

Article

Seeing Like Argus: Multi-Perspective Global–Local Context Learning for Remote Sensing Semantic Segmentation

Hongbing Chen, Yizhe Feng , Kun Wang, Mingrui Liao, Haoting Zhai, Tian Xia, Yubo Zhang , Jianhua Jiao and Changji Wen * 

College of Information Technology, Jilin Agricultural University, Changchun 130118, China; chen hongbing@jlau.edu.cn (H.C.); yeajer@mails.jlau.edu.cn (Y.F.); wangkun@mails.jlau.edu.cn (K.W.); liaomingrui@mails.jlau.edu.cn (M.L.); zhahao ting123@mails.jlau.edu.cn (H.Z.); xiatian@mails.jlu.edu.cn (T.X.); zhangyubo@jlau.edu.cn (Y.Z.); jianhua@mails.jlau.edu.cn (J.J.)

* Correspondence: changjiw@jlau.edu.cn

Highlights

What are the main findings?

- ArgusNet introduces a Hybrid Global–Local Block (HGLB) that synergistically integrates Adaptive Windowed Additive Attention (AWAA) and SS2D-based state-space modeling, enabling complementary fine-grained local detail extraction and global long-range dependency modeling.
- The proposed Macro Guidance Module (MGM) effectively bridges semantic gaps between encoder and decoder features, significantly improving boundary continuity and segmentation consistency on multi-scale remote sensing objects.

What is the implication of the main finding?

- By jointly enhancing global–local context learning with linear computational complexity, ArgusNet improves segmentation accuracy in complex remote sensing scenes containing dense, small objects; large structures; and high intra-class variability.
- The framework demonstrates strong generalization across datasets (LoveDA, Potsdam), suggesting its suitability as a robust backbone for high-resolution land-cover mapping, urban analysis, and other geospatial applications requiring fine-grained semantic understanding.

Abstract

Accurate semantic segmentation of high-resolution remote sensing imagery is crucial for applications such as land cover mapping, urban development monitoring, and disaster response. However, remote sensing data still present inherent challenges, including complex spatial structures, significant intra-class variability, and diverse object scales, which demand models capable of capturing rich contextual information from both local and global regions. To address these issues, we propose ArgusNet, a novel segmentation framework that enhances multi-scale representations through a series of carefully designed fusion mechanisms. At the core of ArgusNet lies the synergistic integration of Adaptive Windowed Additive Attention (AWAA) and 2D Selective Scan (SS2D). Specifically, our AWAA extends additive attention into a window-based structure with a dynamic routing mechanism, enabling multi-perspective local feature interaction via multiple global query vectors. Furthermore, we introduce a decoder optimization strategy incorporating three-stage feature fusion and a Macro Guidance Module (MGM) to improve spatial detail preservation and semantic consistency. Experiments on benchmark remote sensing datasets demonstrate that ArgusNet achieves competitive and improved segmentation performance



Academic Editor: Tomaž Podobnikar

Received: 4 December 2025

Revised: 22 January 2026

Accepted: 3 February 2026

Published: 5 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

compared to state-of-the-art methods, particularly in scenarios requiring fine-grained object delineation and robust multi-scale contextual understanding.

Keywords: additive attention; attention mechanism; dynamic routing; global–local feature fusion; Mamba; multi-scale representation; remote sensing (RS); semantic segmentation

1. Introduction

Semantic segmentation of high-resolution remote sensing imagery (RSI) is a key task in geospatial intelligence analysis, with broad and significant applications in land cover classification [1], disaster assessment [2], and smart city development [3,4]. In recent years, deep learning has achieved remarkable progress in this field. Nevertheless, as illustrated in Figure 1, RSI semantic segmentation remains challenging due to remote sensing-specific issues such as low inter-class contrast (e.g., between Agricultural and Forest regions), boundary blurring caused by shadows or occlusions, and the presence of dense small objects (e.g., vehicles in urban areas), in addition to general challenges like complex spatial structures, significant intra-class variability, and diverse object scales. These unique pain points demand models capable of capturing fine-grained local details while maintaining global context, which existing methods often fail to address comprehensively.

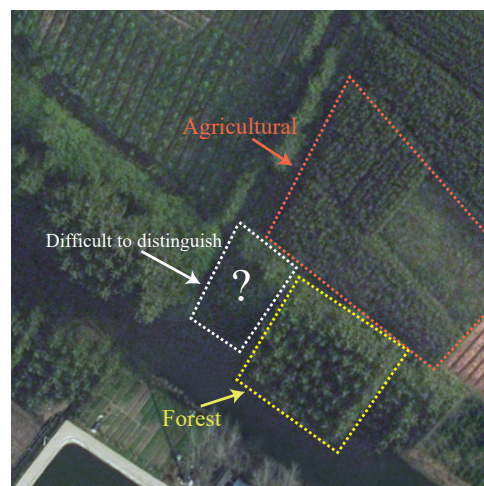


Figure 1. Examples of the challenges in semantic segmentation of RSIs, taken from the LoveDA dataset. Agricultural and Forest exhibit highly similar features, with some transitional zones being difficult to distinguish.

Semantic segmentation methods have gradually evolved from early convolution-based architectures toward more context-aware designs. CNN-based approaches [5–8] and their multi-scale extensions improve local representation and enhance robustness to scale variation, but their intrinsically limited receptive fields restrict the modeling of long-range spatial dependencies in high-resolution remote sensing imagery. As a result, such methods often struggle to resolve ambiguities caused by complex backgrounds, large-scale structures, and strong intra-class variability.

The emergence of the Transformer architecture [9] has brought a new breakthrough to semantic segmentation. Vision Transformers (ViTs) [10,11], based on self-attention mechanisms, establish global pixel-level dependencies and have demonstrated outstanding performance across various vision tasks. However, their limited exploitation of both global and local contextual information leaves the problem of blurred objects and boundaries caused by shadow occlusion unresolved. Furthermore, hybrid CNN–Transformer archi-

textures face inherent limitations: CNNs possess a restricted receptive field, whereas the self-attention mechanism of Transformers incurs high computational cost. Additive attention, as another representative paradigm of attention mechanisms, has recently garnered substantial research interest. In natural language processing (NLP), FastFormer [12] proposed an efficient global context modeling approach based on additive attention, requiring sequential Query–Key–Value interactions. SwiftFormer [13] extended this mechanism to the vision domain, simplifying it to encode only the interaction between global Queries and Keys for efficient global context acquisition, thereby proposing the Efficient Additive Attention (EAA) module. While computationally efficient, EAA still suffers from detail loss in dense small-object scenarios.

