






**Please cite the Published Version**

Ye, Hong, Cai, Jijing , Deng, Jiangtao , Wang, Xiaodong , Bashir, Ali Kashif , Fang, Kai  and Wang, Wei (2025) Efficient Machine Learning-Based Semantic Segmentation Algorithm for Consumer-Grade UAV Remote Sensing. IEEE Transactions on Consumer Electronics. pp. 1-14. ISSN 0098-3063

**DOI:** <https://doi.org/10.1109/TCE.2025.3596058>

**Publisher:** Institute of Electrical and Electronics Engineers (IEEE)

**Version:** Accepted Version

**Downloaded from:** <https://e-space.mmu.ac.uk/641813/>

**Usage rights:**  [Creative Commons: Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)

**Additional Information:** This is an author accepted manuscript of an article published in IEEE Transactions on Consumer Electronics. This version is deposited with a Creative Commons Attribution 4.0 licence [<https://creativecommons.org/licenses/by/4.0/>]. The version of record can be found on the publisher's website.

**Enquiries:**

If you have questions about this document, contact [openresearch@mmu.ac.uk](mailto:openresearch@mmu.ac.uk). Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

# Efficient Machine Learning-Based Semantic Segmentation Algorithm for Consumer-Grade UAV Remote Sensing

Hong Ye, Jijing Cai, Jiangtao Deng, Xiaodong Wang, Ali Kashif Bashir, Kai Fang, and Wei Wang

**Abstract**—The computational complexity of the Transformer model grows quadratically with input sequence length. This causes a sharp increase in computational cost and memory consumption for high-resolution remote sensing images. Consequently, its application in consumer-grade unmanned aerial vehicle remote sensing is limited. To address this issue, we propose an efficient machine learning-based semantic segmentation algorithm (EMLSSA). First, EMLSSA incorporates the hash clustering attention (HCAAttention) mechanism. It employs the locality-sensitive hashing (LSH) algorithm to group similar features into hash buckets, enabling dynamic token clustering. Subsequently, tokens in the same hash bucket are aggregated by weighted summation. This compresses features and reduces the computational complexity of self-attention. Second, EMLSSA incorporates the frequency multi-layer perceptron (FMLP) mechanism. It combines frequency and spatial domain information, enhancing the ability of the Transformer to perceive local features. Experimental results show that EMLSSA-B4 reduces computational cost by 11.7% on FLAME, PWD, EarthVQA, and Potsdam datasets. Furthermore, it maintains comparable segmentation performance to SegFormer-B4.

**Index Terms**—Machine Learning, Consumer-Grade UAV, Remote Sensing, Locally Sensitive Hashing, Dynamic Clustering, Frequency Multi-Layer Perceptron.

## I. INTRODUCTION

Machine learning-driven semantic segmentation algorithms endow unmanned aerial vehicle (UAV) systems with robust environmental perception capabilities. Through pixel-level image classification, UAV systems can identify and segment various ground features, enabling a detailed understanding of complex scenes [1]. The use of consumer-grade UAVs can significantly improve the efficiency of land resource monitoring, agricultural management, and post-disaster rescue. However,

high-accuracy semantic segmentation algorithms are often challenging to deploy on resource-constrained UAV systems due to their high computational complexity [2]. Therefore, developing efficient semantic segmentation algorithms is of great importance for empowering consumer-grade UAV systems.

Convolutional neural networks (CNNs) have made significant progress in semantic segmentation owing to their multi-scale receptive fields and powerful capabilities in modeling local context information [3]. However, the local receptive fields of CNNs limit their ability to capture global context information and long-range dependencies. This limitation can lead to confusion in pixel-level classification of complex scenes. To address the shortcomings of CNNs in global context modeling, researchers have introduced the Vision Transformer (ViT) for semantic segmentation tasks. The global self-attention mechanism of ViT helps to effectively capture long-range dependencies in images. Specifically, ViT first divides the input image into a series of fixed-size patches. It flattens these patches into a sequence and uses them as input to the Transformer encoder. Each image patch is subjected to linear transformation to generate an embedding vector, which is combined with positional encoding to retain spatial information from the original image [4]. Subsequently, the Transformer encoder processes these embedding vectors through a multihead self-attention mechanism and feed-forward neural networks. This allows each image patch to interact with all other patches in the image [5]. This global modeling mechanism enables ViT to establish associations between pixels across the entire image, allowing for a comprehensive understanding of the overall structure and contextual information in complex scenes [6].

Global contextual information is used to model the semantic dependencies between long-range pixels in an image, thereby improving overall semantic consistency and the discriminative ability for object categories. Local contextual information, on the other hand, focuses on capturing spatial relationships between adjacent regions in the image, which is crucial for accurately modeling object boundaries and preserving texture details. Although Transformer models, with their self-attention mechanisms, demonstrate significant advantages in global information modeling, they still exhibit certain limitations in capturing local details. This may result in issues such as blurred object boundaries and missing fine-grained details in semantic segmentation outcomes [7]. Fig. 1 illustrates global and local contextual information. The self-attention mechanism of ViT suffers from quadratic computational complexity growth. Its computational cost scales quadratically

Hong Ye is with the College of Electrical and Information Engineering, Quzhou University, Quzhou 324000, China. (Email: yeh@qzc.edu.cn)

Jijing Cai, Jiangtao Deng, and Kai Fang are with the College of Mathematics and Computer Science, Zhejiang A&F University, Hangzhou 311300, China. (Email: Jijingcai19@gmail.com, dengjiangtao07@gmail.com, and Kaifang@ieee.org)

Xiaodong Wang is with the Zhejiang Jiuzhou Water Control Technology Co., Ltd, Quzhou 324000, China (Email: wangycjq@foxmail.com)

Ali Kashif Bashir is with the Department of Computing and Mathematics, Manchester Metropolitan University, UK; Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, India. (Email: dr.alikashif.b@ieee.org)

Wei Wang is with the Guangdong-Hong Kong-Macao Joint Laboratory for Emotion Intelligence and Pervasive Computing, Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, China, and also with the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China. (Email: ehomewang@ieee.org)

Corresponding author: Xiaodong Wang

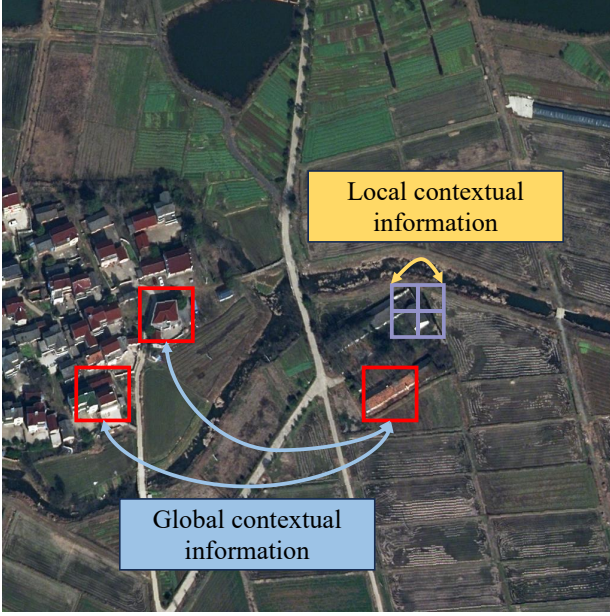


Fig. 1. Illustration of global and local contextual information. Local contextual information is modeled through convolution (yellow). Global contextual information is modeled by random window dependencies (blue). Double-headed arrows represent information exchange.

with the input sequence length, resulting in a sharp increase in computational costs when processing high-resolution images. In addition, memory consumption increases significantly, further limiting its practical applicability in remote sensing scenarios[8]. Therefore, how to effectively integrate local detail features and global contextual information while ensuring computational efficiency has become a critical bottleneck that must be addressed in the field of semantic segmentation.

In light of the above, we propose a new semantic segmentation method that balances computational efficiency and feature representation capability, aiming to enhance remote sensing scene image analysis. Inspired by the SegFormer algorithm, we propose an Efficient Machine Learning-Based Semantic Segmentation Algorithm. This algorithm employs a hash clustering attention (HCAAttention) mechanism, which significantly reduces the computational complexity of the self-attention mechanism by reducing the number of input tokens. Simultaneously, a frequency domain learning module is introduced into the multilayer perceptron. This enhancement improves the ability of the model to extract local detailed information from the input sequence and further boosts its feature representation capability. Our main contributions are as follows:

(1) We propose a lightweight Transformer architecture suitable for remote sensing image semantic segmentation. This model can handle small objects, occluded objects, and complex ground features.

(2) To address the high quadratic complexity issue in Transformer models, we propose the HashAttention mechanism. It maps semantically similar high-dimensional vectors to the same hash bucket using locality-sensitive hashing. By combining this with a token aggregation mechanism to compress vectors within the same hash bucket, we significantly reduce

the complexity of subsequent self-attention calculations.

(3) To overcome the limitation of Transformer models in extracting local details, we introduce the frequency multilayer perceptron module. The module applies feature weighting to local image patches in the frequency domain, assigning higher weights to critical local features. Consequently, the sensitivity of the model to local information is significantly enhanced.

