drones

MDPI

Article

Mamba-UAV-SegNet: A Multi-Scale Adaptive Feature Fusion Network for Real-Time Semantic Segmentation of UAV Aerial Imagery

Longyang Huang, Jintao Tan * and Zhonghui Chen

College of Air Traffic Management, Civil Aviation Flight University of China, Guanghan 618307, China; longyanghuang@cafuc.edu.cn (L.H.); czh@cafuc.edu.cn (Z.C.)
* Correspondence: tjt@cafuc.edu.cn

Abstract: Accurate semantic segmentation of high-resolution images captured by unmanned aerial vehicles (UAVs) is crucial for applications in environmental monitoring, urban planning, and precision agriculture. However, challenges such as class imbalance, small-object detection, and intricate boundary details complicate the analysis of UAV imagery. To address these issues, we propose Mamba-UAV-SegNet, a novel real-time semantic segmentation network specifically designed for UAV images. The network integrates a Multi-Head Mamba Block (MH-Mamba Block) for enhanced multi-scale feature representation, an Adaptive Boundary Enhancement Fusion Module (ABEFM) for improved boundary-aware feature fusion, and an edge-detail auxiliary training branch to capture fine-grained details. The practical utility of our method is demonstrated through its application to farmland segmentation. Extensive experiments on the UAV-City, VDD, and UAVid datasets show that our model outperforms state-of-the-art methods, achieving mean Intersection over Union (mIoU) scores of 71.2%, 77.5%, and 69.3%, respectively. Ablation studies confirm the effectiveness of each component and their combined contributions to overall performance. The proposed method balances segmentation accuracy and computational efficiency, maintaining real-time inference speeds suitable for practical UAV applications.

Keywords: UAV imagery; semantic segmentation; real-time processing; multi-scale feature fusion; boundary enhancement; farmland segmentation

Citation: Huang, L.; Tan, J.; Chen, Z. Mamba-UAV-SegNet: A Multi-Scale Adaptive Feature Fusion Network for Real-Time Semantic Segmentation of UAV Aerial Imagery. *Drones* **2024**, *8*, 671. https://doi.org/10.3390/drones8110671

Academic Editor: Pablo Rodríguez-Gonzálvez

Received: 4 October 2024
Revised: 30 October 2024
Accepted: 3 November 2024
Published: 13 November 2024

1. Introduction

As unmanned aerial vehicle (UAV) technology advances, high-resolution UAV imagery has gained increasing importance in areas like environmental monitoring, agricultural management, disaster assessment, and urban planning. Achieving real-time semantic segmentation of UAV imagery is essential for quickly obtaining actionable insights, thereby enhancing decision making in these fields. However, UAV-acquired images often present challenges such as high resolution, complex backgrounds, variable object scales, and diverse lighting conditions, which make real-time segmentation particularly demanding.

Traditional semantic segmentation methods, such as Fully Convolutional Networks (FCNs), DeepLab, and PSPNet, have made significant progress in segmentation accuracy. However, these models are computationally expensive, limiting their applicability in real-time scenarios. Lightweight segmentation networks, such as BiSeNet and its enhanced version BiSeNetV2, partially address this issue by improving inference speed while maintaining reasonable accuracy. Nevertheless, due to complex scenes and the presence of small objects, capturing fine-grained details, particularly at object boundaries, remains a significant challenge in UAV aerial imagery segmentation.

To address these issues, we propose a multi-scale adaptive feature fusion network named Mamba-UAV-SegNet. Our main contributions are as follows:

- Propose the MH-Mamba Block module: By integrating the Multi-Head 2D Spatial Shift Module and multi-scale convolutions, we enhance the representation capability of intermediate features, improving the model's understanding of complex scenes and its ability to capture details.
- Design a new edge-detail ground truth generation method: By fusing Laplace and Sobel operators to generate edge-detail texture ground truths for the edge-detail auxiliary training branch, we enhance the model's perception of edges and details.
- Introduce the Adaptive Boundary Enhancement Fusion Module (ABEFM): This module effectively fuses high-level semantic information and low-level detailed features, and it strengthens the feature representation of edge regions through a boundary attention mechanism, improving segmentation accuracy.
- To validate the method's effectiveness, extensive experiments conducted on various UAV aerial datasets show that the proposed approach successfully overcomes the limitations of existing methods in UAV aerial scenes, achieving an optimal balance between accuracy and processing speed.

2. Related Work

Semantic segmentation is essential for interpreting and analyzing aerial images captured by unmanned aerial vehicles (UAVs). With applications in areas such as environmental monitoring, urban planning, disaster response, and precision agriculture, achieving accurate and efficient semantic segmentation of UAV imagery holds significant importance [1–5]. Traditional segmentation approaches, which rely on handcrafted features and classical machine learning algorithms, have been largely superseded by deep learning-based methods, particularly Convolutional Neural Networks (CNNs), which have demonstrated superior performance [6–10]. However, UAV imagery presents unique challenges, including high variability in scale and perspective, complex backgrounds, and dynamic environmental conditions, necessitating the development of specialized segmentation techniques [11–15].

2.1. UAV-Based Image Analysis

Unmanned aerial vehicles (UAVs) have revolutionized the field of remote sensing by providing high-resolution, flexible, and real-time data acquisition capabilities [1–5]. The ability to capture detailed aerial imagery from various altitudes and angles enables precise monitoring and analysis of diverse environments, ranging from agricultural fields to urban landscapes [16,17]. However, semantic segmentation of UAV imagery presents unique challenges, including significant variations in object scale, complex and cluttered backgrounds, and the presence of dynamic elements such as moving vehicles or changing weather conditions [11,12]. To address these challenges, researchers have developed specialized techniques that enhance segmentation performance in UAV contexts. Multi-scale feature extraction [13], attention mechanisms [14], and domain adaptation methods [12] have been successfully integrated into segmentation models to improve accuracy and robustness. Additionally, the utilization of advanced datasets like UAVid [18], DOTA [19], and ISPRS Vaihingen [20] has facilitated the benchmarking and advancement of UAV-based segmentation algorithms, providing comprehensive and annotated aerial images that capture the complexity of real-world scenarios.

2.2. Real-Time Segmentation Networks

Real-time semantic segmentation is particularly crucial in unmanned aerial vehicle (UAV) applications, as it requires achieving high-precision segmentation tasks within limited computational resources and time constraints, such as autonomous navigation, real-time monitoring, and obstacle avoidance [10,21–27]. To enhance computational efficiency while maintaining segmentation accuracy, researchers have proposed various lightweight network architectures and optimization techniques. ENet [21] significantly reduces computational complexity by introducing an efficient encoder–decoder structure while preserving high segmentation performance. ERFNet [22] leverages residual modules and factorized

convolutions to achieve real-time segmentation speeds. Additionally, BiSeNet [23] and its subsequent version BiSeNet V2 [24] combine spatial and context paths, incorporating attention mechanisms to effectively balance accuracy and speed. Furthermore, model pruning [25], quantization [26], and knowledge distillation [27] techniques have been applied to optimize real-time segmentation networks, enhancing their operational efficiency on resource-constrained UAV platforms. These advancements have significantly facilitated the feasibility of deploying real-time semantic segmentation models on UAV platforms, providing a solid technical foundation for their widespread application.

2.3. Mamba Framework in Semantic Segmentation

The Mamba framework [28] has emerged as a pivotal tool in the realm of semantic segmentation, offering a versatile and efficient platform for developing and deploying advanced deep learning models. Designed with modularity and scalability in mind, Mamba facilitates the seamless integration of various network architectures, including U-Net, DeepLab, and transformer-based models, thereby enabling researchers to experiment with diverse configurations without extensive reengineering [29]. In the context of aerial imagery segmentation, Mamba has proven instrumental in enhancing both accuracy and computational efficiency. For instance, Zhao et al. [30] leveraged the Mamba framework to implement a real-time semantic segmentation model tailored for UAV-captured images, achieving a commendable balance between high precision and processing speed. Furthermore, Li and Wang [31] extended the framework by incorporating multi-scale feature fusion and attention mechanisms, which significantly improved segmentation performance in complex and cluttered environments. Beyond UAV applications, the adaptability of Mamba has been demonstrated in specialized domains such as medical image segmentation [32] and autonomous driving [33], highlighting its robustness and versatility. These advancements underscore the Mamba framework's capacity to support cutting-edge semantic segmentation research, particularly in scenarios demanding real-time processing and high accuracy.

3. Proposed Method

This paper introduces Mamba-UAV-SegNet, a multi-scale adaptive feature fusion network designed for real-time semantic segmentation of UAV aerial imagery. The network is built upon the lightweight STDC (Short-Term Dense Concatenate) backbone [34], which is known for its efficiency in feature extraction for real-time applications. Our model integrates the MH-Mamba Block module, the Adaptive Boundary Enhancement Fusion Module (ABEFM), and an edge-detail auxiliary training branch to enhance both segmentation accuracy and real-time performance. The overall architecture of the model is depicted in Figure 1.

The architecture of Mamba-UAV-SegNet comprises the following key components:

- STDC Backbone Network: Provides multi-level image features by progressively extracting features through four stages:
  - Stage 1: Extracts low-level features capturing basic edges, textures, and color information from the input image.
  - Stage 2: Captures mid-level features representing more complex patterns and local structures.
  - Stage 3: Extracts high-level semantic features providing abstract representations of objects and scene context.
  - Stage 4: Further refines high-level features, preparing them for subsequent processing in the decoder and additional modules.

  Features from different stages are utilized in subsequent modules to enhance segmentation performance:
- MH-Mamba Block Module: Applied to the feature maps obtained fromStage 2 and Stage 4, this module enhances feature representation by integrating multi-scale convo-

lutions and a Multi-Head 2D State Space Model (Multi-Head 2D-SSM). By processing features from both intermediate and high-level stages, the MH-Mamba Block captures multi-scale contextual information and improves the model's ability to represent both local details and global scene context, which is essential for understanding intricate aerial imagery.

