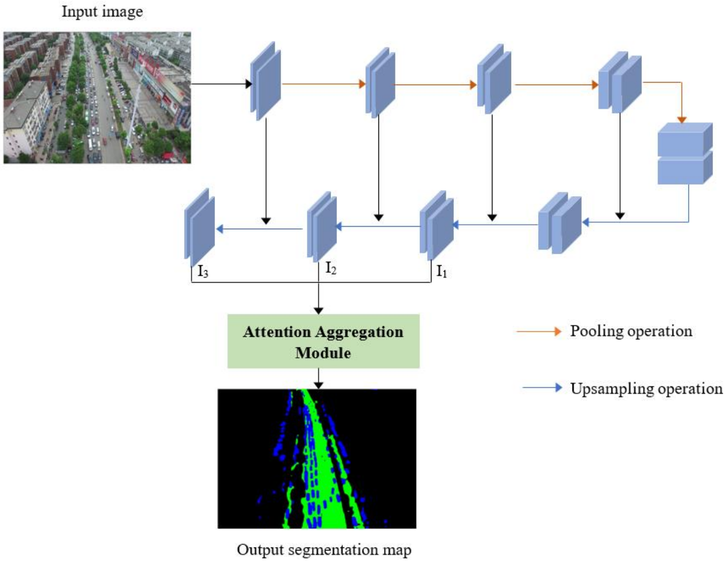


Figure 1. The overall architecture of our approach

To improve the segmentation results, we design an attention aggregation module that integrates a diverse range of deep and shallow features in the decoder branch. The attention aggregation module incorporates both spatial and channel self-attention mechanisms, allowing for adaptive feature refinement at each spatial location while effectively utilizing spatial and channel information during aggregation. This results in an efficiently generated aggregated feature with a rich representation. In particular, the spatial and channel self-attention mechanisms, which employ nonlinear operations, are applied to three distinct feature maps from the final decoder layers to create spatial and channel attention maps. These attention maps are then reassembled, considering the relationships between the input features, to produce the ultimate feature map. This enables the model to capture both global and local contextual information effectively. The subsequent subsections will provide detailed explanations of each module.

2.2. Attention aggregation module

In the process of extracting roads and vehicles from UAV imagery, enhancing the semantic information within deep features is crucial due to the significant scale disparities between road and vehicle targets. As demonstrated Liu *et al*. [27], proved that semantic information gathered from various convolution layers progressively coarsens, leading to the loss of valuable details from earlier convolutional layers in the final layer. The majority of techniques employed for semantic segmentation tasks utilize the decoder's last layer output to create output segmentation maps, consequently diminishing the comprehensiveness and precision of the ultimate output segmentation maps. To fully leverage the semantic details, present in each layer and enhance segmentation results, we have developed an attention aggregation module that employs both spatial and channel attention mechanisms to learn and refine the attention of each input feature, integrating them through nonlinear operations. The attention aggregation module combines three last output layers of the decoder, i.e., I_1, I_2, I_3 where I_1 is the highest resolution feature map and I_3 is the lowest resolution feature map, as shown in Figure 2. For each input feature map, we first compute both spatial and channel attention maps, as illustrated in the next section. Then, we use element-wise multiplication to combine each input feature map with corresponding attention map to generate rich semantic feature map. Finally, we perform hierarchical attentive fusion by fusing the multi-scale features as (1):

$$F_{out} = \sum_{i=1}^{3} I_{i_{cs}} \odot I_i \tag{1}$$

where $I_{i_{cs}}$ is the corresponding attention map generated by spatial and channel attention subnetwork, $\odot$ represents element-wise multiplication with broadcasted unit dimensions.

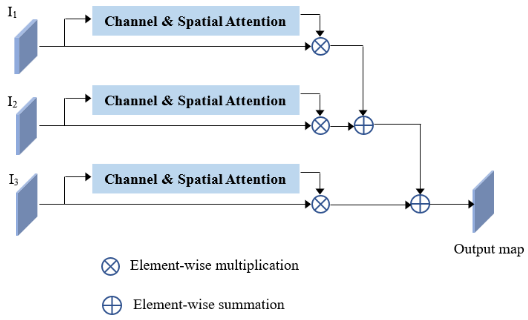

I₁
Channel & Spatial Attention
I₂
Channel & Spatial Attention
I₃
Channel & Spatial Attention
Output map
Element-wise multiplication
Element-wise summation

Figure 2. The pipeline of the attention aggregation module

Through the incorporation of the attention aggregation module, the resulting feature map becomes a confluence of features at various scales, encompassing information spanning from shallow to deep levels. As a result, the model gains increased flexibility to emphasize the aggregation of feature maps from different layers of the network, facilitating the acquisition of more semantic representations.