## II. RELATED WORK

### A. Semantic Segmentation Based on Transformer

Semantic segmentation is a fundamental task in computer vision. It aims to classify each pixel in an image, achieving pixel-level understanding of the image content. Recently, ViT has been widely applied for semantic segmentation tasks. Its advantages in modeling global contextual information have yielded significant results on multiple benchmark datasets. Xie et al. [9] proposed the SegFormer algorithm to address the issues of high computational complexity and poor cross-resolution robustness of Transformer architectures in semantic segmentation. This algorithm employs a position-encoding-free hierarchical Transformer encoder and a lightweight multilayer perceptron decoder designed in collaboration. By employing a multiscale feature fusion mechanism, it can achieve better segmentation results with fewer parameters across multiple benchmark datasets. Gu et al. [10] proposed the HRViT algorithm to address the problem of insufficient multiscale feature representation capability of ViT in dense prediction tasks. This algorithm deeply integrates a high-resolution multi-branch architecture with Transformer, significantly enhancing the ability of ViT to perceive spatial semantic information. Meanwhile, HRViT adopts a heterogeneous branch design and lightweight linear layers to further improve the efficiency and expressive power of the model. To address the problem that convolutional operations struggle to effectively model global context information, He et al. [11] proposed a semantic segmentation algorithm called ST-UNet. This algorithm uses a dual-encoder architecture with parallel Swin Transformers and CNNs. It employs a spatial interaction module to establish pixel-level associations, enhancing the feature representation of occluded objects. Additionally, a feature compression module is used to optimize detail loss during the Transformer's downsampling process. To address the lack of high-level feature structural information in Transformer decoders for semantic segmentation, Shim et al. [12] proposed the FeedFormer algorithm, which aims to improve semantic segmentation performance. This algorithm constructs a cross-level interaction mechanism by employing high-level features as queries and combining low-level features as keys and values. This mechanism enables high-level semantic features to effectively utilize the fine structural information contained in the lowest-level features, thereby significantly enhancing the completeness of high-level semantic features. To tackle the challenge of balancing global context modeling and computational efficiency in MetaFormer architectures for semantic segmentation tasks, Kang et al. [13] introduced the MetaSeg algorithm. This algorithm employs a CNN backbone network based on MetaFormer blocks and a novel self-attention decoder

architecture. By employing a Channel Reduction Attention module, queries and keys are compressed to one dimension. This reduces computational complexity while enabling global context modeling and maintaining efficiency. Chen et al. [14] proposed a Transformer segmentation algorithm with a hybrid attention mechanism. Thus, it addresses the challenge of convolutional local features struggling to capture global contextual information in remote sensing image semantic segmentation. The algorithm employs ResNet50 as an encoder to extract local image features. The decoder consists of a Channel-Spatial Transformer (CST) module, designed to capture global contextual information. Within the CST module, an adaptive channel re-weighting mechanism is introduced to dynamically enhance dependencies between different channels, thereby improving feature representation capabilities. Chen et al. [15] proposed the Hierarchical Spatial Perception Transformer to address the lack of spatial reasoning and attention drift in Transformer-based semantic segmentation for dynamic driving scenarios. This method consists of two key components. The Spatial Depth Perception Auxiliary Network performs multiscale feature extraction and multilayer depth map prediction. The Hierarchical Pyramid Transformer Network uses depth estimation as learnable position embeddings, forming spatially correlated semantic representations and generating global contextual information.

### B. Dynamic Token Generation

The ViT architecture has demonstrated remarkable performance in various visual tasks. However, its encoder, when performing self-attention calculations, experiences a quadratic growth in computational complexity and parameter count with the increase in input sequence length. This significant computational burden poses a challenge to real-time visual tasks, making it difficult to satisfy real-time requirements. To address the problem of low computational efficiency caused by redundant tokens in visual transformers, Rao et al. [16] proposed a dynamic token sparsification framework. The algorithm employs a lightweight prediction module and a hierarchical pruning mechanism, using an attention masking strategy to block the interaction of redundant tokens for differentiated pruning. Simultaneously, it optimizes the hierarchical token importance scores in an end-to-end manner, enabling the model to dynamically retain critical token subsets based on input content and significantly reduce computational complexity. To address high computational costs and the difficulty of adapting a fixed number of tokens to different input images in visual transformers, Fayyaz et al. [17] proposed an adaptive token sampler algorithm, which uses a differentiable dynamic token scoring mechanism for adaptive downsampling. It evaluates and selects highly significant tokens layer by layer, transforming the traditional Transformer into a variable-length token processing architecture. Grainger et al. [18] proposed the PaCa-ViT algorithm to address the inherent quadratic computational complexity bottleneck of traditional block-to-block attention mechanisms in ViTs. The core of this algorithm is the introduction of a key-value pair attention mechanism based on cluster centers. Through an end-to-end joint optimization strategy, it synchronously adjusts cluster centers

and attention maps, thereby reducing the original quadratic complexity to linear complexity. To address the issue of limited cross-task transferability exhibited by traditional Transformer models in heterogeneous visual tasks, Liang et al. [19] proposed a general visual algorithm called ClusterFormer. This algorithm innovatively employs a recurrent cross-attention clustering mechanism and a dual feature dispatch architecture. The recurrent cross-attention clustering mechanism enhances representation learning by dynamically updating cluster centers. Meanwhile, the feature dispatch architecture reorganizes features based on similarity metrics, creating an interpretable unified visual modeling paradigm. Zeng et al. [20] proposed the TCFormer architecture to address the problem of the fixed grid tokenization methods of the traditional Transformer models, ignoring the inherent semantic associations of images. TCFormer can effectively capture contextual information and long-range dependencies in images by adaptively aggregating regions with similar semantic information into shared visual tokens through a semantic-aware process. Chao et al. [21] proposed a framework called multimodal alignment-guided dynamic token pruning (MADTP) to address the high computational cost in ViT. The core of this method is the Multimodality Alignment Guidance module, which aligns features of the same semantic concepts from different modalities. This ensures that the pruned tokens are irrelevant across all modalities. Additionally, the method includes a dynamic token pruning module, which adaptively adjusts the token compression rate for each layer based on different input samples. To reduce the high computational demands of Transformer models for UAV tracking, Du et al. [22] introduced a dynamic token sampling method that improves visual representation by scoring and dynamically selecting tokens, thereby allowing for a flexible token count.

Although current research has significantly reduced the computational complexity of Transformer models, establishing a balance between lightweight nature and optimum performance remains a challenge. This significantly limits their extensive use in analyzing remote sensing images. Therefore, this study introduces a new efficient Transformer model for remote sensing images. This model employs a hashing clustering algorithm to improve the traditional self-attention mechanism. Thus, similar vectors can be mapped to the same hash buckets and subsequently grouped together. This significantly reduces the computational cost of the model while maintaining a strong global modeling ability of self-attention. Additionally, the model incorporates frequency domain operations in its multilayer perceptron. This aims to enhance the ability of the model to precisely extract local features. With the combined advantages of global and local features, this model can significantly improve the performance of remote sensing image analysis.

## III. EMLSSA

Fig. 2(a) shows the model structure of EMLSSA, which consists of an encoder and a decoder. In the encoder, the OverlapPatchEmbed module first transforms the input feature map into image patches with overlapping regions. These

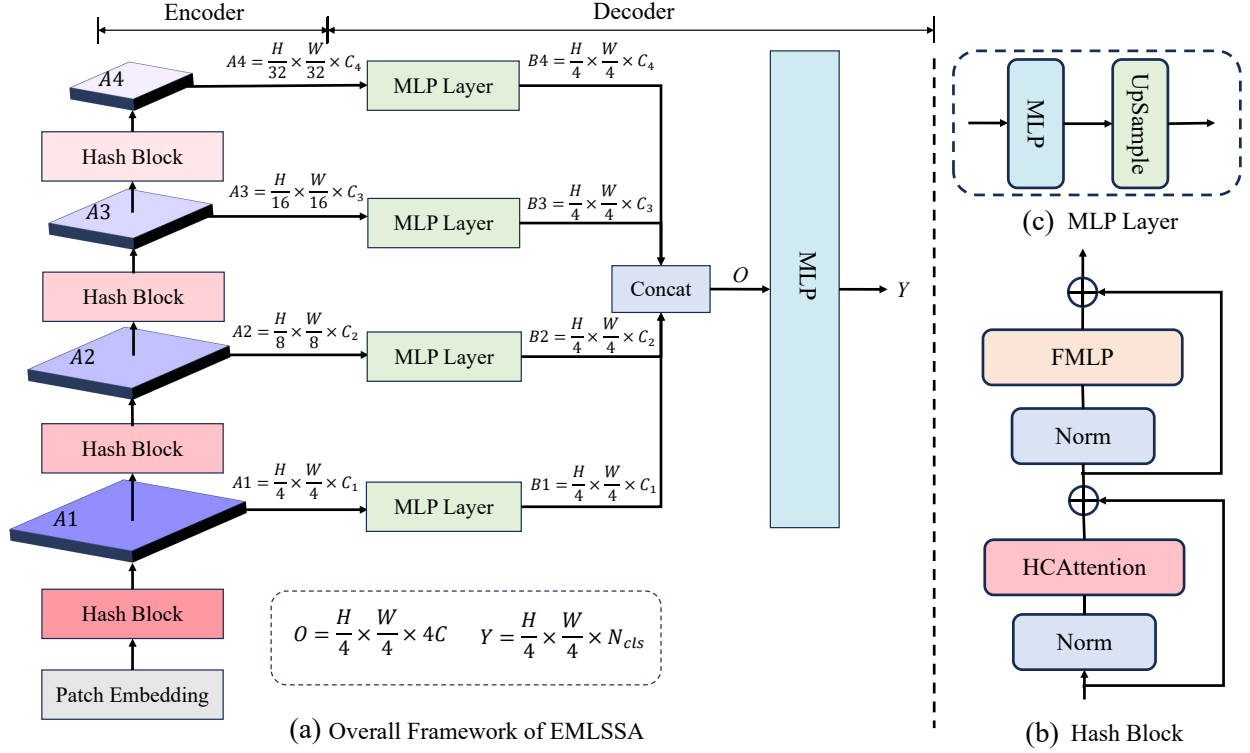


Fig. 2. (a) EMLSSA network architecture. (b) Hash Block module. (c) MLP Layer module