- Adaptive Boundary Enhancement Fusion Module (ABEFM): Fuses the features processed by the MH-Mamba Block Module. Specifically, it takes the enhanced feature maps from the MH-Mamba Block applied to Stage 2 and Stage 4, and effectively combines them. The ABEFM employs a boundary attention mechanism to strengthen feature representation in edge regions, enhancing the accuracy of object boundaries in the segmentation output. This fusion allows the model to integrate detailed local features with rich semantic information, improving segmentation performance across varied object scales.
- Edge-Detail Auxiliary Training Branch: Enhances the model's perception of edges and fine-grained details through auxiliary supervision. It utilizes the edge ground truth generated by combining Sobel and Laplacian operators, as described in Section 4, to guide the network in learning detailed edge features and improving boundary precision.

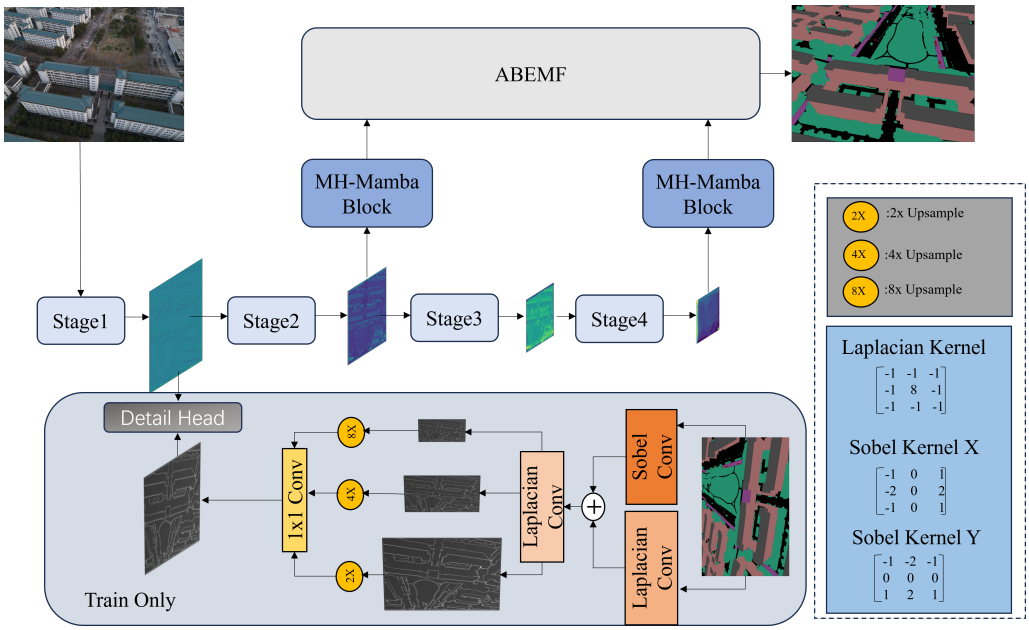

Figure 1. Overall architecture of Mamba-UAV-SegNet. The network consists of four stages within the STDC backbone, corresponding to different levels of feature extraction. The MH-Mamba Block is applied to feature maps from Stage 2 and Stage 4. The Adaptive Boundary Enhancement Fusion Module (ABEFM) fuses features processed by the MH-Mamba Block. The gray block represents the edge detail auxiliary training branch.

These components work synergistically to address the challenges inherent in UAV imagery, such as small-object detection, class imbalance, complex boundaries, and varied object scales. By applying the MH-Mamba Block Module to both intermediate and high-level features (from Stage 2 and Stage 4), the network effectively captures multi-scale information, enhancing both detail preservation and semantic understanding. The ABEFM further refines these features by focusing on boundary regions, ensuring precise segmentation outputs.

In the following subsections, we provide detailed descriptions of each component, explaining how they contribute to the overall performance of Mamba-UAV-SegNet.

STDC Backbone Network

The STDC backbone [34] is designed for real-time semantic segmentation, offering a balance between accuracy and efficiency. It employs short-term dense concatenation to effectively reuse features and reduce computational complexity. The four stages of the STDC backbone progressively extract features at different semantic levels:

- **Stage 1**: Processes the input image with initial convolutional layers, capturing fine-grained details and preserving spatial resolution. This stage is crucial for detecting edges and textures.
- **Stage 2**: Extracts mid-level features through additional convolutional layers. This stage focuses on capturing local patterns and structures, providing richer representations than the initial stage.
- **Stage 3**: Further abstracts the features, extracting high-level semantic information that represents objects and their relationships within the scene.
- **Stage 4**: Produces the most abstract and semantically rich features, essential for accurate classification and understanding of complex scenes.

By leveraging the hierarchical feature extraction capabilities of the STDC backbone, our network effectively balances computational efficiency with the need for detailed feature representations.

4. MH-Mamba Block Module

In remote sensing image segmentation tasks, effectively capturing both fine local details and broader contextual information is critical for improved segmentation accuracy. Conventional Convolutional Neural Networks (CNNs) are proficient in handling local features but struggle with modeling long-range global dependencies. To overcome this challenge, we introduce an innovative module, the MH-Mamba Block, which integrates multi-scale convolutions and a Multi-Head 2D State Space Model (Multi-Head 2D-SSM) to capture comprehensive feature information efficiently in remote sensing imagery.

The core design philosophy of the MH-Mamba Block lies in extracting diverse local features through multi-scale convolutions while leveraging the MultiHead 2D-SSM to capture global dependency relationships within the image. This combination ensures a balanced perception of both local details and global context, thereby improving the overall feature representation. The operational workflow of the MH-Mamba Block is illustrated in Figure 2.

4.1. Multi-Scale Convolution

The input feature map X is initially processed by convolution operations with kernels of different sizes—3×3, 5×5, and 7×7—to capture local features with varying receptive fields:

$$X_{\text{conv3}} = \text{Conv}_{3\times3}(X), \quad X_{\text{conv5}} = \text{Conv}_{5\times5}(X), \quad X_{\text{conv7}} = \text{Conv}_{7\times7}(X) \tag{1}$$

where $X_{\text{conv3}}, X_{\text{conv5}}, X_{\text{conv7}}$ represent the feature maps produced by convolutions with kernel sizes of 3×3, 5×5, and 7×7, respectively. This multi-scale processing enhances the model's sensitivity to features at multiple scales, providing a more comprehensive representation of the local information within the input image.

4.2. Feature Concatenation

The outputs from the multi-scale convolutions are concatenated to integrate features extracted at different scales, resulting in a richer feature representation:

$$X_{\text{concat}} = \text{Concat}(X_{\text{conv3}}, X_{\text{conv5}}, X_{\text{conv7}}) \tag{2}$$

where X_{concat} denotes the concatenated feature map after multi-scale convolutions. This concatenation effectively merges the information captured at various scales, laying a solid foundation for subsequent processing steps.

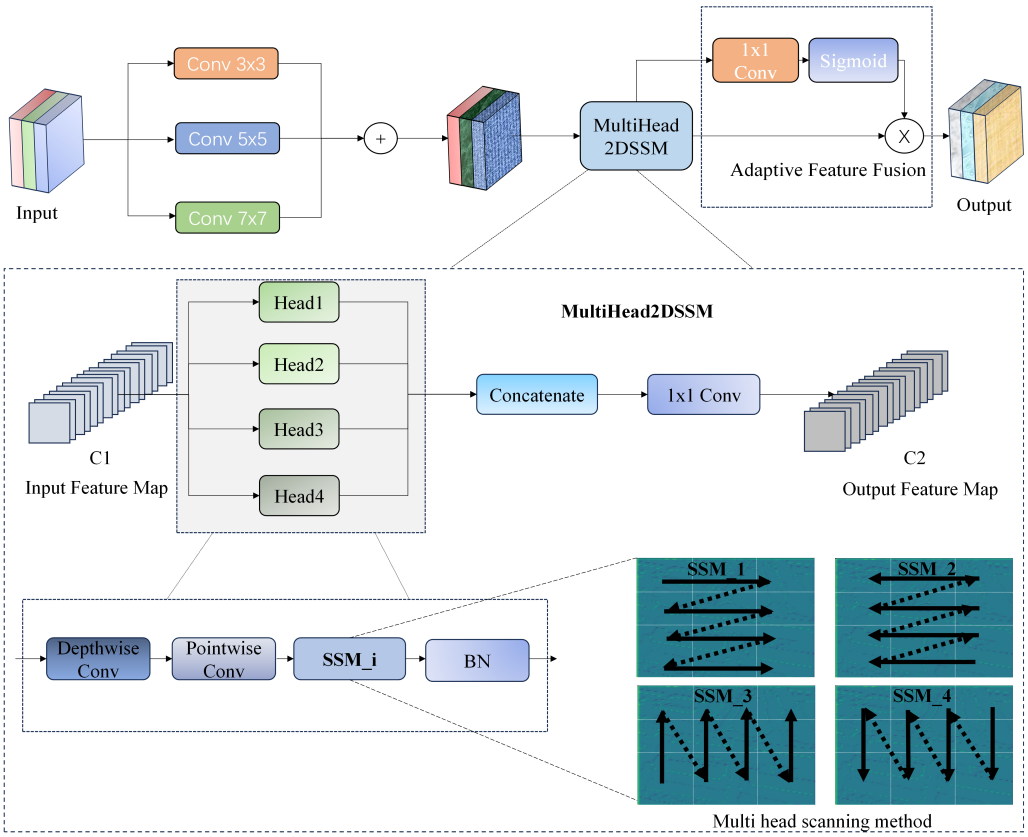

Figure 2. Architecture of the MH-Mamba Block. This module integrates multi-scale convolutions and a Multi-Head 2D State Space Model (Multi-Head 2D-SSM) for enhanced feature extraction. Multi-scale convolutions capture local features, while the Multi-Head 2D-SSM applies four directional scans to model long-range dependencies. The outputs are concatenated and compressed, with Adaptive Feature Fusion selectively enhancing critical features for improved representation.