2.3. Spatial and channel attention

Due to the substantial scale disparities between road and vehicle targets encountered in road and vehicle segmentation tasks, it becomes crucial to direct attention towards different target objects within varying scale contexts. Taking inspiration from SCA-CNN [28] and CBAM [29], which use channel and spatial self-attention to perform adaptive feature refinement and improves performance in image classification, object detection and image captioning tasks, we design a spatial and channel attention subnetwork to generate attention map at each input layer of the attention aggregation module, as shown in Figure 3. For each input feature map $I_i \in \mathbb{R}^{C\times H\times W}$, we use two 3×3 convolution layers followed by a sigmoid activation to generate the spatial attention map $I_{i_s} \in \mathbb{R}^{1\times H\times W}$. At the same time, we apply average pooling and max pooling followed by two 1×1 convolution layers and a sigmoid function to get channel-wise attention maps $I_{i_c} \in \mathbb{R}^{C\times 1\times 1}$. We then apply element-wise multiplication with equal weighting on these maps to generate output attention map $I_o \in \mathbb{R}^{C\times H\times W}$. Since the channel and spatial self-attention subnetwork is a lightweight module, it can effectively perform adaptive feature refinement without much additional computational overhead.

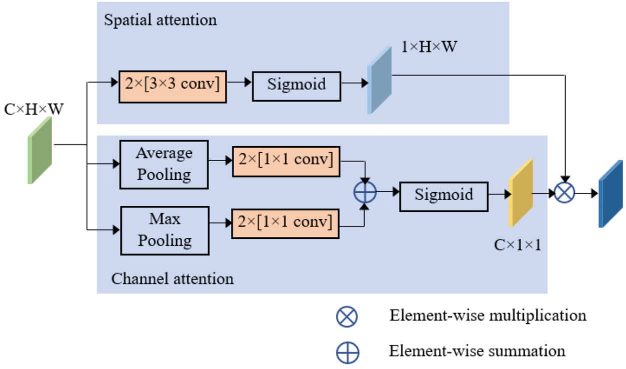


Figure 3. The detailed architecture of the spatial and channel attention

2.4. Loss function

Most recent semantic segmentation approaches employ cross-entropy loss during the training process. This loss function primarily emphasizes the accuracy of pixel classification, assigning equal weight to pixels across different regions. However, there exist huge differences in the number of pixels between the segmentation targets and the background. Therefore, we employ the dice loss [30] along with cross-entropy loss for optimization. The total loss is illustrated as (2):

$$L_{total} = L_{ce} + L_{dice} \tag{2}$$

where

$$L_{ce}(y, \hat{y}) = -\sum_{i=1}^{N} y_i log(\hat{y}_i) \tag{3}$$

and

$$L_{dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i{}^2 + \sum_{i=1}^{N} \hat{y}_i{}^2} \tag{4}$$

Here, L_{ce} is the categorical cross-entropy loss, L_{dice} is the dice loss, y_i refers to the predicted probability value associated with the i_{th} pixel, $\hat{y}_i$ refers to the ground truth value of the i_{th} pixel, and N refers to the overall pixel count of the image. The dice loss is capable of addressing the issue of class imbalance in input data. Therefore, the combination loss proves advantageous for segmenting a smaller foreground against a larger background, while simultaneously promoting smooth training by employing binary cross-entropy loss.

3. RESULTS AND DISCUSSION

This section illustrates the experimental results of our proposed road and vehicle segmentation method from UAV imagery and compare its performance with three well-established methods: U-Net [9], DeepLabv^{3+} [31], and SegNet [11]. We evaluate the segmentation accuracy of each method using common evaluation metrics, including F1-score, intersection over union (IoU), and category mean pixel accuracy (MPA). The experiments were conducted on the UAVid semantic segmentation dataset [32], which is an openly accessible dataset comprising high-resolution UAV images of various urban and suburban scenes.