patches are then embedded into a high-dimensional space to capture rich feature information. Subsequently, the global and local contextual information of the image is deeply extracted through stacked four-layer HashBlock modules. The resolution of the feature map gradually decreases during this process to achieve information abstraction and compression. The model structure of the HashBlock module is shown in Fig. 2(b). The principle of the HashBlock module is to efficiently divide the input sequence into multiple hash buckets using the LSH algorithm. It performs weighted aggregation on tokens within each bucket, reducing the number of tokens for processing. This lowers computational complexity while retaining key information. Additionally, the frequency-domain multilayer perceptron module (FMLP) in HashBlock captures local information by learning in the frequency domain. This enhances the representation capability of the model. In the decoder part, the MLP Layer module gradually restores the feature map to the original resolution through multilayer perceptron and then performs upsampling operations, thereby effectively reconstructing feature information. Its module structure is shown in Fig. 2(c). After upsampling, the feature map undergoes a linear mapping through the MLP, and the final prediction result is output.

#### A. HCAAttention

As stated earlier, for high-resolution remote sensing images, traditional Transformer models face a quadratic increase in computational complexity with input sequence length. This considerably limits their practical application on resource-constrained drone platforms. Although existing efficient attention mechanisms can effectively reduce the computational

load of the models, they struggle to balance computational efficiency with global modeling capabilities [23]. Therefore, we propose an innovative HCAAttention mechanism aimed to address the computational bottleneck in remote sensing image semantic segmentation tasks. This mechanism preserves global modeling capabilities while significantly reducing computational complexity. Furthermore, it enables precise capture of long-range dependencies between ground objects in remote sensing images, achieving high-precision semantic segmentation.

Fig. 3 shows the model structure of the HCAAttention self-attention mechanism, which consists of three core components: hash operation [24], merge aggregation process, and attention calculation. First, the HCAAttention module performs a hash operation on the input sequence, mapping semantically similar high-dimensional vectors in the input to the same hash bucket. Subsequently, the merge aggregation operation weights and combines similar vectors within the same hash bucket, generating representative feature representations. Finally, these compressed token features are used for attention calculation to further extract global contextual information. In this manner, the representation of the feature map is effectively compressed, thereby significantly reducing the complexity of subsequent attention calculations.

**Computational Complexity Analysis.** The self-attention mechanism of traditional Transformer models has a computational complexity of  $O(N^2C)$  [25] when the input sequence length is  $N$  and the feature dimension is  $C$ . The HCAAttention mechanism, utilizing the idea of locality-sensitive hashing, maps the input  $N$  tokens to  $K$  hash buckets through random projection, achieving dynamic clustering. The formula for

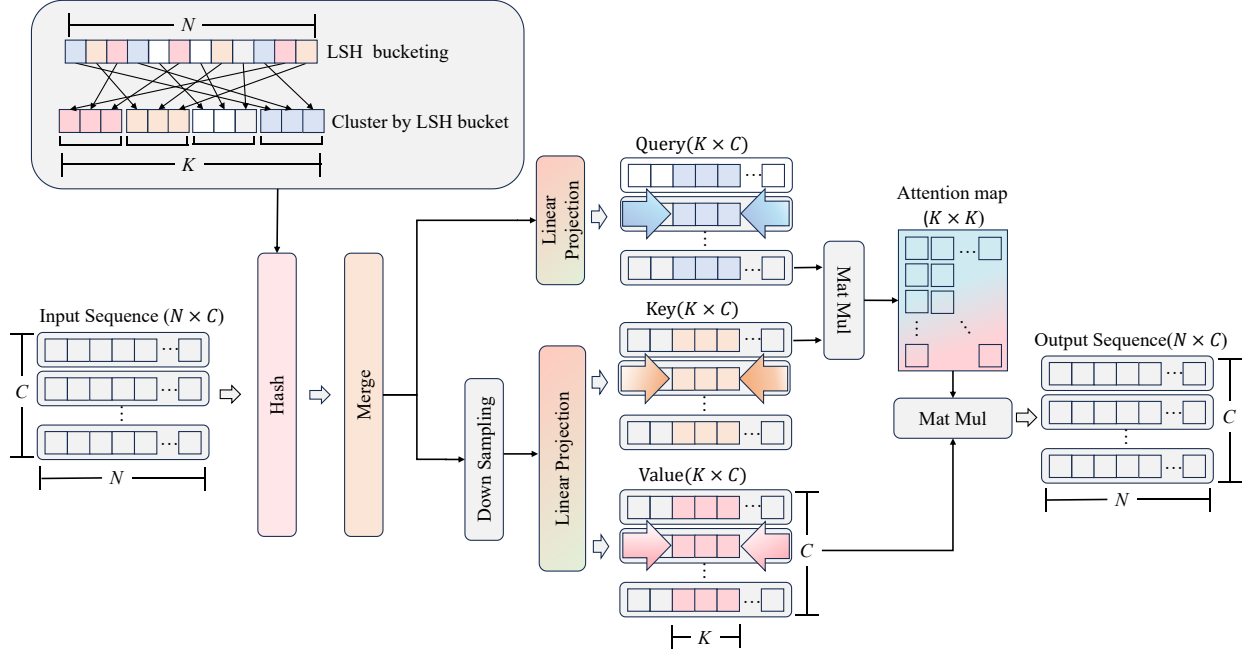


Fig. 3. Structure of HCAAttention module.

calculating the number of hash buckets can be expressed as

$$K = \max(\text{Ceil}(N * Sr), 1) \quad (1)$$

where  $\max$  represents the maximum value,  $\text{Ceil}$  represents rounding up, and  $Sr$  represents the sampling rate,  $K \ll N$ .

By performing token aggregation operations to weighted aggregate, the tokens within the same hash bucket are fused into a representative token, thereby effectively reducing the complexity of subsequent attention calculations. After token hashing and token aggregation operations, the number of input tokens changes from  $N$  to  $K$  representative tokens, and the subsequent self-attention calculation complexity becomes  $O(K^2C)$ .

**Token Hashing.** Token hashing maps semantically similar high-dimensional features to the same hash bucket and generates hash bucket indices for subsequent feature aggregation. Let the input vector be  $V \in \mathbb{R}^{B \times N \times D}$ , where  $B$  is the batch size,  $N$  is the sequence length, and  $D$  is the feature dimension. A random rotation matrix  $R$  is generated, following a standard normal distribution.

$$R \sim N(0, 1) \in \mathbb{R}^{B \times D \times H \times K/2} \quad (2)$$

where  $H$  is the number of hash iterations and  $K$  is the number of hash buckets. The product of the input vector  $V$  and the random rotation matrix  $R$  is computed using Einstein summation convention, yielding the projected vector  $P$ .

$$P_{B,H,N,K} = \sum_{d=1}^D V_{B,N,D} \cdot R_{B,D,H,K/2} \quad (3)$$

Through the projection operation, vectors in high-dimensional space are mapped into low-dimensional hash buckets, such that similar vectors have a higher probability

of being mapped to the same hash bucket. To eliminate projection direction bias and enhance feature discriminability, the projected vector is concatenated with its negative value, constructing a symmetric projection matrix  $Q$ .

$$Q = [P \parallel -P] \in \mathbb{R}^{B \times H \times N \times K/2} \quad (4)$$

where  $\parallel$  denotes the concatenation operation along the last dimension.

Based on the above derivation, the specific form of the hash function is defined as:

$$H(V) = \arg\max_{k \in [0, K)} [V \cdot R_j, -V \cdot R_j] \quad (5)$$

where  $V \cdot R_j$  represents the projection value of the input vector  $V$  onto the direction  $R_j$ , which measures the similarity between  $V$  and  $R_j$ . A larger dot product indicates higher similarity. The  $\arg\max$  operation selects the index of the direction with the highest projection value among all directions, which is then used as the hash bucket index. It is formally defined as:

$$\arg\max f(x) = \{x^* \in Z \mid f(x^*) \geq f(x), \forall x \in Z\} \quad (6)$$

where  $Z$  represents the domain of the input  $x$ ,  $f(x)$  is the objective function, and  $x^*$  is the input that maximizes  $f(x)$ . The condition  $f(x^*) \geq f(x), \forall x \in Z$  ensures that  $x$  corresponds to the global maximum.