4.3. Depthwise Convolution and Pointwise Convolution

The concatenated feature map X_{concat} is first processed by a Depthwise Convolution (DWConv) to further extract local spatial features:

$$X_{depth} = \text{DWConv}(X_{concat}) \tag{3}$$

where X_{depth} represents the feature map obtained after depthwise convolution. Subsequently, a Pointwise Convolution (PWConv) is applied to compress the channel dimensions and facilitate cross-channel information fusion:

$$X_{point} = \text{PWConv}(X_{depth}) \tag{4}$$

where X_{point} denotes the feature map after pointwise convolution. This combination not only reduces computational complexity but also promotes effective feature fusion across channels.

4.4. Multi-Head 2D State Space Model (Multi-Head 2D-SSM)

The feature map X_{point} is then fed into the Multi-Head 2D-SSM module. This module employs a multi-head scanning mechanism, with each head applying a distinct scanning pattern to capture directional dependencies:

- Top-left to bottom-right;
- Bottom-right to top-left;
- Top-right to bottom-left;
- Bottom-left to top-right.

These scanning patterns capture comprehensive global information from different directions, forming directional feature representations:

$$X_{\mathrm{SSM}_i} = \mathrm{SSM}_i(X_{\mathrm{point}}), \quad i \in \{1, 2, 3, 4\} \tag{5}$$

where X_{SSM_i} denotes the feature map produced by the ith scanning direction.

By incorporating four directional scans in the SSM, the model establishes long-range dependencies across different regions in the image. Given the complex geographic and object information typically present in remote sensing images, this long-range modeling is particularly crucial. It enables spatially distant regions to influence each other, thereby enhancing the model's overall contextual understanding. The multi-head scanning mechanism captures global contextual information from multiple perspectives, augmenting the model's global awareness. While this approach increases computational complexity, the parallel processing capabilities of the multi-head structure help mitigate the impact on processing speed, allowing the network to achieve efficient feature extraction without sacrificing accuracy.

4.5. Feature Fusion and Compression

The feature maps obtained from the four directional scans $X_{\mathrm{SSM}_1}, X_{\mathrm{SSM}_2}, X_{\mathrm{SSM}_3}, X_{\mathrm{SSM}_4}$ are concatenated and subsequently passed through a 1×1 convolution to compress the channel dimensions. This step ensures that the output feature map maintains the same number of channels as the input, thereby preventing information redundancy:

$$X_{\mathrm{concat2}} = \mathrm{Conv}_{1\times1}\big(\mathrm{Concat}\big(X_{\mathrm{SSM}_1}, X_{\mathrm{SSM}_2}, X_{\mathrm{SSM}_3}, X_{\mathrm{SSM}_4}\big)\big) \tag{6}$$

where X_{concat2} denotes the feature map obtained after concatenation and 1×1 convolution compression. This step merges directional information without increasing channel dimensions, thereby avoiding redundancy.

4.6. Adaptive Feature Fusion

Finally, an Adaptive Feature Fusion module performs weighted processing on the concatenated features. This module generates weights using a sigmoid activation function and multiplies these weights element-wise with the original feature map, emphasizing important features:

$$X_{\mathrm{fused}} = \sigma(\mathrm{Conv}_{1\times1}(X_{\mathrm{concat2}})) \odot X_{\mathrm{concat2}} \tag{7}$$

where σ represents the sigmoid activation function, and $\odot$ denotes element-wise multiplication. Through adaptive feature fusion, the model selectively enhances important features, further improving feature representation capabilities.

4.7. Adaptive Boundary Enhancement Fusion Module (ABEFM)

The ABEFM is designed to fuse high-level semantic features with low-level detail features while reinforcing feature representation in boundary regions. The module consists of three components (Figure 3):

- Feature Enhancement Module (FE): Applies channel attention to both high-level and low-level features to emphasize important features.
- Feature Fusion Module (FF): Aligns the spatial dimensions and channels of the enhanced features before fusing them.
- Boundary Attention Module (BA): Generates boundary attention maps using a learnable edge detector to strengthen feature representation in boundary areas.

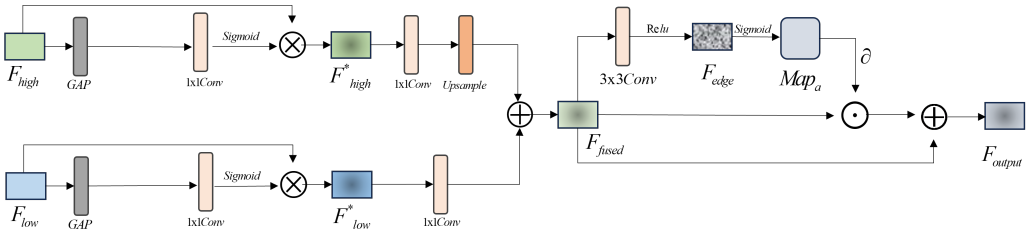

Figure 3. Adaptive boundary enhancement fusion module architecture diagram.

4.7.1. Feature Enhancement Module (FE)

In this context, F_{high} represents the high-level semantic features extracted from Stage 4 of the backbone network. These features encapsulate abstract semantic information and provide a global contextual understanding of the image, which is critical for the accurate object classification and recognition of complex patterns in UAV imagery. Conversely, F_{low} represents the low-level detail features extracted from Stage 2 of the backbone network. These features preserve high spatial resolution and contain rich edge and texture information, making them essential for precise localization and delineation of object boundaries.

Given high-level features $\mathbf{F}_{\text{high}}$ and low-level features $\mathbf{F}_{\text{low}}$, channel attention mechanisms are first applied for adaptive enhancement:

- Channel Weight Generation: For each feature, global average pooling (GAP) is performed, followed by a 1×1 Conv and a Sigmoid activation function to generate channel weights:

$$\mathbf{w}_{\text{high}} = \sigma\Big(\text{Conv}\Big(\text{GAP}\Big(\mathbf{F}_{\text{high}}\Big)\Big)\Big) \tag{8}$$

$$\mathbf{w}_{\text{low}} = \sigma(\text{Conv}(\text{GAP}(\mathbf{F}_{\text{low}}))) \tag{9}$$

  where σ represents the Sigmoid activation function.
- Feature Enhancement: The generated channel weights are applied to the corresponding features through element-wise multiplication:

$$\mathbf{F}'_{\text{high}} = \mathbf{w}_{\text{high}} \odot \mathbf{F}_{\text{high}} \tag{10}$$

$$\mathbf{F}'_{\text{low}} = \mathbf{w}_{\text{low}} \odot \mathbf{F}_{\text{low}} \tag{11}$$

  Here, $\odot$ denotes element-wise multiplication.

4.7.2. Feature Fusion Module (FF)

- Spatial and Channel Alignment: A 1×1 convolution is used to adjust the number of channels, and upsampling operations align the spatial dimensions of the high-level and low-level features.
- Feature Fusion: The aligned features are fused by element-wise addition:

$$\mathbf{F}_{\text{fused}} = \text{U}psample\Big(\text{Conv}_{1\times1}\Big(\mathbf{F}'_{\text{high}}\Big)\Big) + \text{Conv}_{1\times1}(\mathbf{F}'_{\text{low}}) \tag{12}$$

4.7.3. Boundary Attention Module (BA)

- Edge Feature Extraction: A learnable 3×3 convolution, followed by a ReLU activation function, extracts edge features from the fused features:

$$\mathbf{E} = \text{ReLU}(\text{Conv}_{3\times3}(\mathbf{F}_{\text{fused}})) \tag{13}$$

- Boundary Attention Map Generation: A 1×1 convolution and a Sigmoid activation function generate the boundary attention map $\mathbf{A}$:

$$\mathbf{A} = \sigma(\text{Conv}_{1\times1}(\mathbf{E})) \tag{14}$$

- Boundary Reinforcement: The boundary attention map is used to weight the fused features, enhancing feature representation in boundary regions:

$$\mathbf{F}_{\text{output}} = \mathbf{F}_{\text{fused}} + \alpha \cdot (\mathbf{F}_{\text{fused}} \odot \mathbf{A}) \tag{15}$$

where α is a learnable weight parameter.

4.8. Detail Ground Truth Generation Method

In our enhanced approach, we have designed a refined detail ground truth generation module to bolster the neural network's capability in capturing boundary and fine-grained details in semantic segmentation tasks. The specific process is outlined as follows:

- Initial Label Processing:
  The input semantic segmentation labels are first processed separately using Sobel convolution and Laplacian convolution.
  Sobel convolution calculates the gradients of the label image in both the X and Y directions, capturing the preliminary edge information that outlines the boundaries of objects.
  Laplacian convolution, being a second-order derivative operator, further extracts and enhances high-frequency information in the labels, focusing on finer edge details.
  After extracting initial edge features, we perform a feature fusion of the results from Sobel and Laplacian convolutions. This fusion step integrates the directional edge information captured by Sobel convolution with the high-frequency details enhanced by Laplacian convolution, forming a rich edge feature map.
  Next, the fused feature map undergoes a second Laplacian convolution to further refine the edge features. This step helps in sharpening the fused edges and eliminating potential noise, thus generating a more detailed and fine-grained edge feature map, which is crucial for capturing subtle edge details not easily detected in a single pass.
- Multi-Scale Convolution and Upsampling:
  To ensure the accuracy of edge features across different scales, we apply multi-scale convolution. The fused edge features are processed with different strides (stride = 2, 4, 8) to capture both global and local edge information. Following the convolution, each scale's feature map is subjected to corresponding upsampling (2x, 4x, 8x) to restore the original resolution. This multi-scale processing ensures that edge features at various resolutions contain adequate detail.
- Feature Fusion and Output:
  Finally, the upsampled multi-scale feature maps are fused to produce the final edge ground truth. This fusion combines the global edge structures from low-resolution features with the fine local details from high-resolution features, resulting in a high-resolution edge map that contains rich edge details.