Recent advances in state space models (SSMs) have opened new possibilities for long-range dependency modeling. The Mamba model [14] employs a selective state space mechanism to achieve linear-complexity long-sequence modeling. However, applying SSMs to vision tasks remains challenging due to the need for two-dimensional structural modeling. To address this, several studies [15,16] have proposed various 2D scanning strategies to bridge the gap between 1D sequence processing and two-dimensional visual understanding. These developments have promoted the adoption of SSMs in remote sensing semantic segmentation. For example, GLVMamba [17] enhances multi-scale modeling through global–local feature interaction, while RS-Mamba [18] introduces optimized structures for dense prediction tasks in remote sensing. Despite these advancements, most Mamba-based models overlook the limitations of 2D Selective Scan (SS2D) in remote sensing: difficulty in preserving spatial adjacency, leading to insufficient local detail extraction and, consequently, suboptimal coordination between local and global context modeling.

Overall, effective remote sensing semantic segmentation requires models that can simultaneously preserve fine-grained local structures and capture long-range global context under affordable computational cost. Existing CNN-, Transformer-, and Mamba-based approaches address only part of this requirement. To bridge this gap, we propose a novel semantic segmentation framework, termed ArgusNet.

The model is named after the “hundred-eyed giant” in Greek mythology, symbolizing its multi-perspective and multi-scale perception capability. The encoder of ArgusNet is constructed by stacking multiple Hybrid Global–Local Blocks (HGLBs), each integrating two complementary mechanisms: Adaptive Window Additive Attention (AWAA) and a VSS module based on SS2D. AWAA combines window-based attention with a dynamic routing mechanism to achieve multi-perspective local feature modeling, specifically targeting the recognition of dense small objects commonly found in remote sensing imagery (e.g., vehicles or buildings in high-resolution scenes). This design enhances local detail extraction while maintaining computational efficiency, addressing a key limitation in existing methods for RS applications. Furthermore, the window size in AWAA is significantly larger than that of conventional convolution kernels, providing an expanded receptive field and delivering higher-quality inputs for the local scanning process of the VSS module. The VSS module aggregates global contextual information to supplement the long-range dependencies for AWAA. Working synergistically, these two modules effectively mitigate the fragmentation of local features and the blurring of object boundaries. In addition, this study introduces the Macro Guidance Module (MGM), which dynamically injects high-level semantic information into the decoder, thereby enhancing boundary coherence and overall segmentation accuracy. The main contributions of this work are as follows:

1. We propose ArgusNet, a hybrid semantic segmentation network that employs HGLBs, which are composed of AWAA and VSS modules, as the encoder and integrates a lightweight convolutional decoder. This architecture effectively leverages the global

modeling capacity of Mamba together with the local modeling strengths of AWAA, thereby achieving substantial improvements in segmentation performance.

2. Designed the AWAA module, which integrates window-based attention and a dynamic routing mechanism to realize multi-perspective local perception. This design effectively improves dense small-object recognition while maintaining far lower computational complexity than conventional self-attention mechanisms.
3. Optimized the decoder architecture by introducing a three-stage feature fusion strategy to preserve fine-grained spatial details, while incorporating the MGM to dynamically guide the injection of deep semantic features, thereby alleviating discontinuities in land-cover segmentation and further enhancing accuracy.
4. Conducted extensive experiments on two publicly available remote sensing datasets, LoveDA and Potsdam, demonstrating that ArgusNet consistently outperforms or matches mainstream methods such as UNetMamba and SegFormer in segmentation accuracy, supporting the effectiveness of the proposed model.

2. Related Works

2.1. RS Semantic Segmentation Based on CNNs

Early CNN-based remote sensing segmentation methods [5,19] mainly rely on convolutional receptive fields and hierarchical feature fusion to model spatial context. While these approaches exhibit strong robustness in structured scenes and limited-data regimes, their intrinsic locality restricts long-range dependency modeling, often leading to misclassification under complex backgrounds and large-scale variations.

The introduction of U-Net [20] marked a significant milestone in semantic segmentation. Its encoder–decoder architecture and multi-scale feature fusion capability have been inherited and extended by numerous subsequent methods. For instance, GCDB-UNet [21] incorporated a global context dense block to improve thin cloud detection accuracy; ARD-UNet [22] employed depthwise separable residual blocks for crack segmentation to mitigate fine-detail loss; and CFAMNet [23] enhanced inter-class correlations through a class-feature attention mechanism, thereby improving segmentation accuracy. These CNN-based models and their variants continue to play an important role in RS semantic segmentation, particularly demonstrating stability in small-sample scenarios and in structured, regular landscapes.

2.2. Transformer- and Mamba-Based Semantic Segmentation

Transformer-based methods [10,24,25] introduce self-attention mechanisms to explicitly model global dependencies in remote sensing imagery, significantly improving semantic consistency for large-scale objects. However, their quadratic computational complexity and limited local detail preservation restrict their applicability in high-resolution RS scenarios with dense small objects.

To mitigate the quadratic computational complexity $O(n^2)$ inherent in self-attention mechanisms, SegFormer [26] reduced computation by shortening the Key sequence length, GCViT [27] proposed a hybrid local–global attention approach, EdgeNeXt [28] introduced transposed self-attention along the channel dimension to achieve linear complexity, Reformer [29] utilized locality-sensitive hashing (LSH) to reduce attention computation cost. EfficientFormer [30] incorporated multi-head self-attention (MHSA) only in the final network stages, enhancing context modeling with minimal impact on inference time. FastFormer [12] replaces the scaled dot-product attention, which models pairwise token interactions, with additive attention for global context modeling, and subsequently performs token-wise interaction with the global context to efficiently achieve linear complexity. Building upon this, SwiftFormer [13] removed the Key–Value interaction in FastFormer

and replaced it with simple linear transformations to design a highly efficient Transformer variant successfully applied to vision tasks. Furthermore, unsupervised methods, such as STEGO [31] and Cut-LER [32], along with unified multitask frameworks handling instance, semantic, and panoptic segmentation, provide effective solutions for small-sample and weakly supervised RS image segmentation. Large-scale pretrained models like SAM [33] and DINOv2 [34], leveraging massive datasets to learn general visual representations, exhibit superior cross-domain segmentation generalization capabilities.

