Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 18, No. 3, 2025

LMANet-Efficient Multi-Scale Attention Aggregation Network for Semantic Segmentation of UHR UAV Imagery

Dan Wang, Jianying Shen, Yiwen Li

College of Information Engineering, Xizang Minzu University, Xianyang Shaanxi, 712082, China

Abstract: Unmanned Aerial Vehicle (UAV) semantic segmentation is vital for applications like urban management and agricultural monitoring. However, segmenting ultra-high-resolution (UHR) UAV imagery, characterized by dense small objects and complex textures, presents a significant challenge in balancing accuracy, speed, and model size. To address the limitations of existing methods in small-object segmentation and computational efficiency, this paper proposes a Lightweight Multi-scale Attention Network (LMANet). Our framework integrates a Multi-Scale Window Attention Aggregation (MS-WAA) encoder, a Depthwise Separable Hybrid Dilated Convolution (DHC) module, and a lightweight decoder (LADecoder) to achieve efficient and precise feature extraction and reconstruction. Extensive experiments on the Vaihingen and Potsdam datasets demonstrate that LMANet achieves superior segmentation performance, particularly for small objects, while maintaining significantly fewer parameters and higher inference speed compared to state-of-the-art models.

Keywords: UAV Remote Sensing Imagery; Semantic Segmentation; Lightweight Network; Multi-Scale Attention; Hybrid Dilated Convolution; Dense Upsampling.

1. Introduction

1.1. Background

In recent years, with the rapid development of UAV systems, their onboard visual perception capabilities have become a crucial component for task execution. With advances in computing, navigation, and sensing technologies, modern UAVs have evolved from early models that could only fly along fixed routes to advanced systems capable of autonomous perception, intelligent decision-making, and multi-task collaboration.

Semantic segmentation from the UAV perspective places greater emphasis on accurately identifying small targets, complex textures, and the geometric shapes of objects. UAV-based semantic segmentation also needs to cope with a range of interferences that may arise in low-altitude flight environments, such as wind, lighting variations, and occlusions. Research on UAV semantic segmentation technology is not only crucial for enhancing the autonomous operation capabilities of UAVs but also holds significant value for applications such as emergency response, urban management, and environmental monitoring.

UAV remote sensing imagery features extremely high resolution and a wide field of view, which gives rise to a unique scale imbalance problem: within the same scene, it may contain buildings occupying thousands of pixels as well as vehicles and pedestrians occupying only tens of pixels. Traditional convolutional neural networks (CNNs)[13], limited by their fixed receptive fields, struggle to simultaneously capture long-range contextual information and the geometric details of tiny objects. Although attention-based methods such as the Swin Transformer excel at global modeling, their computational complexity makes them difficult to deploy on energy-constrained airborne edge devices. Most current deep learning approaches for semantic segmentation pursue higher accuracy through complex and computationally intensive network architectures, which makes it challenging to meet real-time requirements. Conversely, models optimized for real-time performance often suffer from insufficient accuracy. Moreover, datasets related to low-altitude UAV vision are limited, and existing research largely focuses on simple object detection or coarse-grained segmentation. There is a scarcity of studies on real-time semantic segmentation[14] specifically for UAV scenarios, as well as a lack of systematic analysis of algorithm inference speed and deployment capability.

In practical UAV applications, algorithms are required to achieve a balance between high accuracy and real-time performance on onboard or edge devices. Developing a semantic segmentation model that delivers both high segmentation precision and real-time operation under the complex and dynamic perspectives of UAVs is therefore a critical problem that urgently needs to be addressed.

1.2. Related Work

As a core task in computer vision, semantic segmentation has rapidly evolved from traditional CNNs to Transformers. Fully Convolutional Networks (FCNs)[7] were the first to propose an end-to-end framework for pixel-level prediction, replacing fully connected layers with convolutional layers. This design enables the network to handle images of arbitrary sizes and produce semantic segmentation results of corresponding dimensions, laying the foundation for modern semantic segmentation methods. Subsequently, the DeepLab series[4][5][6] proposed by Google Brain introduced key techniques such as atrous (dilated) convolution, multi-scale pooling, and conditional random fields (CRFs). These innovations effectively expanded the receptive field and improved boundary detail quality, addressing the issues of low resolution and coarse edges commonly found in traditional CNNs.