4.9. Loss Function Design

Due to the relatively small proportion of detail pixels in images, traditional loss functions may underperform in handling class imbalance issues. To enhance the training efficacy of the network, we have designed and improved the loss function as follows:

4.9.1. Binary Cross-Entropy Loss (L_{bce})

Binary cross-entropy loss is commonly employed to quantify the difference between the predicted detail map and the ground truth. This loss function effectively calculates

the deviation between each pixel's predicted probability and its corresponding ground truth value:

$$\mathcal{L}_{\text{bce}} = -\frac{1}{N}\sum_{i=1}^{N}[g_i \log(p_i) + (1 - g_i)\log(1 - p_i)] \tag{16}$$

In this equation, $\mathcal{L}_{\text{bce}}$ denotes the binary cross-entropy loss value, N is the total number of pixels in the image, g_i represents the ground truth label of the ith pixel (1 for edge pixel, 0 for non-edge pixel), p_i is the predicted probability that the ith pixel belongs to the edge class, and log denotes the natural logarithm.

4.9.2. Dice Loss ($\mathcal{L}_{\text{dice}}$)

Dice Loss is employed to quantify the overlap between the predicted detail map and the ground truth map. It is particularly effective in addressing class imbalance, since its loss value remains unaffected by the number of foreground and background pixels. The formula for Dice Loss is as follows:

$$\mathcal{L}_{\text{dice}}(p_d, g_d) = 1 - \frac{2\sum_i p_{d_i} g_{d_i} + \epsilon}{\sum_i p_{d_i}^2 + \sum_i g_{d_i}^2 + \epsilon} \tag{17}$$

where p_d and g_d represent the predicted and ground truth detail maps, respectively, i denotes the pixel index, and ϵ is a small constant to prevent division by zero.

4.9.3. Edge-Aware Loss

To further enhance the learning of boundary information, we introduce an Edge-Aware Loss. This loss function emphasizes the accuracy of edge pixels, compelling the network to better learn details at the boundaries. Specifically, the edge loss is computed by extracting edge features from images using our proposed edge-detail ground truth generation method. The edge loss is defined as:

$$\mathcal{L}_{\text{edge}} = \sum_i \left| E_i - \hat{E}_i \right|^2 \tag{18}$$

where E_i is the extracted edge feature, and $\hat{E}_i$ is the corresponding ground truth edge-detail feature, with i denoting the pixel index. By calculating the mean squared error between the edge features and the ground truth features, we strengthen the network's learning of boundary information, thus improving fine-grained segmentation performance.

4.9.4. Final Detail Loss Function

Considering all the above components, our final detail loss function is formulated as follows:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{edge}} + \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{bce}} \tag{19}$$

This comprehensive loss function not only enhances the network's ability to capture details but also effectively mitigates class imbalance issues, thereby improving overall segmentation accuracy.

5. Experimental Results

This study evaluates the proposed method using three UAV aerial image datasets: VDD [35], UAV-city, and UAVid.

5.1. Datasets

UAVid Dataset: UAVid is a widely adopted dataset for semantic segmentation in UAV applications. It comprises 30 video sequences captured from a tilted viewpoint with 4K high-resolution imagery. The dataset includes a total of 300 images, with each

densely annotated across eight categories: Building, Road, Static Car, Tree, Low Vegetation, Humans, Moving Car, and Background Clutter. The images have resolutions of either 4096 × 2160 pixels or 3840 × 2160 pixels.

VDD Dataset: The Varied Drone Dataset (VDD) consists of 400 high-resolution images covering seven distinct categories. This dataset encompasses a variety of scene types, including urban, industrial, rural, and natural environments. The images have been captured from multiple camera angles and under varying lighting conditions. Released by RussRobin, the VDD aims to advance research in semantic segmentation for UAV imagery by focusing on the effects of long-tail distributions and out-of-distribution scenarios on segmentation algorithms. With its diverse, large-scale, and high-resolution data, the VDD effectively addresses the data scarcity issue in aerial image processing, providing a comprehensive range of visual information.

UAV-city Dataset: The UAV-city dataset includes 600 images from various scenes, annotated on a pixel-wise basis using the LabelMe tool. Each image has a resolution of 1280 × 720 pixels and is divided into training, validation, and test sets in an 8:1:1 ratio. Dominant classes such as Tree, Road, and Building occupy the majority of pixels, whereas minority classes like Car, Horizontal Roof, Horizontal Ground and Lawn, River, Obstacle, and Plant are underrepresented. Notably, the Human class constitutes only 0.103% of the total pixels, posing substantial challenges for segmenting small objects due to their limited pixel count and smaller sizes.

5.2. Implementation Details

Training: In this study, we employed the MMsegmentation (MMseg) framework for model training, utilizing its standard configurations. We selected the AdamW optimizer with an initial learning rate of 0.00006, momentum parameters set to (0.9, 0.999), and a weight decay coefficient of 0.01. The learning rate scheduling strategy comprised a linear warm-up for the first 1500 iterations, during which the learning rate increased gradually from 1×10^{-6} to the initial value, followed by a polynomial decay until reaching the maximum number of iterations. For the VDD dataset, we used a batch size of 16 and trained the model for up to 80,000 iterations. Both the UAVid and UAV-city datasets were trained with a batch size of 8 and 16, respectively, each for a maximum of 10,000 iterations, maintaining the same initial learning rate of 0.00006 across all datasets. All training experiments were conducted on a workstation equipped with an NVIDIA RTX 2080Ti GPU utilizing CUDA 11.6 and cuDNN 8.5.0. The training process was implemented using PyTorch version 1.12.1 within an Anaconda environment configured with the necessary dependencies. Through these configurations, we systematically trained and evaluated the proposed method on three UAV aerial image datasets: VDD, UAVid, and UAV-city.

Evaluation Metric: In this study, we employed the Mean Intersection over Union (Miou) as the primary evaluation metric for our multi-class semantic segmentation models. The Miou is calculated by averaging the Intersection over Union (IoU) across all classes. For a given class c, the IoU is defined as follows:

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c} \tag{20}$$

where TP_c, FP_c, and FN_c represent the true positives, false positives, and false negatives for class c, respectively. The overall Miou is then computed as follows:

$$\mathrm{Miou} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{IoU}_c \tag{21}$$

where C is the total number of classes. This metric provides a comprehensive evaluation of the model's ability to accurately segment each class in the UAV aerial image datasets,

ensuring a robust and quantitative assessment of our proposed method's performance across the VDD, UAVid, and UAV-city datasets.

5.3. Comparison with Mainstream Methods

On the UAV-City,VDD, and UAVid datasets, we compare our approach with mainstream methodologies in this section.

To comprehensively evaluate the performance of the proposed Mamba-UAV-SegNet in semantic segmentation tasks, we selected several representative baseline models for comparison, including the classic FCN-8s and DeepLabV3+, lightweight models such as BiSeNetV2 and EfficientFormer, and modern Transformer-based models like SegFormer, Mask2Former, and SegNeXt. These models were chosen because they encompass different types within the semantic segmentation domain: from classic methods to efficient real-time models to advanced Transformer-based methods. This allowed us to thoroughly assess the comprehensive performance of Mamba-UAV-SegNet across different scenarios, especially in terms of accuracy, boundary processing, and multi-scale feature extraction. All models were trained under the same hardware and experimental conditions to ensure fair comparisons.

As shown in Table 1, Mamba-UAV-SegNet outperformed the classic FCN-8s and the lightweight BiSeNetV2 models across all datasets, especially in complex scenarios like the VDD dataset, where the mIoU reached 77.5%, significantly higher than the other baseline models. Compared to modern Transformer-based models like SegFormer and Mask2Former, Mamba-UAV-SegNet demonstrated strong boundary processing capabilities and effective multi-scale feature extraction, achieving an average mIoU of 72.7%, indicating competitive accuracy.

Table 1. Comparison of different models on three datasets (UAVid, VDD, UAV-City) based on mIoU. UAVid mIoU represents the mean Intersection over Union (mIoU) on the UAVid dataset, VDD mIoU represents the mIoU on the VDD dataset, UAV-City mIoU represents the mIoU on the UAV-City dataset, and Avg. mIoU represents the average mIoU across all three datasets.

Model	UAVid mIoU (%)	VDD mIoU (%)	UAV-City mIoU (%)	Avg. mIoU (%)
FCN-8s [6]	62.4	61.4	63.4	62.4
DeepLabV3+ [36]	67.0	66.8	64.2	66.0
BiSeNetV2 [24]	59.7	67.0	65.5	64.1
UNetFormer [37]	67.8	68.7	67.9	68.1
SCTNet [38]	68.4	72.5	67.9	69.6
SegFormer [39]	67.2	74.3	70.1	70.5
HRNet [40]	63.8	64.9	68.5	65.7
Mask2Former [41]	68.5	75.0	70.2	71.2
EfficientFormer [42]	67.8	70.1	69.8	69.2
SegNeXt [43]	68.2	72.8	70.2	70.4
Ours	69.3	77.5	71.2	72.7

By conducting a consistent comparison with the same set of baseline models, we can more fairly assess the comprehensive performance of Mamba-UAV-SegNet across different datasets and scenarios, demonstrating the advantages of the proposed method in boundary detection, small-object recognition, and multi-scale feature extraction.

Results on UAVid

To confirm the effectiveness of our proposed approach, we performed comparative experiments with several well-known semantic segmentation models on the UAVid dataset. The findings are summarized in Table 2.

Table 2. Presents a comparison of several mainstream semantic segmentation models on the UAVid dataset. Each column represents the Intersection over Union (IoU) performance of the models on different classes, such as "Clutter", "Tree," and "Mov. Car", while the mIoU column indicates the mean IoU across the entire dataset. These metrics provide a clear evaluation of each model's segmentation performance across various target types.