Compared to Transformers, SSMs have recently emerged as promising alternatives for modeling long-range dependencies with linear computational complexity. Vision Mamba [15] achieves 2D global context modeling through a bidirectional scanning mechanism, while VMamba [16] adopts four-directional scanning to alleviate the issue of direction sensitivity. For RS scenarios, RS-Mamba [18] introduces a position-sensitive dynamic multipath activation mechanism to enhance spatial information modeling; RS3Mamba [35] employs a dual-branch architecture to balance long-range dependencies and local details; RTMamba [36] combines Mamba with an iterative token pruning (ITP) module to improve boundary segmentation accuracy. In cross-modal segmentation tasks, Sigma [37] utilizes a multimodal adaptation mechanism to efficiently fuse global features in RGB-Thermal applications. In medical imaging and hyperspectral domains, VM-UNet [38], Mamba-in-Mamba [39], and DHM [40] further demonstrate the transferability and competitiveness of SSMs across diverse scenarios. These studies reveal complementary advantages: Transformers excel in global feature interaction and semantic richness, whereas Mamba achieves superior computational efficiency and scalability in high-resolution tasks owing to its linear complexity.

Despite these achievements, most Mamba-based methods primarily focus on global context modeling, incorporating local feature information only implicitly through convolutions or positional embeddings. For instance, RSMamba enhances orientation robustness with omnidirectional scanning, yet its local features are still implicitly captured by the SSM backbone. To address this limitation, our ArgusNet explicitly separates local and global modeling via parallel branches: AWAA refines local features within windows, while VSS captures global semantic dependencies. This parallel design allows for simultaneous enhancement of fine-grained structures and overall context, improving segmentation accuracy in remote sensing scenes characterized by varying object scales and high intra-class variability. Consequently, while Transformers excel in global feature interaction and semantic richness, Mamba-based SSMs provide superior computational efficiency and scalability, particularly in high-resolution scenarios, with ArgusNet bridging the gap between local and global feature modeling.

2.3. Multi-Scale Feature Fusion in RS

RSI often suffer from low target-to-background contrast due to complex background interference, which traditional downsampling methods struggle to address effectively. Consequently, multi-scale feature fusion techniques have gradually become a focal point of research. The Feature Pyramid Network (FPN) [41], originally developed for object detection, employs hierarchical feature fusion to effectively handle multi-scale variations. Wang et al. [42] introduced FPN into RS scene classification, significantly improving the classification accuracy of small objects (e.g., buildings, roads). The subsequent A-FPN [43] adopts a top-down pyramid structure to construct key feature layers and demonstrates superior suppression of background interference in ship detection tasks. In parallel architectures, the ASPP module proposed by DeepLabv3+ [44] achieves multi-scale feature fusion but exhibits deficiencies in handling irregular object boundaries in large-scale RS image segmentation. To address this limitation, Liu et al. [45] innovatively combined a

dual-attention mechanism with ASPP, enhancing channel feature correlations to improve classification performance and effectively mitigating boundary blurring and hole artifacts in large-scale object segmentation. Additionally, MFVNet [46] designs a pyramid sampling strategy tailored to the large field-of-view (FOV) characteristic of RS images, while Ma et al. [47] propose the MSPM framework, which integrates superpixel segmentation with multi-scale feature extraction to enhance hyperspectral image classification.

Despite these advancements alleviating some challenges posed by complex backgrounds, significant limitations remain. High-level semantic features extracted by deep networks are highly abstract, and the mainstream fusion strategies, such as simple addition or concatenation, may cause the loss of shallow spatial detail information. An ideal solution should establish a cooperative mechanism between global semantics and local details, where high-level features guide the refinement of low-level features to achieve substantial improvements in segmentation accuracy.

2.4. Large-Scale Remote Sensing Datasets and Multimodal Benchmarks

In addition to the architecture-focused research on segmentation models, recent work in remote sensing has highlighted the increasing importance of large-scale datasets and multimodal benchmarks that target higher-level semantic understanding, cross-modal reasoning, and real-world decision-making tasks. One early example of this trend is the Remote Sensing Visual Question Answering (RSVQA) task, which was introduced to enable natural language Queries about high-level content in remote sensing imagery, enabling interactive access to scene information by combining image understanding with language reasoning [48]. A variant of this approach, RSIVQA, was proposed with a dataset and network tailored for visual question answering on remote sensing images [49].

More recent large-scale benchmarks have been developed explicitly for evaluating vision–language models in remote sensing. For example, the RSVLM-QA dataset integrates segmentation and object detection imagery into a content-rich visual question answering benchmark with a diverse set of more than 162k question–answer pairs, enabling systematic evaluation of vision–language reasoning capabilities [50]. Another emerging benchmark, DisasterM3, is designed to support disaster damage assessment and response by providing multi-sensor, multi-task instruction pairs tied to satellite imagery across diverse hazard scenarios, reflecting real-world challenges for vision–language models in complex remote sensing tasks [51]. Also, large-scale vision–language datasets such as SkyScript have been proposed to facilitate pre-training and transfer learning for remote sensing vision–language models by linking millions of image–text pairs with rich semantic diversity [52].

In parallel with multimodal resources, broader surveys and reviews have documented the rise of foundation models and large-scale training data in the remote sensing domain. Comprehensive reviews highlight the trend toward foundation models that can generalize across tasks and modalities, and they discuss the role of increasingly large and semantically rich datasets for both unimodal and multimodal learning in Earth observation [53,54]. Although these works emphasize tasks beyond pixel-wise prediction, we note that the effectiveness of such high-level benchmarks depends fundamentally on robust visual representations, including accurate global–local segmentation and multi-scale context understanding. Therefore, advances in dense prediction and semantic segmentation remain essential building blocks for higher-level remote sensing applications supported by vision–language and decision-oriented benchmarks.