To further enhance the representation of both global and local features, PSPNet[8] employs a pyramid pooling module to fuse contextual information at multiple scales,

demonstrating outstanding performance in scene parsing tasks. Mask R-CNN[9] combines object detection with instance segmentation by generating candidate regions through a region proposal network and classifying pixels within these regions, achieving precise segmentation results.

With the growing demand for real-time, high-accuracy semantic segmentation in fields such as autonomous driving and UAV inspection, model lightweighting and inference speed have become focal points. ENet[10], one of the early real-time segmentation networks, reduces both the number of parameters and computational cost, achieving very fast inference. However, its accuracy is relatively low, making it insufficient for handling complex scenes. To balance speed and accuracy, ICNet[11] proposed a multi-resolution cascade architecture that leverages the semantic information from low-resolution branches alongside the detail information from high-resolution branches. This design achieves high accuracy with relatively low computational cost, reaching 30 FPS on 1024×2048 inputs, making it a representative example of an efficient semantic segmentation structure.

However, ultra-high-resolution (UHR) UAV remote sensing imagery exhibits characteristics that differ from those in autonomous driving: dense small targets, large scale variations, rich texture details, and complex background changes, with interference factors such as trees, shadows, and clouds being more pronounced. As a result, traditional real-time segmentation networks struggle to meet the high-precision requirements for small targets and complex boundaries, while large models like DeepLab and PSPNet[14] suffer from high computational costs, making them difficult to deploy on embedded platforms.

In summary, existing semantic segmentation methods still face the following challenges when dealing with UHR UAV imagery:

(1) Small objects are easily overlooked or segmented incompletely;

(2) Edge details are inadequately recovered, resulting in blurriness;

(3) The model's computational cost is too high, making it unsuitable for real-time processing on onboard or mobile devices.

Based on the above limitations, it is necessary to propose a novel network architecture that combines lightweight design, multi-scale modeling capability, and edge detail reconstruction.

2. Methodology

The LMANet method studied in this work mainly consists of three modules: the MS-WAA encoder, the DHC atrous convolution context module, and the lightweight LADecoder decoder.The lightweight multi-scale window attention encoder, MS-WAA Encoder, captures multi-scale features using different window sizes. Its context module, based on Depthwise Separable Hybrid Atrous Convolution (DHC), addresses the gridding effect while providing a large receptive field. The lightweight multi-level feature aggregation decoder, LADecoder, restores relevant details through DUC upsampling.

2.1. Overall Model Architecture

The model consists of three parts: the encoder, the context module, and the decoder.

The MS-WAA is responsible for multi-scale feature extraction, the DHC provides global semantic enhancement, and finally, the decoder restores spatial structure and edge details. The overall architecture of LMANet is shown in Figure 1, adopting a typical Encoder-Decoder structure.

The input image is first processed to extract basic features. The multi-scale weighted attention (MS-WAA) module then captures multi-scale feature information while aggregating local and global context. The DHC module acts as a bridge between the encoder and decoder, integrating features from different layers to enhance feature representation. Through DUC and ECA, low-resolution features are gradually restored to the input image resolution, enhancing important feature channels and improving segmentation accuracy. The decoder progressively recovers spatial information via upsampling, leveraging the encoder's multi-scale features to assist fine-grained segmentation, and ultimately outputs a segmentation mask of the same size as the input image to label different regions.