Model	Class IoU (%)								mIoU (%)
	Clutter	Building	Road	Tree	Low Veg.	Mov. car	Static Car	Human	
FCN-8s [6]	63.9	84.7	76.5	73.3	61.9	65.9	45.5	22.3	62.4
SegNet [10]	65.6	85.9	79.2	78.8	63.7	68.9	52.1	19.3	64.2
BiseNet [23]	64.7	85.7	61.1	78.3	**77.3**	48.6	63.4	17.5	61.5
U-Net [7]	61.8	82.9	75.2	77.3	62.0	59.6	30.0	18.6	58.4
BiSeNetV2 [24]	61.2	81.6	77.1	76.0	61.3	66.4	38.5	15.4	59.7
DeepLabV3+ [36]	68.9	87.6	**82.2**	79.8	65.9	69.9	55.4	26.1	67.0
UNetFormer [37]	68.4	87.4	81.5	80.2	63.5	73.6	56.4	31.0	67.8
BANet [44]	66.6	85.4	80.7	78.9	62.1	69.3	52.8	21.0	64.6
STDC-Seg75 [34]	68.7	86.8	79.4	78.6	65.4	68.1	55.7	24.5	65.9
STDC-CT75 [45]	**69.2**	88.5	80.1	**80.4**	66.3	**73.8**	60.3	28.4	68.4
Ours	64.3	**91.3**	78.2	78.2	68.4	72.5	**64.5**	**36.8**	**69.3**

From Table 2, we can observe that our method achieved the highest mIoU of 69.3% on the UAVid dataset, demonstrating a significant improvement over other methods. Specifically, our method attained the highest IoU in the Building and Static Car categories, with values of 91.3% and 64.5%, respectively. Additionally, in the Human category, our method showed a remarkable enhancement, achieving an IoU of 36.8%, which is substantially higher compared to the other methods.

Compared to the classical methods, FCN-8s and SegNet exhibited relatively lower overall performance due to their limited ability to capture multi-scale features and edge details. Real-time semantic segmentation methods like BiSeNet and BiSeNetV2 offered faster inference speeds but somewhat lacked accuracy. Methods such as DeepLabV3+ and UNetFormer performed well in certain categories but had overall mIoUs slightly lower than ours.

Although the overall improvement in the mIoU compared to other methods, such as the approach of STDC-CT [45], was only 0.9%, our method demonstrated specific advantages in challenging categories like "Static Car" and "Human". These categories often present complex boundaries and varied object scales, which were effectively handled by our proposed MH-Mamba Block module, the Adaptive Boundary Enhancement Fusion Module (ABEFM), and the edge-detail auxiliary training branch. This highlights the strengths of our method in enhancing segmentation accuracy for difficult object types, validating the effectiveness of our approach in addressing UAV image segmentation challenges.

To visually illustrate the segmentation performance of our approach, Figure 4 provides a comparative view of segmentation results on selected samples from the UAVid dataset.

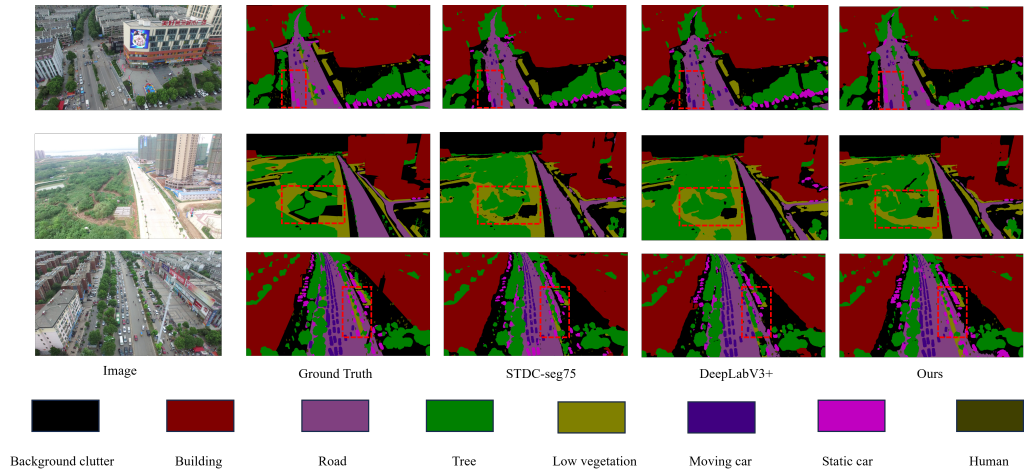

Figure 4. Visualization results on the UAVid dataset.

Results on UAV-City

To further confirm the effectiveness of our proposed approach, we performed comprehensive experiments on the UAV-City dataset. This dataset presents considerable challenges, such as small-object sizes and class imbalance, making it an optimal benchmark for testing semantic segmentation models in UAV aerial imagery.

We evaluated our model against several prominent semantic segmentation approaches, such as U-Net [7], PSPNet [9], DDRNet23 [46], DeepLabV3+ [36], STDC-Seg [34], BiSeNetV3 [47], and SCTNet [38]. This comparison centered on segmentation accuracy, indicated by mean Intersection over Union (mIoU), and inference speed, measured in Frames Per Second (FPS). A summary of the results is provided in Table 3.

Table 3. Comparisons with other mainstream methods on UAV-City.

Model	Resolution	Backbone	mIoU (%)	FPS
U-Net [7]	960 × 540	VGG16	63.4	28.9
PSPNet [9]	960 × 540	ResNet50	54.5	34.5
DDRNet23 [46]	960 × 540	DDRNet	57.4	35.5
DeepLabv3+ [36]	960 × 540	MobileNetV2	64.2	80.5
STDC-Seg [34]	960 × 540	STDC1	65.1	212.3
STDC-CT [45]	960 × 540	STDC1	67.3	196.8
BiseNetV3 [47]	960 × 540	ResNet50	65.5	121.0
SCTNet [38]	960 × 540	CFBlock-Net	67.9	107.2
Ours	960 × 540	STDC1	71.2	109.4

From Table 3, it is evident that our proposed method achieved the highest mIoU of **71.2%**, outperforming all the other compared methods on the UAV-City dataset. Although some methods like STDC-Seg and STDC-CT offer higher FPS due to their lightweight architectures, they fell short in terms of segmentation accuracy. Our method strikes a favorable balance between accuracy and efficiency, achieving competitive FPS while significantly improving mIoU.

To provide a comprehensive analysis, we present the per-class IoU results in Table 4. This detailed evaluation highlights how each method performed across different categories present in the UAV-City dataset.

Table 4. The results of the experiment on the UAV-City dataset.

Model	Class IoU (%)												mIoU (%)
	Hor. roof	Hor. gro.	Hor. lawn	River	Plant	Tree	Car	Hum.	Bui.	Road	Obs.	Back.	
U-Net [7]	63.4	51.5	57.7	81.4	57.9	85.6	56.1	13.5	78.9	80.6	81.5	53.5	63.4
PSPNet [9]	52.1	47.8	55.2	75.3	43.5	80.6	32.2	2.1	71.7	73.2	68.8	51.5	54.5
DDRNet23 [46]	53.8	59.5	64.5	77.5	35.9	83.7	42.1	5.5	75.9	74.9	65.5	49.7	57.4
DeepLabv3+ [36]	65.7	59.7	62.7	82.7	58.7	86.7	58.7	15.7	81.7	82.7	81.7	54.7	64.2
STDC-Seg [34]	64.3	62.8	58.5	80.6	60.5	83.4	58.1	16.1	81.6	79.5	83.9	52.3	65.1
STDC-CT [45]	65.6	62.3	66.2	85.7	59.2	86.7	61.3	18.6	83.1	82.5	82.1	54.3	67.3
BiSeNetV3 [47]	64.0	61.0	61.0	82.0	59.0	84.0	60.0	19.0	81.0	80.0	80.0	55.0	65.5
SCTNet [38]	66.0	63.0	65.0	85.0	61.0	86.0	62.0	20.0	84.0	83.0	83.0	56.0	67.9
Ours	**75.7**	**67.0**	**70.5**	**86.3**	**64.1**	**89.0**	**66.0**	**21.4**	**85.3**	**83.4**	**85.0**	**59.9**	**71.2**

From Table 4, we can observe that our method achieved the highest IoU in most categories. Specifically we present the following results:

- Horizontal Roof, Ground, and Lawn: Our method significantly outperformed others, indicating superior ability in segmenting flat surfaces and open areas.
- River and Plant: Achieving IoUs of **86.3%** and **64.1%**, our model effectively distinguished these classes, which often have similar visual features.
- Tree and Background: With IoUs of **89.0%** and **59.9%**, the model demonstrated strong performance in capturing complex textures and background regions. Car and Human: Notably, our method achieved higher IoUs for small objects like cars (**66.0%**) and humans (**21.4%**), addressing the challenge of segmenting small-scale targets in UAV imagery.

Compared to the other methods, our model consistently showed improved performance across all categories. The enhancements can be attributed to the integration of the MH-Mamba Block module and the Adaptive Boundary Enhancement Fusion Module (ABEFM), which together enhance feature representation and boundary detection.

Results on VDD

To further evaluate the generalization capability and effectiveness of our proposed method, we conducted experiments on the Varied Drone Dataset (VDD). The VDD dataset presents diverse scenes captured from different camera angles and under various lighting conditions, making it a challenging benchmark for semantic segmentation in aerial imagery.

We compared our model with several mainstream semantic segmentation methods, including BiSeNetV1 [23], BiSeNetV2 [24], ENet [21], DFANet [48], STDC-Seg [34], DDRNet [46], and SCTNet [38]. The comparison focused on both segmentation accuracy, measured by mean Intersection over Union (mIoU), and inference speed, measured by Frames Per Second (FPS). The results are summarized in Table 5.

Table 5. Comparisons with other mainstream methods on VDD.