3. Methods

3.1. Overall Architecture

The overall architecture of ArgusNet, as shown in Figure 2, follows the classic encoder–decoder paradigm. The encoder is composed of four stages with progressively reduced spatial resolution. The first stage includes a $4\times$ downsampling PatchEmbed layer followed by multiple HGLBs, while each of the remaining stages consists of a $2\times$ downsampling module and several subsequent HGLBs. Inspired by SegFormer [26], and considering that the encoder stage already possesses a sufficiently large receptive field, we also adopt a highly simplified decoder design. Specifically, except for Stage 4, which employs the PPM, the features from each of the other stages are first processed by a convolutional block and then sequentially upsampled, followed by progressive element-wise addition for feature fusion. Subsequently, the features from Stages 1 to 3 are concatenated along the channel dimension, channel-compressed, and combined with the Stage 4 features as input to the proposed MGM. The output of the MGM is then passed to the segmentation head to generate the final segmentation map. Given the four feature maps $\{F_i\}_{i=1}^4$, where $F_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, output by the backbone at the four stages, the decoder processes them as follows:

$$\hat{F}_i = \begin{cases} \text{Conv}_{3 \times 3}(F_i, C_{out} = D), & i = 1, 2, 3 \\ \text{PPM}(F_i, C_{out} = D), & i = 4 \end{cases} \quad (1)$$

where D denotes the predefined target number of channels (a hyperparameter), $\text{Conv}_{3 \times 3}$ denotes a 3×3 convolution operation, PPM denotes the Pyramid Pooling Module, and C_{out} denotes the number of output channels. Bilinear interpolation is then applied to each processed feature map to match the spatial resolution of the preceding stage:

$$G_i = \begin{cases} \hat{F}_i + \text{BilinearUp}(G_{i+1}, \text{scale} = 2), & i = 1, 2 \\ \hat{F}_i + \text{BilinearUp}(\hat{F}_4, \text{scale} = 2), & i = 3 \end{cases} \quad (2)$$

where $G_i \in \mathbb{R}^{D \times H_i \times W_i}$ denotes the fused feature map at Stage i and scale represents the upsampling factor. Next, G_1, G_2, G_3 are all upsampled to the spatial resolution of G_1 (Stage 1), concatenated along the channel dimension:

$$A = \text{Concat}(\text{BilinearUp}(G_1, \text{scale} = 1), \text{BilinearUp}(G_2, \text{scale} = 2), \text{BilinearUp}(G_3, \text{scale} = 4)) \quad (3)$$

where $A \in \mathbb{R}^{3D \times H_1 \times W_1}$ denotes the concatenated feature map. This concatenation operation, as opposed to element-wise summation, is employed to preserve the distinct characteristics of multi-scale features. While summation inherently assumes feature maps are semantically aligned and may dilute unique information, concatenation retains the complete feature channels from all stages. This provides a richer representation for the subsequent convolutional layer, which can then adaptively learn the optimal fusion strategy by re-calibrating channel-wise importance rather than relying on a fixed, equal-weight combination:

$$\hat{A} = \text{Conv}_{1 \times 1}(A, C_{out} = C) \quad (4)$$

where $\hat{A} \in \mathbb{R}^{C \times H_1 \times W_1}$ denotes the fused feature map. Finally, $\hat{A}$ is fused with the PPM output feature map and fed into the MGM. The output of the MGM is passed through the segmentation head to produce the final segmentation result:

$$Out = \text{SegHead}(\text{MGM}(\hat{A}, \hat{F}_4)) \quad (5)$$

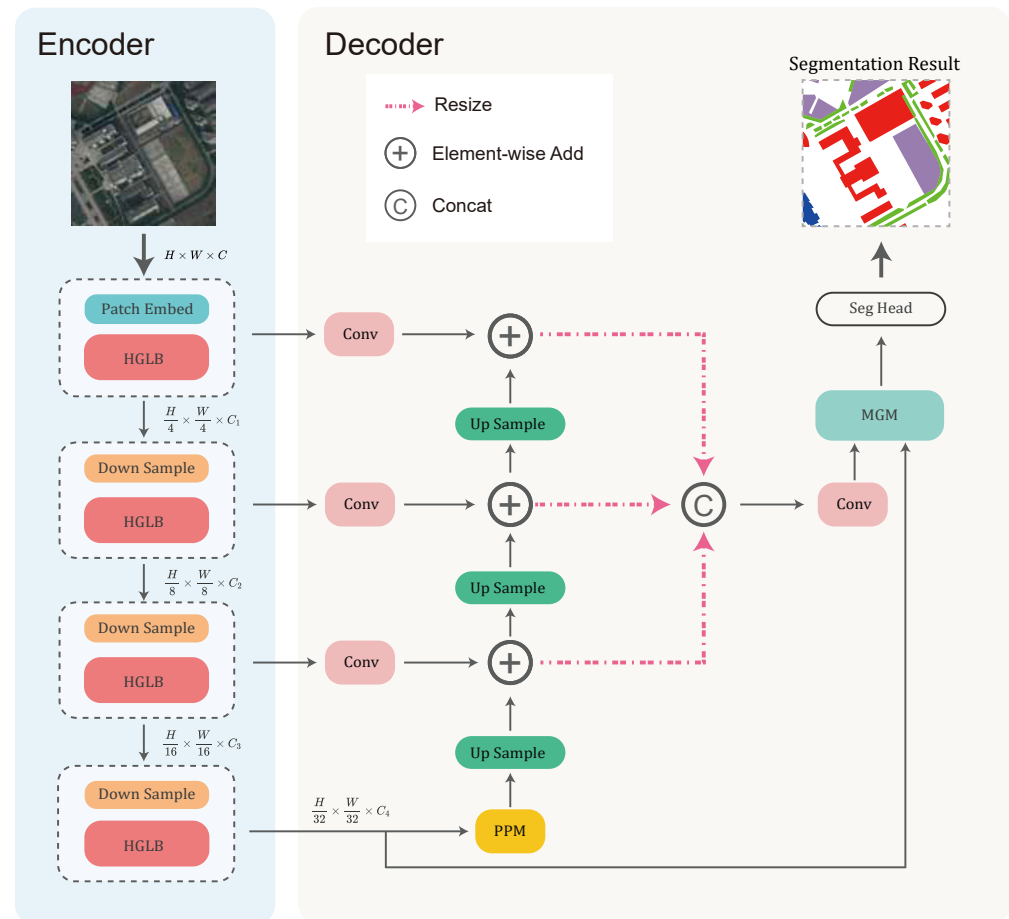


Figure 2. The overall architecture of ArgusNet includes an encoder featuring HGLB and a decoder for multi-scale feature fusion.