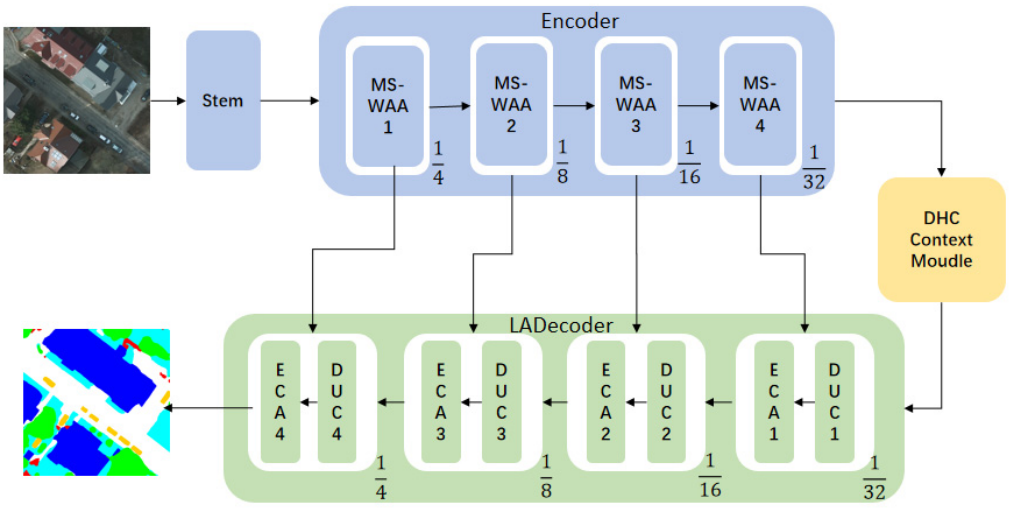


Figure 1. Overall Architecture of the LMANet Model

2.2. Basic Structure: Swin Transformer

The Swin Transformer primarily extracts multi-level feature information from UAV imagery[1]. It employs a hierarchical structure, using Patch Partition (image splitting)[16] and Patch Merging (image block merging)[17] operations to gradually reduce image resolution while

increasing the number of channels. It ultimately outputs four feature maps at different scales and depths, containing both shallow-layer detailed texture information and deep-layer semantic features.

To address the high computational cost of global self-attention in traditional Transformers, the Swin Transformer restricts attention computation to fixed local windows. This reduces computational complexity, enabling more efficient processing of high-resolution UAV imagery. Since fixed windows lack inter-window information exchange, the module introduces the Shifted Window operation. By cyclically shifting the windows[18], previously non-overlapping windows interact in the next layer. This mechanism preserves the efficiency of local computation while enabling global information flow across windows, enhancing the completeness of feature extraction.

2.3. Multi-Scale Window Attention Aggregation (MS-WAA)

MS-WAA addresses challenges in UAV remote sensing imagery, such as the difficulty of segmenting small objects and blurry boundary details. It improves the computational precision of the window self-attention mechanism in the Swin Transformer and overcomes issues caused by fixed windows, including the local effect (objects being split) and boundary effect (lack of contextual information). This module adopts a three-branch design, computing self-attention (W-MSA) within three different window sizes simultaneously, as shown in Figure 2.

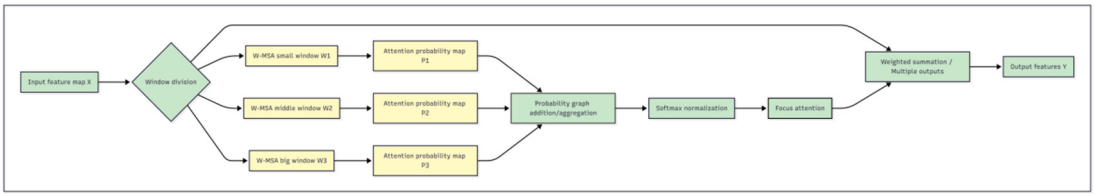


Figure 2. MS-WAA Multi-Scale Window Attention Module

MS-WAA uses the self-attention similarity probability map[19] of the smallest window as a reference, adding the similarity maps of the other two larger windows and then normalizing them via softmax. By aggregating information across multiple window scales, it captures richer contextual information, effectively mitigates the window boundary effect, and reduces the impact of individual objects being split. This makes the module's attention values more precise, enhancing the representation of small objects and object boundaries in UAV imagery. It computes attention simultaneously using three different windows (s_1, s_2, s_3), capturing multi-scale context: the small window emphasizes edge textures, while the larger windows compensate for information lost in local windows.

$$WAA(\mathrm{x}) = soft\max(W_{s1}(x) + W_{s2}(x) + W_{s3}(x)) \tag{1}$$

2.4. DHC Hybrid Atrous Convolution Context Module

Although traditional dilated convolution can expand the receptive field, large dilation rates often cause the gridding effect, leading to discontinuous information and reduced accuracy. The DHC module[2] introduces a combination of coprime dilation rates, ensuring spatially dense coverage of sampling points. This expands the effective receptive field while maintaining feature map continuity, all with minimal parameter overhead. It replaces the original ASPP module with Depthwise Separable Hybrid Dilated Convolution (DHC), decomposing standard convolutions to reduce computation and parameters while maintaining performance. The Hybrid Dilated Convolution (HDC) uses a non-uniform combination of dilation rates to ensure that sampled pixels are continuous, mitigating or even eliminating the gridding effect. Each DHC branch output is equipped with a learnable adaptive coefficient, allowing the network to automatically adjust and weight the importance of different branches, thereby better fusing multi-scale information.