Model	Backbone	GPU	Resolution	mIoU (%)	FPS
BiSeNetV1 [23]	ResNet101	RTX2080Ti	512 × 1024	64.5	45.0
BiSeNetV2 [24]	MobileNetV2	RTX2080Ti	512 × 1024	67.0	70.0
ENet [21]	-	RTX2080Ti	512 × 1024	60.0	160.0
DFANet [48]	-	RTX2080Ti	512 × 1024	69.0	80.0
STDC-Seg [34]	STDC1	RTX2080Ti	512 × 1024	71.0	150.0
DDRNet [46]	DDRNet	RTX2080Ti	512 × 1024	68.0	90.0
SCTNet [38]	-	RTX2080Ti	512 × 1024	72.5	100.0
Ours	STDC1	RTX2080Ti	512 × 1024	77.5	104.0

From Table 5, it is evident that our proposed method achieved the highest mIoU of **77.5%**, outperforming all other compared methods on the VDD dataset. While ENet [21]

demonstrated the highest FPS due to its lightweight architecture, it suffered from lower segmentation accuracy. Our method not only surpassed others in accuracy but also maintained a competitive inference speed of 104.0 FPS, making it suitable for real-time applications.

To qualitatively assess the segmentation performance, we provide visualization results in Figure 5, showcasing the segmentation outputs of different models, alongside the ground truth.

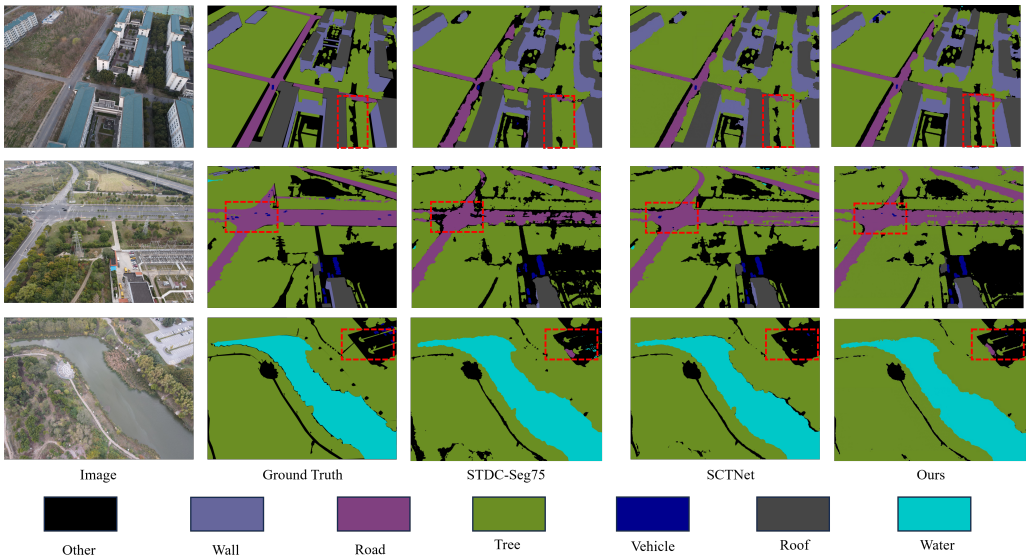

Figure 5. Visualization results on the VDD dataset.

5.4. Ablation Study

We designed several variants of our model by selectively adding or removing specific components:

- Baseline: The base network without any of our proposed modules.
- Baseline + MH-Mamba Block: Incorporating the MH-Mamba Block into the baseline network.
- Baseline + ABEFM: Adding the Adaptive Boundary Enhancement Fusion Module (ABEFM) to the baseline network.
- Baseline + Edge-Detail Auxiliary Training: Including the edge-detail auxiliary training branch with the baseline.
- Baseline + MH-Mamba Block + ABEFM: Combining the MH-Mamba Block and ABEFM with the baseline.
- Baseline + MH-Mamba Block + Edge-Detail Auxiliary Training: Combining the MH-Mamba Block and edge-detail auxiliary training with the baseline.
- Baseline + ABEFM + Edge-Detail Auxiliary Training: Combining the ABEFM and edge-detail auxiliary training with the baseline.
- Full Model: The complete model with all proposed components integrated.

The results of the ablation study on the VDD dataset are summarized in Table 6. The performance of each model variant was measured using the mean Intersection over Union (mIoU) metric.

Table 6. Ablation study results on the VDD dataset. MH-MB refers to the MH-Mamba Block, Edge-Aux. stands for Edge-Detail Auxiliary Training, and the Full Model indicates the complete model incorporating all components.

Model Variant	MH-MB	ABEFM	Edge-Aux.	mIoU (%)
Baseline	×	×	×	69.0
Baseline + MH-MB	✓	×	×	72.5
Baseline + ABEFM	×	✓	×	71.2
Baseline + Edge-Aux.	×	×	✓	70.0
Baseline + MH-MB + ABEFM	✓	✓	×	74.0
Baseline + MH-MB + Edge-Aux.	✓	×	✓	73.2
Baseline + ABEFM + Edge-Aux.	×	✓	✓	72.0
Full Model	✓	✓	✓	**77.5**

Through the ablation study analysis on the VDD dataset, we found that the baseline model achieved an mIoU of 69.0%. Introducing the MH-Mamba Block increased the mIoU to 72.5%, indicating that this module effectively enhances feature representation and captures multi-scale context. Adding the ABEFM module raised the mIoU to 71.2%, demonstrating improved boundary refinement and feature fusion. Incorporating the edge-detail auxiliary training branch resulted in an mIoU of 70.0%, reflecting its effectiveness in capturing fine-grained details.

When different modules were combined, the performance improved further: the combination of the MH-Mamba Block and ABEFM boosted the mIoU to 74.0%, showing a synergistic effect; combining the MH-Mamba Block with edge-detail auxiliary training achieved an mIoU of 73.2%; and the combination of ABEFM and edge-detail auxiliary training yielded an mIoU of 72.0%. Finally, integrating all modules into the full model reached the highest mIoU of 77.5%. This significant improvement confirms the effectiveness of combining our proposed modules.

5.5. Visualization and Analysis of Ablation Results

To gain deeper insights into how each proposed module enhances the model's performance, we conducted qualitative analyses using visualization techniques. These visualizations help illustrate the impact of our modules on feature representation, attention mechanisms, and feature fusion.

Visualization of Low-Level Feature Maps

Figure 6 compares the low-level feature maps extracted from the baseline model and those from the model trained with the edge-detail auxiliary branch. The low-level features are crucial for capturing fine-grained details and edges, which are essential for accurate segmentation, especially of small objects.

In Figure 6, we observe the following:

- Baseline Model: The feature maps are less sharp and lack clear edge definitions. The model struggled to capture fine details, leading to blurred feature representations.
- With Edge-Detail Auxiliary Training: The feature maps exhibit more pronounced edges and finer details. The auxiliary training branch effectively enhanced the model's ability to focus on important low-level features.

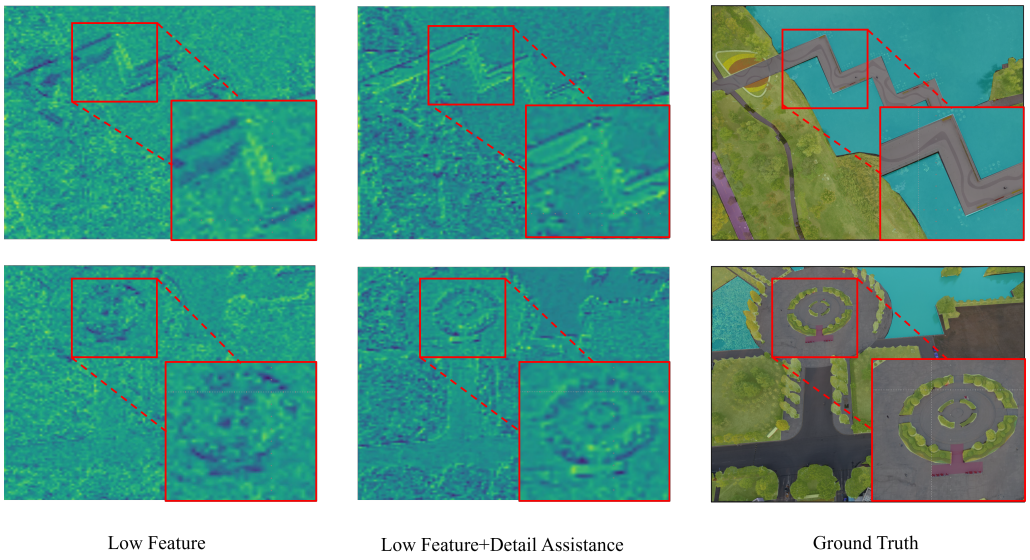

Figure 6. Visualization of low-level feature maps.

This comparison demonstrates that the edge-detail auxiliary training branch significantly improved the extraction of detailed features, contributing to better segmentation performance.

Grad-CAM Visualization for Lake Class

To examine the shifts in model attention following the integration of various modules, we applied Gradient-Weighted Class Activation Mapping (Grad-CAM) to visualize activation areas specific to the lake class. Figure 7 displays Grad-CAM visualizations comparing the baseline model with the model augmented by the MH-Mamba Block.

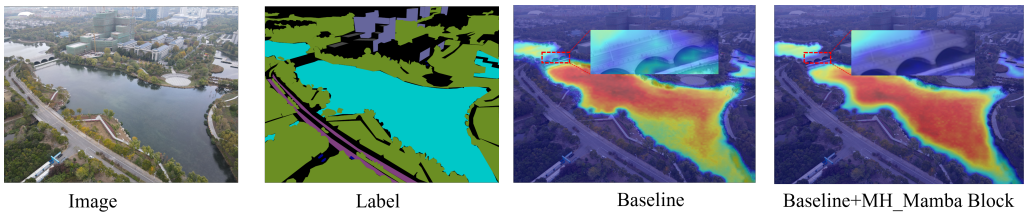

Figure 7. Grad-CAM visualization for lake class.