Within the range of  $K$  hash buckets, a high-dimensional input vector  $V$  is mapped onto  $R_j$  Gaussian random directions and their opposite directions through Gaussian random rotation, resulting in a set of projection values. The direction with the maximum projection value is selected via the  $\arg\max$  operation, and the corresponding index is assigned as the hash bucket number for the vector. Since similar vectors



tend to exhibit similar maximum response directions under random projections, they are more likely to be mapped to the same hash bucket. This property enables efficient clustering of similar tokens and enhances the effectiveness of the hashing mechanism. After token hashing, similar tokens share the same hash bucket index, and a subsequent token aggregation operation is applied to compress the feature representation by aggregating input tokens based on their respective bucket indices.

**Token Aggregation.** The purpose of the token aggregation operation is to merge semantically similar tokens into a single representative token through weighted fusion. Input features are clustered based on the hash bucket index, number of clusters, and batch index. This generates a global index, ensuring unique cluster numbers across different batches. The global index generation process can be expressed as

$$idx = idx\_buckets + idx\_batch \times K \in \mathbb{R}^{B \times N} \quad (7)$$

where  $idx\_batch$  represents the batch index,  $idx\_buckets$  represents the hash bucket index, and  $K$  represents the number of hash buckets for clustering. Subsequently, the sum of all weights within each hash bucket is computed using the global index. The weight of each token is then divided by the total weight of its bucket, producing the normalized weight  $w_j$ .

$$w_j = \frac{e^{p_j}}{\sum_{j \in C_i} e^{p_j}} \quad (8)$$

where  $p_j$  represents the importance score of the  $j$ -th token, obtained through linear mapping, measuring the importance of the token.  $e^{p_j}$  represents the exponential transformation of the importance score, assigning greater weight to tokens with higher importance scores, ensuring that information-rich tokens have a greater impact on the merged result.  $C_i$  represents the new token representation after weighted merging.

The token features within each hash bucket are weighted summed according to the normalized weights, resulting in the representative token feature  $y_i$ .

$$y_i = \sum_{j \in C_i} w_j x_j = \sum_{j \in C_i} \left( \frac{e^{p_j}}{\sum_{j \in C_i} e^{p_j}} \cdot x_j \right) \quad (9)$$

where  $x_j$  is the feature vector for the  $j$ -th token, capturing its semantic information.

The token aggregation operation performs weighted fusion of tokens based on their semantic similarity and importance scores, thereby generating representative dynamic tokens. In remote sensing images, ground object categories are complex and unevenly distributed; however, dynamic tokens can adaptively focus on important regions, significantly improving the ability of the model to perceive detailed features [26]. Furthermore, background regions in remote sensing images usually occupy a large proportion; however, their contribution to segmentation tasks is relatively limited [27]. Therefore, by aggregating tokens from background regions, computational redundancy can be effectively reduced while ensuring that tokens from key regions are retained.

After token hashing and token aggregation operations, the number of tokens in the input sequence is significantly reduced, and the semantic information of the remote sensing image is effectively compressed. The aggregated tokens, serving as queries, keys, and values, are used for subsequent self-attention calculations, expressed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (10)$$

where  $d_k$  represents the dimension scaling factor used to prevent excessive gradients,  $\text{Softmax}$  represents the normalization operation, and  $QK^T$  represents the correlation score used to calculate the similarity between query and key.

### B. Frequency Domain Multilayer Perceptron

Traditional multilayer perceptrons are ideal at capturing global features when processing image data but lack the ability to perceive local detailed information contained in images [28]. However, in remote sensing image analysis, small targets often have limited feature information. Relying solely on global features makes it challenging to model these targets accurately, which can result in the loss of local features and negatively impact recognition accuracy [29]. Therefore, we introduce a frequency domain modulation mechanism into the multilayer perceptron structure to allow the model utilize local features and fully explore the frequency domain information of images, thereby improving the overall feature representation ability of the model [30].

The structure of the FMLP module is illustrated in Fig. 4. The input sequence  $X$  is first linearly projected and then reshaped into the spatial domain. In the spatial domain, a Patch Folding operation divides the feature map into multiple  $P \times P$  image patches, each represented as  $X_{\text{patch}} \in \mathbb{R}^{B \times C \times \frac{H}{P} \times \frac{W}{P}}$ . Subsequently, each image patch undergoes a 2D FFT operation to be transformed into the frequency domain [31]. In the frequency domain, learnable modulation weights  $W$  are applied to adaptively adjust the high- and low-frequency features of each patch. The output  $W$  can be represented as  $W \in \mathbb{R}^{C \times 1 \times 1 \times P \times (P/2+1)}$ , where each channel corresponds to an independent frequency-domain filter. The frequency domain weighting process can be expressed as:

$$Y = \mathcal{F}(P(X)) \odot W \quad (11)$$

where  $Y$  represents the frequency-domain weighted result,  $\mathcal{F}$  denotes the 2D Fourier transform, and  $P$  stands for the patch folding operation applied to the input features.  $\odot$  indicates element-wise multiplication, meaning that frequency-domain modulation is performed on each image patch.

After frequency-domain modulation, the high-frequency contour information of each local image patch is enhanced while the low-frequency detail information is preserved, significantly improving the Transformer model's ability to perceive local detail features. Subsequently, the feature information is transformed back to the spatial domain via a 2D inverse FFT (IFFT), and patch reconstruction techniques are used to restore the image features. Finally, depthwise separable convolution

is applied to further capture local contextual information, followed by activation and linear mapping to output the enhanced feature representation.

$$\begin{aligned} X_f &= P^{-1} (\mathcal{F}^{-1}(Y)) \\ Y &= L(\text{GELU}(\text{DW}(X_f))) \end{aligned} \quad (12)$$

where  $X_f$  represents the features processed in the frequency domain.  $P^{-1}$  represents patch unfolding, restoring the original feature map,  $\mathcal{F}^{-1}$  denotes the inverse Fourier transform.  $DW$  stands for depthwise separable convolution, and  $L$  indicates the linear layer.

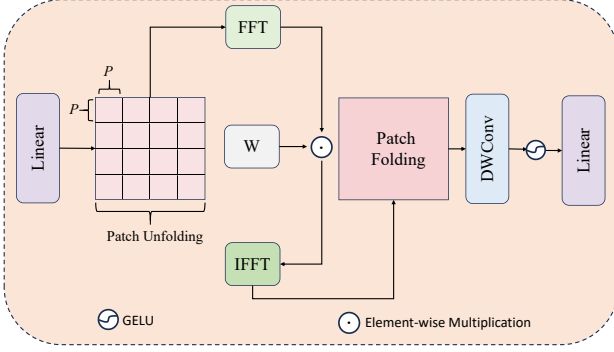


Fig. 4. Structure of FMLP module.

The FMLP module introduces a learnable frequency-domain weighting mechanism to adaptively modulate different frequency components within image features. During training, the module learns the frequency distribution characteristics of image patches, enabling it to enhance high-frequency details while preserving low-frequency structural information. Small objects in remote sensing images often exhibit prominent high-frequency features and sharp edge transitions. The FMLP module, through its frequency-domain enhancement mechanism, can adaptively strengthen high-frequency responses, enabling the model to maintain strong small-object perception even in complex backgrounds. This contributes to improved detection accuracy and edge localization performance in remote sensing imagery.

#### IV. EXPERIMENTS AND RESULTS

##### A. Experimental setup

In this study, we comprehensively evaluated the performance of the EMLSSA model by conducting experiments on three publicly available datasets and one self-collected dataset. The datasets used are as follows:

(1) FLAME Forest Fire Dataset [32]. The dataset consists of 2003 high-resolution fire images, which were acquired by drones during prescribed burning in the pine forest area of Arizona, USA. The RGB images in the dataset provide rich visual information, offering strong support for accurate segmentation of fire areas [33]. However, owing to the dynamic evolution and irregular morphological changes of fires, accurate segmentation of fire areas remains a challenge. The introduction of this dataset not only provides key data support for forest fire monitoring but also further promotes

the application of drone technology in disaster emergency response.

(2) PWD Dataset. We collected 1,106 high-resolution images of infected pine forests within four town-level areas in Longyou County, Quzhou City, Zhejiang Province, China and then constructed a precise segmentation dataset targeting the pathological features of pine wilt disease. The study area is approximately 8 square kilometers, covering typical Masson pine forests and broadleaf evergreen forests, with sampling points mainly distributed in pine forest areas with a high degree of pine wilt disease infection [34, 35]. Specific sampling sites involve six representative forest lands, including Hengshan Town, Shifo Town, and Zhaxi Town, to ensure regional representativeness and diversity of the data. Data annotation was meticulously processed using the X-AnyLabeling semantic segmentation tool and calibrated by experts in pest and disease prevention and control to ensure the accuracy and scientific nature of the annotations. This dataset is useful for training semantic segmentation algorithms. Furthermore, it serves as a reliable benchmark for evaluating model generalization in complex backgrounds and various disease distribution conditions. The study area and sampling sites are shown in Fig. 5. In this figure, red sampling points represent the training set, while yellow points indicate the validation and test sets.