DHC achieves lightweight design by combining depthwise separable convolutions, eliminates the gridding effect using non-uniform hybrid dilation rates, and fuses multi-scale features through adaptive coefficients. This allows more efficient and accurate extraction of multi-scale contextual information from UAV remote sensing images, improving the segmentation of object edges and fine details.

$$DHC(x) = \sum_{i=1}^{3} w_i \cdot DWConv(x; r_i) \tag{2}$$

2.5. LADecoder: Lightweight Multi-Level Decoder

To address the challenges of dense small objects and complex edge details in high-resolution UAV imagery, many researchers have designed lightweight decoders[3]. In this work, we propose the LADecoder, a multi-level lightweight decoder that achieves high-quality spatial detail recovery while maintaining a lightweight design. LADecoder addresses this issue by combining multi-level upsampling with lightweight attention and multi-scale fusion through skip connections, effectively restoring small object boundaries.

The decoder uses Dense Upsampling Convolution (DUC) instead of simple bilinear interpolation. By employing sub-pixel convolution, it efficiently restores spatial resolution and reduces detail loss caused by interpolation.

$$F_{\mathrm{duc}} = PixelShuffle(Conv(F_h)) \tag{3}$$

Here, F_h represents the high-level features output by the encoder,$F_h \in R^{H\times W\times C}$, $F_{up} \in R^{(rH)\times(rW)\times C}$.

$$s = \sigma(Conv_1(GAP(F_{up}))) \tag{4}$$

In this work, a lightweight ECA [12](Efficient Channel Attention) mechanism is applied after the DUC upsampling module. ECA can adaptively adjust the response strength of different channels, enabling the decoder to focus more on salient target regions when restoring spatial details, thereby

effectively improving the segmentation accuracy of small objects and boundary regions. At each upsampling stage, the lightweight attention module enhances key information channels, increasing the model's focus on texture boundaries and small-object areas. Here, $GAP(\cdot)$ denotes global average pooling, $Conv_1$ is a one-dimensional convolution, and σ represents the sigmoid function.

$$F_{att} = s \square F_{up} \tag{5}$$

$$F_{cat} = [F_{aat}, F_{low}] \tag{6}$$

Finally, multi-level skip connections[20][21] are used to fuse the encoder's low-level features with the upsampled features, allowing the model to combine high-level semantics with low-level details and achieve progressive, layer-by-layer refinement of semantic reconstruction.

$$F^k = \Phi(F^{k+1}, F^k_{low}) \tag{7}$$

$$F_{fuse} = Conv(F_{cat}) \tag{8}$$

Here, F^k denotes the features at the k-th decoding layer. The decoder operates in three stages.

$$\Phi(F^{k+1}, F^k_{low}) = Conv([Att(DUC(F^{k+1})), F^k_{low}]) \tag{9}$$

$$P = Soft\max(Conv(F^0)) \tag{10}$$

Here, F^0 represents the features at the final stage of the decoder.

Experiments show that the LADecoder in this model significantly improves the MIoU of small object classes and effectively alleviates edge blurriness.

2.6. Loss Function

This work employs a combined loss function integrating cross-entropy loss[23] and Dice loss[22]. Cross-entropy ensures stable pixel-level classification and improves overall segmentation accuracy, while Dice loss optimizes region overlap, enhancing the prediction of small objects and irregular land cover areas. To further improve object contour and boundary sharpness, a lightweight boundary loss is introduced during the decoding stage, constraining the predicted edge regions and enabling better restoration of boundary details. The final total loss can be expressed as:

$$L = L_{CE} + L_{Dice} + \lambda_b L_{Boundary} \tag{11}$$

Here, λ_b represents the weight of the boundary loss. The combined loss maintains training stability while improving the segmentation performance of small objects and enhancing overall MIoU.

3. Experiments

3.1. Dataset

Vaihingen: The Vaihingen dataset, provided by ISPRS, consists of 33 real-world aerial images collected with advanced airborne sensors and is primarily used for urban classification. Each image contains infrared, red, and green channels, with 16 images accurately annotated into six classes: impervious surfaces, buildings, low vegetation, trees, cars, and background. It is one of the most widely used and classic datasets for semantic segmentation.