From Figure 7, we can observe the following:

- Baseline Model: The attention is diffused, with the model not fully focusing on the lake regions. This led to misclassification and incomplete segmentation of the lake area.
- With MH-Mamba Block: The attention is more concentrated on the lake regions. The MH-Mamba Block enhanced the model's ability to capture contextual information and focus on relevant features.

The improved attention visualization indicates that the MH-Mamba Block effectively helps the model to better understand and segment specific classes by capturing multi-scale contextual information.

Visualization of Feature Fusion

To illustrate the effectiveness of the Adaptive Boundary Enhancement Fusion Module (ABEFM), we compared the feature fusion results of the traditional Feature Fusion Module (FFM) with our proposed ABEFM. Figure 8 shows the visualization of fused features from both modules.

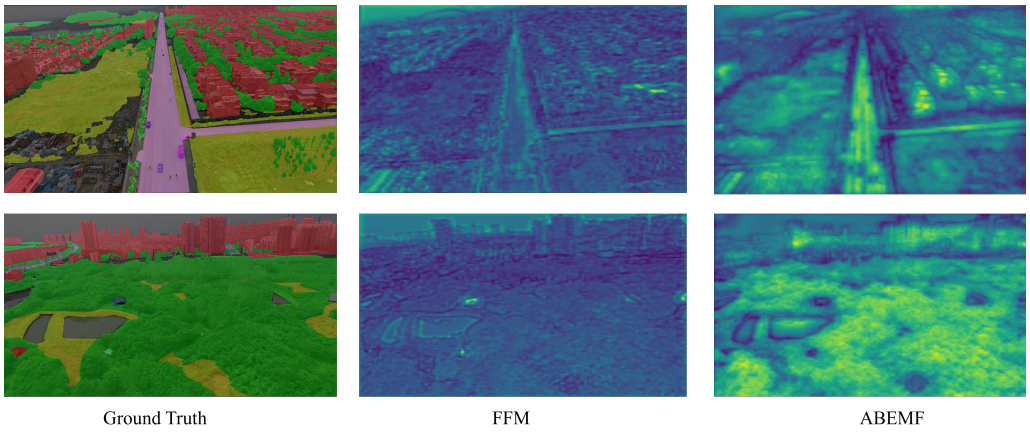

Figure 8. Feature fusion and visualization of feature maps.

In Figure 8, we notice the following:

- FFM: The fused features are less distinct, and boundaries between different objects are not well defined. This can lead to confusion between adjacent classes.
- ABEFM: The fused features exhibit clearer boundaries and more distinct representations of different objects. The ABEFM enhanced the fusion process by emphasizing boundary information.

This comparison demonstrates that ABEFM improves the quality of feature fusion, leading to better preservation of boundary details and improved segmentation accuracy.

Overall, the ablation study—supported by both quantitative results and qualitative visualizations—confirms that each proposed module individually enhances the model's performance. Furthermore, integrating all the modules leads to synergistic effects, resulting in the highest segmentation accuracy observed. Our proposed method effectively captures detailed features, focuses on relevant regions, and preserves boundary information, making it exceptionally well suited for semantic segmentation tasks in UAV imagery. These findings validate the effectiveness of our approach and demonstrate its potential for real-world applications in aerial image analysis.

6. Practical Application in Agricultural Land Segmentation

To further demonstrate the practical utility of the proposed Mamba-UAV-SegNet model, we present a case study of its application in agricultural land segmentation. The method was tested for delineating field boundaries and monitoring agricultural plots, effectively addressing key challenges inherent in agricultural environments, such as small-object detection, complex boundaries, and varying target scales.

The accurate segmentation of agricultural fields is crucial for precision agriculture, enabling farmers and agricultural stakeholders to monitor crop health, optimize irrigation strategies, and manage land resources effectively. Mamba-UAV-SegNet, with its enhanced boundary detection capabilities, adaptive multi-scale feature extraction, and robustness in small-object segmentation, is particularly well suited for this application.

Figure 9 illustrates an example of this application. In the top image, a UAV is prepared for takeoff, equipped with sensors to capture high-resolution aerial imagery of agricultural fields. The lower images depict the raw aerial view of farmland (left) and the corresponding semantic segmentation output produced by Mamba-UAV-SegNet (right). The segmentation map highlights various regions such as pathways, crop areas, and other field features, demonstrating the model's ability to distinguish between different surface types and effectively map complex agricultural landscapes.

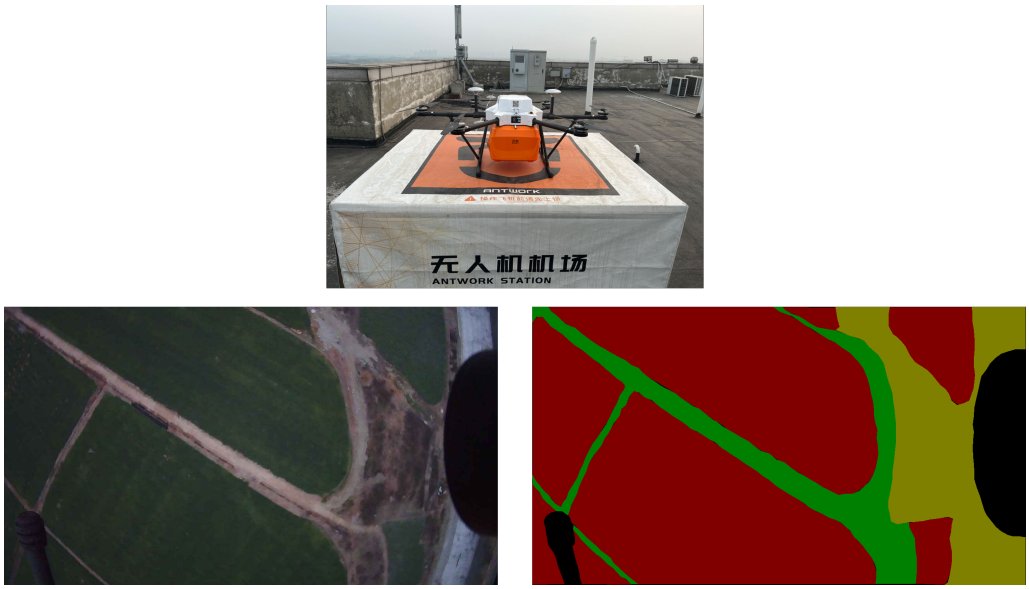

Figure 9. Application of Mamba-UAV-SegNet in agricultural land segmentation. (**Top**) UAV equipped for aerial imaging. (**Bottom left**) Raw aerial image of farmland. (**Bottom right**) Semantic segmentation output produced by Mamba-UAV-SegNet. The segmentation map distinguishes various field features, demonstrating the model's effectiveness in mapping agricultural landscapes.

By employing the Mamba-UAV-SegNet model, precise field boundaries and individual land parcels can be readily identified, supporting activities such as targeted pesticide application, soil analysis, and crop management. The model's robustness under varying conditions—including different times of day, weather conditions, and flight altitudes—ensures consistent and reliable segmentation results, making it a valuable tool in the domain of smart farming. This real-world example not only demonstrates the practical applicability of the proposed method but also highlights its potential to enhance productivity in agriculture. By providing accurate and consistent segmentation results, Mamba-UAV-SegNet addresses the growing demands of sustainable agriculture and precision land management, making it an indispensable tool for modern agricultural practices.

7. Conclusions

In this study, we introduced an advanced semantic segmentation model specifically designed for UAV aerial imagery. By integrating the MH-Mamba Block, the Adaptive Boundary Enhancement Fusion Module (ABEFM), and an edge-detail auxiliary training branch, our model effectively addresses challenges such as class imbalance, small-object detection, and boundary refinement inherent in UAV data. Comprehensive experiments conducted on multiple datasets—including the UAV-City, VDD, and UAVid—validate the superiority of our approach in terms of segmentation accuracy and efficiency, achieving higher mIoU scores and exhibiting strong generalization capabilities.

Our method significantly enhances feature representation by employing the MH-Mamba Block, which improves the model's ability to capture multi-scale features essential for accurately segmenting objects of different sizes in aerial images. The ABEFM module enhances boundary refinement and feature fusion by incorporating boundary information, resulting in more precise segmentation maps with well-defined object boundaries. Additionally, the edge-detail auxiliary training branch allows the model to focus on intricate details, improving the segmentation of small and complex objects. The consistent improvements across the various datasets demonstrate the robustness and generalization capability of our method.

The ablation studies confirm that each module contributes significantly to performance improvement, and visual analyses further support these findings by displaying improved feature maps, focused attention, and better boundary preservation. Our method achieves a

good balance between accuracy and efficiency, maintaining inference speeds suitable for real-time applications.

The findings of this research contribute to the field of semantic segmentation in UAV imagery, providing a robust and efficient solution that can be leveraged in various real-world applications, including agricultural management, environmental monitoring, and crisis management. By addressing key challenges and demonstrating strong performance across multiple datasets, our model offers a valuable tool for advancing UAV-based image analysis.

8. Discussion

While Mamba-UAV-SegNet demonstrates strong performance across multiple UAV datasets, it is important to acknowledge potential limitations and consider areas for future improvement. This section discusses possible challenges the model may face under varying conditions and outlines strategies to enhance its versatility and robustness.

8.1. Performance Under Different Environmental Conditions

Although our experiments cover diverse urban and rural scenes, the datasets used predominantly feature images captured under favorable conditions with consistent lighting, weather, and flight altitudes. In real-world applications, UAV imagery can be subject to significant variability due to factors such as time of day, weather changes, and varying flight heights. These factors can affect image quality and, consequently, the performance of semantic segmentation models.

8.1.1. Time of Day Variations

Changes in illumination throughout the day can lead to shadows, glare, or low-light conditions, potentially impacting the model's ability to accurately segment objects. For instance, strong shadows during sunrise or sunset may obscure features, while low-light conditions at dusk can reduce image contrast.

Potential Impact: The model may experience decreased accuracy in edge detection and object recognition due to altered pixel intensities and contrast levels.