3.1. Dataset and evaluation metrics

We employ the UAVid semantic segmentation dataset to evaluate the proposed model. UAVid is a semantic segmentation dataset designed specifically for aerial imagery captured by UAVs or drones. The dataset focuses on urban scenes and aims to facilitate the development of deep learning models for aerial image understanding. The dataset comprises high-resolution aerial images captured at different altitudes, covering diverse urban scenarios. The images in the dataset have a high resolution of 3840×2160 pixels, which allows for the detailed analysis of urban scenes and the extraction of fine-grained features. Every image in the dataset is annotated with pixel-level semantic labels, offering an extensive collection of ground truth data for training and evaluating semantic segmentation models. The dataset contains multiple object classes, including buildings, roads, trees, vehicles, and pedestrians. Given the large size of the training set images, we initially extract 10,000 non-overlapping small patches of size 512×512 from the training set. Subsequently, we employ 8,000 image patches from this set to train the proposed model. For the task of road and vehicles extraction, we only employ labels of three categories for each pixel, including: road, car, and background.

To assess the performance of semantic segmentation models on the UAVid dataset, we use several evaluation metrics, including F1-score, IoU, and category MPA. The F1-score is a balanced measure of segmentation performance, calculated as the harmonic mean of precision and recall. IoU quantifies the overlap between the predicted segmentation and the ground truth for an object class. MPA represents the percentage of accurately classified pixels in the image.

3.2. Implementation details

All methods were implemented using TensorFlow and trained on an NVIDIA RTX 4080 GPU. The models underwent training for 150 epochs using the Adam optimizer, with a learning rate of 0.0001 and a batch size of 16. Data augmentation techniques such as horizontal flipping, random cropping, and brightness adjustments were applied to prevent overfitting.

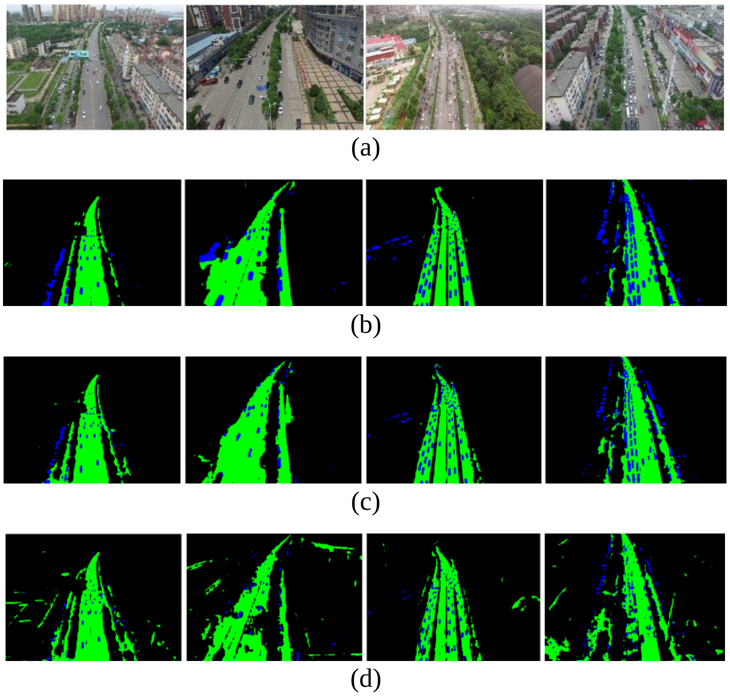
3.3. Experimental results

Table 1 presents the experimental results of our proposed method alongside the comparative methods. For road segmentation, the proposed method achieves an IoU of 84.91%, which is higher than U-Net (82.68%), DeepLabv^{3+} (81.99%), and SegNet (78.93%). This indicates that our method is more effective at delineating road boundaries and identifying connected road networks. The F1-score for roads also demonstrates the superior performance of our method, with a score of 90.46% compared to 90.06% for U-Net, 90.10% for DeepLabv^{3+}, and 88.22% for SegNet. The higher F1-score implies that our method has a better balance between precision and recall in road segmentation tasks. In terms of vehicle segmentation, the proposed method achieves an IoU of 79.66%, outperforming U-Net (74.31%), DeepLabv^{3+} (77.08%), and SegNet (66.37%). This suggests that our method is more capable of accurately identifying individual vehicles, even in congested scenes or when vehicles are partially occluded. The F1-score for vehicles further supports this observation, with our method achieving a score of 87.10%, which is higher than U-Net (85.06%), DeepLabv^{3+} (87.06%), and SegNet (79.78%). This demonstrates that our proposed method maintains a better trade-off between precision and recall for vehicle segmentation tasks. The mean pixel accuracy of our proposed method is 93.89%, which is notably higher than U-Net (90.35%), DeepLabv^{3+} (93.23%), and SegNet (87.58%). This metric quantifies the proportion of correctly classified pixels in relation to the total number of pixels, and the higher value for our method indicates that it is more effective at assigning the correct class labels to individual pixels. This superior pixel-wise classification performance contributes to the overall improved segmentation results for both roads and vehicles. The detailed analysis of the quantitative results demonstrates that our proposed method outperforms U-Net, DeepLabv^{3+}, and SegNet across all evaluation metrics. This superior performance can be attributed to the method's ability to capture both local and global contextual information, as well as the inclusion of an attention mechanism to enhance the model's capacity to focus on relevant regions. Consequently, our proposed method is better suited for road and vehicle segmentation tasks in high-resolution UAV imagery.