Postdam: The Potsdam dataset is a highly representative historical urban dataset, featuring numerous large buildings, narrow roads, and dense settlement structures. It contains 38 images, each with a resolution of 6000×6000 pixels and a spatial resolution of 5 cm. The dataset provides multiple data types, including infrared, red, green, and blue channels, as well as digital surface models (DSM) and normalized digital surface models (nDSM). It covers the most common land cover classes: roads, buildings, grass, trees, vehicles, and background. Like Vaihingen, it is an official ISPRS standard dataset.

3.2. Experimental Environment

This study was implemented on an NVIDIA RTX 3090 (24 GB) platform using PyTorch 2.0. During training, the batch size was set to 8, and the model was trained for 200 epochs. The optimizer used was AdamW, with an initial learning rate of 1×10^{-4}, combined with a Poly learning rate decay strategy (power = 0.9), and a weight decay of 0.01.

4. Experimental Results

The proposed model was evaluated on the Vaihingen and Potsdam datasets. Experimental results show that LMANet achieves optimal performance while maintaining a very low number of parameters.

4.1. Quantitative Results

Table 1. Segmentation accuracy of different methods on the Vaihingen dataset

Model	Backbone	Surface	Building	Vegetation	Tree	Car	MF1	MIoU
LMANet (Ours)	**MobileNetV3**	**92.6**	**96.1**	**83.8**	**90.9**	**89.2**	**90.5**	**84.1**
Swin Transformer	Swin-T	94.1	96.2	85.8	90.4	87.6	90.7	83.2
DeepLabv3+	ResNet50	91.8	95.4	82.1	89.1	87.3	89.1	82.4
UNeTFormer	ConvNeXt-T	93.8	95.5	83.7	90.1	88.5	90.2	83.6
SwinUMamba	Swin-T	92.3	96.8	85.3	90.5	88.9	88.9	84.0
MANet	ResNet50	91.4	93.6	82.4	88.5	88.7	90.4	82.7
MambaUNet	ConvNeXt-S	92.0	96.7	85.2	90.3	89.0	90.3	83.9

On the Vaihingen dataset, LMANet outperforms the comparison models overall, achieving higher scores in major categories such as Surface, Building, and Vegetation. For the Car category, SwinUMamba[26] achieved an MF1 score of 88.9%, MambaUNet[25] 89.0%, and UNetFormer[24] 88.5%, while LMANet reached 89.2%, surpassing all other comparison models. LMANet achieves an MF1 score of 90.5% and an mIoU of 84.1%, both higher than those of the other models. Some experimental results are shown in Figure 3. It can be observed that the segmentation performance of our

model surpasses that of the comparison models, particularly in the areas highlighted with black boxes, where vehicles and buildings are clearly segmented, closely matching the ground truth.

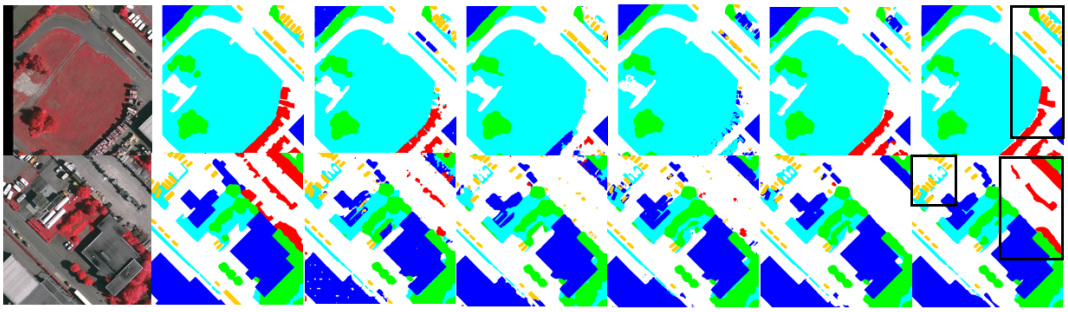


Figure 3. Visual comparison of semantic segmentation for small object features on the Vaihingen datataset. (a) Image, (b) label, (c) DeepLabV3+, (d) mambaunet, (e) swinumamba, (f) unetformer, (g) my