ArgusNet appears overall similar to most networks based on the FPN architecture; however, our MGM, through its unique design, effectively reduces interference from background noise, a finding that is further validated by the subsequent ablation experiments.

3.2. Adaptive Windowed Additive Attention

The conventional additive attention mechanism [12] models global context through a three-step Query–Key–Value interaction. Employing element-wise multiplicative pairwise token operations effectively captures long-range dependencies; however, its sequential computation pattern imposes an efficiency bottleneck. To address this issue, inspired by [13], we remove the Key–Value interaction and encode the Query–Key relationship solely via a linear layer. Specifically, the Query matrix is multiplied by a learnable parameter vector and aggregated into a global query vector, which is then broadcast and multiplied element-wise with the Key matrix to obtain the global context representation. Although this design accelerates computation, its performance is hindered by the fact that all inputs within a layer share the same parameter vector, forcing compromises across diverse features in complex scenes and, thereby, reducing discriminative capability.

To enhance the specialization and feature propagation capacity of the parameter vector, we introduce a window-based design that restricts attention computation to local regions. Furthermore, a dynamic integration mechanism is incorporated: K_s shared vectors are always involved in computation, while T routable vectors are maintained, from which K_e vectors are dynamically selected for task-specific modeling. The shared vectors integrate domain-invariant knowledge across scenes, whereas the routable vectors capture scene-specific patterns. Collaboratively, they enable a diversified and efficient feature propagation mechanism. The detailed workflow of the proposed AWAA is as follows:

$$\hat{x} = \text{SplitWindows}(x) \in \mathbb{R}^{c \times p \times p} \quad (6)$$

where the SplitWindows operation partitions the original input x into a windowed representation $\hat{x}$, p denotes the window size. For each routing weight V_e^j , the correlation score between V_e^j and the windowed input $\hat{x}$ is computed as follows:

$$g_j = \text{Linear}(\text{GlobalAvgPool}(\hat{x})) \quad (7)$$

where g_j denotes the correlation score between the input $\hat{x}$ and the routing weight vector V_e^j . A combination of the Softmax function and a topK selection strategy is then applied to identify the K_e most relevant routed vectors:

$$S = \text{topK}(\text{Softmax}(g_1, \dots, g_T), k = K_e) \quad (8)$$

where S denotes the index set of the selected vectors. The normalized weights are subsequently recomputed as follows:

$$\omega = \begin{cases} \frac{\exp(g_j)}{\sum_{k \in S} \exp(g_k)} & , j \in S \\ 0 & , \text{otherwise} \end{cases} \quad (9)$$

Next, two separate linear transformations are employed to project each windowed input into the Query (Q) and Key (K) matrices ($Q, K \in \mathbb{R}^{n \times c}$). A multi-perspective attention mechanism is then constructed using both the shared vectors V_s ($\forall V_s^i \in \mathbb{R}^{1 \times d}, i = 1, \dots, K_s$) and the routed vectors V_e ($\forall V_e^j \in \mathbb{R}^{1 \times d}, j \in S$). The shared vectors yield stable, domain-invariant feature representations, while the routed vectors dynamically adapt to the characteristics of the input:

$$\begin{aligned} \alpha_e &= \left\{ \text{Softmax}\left(\frac{QV_e^i}{\sqrt{C}}\right) \mid i \in S \right\}, \\ \alpha_s &= \left\{ \text{Softmax}\left(\frac{QV_s^j}{\sqrt{C}}\right) \mid j = 1, \dots, K_s \right\}. \end{aligned} \quad (10)$$

where α_e and α_s denote the sets of routed global-attention query vectors and shared global-attention query vectors, respectively. Based on the learned attention weights α_e and α_s , multiple sets of global query vectors q_e and q_s are generated:

$$\begin{aligned} q_e &= \left\{ \sum_{i=1}^n \beta_e^i \odot Q_i \mid \beta_e \in \alpha_e \right\}, \\ q_s &= \left\{ \sum_{j=1}^n \beta_s^j \odot Q_j \mid \beta_s \in \alpha_s \right\} \end{aligned} \quad (11)$$

Next, each global query vector q interacts with the Key matrix $K \in \mathbb{R}^{c \times d}$ via element-wise multiplication to produce multiple sets of global context representations $\mathbb{R}^{n \times c}$. These representations are then integrated through weighted aggregation, where the outputs associated with the shared weights are directly involved in the fusion process. The resulting matrix is analogous to the attention matrix in multi-head self-attention (MHSA), enabling the capture of token-level information and flexible learning of correlations within the input sequence. However, compared with MHSA, the proposed approach achieves a lower computational cost, exhibiting linear complexity with respect to token length. Finally, a linear transformation layer is applied to the Query–Key interactions to learn the latent relationships among tokens. The output of the proposed additive attention mechanism, denoted as $\hat{x}_{out}$, can be expressed as

$$\hat{x}_{out} = \text{Linear} \left(\hat{Q} + \text{Linear} \left(\sum_{i=1}^{K_s} q_s^i \odot K + \sum_{j \in S} \omega_j q_e^j \odot K \right) \right) \quad (12)$$

where $\hat{Q}$ denotes the normalized query matrix. Overall, AWAA performs window-level additive attention with dynamic routing, which enables dynamic multi-perspective local context aggregation while maintaining linear computational complexity (Algorithm 1).

Algorithm 1 Adaptive Windowed Additive Attention (AWAA)

Require: Input feature map $x \in \mathbb{R}^{C \times H \times W}$; Window size p ; Number of shared vectors K_s ; Number of routable vectors T ; Number of selected routable vectors K_e .