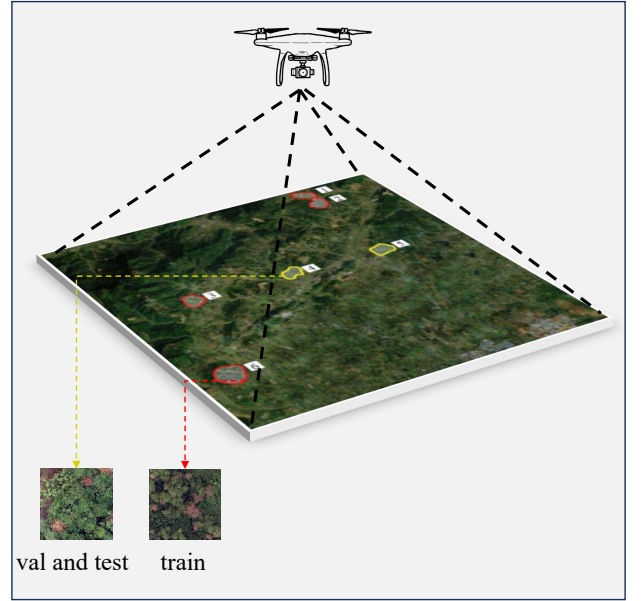


Fig. 5. Study area and sampling site.

(3) EarthVQA Land Cover Dataset [36]. The dataset consists of 6,000 high-resolution remote sensing images, collected from Nanjing, Changzhou, and Wuhan, China, aiming to explore the differences between urban and rural geographical environments. The EarthVQA dataset includes seven land cover categories, namely, buildings, roads, water bodies, wasteland, forests, agriculture, and playgrounds, providing high-quality remote sensing data support for multi-category land cover classification tasks. This dataset presents several challenges for researchers. Remote sensing images contain target objects



of different scales, requiring models to have strong multiscale information processing capabilities to adapt to the morphological and scale changes of land cover. Meanwhile, there are significant differences in the category distribution between urban and rural scenes. Models need strong generalization capabilities to manage changes in ground object features across different environments.

(4) Potsdam Dataset [37]. This dataset is a benchmark dataset, specifically designed for remote sensing image analysis tasks. It was acquired in Potsdam, Germany, through a high-precision aerial platform, and contains 38 orthophoto tile slices with a ground sampling distance of 5 cm: each image with a size of 6000×6000 pixels. We processed the Potsdam dataset to generate 3804 high-resolution images. The dataset mainly covers targets such as buildings, roads, trees, low vegetation, and cars. The dataset contains diverse scenes and object categories, which can effectively test the generalization ability of algorithms in complex environments. Additionally, there is visual similarity between low vegetation and tree categories, making it difficult for the model to distinguish between them. This challenge helps assess the ability of the model to recognize subtle differences.

**Implementation Details:** We developed the EMLSSA semantic segmentation algorithm based on the mmsegmentation framework. In addition, we trained it for 80k iterations on four remote sensing datasets using an RTX 4090 GPU. This ensured full convergence in an efficient computing environment. The training parameters and experimental environment during the experiment are specified in Tables I and II, respectively.

TABLE I  
TRAINING PARAMETERS

Training Parameters	Details
Iterations	80k
Input size	512 × 512
Batch size	4
Workers	8
Optimizer	SGD
Learning rate	0.01
Weight decay	0.0005

TABLE II  
EXPERIMENTAL ENVIRONMENT

Experimental Environment	Details
Operating System	Ubuntu 20.04
Develop Framework	MMsegmentation
MMCV	1.7.2
Deep Learning Framework	PyTorch 1.11.0
GPU	NVIDIA GeForce RTX 4090

During the training process, the model is pre-trained on ImageNet-1K. The primary evaluation metric used is mIoU (mean intersection over union), which comprehensively measures the segmentation accuracy across different categories. This ensures the performance stability and robustness of the

model in remote sensing image semantic segmentation tasks. The calculation formula for mIoU is as follows.

$$mIoU = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FP_i + FN_i} \quad (13)$$

where  $N$  represents the number of categories in the remote sensing image,  $TP_i$  represents the number of correctly predicted pixels for category  $i$ ,  $FP_i$  represents the number of false positive pixels for category  $i$ , and  $FN_i$  represents the number of false negative pixels for category  $i$ .

### B. Ablation Experiment

To verify the effectiveness of each module of the proposed EMLSSA semantic segmentation algorithm, we performed ablation analysis using the FLAME Forest Fire Dataset. Through this experiment, we evaluated the performance of EMLSSA under different module combinations to analyze its adaptability and advantages in remote sensing image scenarios. In Table III, we sequentially added the HCAAttention module and the FMLP module to the SegFormer model to demonstrate the performance gain of different module optimizations to the model.

First, the HCAAttention mechanism is introduced to achieve significant model compression through a hash-driven dynamic clustering strategy. Specifically, the computational load of the model is reduced by 11.7%, while the mIoU value is increased by 0.51%. The HCAAttention module utilizes a locality-sensitive hashing algorithm to map high-dimensional input data to a low-dimensional hash space, thereby mapping similar feature vectors into the same hash bucket. Thereafter, the token aggregation operation is used to perform weighted aggregation of tokens within the same hash bucket, effectively reducing the computational complexity of the self-attention mechanism.

Subsequently, the FMLP module is incorporated, resulting in a slight increase in the parameters load of the model as well as an improvement in the mIoU value by 0.66%, reaching 91.25%. The FMLP module reconstructs the feature tensor into  $8 \times 8$  local patches and uses fast Fourier transform to convert spatial features into frequency features. In the frequency domain, FMLP applies weighted learning to local patches, making the model sensitive to local feature changes. This effectively enhances the overall performance of the model.

### C. Comparison with State-of-the-art Methods

To verify the superiority of the EMLSSA model to other advanced methods, we conducted systematic comparative experiments on four remote sensing datasets: FLAME, PWD, EarthVQA, and Potsdam. Tables IV, V, and VI present the test results of each model on different datasets. They include metrics such as model parameters, computational complexity, and mIoU values.

The experimental results in Table IV show that on the FLAME dataset, the EMLSSA model reduces computational load by 11.7% compared to SegFormer and 56.6% compared to MMLN. Meanwhile, it increases mIoU by 1.17% and 0.03%, respectively. In the PWD dataset, the mIoU value of the

TABLE III  
ABLATION EXPERIMENT

Model	Backbone	Input Size	Params(M)	GFLOPs	mIoU(%)
SegFormer	MIT-B4	512×512	64.13	95.76	90.08
+HCAttention	MIT-B4	512×512	64.13	84.53	90.59
+HCAttention+FMLP(EMLSSA)	MIT-B4	512×512	65.83	84.53	91.25

TABLE IV  
COMPARISON WITH STATE-OF-THE-ART METHODS ON FLAME AND PWD DATASETS

Method	Backbone	Params (M)	GFLOPs	FLAME mIoU (%)	PWD mIoU (%)
Unet [38]	ResNet50	43.93	91.7	87.26	86.12
PSPNet [39]	ResNet50	49.08	178.76	87.62	84.20
DeepLabV3+ [40]	ResNet50	43.69	177.46	87.86	84.56
SegFormer	MiT-B4	64.13	95.76	90.08	87.13
Mask2Former [41]	Swin-B	102	195	91.16	87.23
SSformer [42]	Swin-T	87.5	91.01	90.88	86.37
MMLN [43]	Swim-S	62.2	194.9	91.22	87.23
EMLSSA (Ours)	MiT-B4	65.83	84.53	91.25	87.59

TABLE V  
COMPARISON WITH STATE-OF-THE-ART METHODS ON EARTHVQA DATASET

Method	Background	Building	Road	Water	Barren	Forest	Agriculture	Playground	mIoU(%)
Unet	99.19	47.89	57.78	54.73	69.34	20.93	46.87	55.41	56.52
PSPNet	99.35	46.21	60.34	58.23	71.02	25.08	44.9	57.23	57.80
DeepLabV3+	99.32	45.16	61.51	58.55	70.98	20.05	45.09	56.04	57.09
SegFormer	99.49	47.56	63.95	58.74	72.52	30.05	49.59	64.15	60.75
Mask2Former	99.5	47.23	61.8	61.37	73.68	32.74	50.68	60.19	60.90
SSformer	98.7	43.45	46.44	43.85	63.15	18.74	43.24	55.05	51.57
MMLN	99.36	48.81	51.34	47.49	73.98	43.26	63.03	65.33	61.58
EMLSSA	99.55	47.92	65.34	60.91	74.18	31.79	50.13	65.58	61.93

EMLSSA model is 0.46% higher than that of the SegFormer model. These data demonstrate the excellent performance of the EMLSSA model in binary classification tasks, providing a solid theoretical basis for its promotion in remote sensing applications, such as forest fire and pest and disease detection.