Table 1. Segmentation results on the UAVid semantic segmentation dataset

Method	F1-score (roads)	F1-score (vehicles)	IoU (roads)	IoU (vehicles)	MPA
U-Net [9]	90.06	85.06	82.68	74.31	90.35
DeepLabv^{3+} [31]	90.10	87.06	81.99	77.08	93.23
SegNet [11]	88.22	79.78	78.93	66.37	87.58
Our method	90.46	87.10	84.91	79.66	93.89

Figure 4 shows visual segmentation results of our proposed method and the comparative methods. Visual inspection of the segmentation results from input images (Figure 4(a)) further demonstrates the effectiveness of our proposed model, as illustrated in Figure 4(b). In comparison to DeepLabv^{3+}(Figure 4(c)) and U-Net (Figure 4(d)), the proposed model generates more accurate and coherent segmentations of roads and vehicles. The comparative methods tend to suffer from issues such as over-segmentation and misclassification, particularly in complex scenes with occlusions and shadows. In contrast, our method can better handle such challenges and produce more accurate and robust segmentation results.



(a)

(b)

(c)

(d)

Figure 4. Segmentation results with; (a) input image, (b) segmentation output of our model, (c) DeepLabv3+, and (d) U-Net

4. CONCLUSION

In this study, we proposed a novel deep learning methodology aimed at overcoming the complexities involved in the segmentation of roads and vehicles within high-resolution UAV imagery. Our strategy capitalized on the advantages offered by U-Net architectures, integrating them with an attention aggregation module designed to incorporate a diverse spectrum of deep and shallow features in the decoder section. Our method's unique dual attention mechanism and effective multi-layer feature aggregation scheme allowed for the successful capture of both local and global contextual data, thus enabling more accurate, consistent segmentation of roads and vehicles in intricate urban landscapes. This not only advances our understanding of the complex features in urban UAV imagery but also paves the way for further advancements in high-resolution image segmentation technologies. Comparatively, our approach significantly outperformed current leading methods such as U-Net, DeepLabv^{3+}, and SegNet in relation to segmentation accuracy and efficiency. This superior performance underscores the potential of our method as a more effective solution for urban scene analysis and interpretation, particularly in applications where precision and speed are paramount. The findings of this study have profound implications. By enabling more accurate segmentation, they improve our understanding of urban environments as seen from UAVs, which can be invaluable in fields such as urban planning, traffic management, and environmental monitoring. Looking ahead, we plan to extend our segmentation technique to include other object classes commonly seen in UAV imagery, such as buildings, pedestrians, and vegetation. This expansion aims to provide a more holistic understanding of urban landscapes, significantly broadening the potential applications of our method. Future research will further investigate the implications of these findings, shaping a new direction for the application of deep learning in image segmentation and urban analysis.

REFERENCES

[1] J. Shahmoradi, E. Talebi, P. Roghanchi, and M. Hassanalian, “A Comprehensive Review of Applications of Drone Technology in the Mining Industry,” *Drones*, vol. 4, no. 3, p. 34, Jul. 2020, doi: 10.3390/drones4030034.

[2] F. Al-Turjman, H. Zahmatkesh, I. Al-Oqily, and R. Daboul, “Optimized Unmanned Aerial Vehicles Deployment for Static and Mobile Targets’ Monitoring,” *Computer Communications*, vol. 149, pp. 27–35, Jan. 2020, doi: 10.1016/j.comcom.2019.10.001.

[3] M. Hegarty-Craver *et al.*, “Remote Crop Mapping at Scale: Using Satellite Imagery and UAV-Acquired Data as Ground Truth,” *Remote Sensing*, vol. 12, no. 12, p. 1984, Jun. 2020, doi: 10.3390/rs12121984.