Ensure: Refined feature map x_{out} .

```

1: Step 0: Parameter Initialization
2: Initialize Shared Pool  $\mathcal{V}_s = \{V_s^1, \dots, V_s^{K_s}\} \sim \text{Xavier\_Normal}$ 
3: Initialize Routable Pool  $\mathcal{V}_e = \{V_e^1, \dots, V_e^T\} \sim \text{Xavier\_Normal}$ 
4: Step 1: Window Partitioning
5:  $\hat{x} \leftarrow \text{SplitWindows}(x, p)$  ▷  $\hat{x} \in \mathbb{R}^{L \times C}$ , where  $L = p \times p$ , Equation (6)
6: Step 2: Dynamic Routing Mechanism
7:  $g \leftarrow \text{Linear}(\text{GlobalAvgPool}(\hat{x}))$  ▷ Compute correlation scores  $g \in \mathbb{R}^T$ , Equation (7)
8:  $S \leftarrow \text{TopK}(\text{Softmax}(g), K_e)$  ▷ Select top  $K_e$  vector indices, Equation (8)
9:  $\omega \leftarrow \text{ReNormalize}(\text{Softmax}(g), S)$  ▷ Routing weights for selected vectors, Equation (9)
10: Step 3: Linear Projections
11:  $Q \leftarrow \text{Linear}_Q(\hat{x}), K \leftarrow \text{Linear}_K(\hat{x})$  ▷  $Q, K \in \mathbb{R}^{L \times C}$ 
12:  $Y \leftarrow \mathbf{0} \in \mathbb{R}^{L \times C}$  ▷ Initialize output matrix as zero tensor
13: Step 4: Shared Context Interaction
14: for  $j = 1$  to  $K_s$  do
15:    $\alpha_s \leftarrow \text{Softmax}(Q \cdot V_s^j / \sqrt{c})$  ▷ Shared attention weights, Equation (10)
16:    $q_s \leftarrow \sum_{i=1}^L (\alpha_s^i \cdot Q_i)$  ▷ Shared global query vector, Equation (11)
17:    $Y \leftarrow Y + (q_s \odot K)$  ▷ Aggregate shared global context, Equation (11)
18: end for
19: Step 5: Routed Context Interaction
20: for  $i \in S$  do
21:    $\alpha_e \leftarrow \text{Softmax}(Q \cdot V_e^i / \sqrt{c})$  ▷ Routed attention weights, Equation (10)
22:    $q_e \leftarrow \sum_{i=1}^L (\alpha_e^i \cdot Q_i)$  ▷ Routed global query vector, Equation (11)
23:    $Y \leftarrow Y + \omega_i \cdot (q_e \odot K)$  ▷ Weighted routed global context, Equation (11)
24: end for
25: Step 6: Feature Fusion
26:  $\hat{x}_{out} \leftarrow \text{Linear}_{out}(Q + \text{Linear}_{ctx}(Y))$  ▷ Context fusion and projection, Equation (12)
27:  $x_{out} \leftarrow \text{MergeWindows}(\hat{x}_{out})$  ▷ Restore to spatial resolution  $H \times W$ 
28: return  $x_{out}$ 

```

3.3. Hybrid Global–Local Block

In RS semantic segmentation, the low-contrast characteristic of target objects renders boundary delineation based solely on local texture insufficiently accurate, necessitating the integration of global contextual information for comprehensive judgment. Although our proposed AWAA effectively models local features, its window-based design compromises

the ability to capture global contextual information, thus necessitating the incorporation of additional methods to mitigate this limitation. However, Transformers, which excel at capturing long-range dependencies, face the issue that their quadratic computational complexity limits efficiency when applied to large-scale RS images. To address this, SSMs, characterized by linear complexity, emerge as a viable alternative. The SS2D architecture proposed by VMamba [16] employs multi-directional scanning via scan expanding, followed by global correlation modeling through the SSM (S6) [14], and, finally, restores spatial structure using scan merging operations (Figure 3). This approach significantly reduces computational cost while preserving modeling capability. The core of S6 lies in representing system state evolution via first-order differential/difference equations, mapping the input sequence $x(t) \in \mathbb{R}^L$ through a hidden state $h(t) \in \mathbb{R}^N$ to the output $y(t) \in \mathbb{R}^L$. This process is formalized as

$$h_t = Ah_{t-1} + Bx_t \quad (13)$$

$$y_t = Ch_t + Dx_t \quad (14)$$

where $A \in \mathbb{R}^{N \times N}$, $B, C, D \in \mathbb{R}^N$ denote system matrices governing dynamic behavior and output response (with N indicating the number of partitions and L representing the number of images per group, as illustrated in Figure 3, where $N = 4$ and $L = 9$). Intuitively, A acts as a memory controller, governing the evolution of the global context within the hidden state. B and C function as encoders and decoders for the state, respectively, modulating how current visual input interacts with the stored context. Specifically, in the selective mechanism (S6), B and C are input-dependent, allowing the model to adaptively ‘focus’ on or ‘ignore’ specific spatial information based on the content, which is crucial for capturing precise boundaries in semantic segmentation. For detailed derivations, refer to [14]. In this work, we employ the VSS module from VMamba [16] to implement this mechanism. Subsequently, the proposed AWAA captures local features while compensating for S6’s deficiency in modeling fine-grained details. The integration of these modules maximizes the utilization of both local and global information. The overall architecture is depicted in Figure 4. Specifically, the input $x \in \mathbb{R}^{C \times H \times W}$ is first processed by layer normalization and then fed into two parallel branches. In the first branch, the normalized features are passed through the VSS module to capture long-range dependencies. In the second branch, the input is processed by a 3×3 convolution layer to embed positional information, followed by the AWAA module to learn local features. The procedure can be expressed as:

$$\hat{x} = \text{LN}(x) \quad (15)$$

$$x_1 = \text{VSS}(\hat{x}), x_2 = \text{AWAA}(\text{Conv}_{3 \times 3}(\hat{x})) \quad (16)$$