Table V presents the test results on the EarthVQA dataset. The EMLSSA model has the highest mIoU of 61.93%, outperforming the traditional SegFormer model by 1.18%, demonstrating its performance advantage. The EMLSSA model has a slightly larger parameter size than UNet, PSPNet, DeepLabV3+, and MMLN. However, its computational complexity is significantly lower, achieving superior performance with only half the computational load. While its segmentation performance for categories, such as water bodies, forests, and farmland, is slightly insufficient, it achieves the highest overall mIoU.

The results in Table VI show that the EMLSSA model outperforms SegFormer and Mask2Former, increasing mIoU by 0.34% and 3.64%, respectively. Compared to UNet, PSPNet,

and DeepLabV3+, it achieves mIoU improvements of 7.66%, 2.64%, and 3.04%, respectively. These results further confirm the advantages of EMLSSA in complex remote sensing image segmentation tasks.

To comprehensively evaluate the practical value of the proposed method and highlight the performance advantages of the EMLSSA module under lightweight constraints, we selected four representative lightweight semantic segmentation algorithms for comparative analysis. The experimental results are shown in Table VII.

Although the computational cost of the EMLSSA model is higher than that of some lightweight algorithms, unlike these methods which suffer from significant accuracy degradation, the proposed approach introduces the HCAttention module to effectively reduce computation while maintaining high segmentation performance. This method demonstrates notable advantages in edge computing scenarios and is better suited for real-world deployment needs.

Experimental results demonstrate that the EMLSSA model significantly reduces computational overhead compared to

TABLE VI  
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE POTSDAM DATASET

Method	Background	Surfaces	Low Vegetation	Tree	Car	Building	mIoU(%)
Unet	89.81	66.54	70.71	41.56	78.47	80.94	71.34
PSPNet	93.14	74.19	75.70	54.19	77.16	83.76	76.36
DeepLabV3+	92.37	72.97	75.14	53.37	78.70	83.23	75.96
SegFormer	93.84	74.91	77.57	57.26	82.84	85.52	78.66
Mask2Former	91.86	73.32	73.56	46.55	81.28	85.63	75.36
SSformer	91.77	71.97	76.33	48.16	75.56	83.11	74.48
MMLN	91.96	72.98	76.71	62.64	81.76	84.81	78.48
EMLSSA	93.85	75.64	77.81	58.05	82.72	85.91	79.0

TABLE VII  
COMPARISON OF LIGHTWEIGHT MODEL PERFORMANCE

Method	Params (M)	GFLOPs	FLAME mIoU (%)	PWD mIoU (%)	EARTHVQA mIoU (%)	POTSDAM mIoU (%)
EMLSSA	65.83	84.53	91.25	87.59	61.93	79.00
Fast-SCNN	1.14	0.78	86.74	82.83	54.25	68.48
ESPNet	0.35	1.78	83.60	81.73	56.97	72.66
Paca-ViT	23.41	47.25	87.60	83.06	45.59	73.73
MetaSeg	29.60	30.40	88.12	85.90	49.43	69.63

other baseline models, while also achieving superior mIoU scores across various segmentation tasks. This performance advantage is primarily attributed to the introduction of the HCAAttention self-attention mechanism. This mechanism performs dynamic aggregation of the input sequence through a hash clustering operation, effectively reducing the number of redundant tokens and thus substantially lowering computational complexity. The tokens generated after aggregation are more representative and retain critical semantic information. When combined with the global modeling capability of the self-attention mechanism, these representative token features further enhance the model’s perception of structural information in remote sensing images.

Additionally, to compensate for potential loss of local features during the token aggregation process, EMLSSA incorporates the FMLP module. This module dynamically learns weighted representations of local image patches in the frequency domain, thereby enhancing the model’s ability to capture fine-grained local details. As a result, the model achieves improved recognition and segmentation accuracy at the detail level, while maintaining computational efficiency.

#### D. Qualitative Results

Through qualitative analysis, the segmentation effect of the model at object edge details and its ability to manage complex backgrounds can be intuitively observed [44, 45]. To evaluate the performance advantages of the EMLSSA model in remote sensing image segmentation tasks, we conducted qualitative analysis on four representative datasets: FLAME, PWD, EarthVQA, and Potsdam. We examined its improvements by comparing it to the SegFormer model. The qualitative results are shown in Figs. 6, 8, 10, and 11. These figures compare the

original input images, ground truth annotations, and predicted segmentation masks, revealing the performance differences of each model in various scenarios.

The qualitative results on the FLAME dataset are shown in Fig. 6. Compared with the ground truth masks, the prediction results of SegFormer and EMLSSA are almost identical, making them difficult to distinguish with the naked eye. Therefore, we calculated the segmentation areas of the prediction results of the two models and the ground truth masks. As shown in Fig. 7, in different predicted images, the segmentation area of the EMLSSA model is closest to the ground truth masks, indicating that it contains richer detail information and can accurately segment object boundaries.

Fig. 8 shows the qualitative results of different models on the PWD dataset. In the pine wilt disease segmentation task, a detailed comparison of the segmentation areas in Fig. 8(e) shows that the EMLSSA model has clearer and more accurate segmentation boundaries. It effectively distinguishes between infected and uninfected areas. Meanwhile, the SegFormer model exhibits obvious false negatives and false positives, as shown in the red boxes in Fig. 8(f) and 8(g). In Fig. 8(f), the SegFormer model fails to accurately segment all infected areas. In Fig. 8(g), the SegFormer model mistakenly identifies land as infected areas, resulting in false positives. However, the EMLSSA model can more accurately segment the contours and details of the infected areas. We calculated the segmentation areas of the prediction results of the two models and the ground truth masks to further verify the segmentation effect of the EMLSSA model. In Fig. 9, it can be observed that the segmentation area of the EMLSSA model is closer to the ground truth mask area in different images. This improvement is mainly owing to the ability of the EMLSSA model to effectively model long-range dependencies in images,

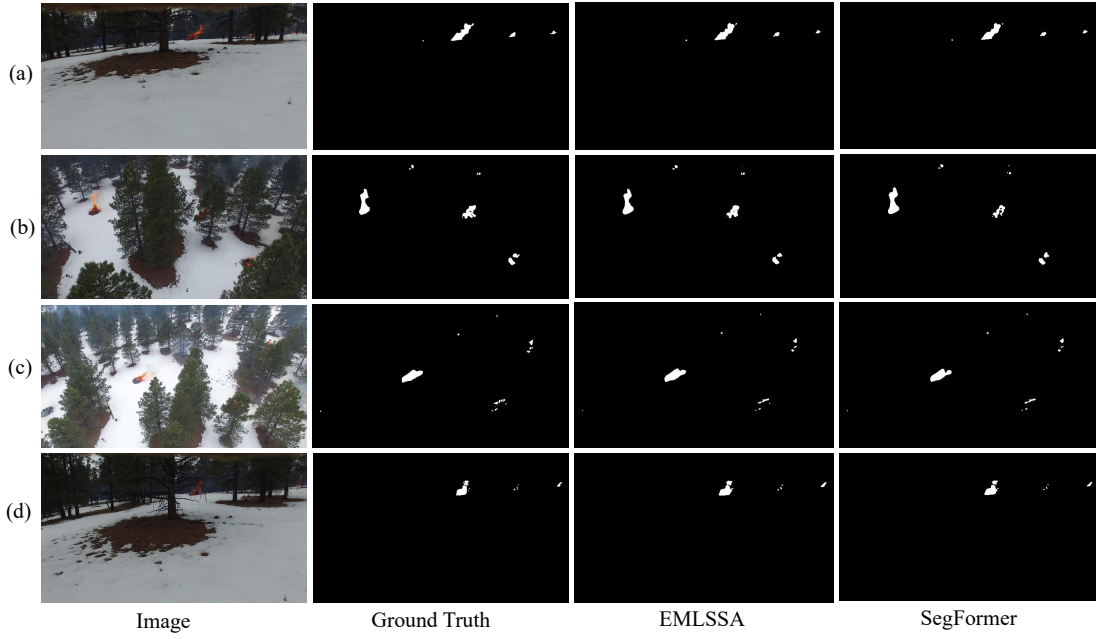


Fig. 6. Qualitative results on FLAME dataset.

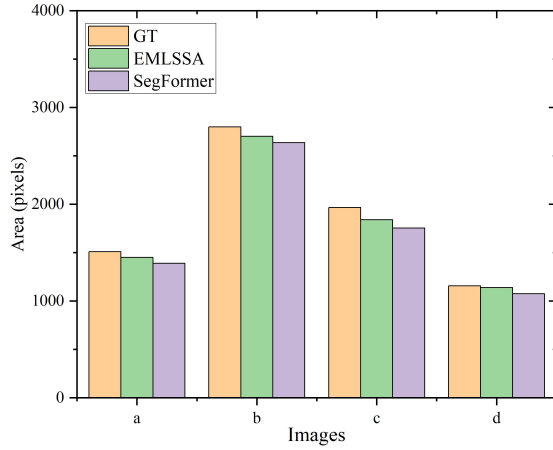


Fig. 7. Comparison results of segmentation area in FLAME dataset.