Table 2. Segmentation accuracy of different methods on the postdam dataset

Model	Backbone	Surface	Building	Vegetation	Tree	Car	MF1	MIoU
LMANet(Ours)	**MobileNetV3**	**92.9**	**97.4**	**85.6**	**91.7**	**96.1**	**92.8**	**85.3**
Swin Transformer	Swin-T	90.8	97.1	86.3	90.6	95.4	92.0	85.1
DeepLabv3+	ResNet50	92.0	96.8	84.1	91.8	95.4	92.5	83.9
UNeTFormer	ConvNeXt-T	92.5	97.4	85.3	91.0	95.8	92.6	85.7
SwinUMamba	Swin-T	92.2	97.5	86.7	91.3	96.0	92.6	86.1
MANet	ResNet50	91.7	96.3	84.3	90.1	94.6	91.4	83.2
MambaUNet	ConvNeXt-S	93.9	97.5	85.7	90.2	95.5	91.9	85.9

As shown in Table 2, LMANet also outperforms other comparison models overall on the Potsdam dataset, achieving higher metrics in the Surface, Building, Vegetation, and Car categories. In terms of overall performance, LMANet attains an MF1 of 92.8% and an mIoU of 85.3%, surpassing most of the other models. Some experimental results are shown in Figure 4. It can be seen that our model's segmentation performance surpasses that of the comparison models, particularly in the areas highlighted with black boxes, where objects in the lower-left corner and the orange object in the center of the image are accurately segmented, closely matching the ground truth.

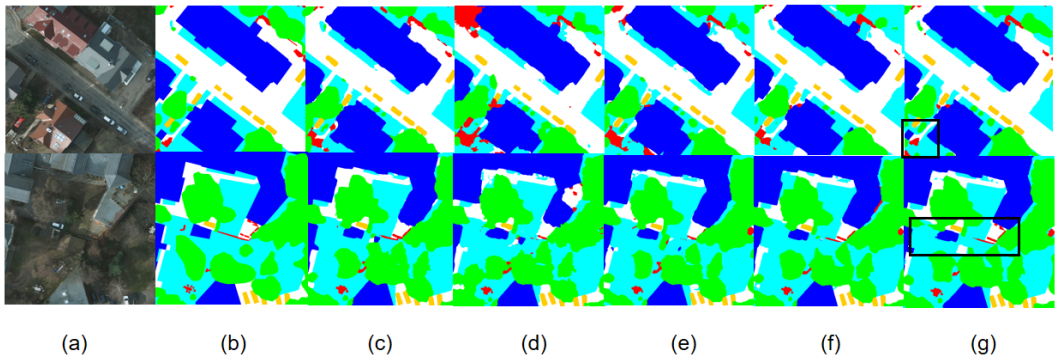


Figure 4. Visual comparison of semantic segmentation for small object features on the Postdam datataset. (a) Image, (b) label, (c)DeepLabV3+, (d) mambaunet, (e) swinumamba, (f) unetformer, (g) my

Table 3. Accuracy–Efficiency comparison on Vaihingen and Postdam

Model	MIoU (Vaihingen)	MIoU (Postdam)	Params (M)	FLOPs (G)	FPS
LMANet (Ours)	**84.1**	**85.3**	**8.5**	**45**	**20**
Swin Transformer	83.2	85.1	59.0	210	6
MANet	82.7	83.2	42.3	160	7
UNeTFormer	83.6	85.7	32.0	120	9
MambaUNet	83.9	85.9	26.0	95	12
SwinUMamba	84.0	86.1	60.0	210	6
DeepLabv3+	82.4	83.9	42.0	150	8

Among the land cover classes, vehicles represent typical small objects: they occupy a small area, are sparsely distributed, and are easily affected by shadows, tree canopies, and other occlusions. This places higher demands on the model's ability to capture fine-grained features.The experimental results show that the proposed LMANet

performs particularly well in the Car category, achieving an MF1 of 89.2% on the Vaihingen dataset and 96.1% on the Potsdam dataset. In comparison, UNeTFormer, MambaUNet, and SwinUMamba also achieve relatively high MF1 scores for the Car class, but none manage to attain a comparable balance of performance under lightweight conditions like LMANet.

As shown in Table 3, despite significantly reducing model size and computational cost, with only 8.5M parameters, LMANet maintains segmentation accuracy at the level of mainstream models or even better. It achieves an mIoU of 84.1% on Vaihingen and 85.3% on Potsdam, comparable to or slightly surpassing larger models such as MambaUNet and SwinUMamba.