Future Improvement Strategies: Incorporating data augmentation techniques that simulate different lighting conditions can help the model generalize better. Additionally, training with datasets captured at various times can enhance robustness to illumination changes.

8.1.2. Weather Conditions

Adverse weather conditions such as rain, fog, or snow introduce noise and distortions in aerial images. Rain droplets can blur images, fog can obscure details, and snow can alter the appearance of surfaces.

Potential Impact: The presence of weather-induced artifacts may lead to misclassification or missed detections, particularly for small objects or subtle features.

Future Improvement Strategies: Developing preprocessing methods to mitigate weather effects, such as image dehazing or denoising algorithms, can improve image quality before segmentation. Training the model on weather-diverse datasets can also enhance its adaptability.

8.1.3. Flight Altitude Variations

Variations in flight altitude affect the ground sampling distance (GSD) and the scale of objects in the image. Higher altitudes result in lower-resolution images with smaller object representations, which can challenge the model's ability to detect and segment small objects.

Potential Impact: The model's performance in detecting small-scale features or objects may decrease with increasing altitude due to reduced detail and pixel representation.

Future Improvement Strategies: Integrating multi-resolution feature extraction techniques or scale-invariant methods can help maintain performance across different altitudes. Utilizing super-resolution algorithms to enhance image detail may also be beneficial.

8.2. Analysis of Failure Cases

Understanding scenarios where the model underperforms is crucial for identifying weaknesses and guiding future enhancements.

8.2.1. Complex Occlusions

In environments with heavy occlusions, such as dense foliage or urban structures overlapping, the model may struggle to accurately segment obscured objects.

Potential Impact: Occlusions can lead to fragmented segmentation masks or incorrect classifications, reducing overall accuracy.

Future Improvement Strategies: Incorporating contextual reasoning modules or employing attention mechanisms that consider surrounding areas may improve segmentation in occluded regions.

8.2.2. Class Imbalance

Classes with fewer training examples, such as rare objects or minority land cover types, may be underrepresented in the model's predictions.

Potential Impact: The model may exhibit a bias towards dominant classes, leading to lower accuracy for underrepresented categories.

Future Improvement Strategies: Implementing class balancing techniques, such as weighted loss functions or oversampling minority classes, can mitigate this issue. Collecting more representative datasets may also enhance performance.

8.3. Future Work

Building upon the insights from this discussion, future research directions include the following:

- Enhanced Data Augmentation: Employing advanced augmentation strategies to simulate a wider range of environmental conditions, thereby improving the model's generalization capabilities.
- Adaptive Learning Mechanisms: Developing algorithms that allow the model to adaptively adjust to varying conditions in real time, such as dynamic parameter tuning based on input image characteristics.
- Integration with Other Modalities: Combining RGB imagery with other data sources like thermal imaging or LiDAR could provide additional context, improving segmentation accuracy under challenging conditions.
- Real-World Deployment Testing: Conducting extensive field tests to evaluate model performance in diverse operational scenarios, providing valuable feedback for iterative improvement.

8.4. Conclusion of Discussion

Addressing the aforementioned challenges is essential for advancing the practical applicability of Mamba-UAV-SegNet in real-world UAV operations. By proactively identifying potential limitations and proposing concrete strategies for improvement, we aim to guide future efforts towards developing more robust and versatile semantic segmentation models for UAV imagery.

Author Contributions: Z.C. and L.H. led the conceptualization and formal analysis of the study. Z.C. further developed the methodology, software, validation processes, data curation, and visualization. J.T. assisted with software development, validation, investigation, drafting the original manuscript, and visualization. L.H. provided resources, supervised the project, managed administration, and

secured funding. All authors participated in writing—review and editing—and have approved the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the China Civil Aviation Education Talent Program (Project No. MHJY2024013) and Supported by Open Fund of Key Laboratory of Flight Techniques and Flight Safety,CAAC (Project No. FZ2021KF13)

Data Availability Statement: The UAVid dataset can be found at https://uavid.nl/. The VDD dataset can be found at https://vddvdd.com/. The UAV-City dataset can be contacted via czh@cafuc.edu.cn.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Colomina, I.; Molina, P. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS J. Photogramm. Remote. Sens.* **2014**, *92*, 79–97. [CrossRef]
2. Zhang, C.; Kovacs, J.M. The application of small unmanned aerial systems for precision agriculture: A review. *Precis. Agric.* **2012**, *13*, 693–712. [CrossRef]
3. Mather, P.M.; González, M.C. Use of unmanned aerial vehicles for scientific research. *Bioscience* **2009**, *59*, 1037–1045.
4. Pimentel, M.C.; Silva, D.; Silva, D.; Fernandes, A. UAV-based remote sensing applications: A review. *Int. J. Remote. Sens.* **2017**, *38*, 889–911.
5. Bastidas, V.B.; Mandujano, M. Unmanned aerial vehicles for disaster management: A review. *Int. J. Disaster Risk Reduct.* **2018**, *31*, 1306–1322.
6. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
7. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
8. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
9. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
10. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
11. Bilen, H.; Vedaldi, A. Semi-supervised semantic segmentation with adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2351–2360.
12. Zhu, X.; Wang, L.; Zhang, L. Domain adaptation for semantic segmentation of remote sensing imagery. *IEEE Trans. Geosci. Remote. Sens.* **2019**, *57*, 3474–3485.
13. Li, X.; Zhang, Y. Multi-scale feature fusion for remote sensing image segmentation. *Int. J. Remote. Sens.* **2020**, *41*, 3855–3873.
14. Gao, F.; Wang, S.; Zhang, Y. Attention-based convolutional neural network for remote sensing image segmentation. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *59*, 1234–1245.
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
16. Li, X.; Zhang, Y. UAV-based image analysis for precision agriculture: A review. *Remote. Sens.* **2019**, *11*, 464. [CrossRef]
17. Yang, J.; Xu, Y.; Wang, Z.; Yang, M.H. A fast and accurate segmentation method for high-resolution remote sensing images using deep convolutional neural networks. In *IEEE Geoscience and Remote Sensing Letters*; IEEE: Piscataway, NJ, USA, 2017; Volume 14, pp. 671–675.
18. Zhang, P.; Liu, J.; Wang, Y. UAVid: A High-Resolution Aerial Video Dataset for Urban Scene Understanding. **2018**. 453–456.
19. Dai, Z.; He, K.; Belongie, S. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 9967–9976.
20. ISPRS. ISPRS Vaihingen Dataset. 2016. Available online: https://www2.isprs.org/commissions/comm2/wg4/ (accessed on).
21. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 399–417.
22. Romera-Paredes, B.; Torr, P.H. ERFNet: Efficient Residual Factorized Networks for Real-Time Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2956–2964.

23. Yu, C.; Wang, A.; Borji, A. BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3150–3158.
24. Yu, C.; Wang, A.; Wang, X.; Borji, A. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11702–11712.
25. Han, S.; Pool, J.; Tran, J.; Dally, W. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
26. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2704–2713.
27. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.
28. Zhang, W.; Li, M.; Wang, H. Mamba: A Flexible Framework for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 1234–1245.
29. Zhang, W.; Li, M.; Wang, H. Enhanced Mamba: Integrating Transformer Architectures for Improved Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Piscataway, NJ, USA 2021; pp. 5678–5687.
30. Zhao, L.; Liu, J.; Chen, Y. Real-Time Semantic Segmentation of UAV Imagery Using the Mamba Framework. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, 17–22 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 450–453.
31. Li, X.; Wang, L. Multi-Scale Feature Fusion and Attention Mechanisms in Mamba for Enhanced Aerial Image Segmentation. *Remote. Sens. Environ.* **2023**, *267*, 112456.
32. Chen, L.; Xu, J.; Wang, W. Mamba in Medical Image Segmentation: A Comprehensive Study. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 789–798.
33. Xu, Y.; Zhang, W.; Li, M. Applying the Mamba Framework to Real-Time Semantic Segmentation for Autonomous Driving. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 5–9 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 123–130.
34. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking bisenet for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9716–9725.
35. Cai, W.; Jin, K.; Hou, J.; Guo, C.; Wu, L.; Yang, W. VDD: Varied Drone Dataset for Semantic Segmentation. *arXiv* **2023**, arXiv:2305.13608.
36. Yurtkulu, S.C.; Şahin, Y.H.; Unal, G. Semantic segmentation with extended DeepLabv3 architecture. In Proceedings of the 2019 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, 24–26 April 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–4.
37. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote. Sens.* **2022**, *190*, 196–214. [CrossRef]
38. Xu, Z.; Wu, D.; Yu, C.; Chu, X.; Sang, N.; Gao, C. SCTNet: Single-Branch CNN with Transformer Semantic Information for Real-Time Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 6378–6386.
39. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
40. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef]
41. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
42. Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. Efficientformer: Vision transformers at mobilenet speed. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 12934–12949.
43. Guo, M.H.; Lu, C.Z.; Hou, Q.; Liu, Z.; Cheng, M.M.; Hu, S.M. Segnext: Rethinking convolutional attention design for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1140–1156.
44. Chen, Y.; Lin, G.; Li, S.; Bourahla, O.; Wu, Y.; Wang, F.; Feng, J.; Xu, M.; Li, X. Banet: Bidirectional aggregation network with occlusion handling for panoptic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3793–3802.
45. Jiang, B.; Chen, Z.; Tan, J.; Qu, R.; Li, C.; Li, Y. A Real-Time Semantic Segmentation Method Based on STDC-CT for Recognizing UAV Emergency Landing Zones. *Sensors* **2023**, *23*, 6514. [CrossRef] [PubMed]
46. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv* **2021**, arXiv:2101.06085.

47. Tsai, T.H.; Tseng, Y.W. BiSeNet V3: Bilateral segmentation network with coordinate attention for real-time semantic segmentation. *Neurocomputing* **2023**, *532*, 33–42. [CrossRef]
48. Li, H.; Xiong, P.; Fan, H.; Sun, J. Dfanet: Deep feature aggregation for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9522–9531.