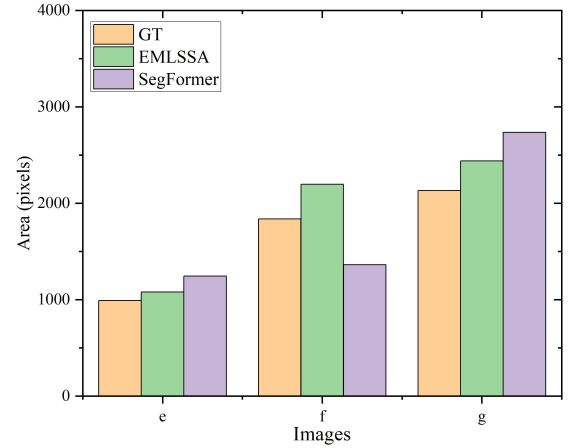


Fig. 9. Comparison results of segmentation area in PWD dataset.

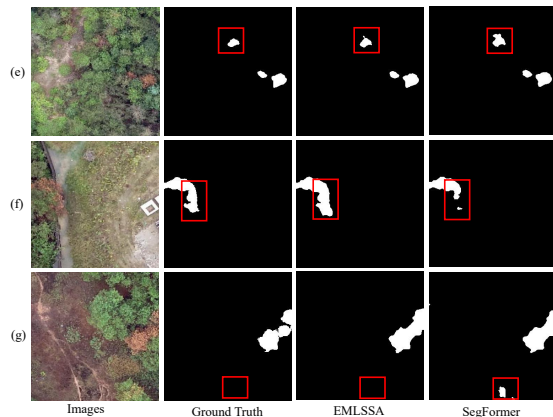


Fig. 8. Qualitative results on PWD dataset.

enhancing its perception of global information. Additionally, the FMLP module precisely extracts local detailed features, allowing the EMLSSA model to capture key feature information more accurately when managing small and occluded targets commonly found in remote sensing images.

Fig. 10 shows the qualitative results on the EarthVQA dataset. The colors represent different categories. Red boxes highlight the differences between the predictions of different models and the true masks. As shown in Fig. 10, our EMLSSA model performs significantly better than the SegFormer model in terms of segmentation boundary clarity. This is true for categories such as roads and barren land. EMLSSA can generate sharper and more accurate edges. Therefore, EMLSSA can capture the actual range and shape of these features more precisely and reduces blurry or jagged segmentation. Furthermore, in complex areas with multiple overlapping or closely adjacent categories, the EMLSSA model shows a

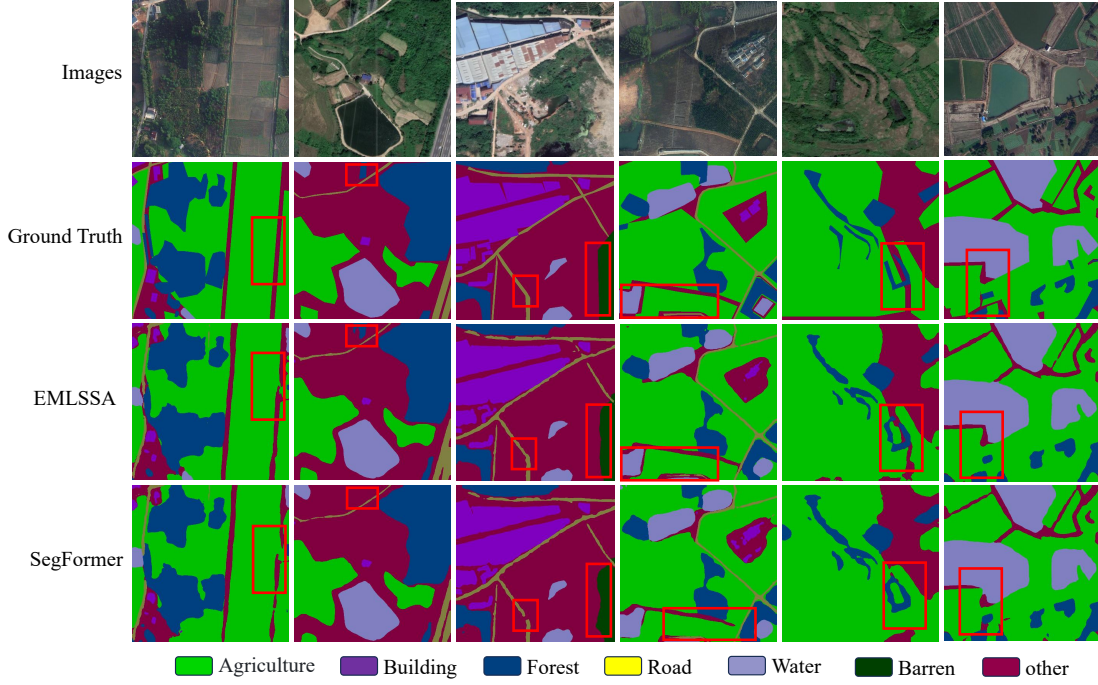


Fig. 10. Qualitative results on EarthVQA dataset.

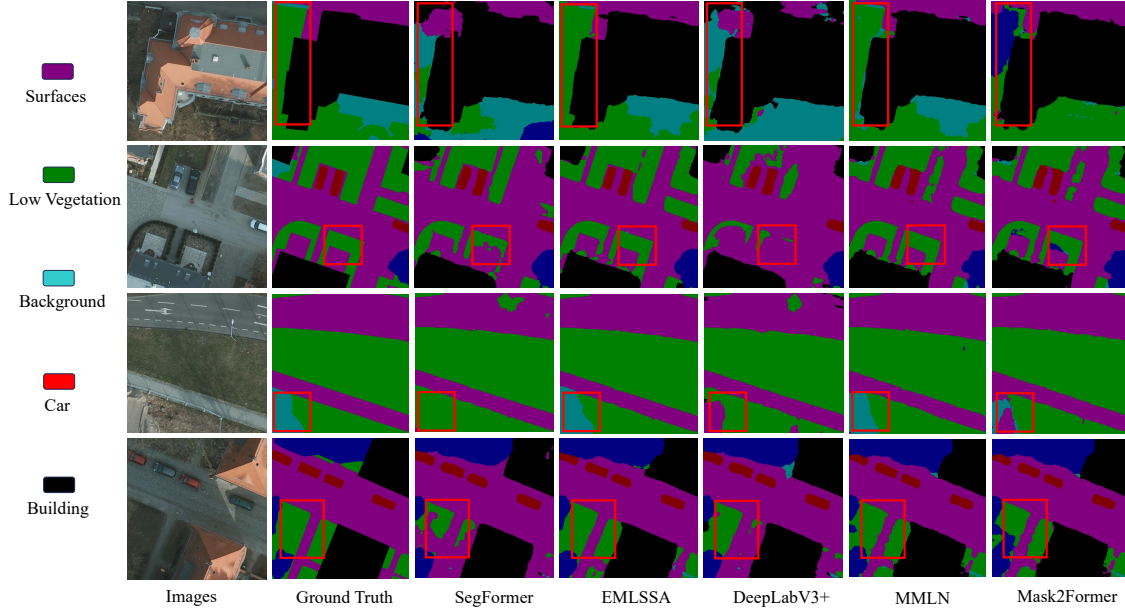


Fig. 11. Qualitative results on the Potsdam dataset.

stronger ability to distinguish between them. It can accurately identify and differentiate objects. However, the SegFormer model often shows category confusion in these areas.

In the land cover segmentation task on the Potsdam dataset, the performance of various models varies significantly across different scenarios. As shown in Fig. 11, the SegFormer model struggles to accurately segment land cover features and even exhibits omissions. While the DeepLabV3+ and Mask2Former models can identify most land cover categories, they show noticeable false detections in their predictions. The segmentation results of the MMLN model are similar to those

of the EMLSSA model, but some boundaries remain unclear. However, the EMLSSA model demonstrates superior segmentation performance and stronger robustness when managing complex land cover scenarios in the Potsdam dataset.

The performance of the EMLSSA model improved because of its precise capture of global contextual information and local detailed features. This combination enables the model to comprehensively understand complex scenes in remote sensing images. Specifically, the introduction of the HCA-ttention self-attention mechanism effectively guides the model to focus on key target regions through dynamic clustering



while maintaining its lightweight design. This mechanism reduces interference from irrelevant information and allows the model to accurately define boundaries between different land covers, thereby significantly improving segmentation accuracy. Additionally, the optimization of the FMLP enhances the capability of the model to extract edge detail information. Edge details often contain important land cover information, such as the precise contours of buildings, roads, and vegetation. The FMLP optimization enables the model to capture land cover features in complex scenarios more effectively, demonstrating excellent performance in small object recognition, occluded object detection, and understanding complex multi-object scenes. According to this qualitative analysis, the EMLSSA model has significant advantages in remote sensing image analysis tasks, particularly in challenging complex scenarios. Its precise capture of global and local features, along with the enhancements from HCAttention and FMLP, collectively contribute to improved model performance.