$$x_{\text{fuse}} = \text{Linear}(\text{Concat}(x_1, x_2)) \quad (17)$$

where $\text{Conv}_{3 \times 3}$ denotes the 3×3 convolution operation, and LN denotes the layer normalization operation. The reason for adopting the parallel + concatenation fusion strategy is that it avoids imposing an explicit ordering between long-range dependency modeling and fine-grained local interaction, which could otherwise bias the representation toward a particular information flow. By allowing the two branches to process the same normalized input independently, HGLB preserves the heterogeneity of global and local features during feature extraction. The outputs of both branches are concatenated and projected through a linear layer for feature fusion. Rather than enforcing a fixed or equal contribution, this operation enables the network to adaptively re-calibrate channel-wise importance during training, effectively learning how to balance global semantics and local details based on data. Compared with simple additive fusion, concatenation better retains complementary information from each branch and provides greater flexibility for subsequent feature re-

finement, while maintaining architectural simplicity and linear computational complexity. Finally, x_{fuse} is processed by layer normalization and fed into a feed-forward network (FFN) to obtain the module output:

$$x_{out} = \text{FFN}(\text{LN}(x_{fuse})) \quad (18)$$

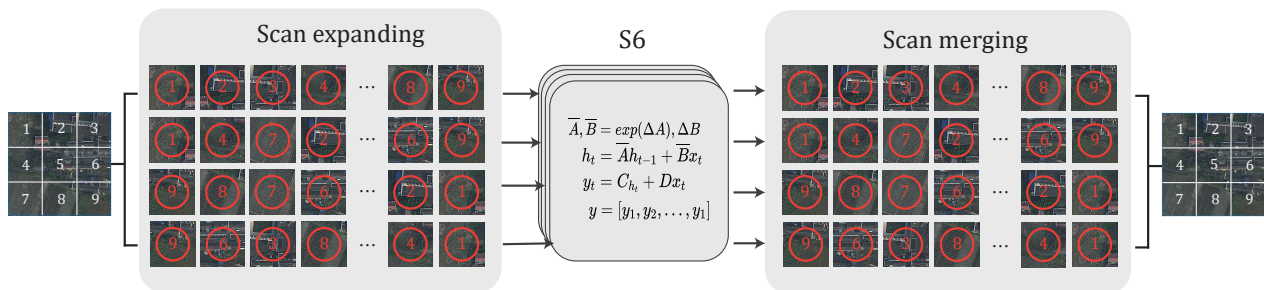


Figure 3. The SS2D framework consists of three key phases—scan expanding, selective state-space models (S6), and scan merging. In the scan expanding step, a two-dimensional feature map is unfolded into one-dimensional sequences along four orientations, where the numbers in the illustration denote the indices of image patches after unfolding, enabling each pixel to aggregate contextual information from multiple directions and thus provide richer perspectives for subsequent computation. S6 then applies the selective scan mechanism to these four sequences, adaptively tuning parameters so that crucial signals are preserved while redundant ones are suppressed. Finally, the scan merging step reconstructs a global receptive field by combining the directional 1-D representations back into a two-dimensional map of the same resolution as the original input via an inverse transformation.

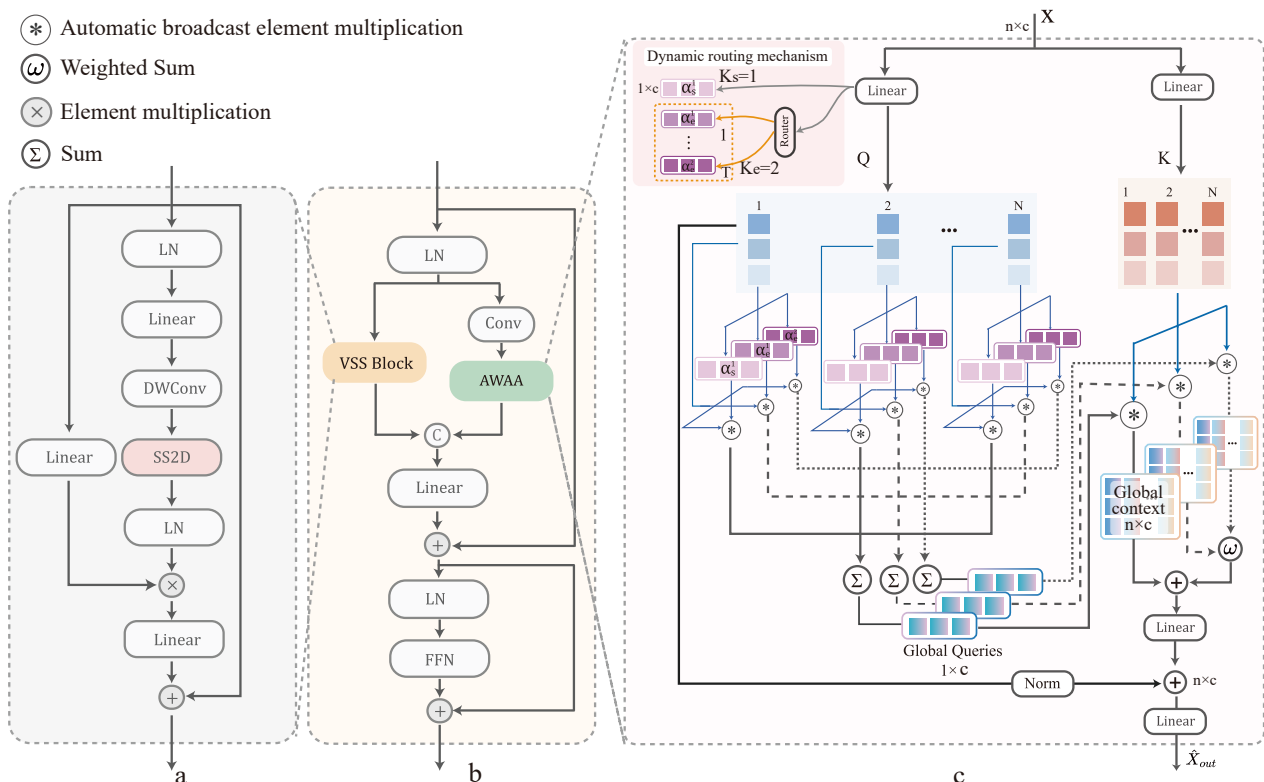


Figure 4. Structure of the HGLB. (a) Overall structure of the original VSS block. (b) Overall structure of the HGLB. (c) Overall structure of the AWAA, data flows from different vectors are distinguished by arrows with different styles. The routing mechanism selects multiple learnable parameter vectors based on input features. These vectors remain isolated and independently participate in computations, with their results being weighted and combined at the final stage to achieve multi-perspective semantic modeling.