4.2. Ablation Study

Table 4. Ablation study results of LMANet

Method	MS-WAA	DHC	LADecoder	MF1	MIoU
Baseline (MobileNetV3 + Decoder)	×	×	×	88.7	81.2
Baseline + LADecoder	×	×	✓	89.2	81.8
Baseline + DHC	×	✓	×	89.6	82.7
Baseline + MS-WAA	✓	×	×	90.1	83.4
Baseline + MS-WAA + DHC	✓	✓	×	90.9	84.0
Baseline + MS-WAA + LADecoder	✓	×	✓	90.6	83.8
Baseline + DHC + LADecoder	×	✓	✓	90.0	84.6
Baseline + MS-WAA + DHC + LADecoder (LMANet)	✓	✓	✓	90.5	84.1

Ablation experiments, as shown in Table 4, demonstrate that each key module in LMANet contributes positively to performance improvement. MS-WAA and DHC are the primary sources of performance gains, while ECA provides lightweight enhancement. The three modules work synergistically, achieving the best overall results.

5. Conclusion

This work addresses key challenges in UAV ultra-high-resolution image semantic segmentation, such as dense small objects, significant scale variations, complex edge details, and high computational costs. We propose LMANet, a lightweight and deployable network model with excellent multi-scale modeling capability. LMANet achieves a balance between global semantic understanding and local detail reconstruction by incorporating an efficient multi-scale attention mechanism, lightweight convolutional structures, and a multi-level decoder, all while maintaining overall model accuracy and robustness.

The proposed MS-WAA multi-scale window attention encoder overcomes the limitations of fixed window sizes in the traditional Swin Transformer. By introducing parallel attention mechanisms across different window scales, the model can simultaneously capture fine-grained textures, structural boundaries, and long-range semantic dependencies, resulting in improved performance in small object recognition, complex road structure analysis, and shadow-occluded regions.

The DHC depthwise separable hybrid dilated convolution module effectively expands the receptive field while maintaining a lightweight design. By combining multiple dilation rates, it alleviates the gridding effect of traditional dilated convolutions and enhances semantic aggregation for large-scale objects. This module enables the model to achieve stronger global representation without significantly increasing computational cost.

The LADecoder multi-level lightweight decoder integrates dense upsampling convolution with a lightweight attention mechanism during the decoding stage. Through multi-step high-precision upsampling and multi-level skip connections, it progressively restores resolution and enhances spatial details, resulting in segmentation outputs with improved detail integrity and boundary accuracy.

Experimental results on the Vaihingen and Potsdam datasets demonstrate that, despite having a very low overall parameter count, LMANet still outperforms mainstream methods such as Swin Transformer and DeepLabv3+ in segmentation accuracy, particularly for small object classes like pedestrians and vehicles. This indicates that LMANet has strong practical potential for UAV intelligent perception and can be deployed on terminal devices or platforms with limited computational resources.

In the future, we plan to further explore dynamic window attention, adaptive dilation rate selection, real-time inference optimization, and integration with multi-modal data (e.g., LiDAR, infrared) to enhance the model's generalization and real-time performance in more complex UAV tasks.

Acknowledgments

I would like to acknowledge the financial support provided by the Scientific Research Innovation and Practice Project of Xizang Minzu University (Grant No. Y2025130), which facilitated the research conducted in this dissertation.

References

[1] Li, J., Yi, S., He, R., & Liu, Q. (2024). Semantic Segmentation Method of UAV Image Based on Swin Transformer with Window Attention Aggregation. Computer Engineering and Applications, 60(15), 198–210.

[2] Sun, G., Luo, X., & Zhang, K. (2024). DeepLabV3_DHC: Semantic Segmentation of Urban UAV Remote Sensing Images. Laser & Optoelectronics Progress, 61(04), 394–403.

[3] R. Guan, M. Wang, L. Bruzzone, H. Zhao and C. Yang, "Lightweight Attention Network for Very High-Resolution Image Semantic Segmentation," in IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-14, 2023, Art no. 4403514, doi: 10.1109/TGRS.2023.3272614.

[4] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834-848, 1 April 2018, doi: 10.1109/TPAMI.2017.2699184.