[4] T. Chowdhury, M. Rahnemoonfar, R. Murphy, and O. Fernandes, “Comprehensive Semantic Segmentation on High Resolution UAV Imagery for Natural Disaster Damage Assessment,” in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2020, pp. 3904–3913, doi: 10.1109/BigData50022.2020.9377916.

[5] S. A. Ahmed, H. Desa, and A.-S. T. Hussain, “Classification of semantic segmentation using fully convolutional networks based unmanned aerial vehicle application,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 2, p. 641, Jun. 2023, doi: 10.11591/ijai.v12.i2.pp641-647.

[6] W. N. S. Rahimi, M. A. H., and M. S. A. M. Ali, “Ananas comosus crown image thresholding and crop counting using a colour space transformation scheme,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 5, p. 2472, Oct. 2020, doi: 10.12928/telkomnika.v18i5.13895.

[7] M. F. E. Purnomo, V. Kusumasari, E. Supriana, R. Ambarwati, and A. Kitagawa, “Development of triangular array eight patches antennas for circularly-polarized synthetic aperture radar sensor,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, p. 631, Apr. 2020, doi: 10.12928/telkomnika.v18i2.14759.

[8] E. Shelhamer, J. Long, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, Apr. 2017, doi: 10.1109/TPAMI.2016.2572683.

[9] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[10] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *Computer Vision – ECCV 2014*, 2014, pp. 818–833, doi: 10.1007/978-3-319-10590-1_53.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[12] G. Y. Abbass and A. F. Marhoon, “Car license plate segmentation and recognition system based on deep learning,” *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 4, pp. 1983–1989, Aug. 2022, doi: 10.11591/eei.v11i4.3434.

[13] O. A. Boraik, M. Ravikumar, and M. A. N. Saif, “Characters Segmentation from Arabic Handwritten Document Images: Hybrid Approach,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, 2022, doi: 10.14569/IJACSA.2022.0130447.

[14] Marcellino, T. W. Cenggoro, and B. Pardamean, “UNet^{++} with Scale Pyramid for Crowd Counting,” *ICIC Express Letters*, 2022.

[15] N. A. Sehree and A. M. Khidhir, “Olive trees cases classification based on deep convolutional neural network from unmanned aerial vehicle imagery,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 1, pp. 92–101, Jul. 2022, doi: 10.11591/ijeecs.v27.i1.pp92-101.

[16] P. Vasavi, A. Punitha, and T. V. N. Rao, “Crop leaf disease detection and classification using machine learning and deep learning algorithms by visual symptoms: a review,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, pp. 2079–2086, Apr. 2022, doi: 10.11591/ijece.v12i2.pp2079-2086.

[17] N. S. Ibrahim, S. M. Sharun, M. K. Osman, S. B. Mohamed, and S. H. Y. S. Abdullah, “The application of UAV images in flood detection using image segmentation techniques,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 2, pp. 1219–1226, Aug. 2021, doi: 10.11591/ijeecs.v23.i2.pp1219-1226.

[18] Z. Zhang, Q. Liu, and Y. Wang, “Road Extraction by Deep Residual U-Net,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, May 2018, doi: 10.1109/LGRS.2018.2802944.

[19] J. Wan, Z. Xie, Y. Xu, S. Chen, and Q. Qiu, “DA-RoadNet: A Dual-Attention Network for Road Extraction From High Resolution Satellite Imagery,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 6302–6315, 2021, doi: 10.1109/JSTARS.2021.3083055.

[20] Y. Ren, Y. Yu, and H. Guan, “DA-CapsUNet: A Dual-Attention Capsule U-Net for Road Extraction from Remote Sensing Imagery,” *Remote Sensing*, vol. 12, no. 18, p. 2866, Sep. 2020, doi: 10.3390/rs12182866.

[21] R. Kestur, S. Farooq, R. Abdal, E. Mehraj, O. Narasipura, and M. Mudigere, “UFCN: a fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle,” *Journal of Applied Remote Sensing*, vol. 12, no. 01, p. 1, Feb. 2018, doi: 10.1117/1.JRS.12.016020.

[22] N. Varia, A. Dokania, and J. Senthilnath, “DeepExt: A Convolution Neural Network for Road Extraction using RGB images captured by UAV,” in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, Nov. 2018, pp. 1890–1895, doi: 10.1109/SSCI.2018.8628717.

[23] Y. Qian, J. M. Dolan, and M. Yang, “DLT-Net: Joint Detection of Drivable Areas, Lane Lines, and Traffic Objects,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4670–4679, Nov. 2020, doi: 10.1109/TITS.2019.2943777.