## V. CONCLUSION

The proposed EMLSSA efficient semantic segmentation model effectively reduces computational complexity, making it ideal for remote sensing image semantic segmentation tasks. This model innovatively introduces the HCAttention self-attention mechanism to compress input token features, thereby improving computational efficiency. To enhance the ability of the model to capture local detailed textures, this study incorporates the FMLP module. This module performs feature weighting on local image patches in the frequency domain, effectively strengthening the representation of local features. Experimental results show that the EMLSSA model demonstrates outstanding performance across multiple public remote sensing image semantic segmentation datasets, validating its effectiveness and superiority in remote sensing image segmentation tasks.

## REFERENCES

- [1] X. Xiao, X. Wang, and W. Lin, "Joint aoi-aware uavs trajectory planning and data collection in uav-based iot systems: A deep reinforcement learning approach," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 4, pp. 6484–6495, 2024.
- [2] J. Li, S. Zhang, Y. Sun, Q. Han, Y. Sun, and Y. Wang, "Frequency-driven edge guidance network for semantic segmentation of remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
- [3] C. Xie, X. Zhai, H. Chi, W. Li, X. Li, Y. Sha, and K. Li, "A novel fusion pruning-processed lightweight cnn for local object recognition on resource-constrained devices," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 4, pp. 6713–6724, 2024.
- [4] L. Wang, X.-s. Tang, and K. Hao, "Gfpe-vit: vision transformer with geometric-fractal-based position encoding," *The Visual Computer*, pp. 1–16, 2024.
- [5] S. Reza, M. C. Ferreira, J. J. Machado, and J. M. R. Tavares, "A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks," *Expert Systems with Applications*, vol. 202, p. 117275, 2022.
- [6] X. Dong, Q. Wang, H. Deng, Z. Yang, W. Ruan, W. Liu, L. Lei, X. Wu, and Y. Tian, "From global to hybrid: A review of supervised deep learning for 2d image feature representation," *IEEE Transactions on Artificial Intelligence*, 2025.
- [7] K. Fang, J. Deng, C. Dong, U. Naseem, T. Liu, H. Feng, and W. Wang, "Mocfl: Mobile cluster federated learning framework for highly dynamic network," *arXiv preprint arXiv:2503.01557*, 2025.
- [8] H. Su, L. Liu, G. Jeon, Z. Wang, T. Guo, and M. Gao, "Remote sensing image dehazing based on dual attention parallelism and frequency domain selection network," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 5300–5311, 2024.
- [9] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.
- [10] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y.-H. Chen, L. Lai, V. Chandra, and D. Z. Pan, "Multi-scale high-resolution vision transformer for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 094–12 103.
- [11] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding unet for remote sensing image semantic segmentation," *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–15, 2022.
- [12] J.-h. Shim, H. Yu, K. Kong, and S.-J. Kang, "Feedformer: Revisiting transformer decoder for efficient semantic segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 2263–2271.
- [13] B. Kang, S. Moon, Y. Cho, H. Yu, and S.-J. Kang, "Metaseg: Metaformer-based global contexts-aware network for efficient semantic segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 434–443.
- [14] Y. Chen, Q. Dong, X. Wang, Q. Zhang, M. Kang, W. Jiang, M. Wang, L. Xu, and C. Zhang, "Hybrid attention fusion embedded in transformer for remote sensing image semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 4421–4435, 2024.
- [15] S. Chen, T. Han, C. Zhang, J. Su, R. Wang, Y. Chen, Z. Wang, and G. Cai, "Hspformer: Hierarchical spatial perception transformer for semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, 2025.
- [16] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "Dynam-icvit: Efficient vision transformers with dynamic token sparsification," *Advances in neural information processing systems*, vol. 34, pp. 13 937–13 949, 2021.
- [17] M. Fayyaz, S. A. Koohpayegani, F. R. Jafari, S. Sengupta, H. R. V. Joze, E. Sommerlade, H. Pirsiavash, and J. Gall, "Adaptive token sampling for efficient vision transformers," in *European Conference on Computer Vision*. Springer, 2022, pp. 396–414.
- [18] R. Grainger, T. Paniagua, X. Song, N. Cuntoor, M. W. Lee, and T. Wu, "Paca-vit: learning patch-to-cluster attention in vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 568–18 578.
- [19] J. Liang, Y. Cui, Q. Wang, T. Geng, W. Wang, and D. Liu, "Clusterfomer: clustering as a universal visual learner," *Advances in neural information processing systems*, vol. 36, pp. 64 029–64 042, 2023.
- [20] W. Zeng, S. Jin, L. Xu, W. Liu, C. Qian, W. Ouyang, P. Luo, and X. Wang, "Tcformer: Visual recognition via token clustering transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [21] J. Cao, P. Ye, S. Li, C. Yu, Y. Tang, J. Lu, and T. Chen, "Madtp: Multi-modal alignment-guided dynamic token pruning for accelerating vision-language transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 15 710–15 719.
- [22] G. Du, P. Zhou, N. Yadikar, A. Aysa, and K. Ubul, "Ddtrack dynamic token sampling for efficient uav transformer tracking," in *International Conference on Pattern Recognition*. Springer, 2025, pp. 129–144.
- [23] J. Ruan, S. Xiang, M. Xie, T. Liu, and Y. Fu, "Malunet: A multi-attention and light-weight unet for skin lesion segmentation," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 1150–1156.
- [24] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *arXiv preprint arXiv:2001.04451*, 2020.
- [25] Q. Ge, J. Li, X. Wang, Y. Deng, K. Zhang, and H. Sun, "Litetransnet:

- An interpretable approach for landslide displacement prediction using transformer model with attention mechanism,” *Engineering Geology*, vol. 331, p. 107446, 2024.
- [26] O. Ghozatlou, M. Datu, A. Focsa, M. H. Conde, and S. L. Ullo, “A review and a perspective of deep active learning for remote sensing image analysis: Enhanced adaptation to user conjecture,” *IEEE Geoscience and Remote Sensing Magazine*, 2024.
- [27] K. K. Brar, B. Goyal, A. Dogra, M. A. Mustafa, R. Majumdar, A. Alkhayyat, and V. Kukreja, “Image segmentation review: Theoretical background and recent advances,” *Information Fusion*, p. 102608, 2024.
- [28] N. Sharma and R. K. Sunkaria, “The enigmatic u-wave delineation and classification based on hybrid feature fusion using stationary wavelet transform and ensemble machine learning algorithm,” *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 5286–5299, 2024.
- [29] X. Yuan, J. Shi, and L. Gu, “A review of deep learning methods for semantic segmentation of remote sensing imagery,” *Expert Systems with Applications*, vol. 169, p. 114417, 2021.
- [30] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, “Efficient frequency domain-based transformers for high-quality image deblurring,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5886–5895.
- [31] P. Li, R. Zhou, J. He, S. Zhao, and Y. Tian, “A global-frequency-domain network for medical image segmentation,” *Computers in Biology and Medicine*, vol. 164, p. 107290, 2023.
- [32] A. Shamsoshoara, F. Afghah, A. Razi, L. Zheng, P. Z. Fulé, and E. Blasch, “Aerial imagery pile burn detection using deep learning: The flame dataset,” *Computer Networks*, vol. 193, p. 108001, 2021.
- [33] H. Feng, J. Qiu, L. Wen, J. Zhang, J. Yang, Z. Lyu, T. Liu, and K. Fang, “U3unet: An accurate and reliable segmentation model for forest fire monitoring based on uav vision,” *Neural Networks*, p. 107207, 2025.
- [34] J. Yuan, L. Wang, T. Wang, A. K. Bashir, M. M. Al Dabel, J. Wang, H. Feng, K. Fang, and W. Wang, “Yolov8-rd: High-robust pine wilt disease detection method based on residual fuzzy yolov8,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
- [35] L. Wang, J. Cai, T. Wang, J. Zhao, T. R. Gadekallu, and K. Fang, “Pine wilt disease detection based on uav remote sensing with an improved yolo model,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
- [36] J. Wang, Z. Zheng, Z. Chen, A. Ma, and Y. Zhong, “Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5481–5489.
- [37] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, “Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.
- [38] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.
- [39] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.
- [40] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.
- [41] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.
- [42] W. Shi, J. Xu, and P. Gao, “Ssformer: A lightweight transformer for semantic segmentation,” in *2022 IEEE 24th international workshop on multimedia signal processing (MMSP)*. IEEE, 2022, pp. 1–5.
- [43] H. Sun, Y. Xie, D. Ren, F. Wen, L. Tong, and L. Chang, “Mmln: Multi-directional and multi-constraint learning network for remote sensing imagery semantic segmentation,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
- [44] J. Liang, Q. Zhang, and X. Gu, “Lightweight convolutional neural network driven by small data for asphalt pavement crack segmentation,” *Automation in Construction*, vol. 158, p. 105214, 2024.
- [45] Z. Guo, D. Cai, Z. Jin, T. Xu, and F. Yu, “Research on unmanned aerial vehicle (uav) rice field weed sensing image segmentation method based on cnn-transformer,” *Computers and Electronics in Agriculture*, vol. 229, p. 109719, 2025.