[5] Chen L C , Papandreou G, Kokkinos I, et al.Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J].Computer Science, 2014(4):357-361. DOI: 10. 1080/17476938708814211.

[6] Chen L C , Papandreou G , Schroff F ,et al.Rethinking Atrous Convolution for Semantic Image Segmentation[J]. 2017. DOI: 10. 48550/arXiv.1706.05587.

[7] Long J , Shelhamer E , Darrell T .Fully Convolutional Networks for Semantic Segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4):640-651.DOI:10.1109/CVPR.2015.7298965.

[8] Zhao H, Shi J, Qi X ,et al. Pyramid Scene Parsing Network [J]. IEEE Computer Society, 2016.DOI: 10.1109/ CVPR. 2017.660.

[9] He K, Gkioxari G, Piotr Dollár,et al.Mask R-CNN[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017.DOI:10.1109/TPAMI.2018.2844175.

[10] Paszke A, Chaurasia A , Kim S ,et al. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation [J]. 2016.DOI:10.48550/arXiv.1606.02147.

[11] Zhao H, Qi X, Shen X ,et al.ICNet for Real-Time Semantic Segmentation on High-Resolution Images[C]//Springer, Cham. Springer, Cham, 2018.DOI:10.1007/978-3-030-01219-9_25.

[12] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 11531-11539, doi:10.1109/CVPR42600.2020.01155.

[13] Chua LO. CNN: A vision of complexity. International Journal of Bifurcation and Chaos. 1997 Oct;7(10):2219-425.

[14] He, J., Chen, H., & Luo, D. (2023). Review of Real-Time Semantic Segmentation Algorithms Based on Deep Learning. Journal of Computer Engineering & Applications, 59(8).

[15] Yan L, Liu D, Xiang Q, Luo Y, Wang T, Wu D, Chen H, Zhang Y, Li Q. PSP net-based automatic segmentation network model for prostate magnetic resonance imaging. Computer Methods and Programs in Biomedicine. 2021 Aug 1; 207:106211.

[16] Campen M. Partitioning surfaces into quadrilateral patches: A survey. InComputer graphics forum 2017 Dec (Vol. 36, No. 8, pp. 567-588).

[17] Li J, Guan W. Patch merging refiner embedding UNet for image denoising. Information Sciences. 2023 Sep 1;641: 119 123.

[18] Zhai G, Zhang W, Yang X, Lin W, Xu Y. Efficient image deblocking based on postfiltering in shifted windows. IEEE Transactions on Circuits and Systems for Video Technology. 2008 Jan 28;18(1):122-6.

[19] Gao X, Zhang Z, Mu T, Zhang X, Cui C, Wang M. Self-attention driven adversarial similarity learning network. Pattern Recognition. 2020 Sep 1; 105:107331.

[20] Orhan AE, Pitkow X. Skip connections eliminate singularities. arXiv preprint arXiv:1701.09175. 2017 Jan 31.

[21] Drozdzal M, Vorontsov E, Chartrand G, Kadoury S, Pal C. The importance of skip connections in biomedical image segmentation. InInternational workshop on deep learning in medical image analysis 2016 Sep 27 (pp. 179-187). Cham: Springer International Publishing.

[22] Zhao R, Qian B, Zhang X, Li Y, Wei R, Liu Y, Pan Y. Rethinking dice loss for medical image segmentation. In2020 IEEE international conference on data mining (ICDM) 2020 Nov 17 (pp. 851-860). Ieee.

[23] Mao A, Mohri M, Zhong Y. Cross-entropy loss functions: Theoretical analysis and applications. InInternational conference on Machine learning 2023 Jul 3 (pp. 23803-23828). pmlr.

[24] Wang L, Li R, Zhang C, Fang S, Duan C, Meng X, Atkinson PM. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing. 2022 Aug 1; 190: 196-214.

[25] Wang Z, Zheng JQ, Zhang Y, Cui G, Li L. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. arXiv preprint arXiv:2402.05079. 2024 Feb 7.

[26] Liu J, Yang H, Zhou HY, Xi Y, Yu L, Li C, Liang Y, Shi G, Yu Y, Zhang S, Zheng H. Swin-umamba: Mamba-based unet with imagenet-based pretraining. InInternational conference on medical image computing and computer-assisted intervention 2024 Oct 3 (pp. 615-625). Cham: Springer Nature Switzerland.