[24] L. C. Lo Bianco, J. Beltrán, G. F. López, F. García, and A. Al-Kaff, “Joint semantic segmentation of road objects and lanes using Convolutional Neural Networks,” *Robotics and Autonomous Systems*, vol. 133, p. 103623, Nov. 2020, doi: 10.1016/j.robot.2020.103623.

[25] M. Teichmann, M. Weber, M. Zollner, R. Cipolla, and R. Urtasun, “MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving,” in *2018 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, Jun. 2018, pp. 1013–1020, doi: 10.1109/IVS.2018.8500504.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[27] Y. Liu *et al.*, “Richer Convolutional Features for Edge Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1939–1946, Aug. 2019, doi: 10.1109/TPAMI.2018.2878849.

[28] L. Chen *et al.*, “SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 6298–6306, doi: 10.1109/CVPR.2017.667.

[29] S. Woo, J. Park, J. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” *Eccv*, p. 17, 2018, doi: 10.1007/978-3-030-01234-2_1.

[30] S. A. Taghanaki *et al.*, “Combo loss: Handling input and output imbalance in multi-organ segmentation,” *Computerized Medical Imaging and Graphics*, vol. 75, pp. 24–33, Jul. 2019, doi: 10.1016/j.compmedimag.2019.04.005.

[31] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation,” in *Computer Vision – ECCV*, 2018, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.

[32] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, “UAVid: A semantic segmentation dataset for UAV imagery,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 165, pp. 108–119, Jul. 2020, doi: 10.1016/j.isprsjprs.2020.05.009.

BIOGRAPHIES OF AUTHORS

Trung Dung Nguyen was born in 1976 in Vietnam. He received a Master's degree in Electrical Engineering from Ho Chi Minh City University of Technology and Education, Vietnam, in 2017. He is currently a lecturer at the Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam. His research focuses on applications of metaheuristic algorithms in power system optimization, optimal control, and model predictive control. He can be contacted at email: nguyentrungdung@iuh.edu.vn.

Trung Kien Pham was born in 1976 in Vietnam. Between 1993 and 1998, he pursued studies in Electrical and Electronic Engineering Technology and graduated from The Ho Chi Minh City University of Technology. After that, he enrolled at the Ho Chi Minh City University of Technology from 2000 to 2003 and received a master's degree in Power Systems. He worked as an electrical engineer at Phu My Power Plant from 1998 to 2000. From 2002 to 2012, he held the position of Vice Dean for the Faculty of Electrical Technology at The Industrial University of Ho Chi Minh City. He served as the Dean of the Faculty of Electrical Technology at The Industrial University of Ho Chi Minh City from 2012 to 2020. From 2020 to the present, he has been the Head of the Office of Human Resources and Administration at The Industrial University of Ho Chi Minh City. His research focuses on applications of control algorithms in power system optimization, optimal control, and model predictive control. He can be contacted at email: phamtrungkien@iuh.edu.vn.

Chi Kien Ha received M.Sc. degrees from the Ho Chi Minh City University of Technology, Vietnam, in 2013. Currently, he is a lecturer with the Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Vietnam. His main research interests include automated guided vehicles (AGV) and automatic systems. He can be contacted at email: hachikien@iuh.edu.vn.

Long Ho Le received M.Sc. degrees from the Ho Chi Minh City University of Technology, Vietnam, in 2013. Currently, he is a lecturer with the Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Vietnam. His main research interests include AGV and automatic systems. He can be contacted at email: lelongho@iuh.edu.vn.

Thanh Quyen Ngo received a B.S. degree in electrical engineering from HCMC University and an M.S. degree in automatic control from HCMC University of Transport. He obtained his Ph.D. in Control Science and Engineering from Hunan University, China. Currently, he is a lecturer in the Department of Electrical Engineering at the Industrial University of Ho Chi Minh City, Vietnam. His key research interests include fuzzy logic, neural networks (NNs), cerebellar model articulation controllers (CMACs), adaptive control, and intelligent control. He can be contacted at email: ngothanhquyen@iuh.edu.vn.

Hoanh Nguyen received a Bachelor's and Master's degree in Automation at Ho Chi Minh City University of Technology in 2010 and 2015, respectively. Currently, he is pursuing a doctoral program at Ho Chi Minh City University of Technology. He is also a lecturer at the Department of Electrical Engineering at the Industrial University of Ho Chi Minh City, Vietnam. His research focuses on deep learning, intelligent control, computer vision, and smart transportation. He can be contacted at email: nguyenhoanh@iuh.edu.vn.