The proposed HGLB further refines feature maps through the cooperative integration of local and global representations. In contrast to Mamba-based global–local attention frameworks like UNetMamba, HGLB differs significantly in its fundamental design philosophy. UNetMamba primarily relies on the linear complexity of the Mamba module to construct the decoder, where its Locally Supervised Module (LSM) enhances local detail perception via convolutional branches. However, the global–local feature interaction still adopts serial or parallel multi-path fusion.

In contrast, the core innovation of HGLB lies in its synergistic complementary mechanism: AWAA achieves multi-perspective local feature extraction through adaptive window routing, while SS2D is responsible for modeling global long-range dependencies. These two components enable dynamic coupling of local details and global context during feature interaction. Moreover, since both SS2D and the additive attention algorithm exhibit linear computational complexity with respect to sequence length, HGLB maintains linear complexity without introducing additional local supervision structures. Instead, it strengthens local perception through AWAA’s dynamic routing and window attention itself, thereby achieving more efficient endogenous fusion of global and local information.

3.4. Macro Guidance Module

To address the semantic gap issue in existing methods, where feature fusion is often performed using simple skip connections or element-wise addition, we propose the MGM, which dynamically modulates low-level detail features using high-level semantic features (architecture shown in Figure 5). The core of MGM is a global context-guided mechanism: high-level features $A \in \mathbb{R}^{C \times H' \times W'}$ are used to generate a spatial attention map that adaptively enhances low-level features $X \in \mathbb{R}^{C \times H \times W}$, effectively mitigating boundary blurring and segmentation discontinuities caused by low contrast in remote sensing imagery. Unlike simple skip connections or element-wise addition, which implicitly assume equal and spatially uniform contributions from encoder and decoder features, MGM introduces a content-adaptive modulation mechanism that allows high-level semantics to selectively emphasize or suppress low-level responses at each spatial location. This is particularly crucial for classes with high inter-class similarity, such as Agricultural and Forest. The implementation begins by projecting 1×1 convolution, followed by layer normalization:

$$A' = \sigma(\text{LN}(\text{Conv}_{1 \times 1}(A))) \quad (19)$$

where σ denotes the SiLU activation function. The resulting feature A' is then upsampled to match the spatial resolution of X via bilinear interpolation, after which a depthwise separable convolution is applied to extract spatial features:

$$\hat{A} = \text{DWConv}_{7 \times 7}(A') \quad (20)$$

where $\text{DWConv}_{7 \times 7}$ represents a depthwise separable convolution with a 7×7 kernel. For brevity, the upsampling operation is omitted in the notation. Next, feature modulation is performed via element-wise multiplication:

$$F = X \odot \hat{A} \quad (21)$$

where F denotes the fused feature. In this step, the multiplication operation acts as a gating mechanism, forcing the network to learn how to selectively control the flow of information, thereby preserving critical features while suppressing non-essential ones. To

enhance multi-scale information integration, a residual connection is established, and a squeeze-and-excitation (SE) module is applied for channel attention weighting:

$$\hat{F} = \text{SE}(F + X) \quad (22)$$

Finally, a 3×3 convolution is used to further refine the feature representation:

$$\text{out} = \hat{F} + \text{Conv}_{3 \times 3}(\hat{F}) \quad (23)$$

This design achieves multi-stage feature interaction, where high-level semantic guidance is fused with low-level details, thereby significantly improving semantic consistency in feature fusion and enhancing the continuity of land-cover segmentation. It is worth emphasizing that, although MGM is deployed at the decoder stage and shares a superficial similarity with existing feature fusion strategies, its underlying design philosophy is fundamentally different from commonly used decoder fusion modules. Specifically, conventional approaches such as skip connections, FPN-style summation, or concatenation-based fusion typically assume a symmetric or uniform contribution between encoder and decoder features and lack explicit mechanisms to resolve semantic ambiguity in low-contrast remote sensing scenes.

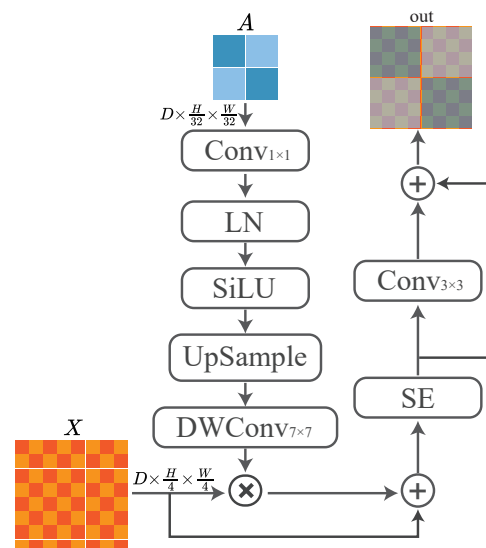


Figure 5. Structural Diagram of the MGM. Different colors indicate feature maps at different spatial resolutions, where denser grids represent higher-resolution features and sparser grids represent lower-resolution features. The output feature map calculate features from two resolutions.

In contrast, MGM explicitly formulates high-level features as a macro-level semantic guidance signal rather than a fusion counterpart and uses them to generate spatially adaptive modulation weights that selectively regulate low-level responses at each pixel. This asymmetric, content-adaptive guidance mechanism enables MGM to suppress background interference while reinforcing semantically consistent structures, particularly along object boundaries. As a result, MGM is not merely an attention-based refinement module but a dedicated semantic guidance framework tailored to alleviate boundary blurring and segmentation discontinuities in remote sensing imagery.

3.5. Architecture Variants

For a fair comparison and drawing inspiration from the setup in [11], we construct three variants of the ArgusNet model based on the proposed module, corresponding to different scales: Tiny (T), Small (S), and Base (B). The default window size p is set to

8, and the expansion ratio of each FFN layer is 4 for all experiments. The architectural hyper-parameters of these model variants are

- ArgusNet-T: $C = \{64, 128, 320, 512\}$, layer numbers = $\{2, 2, 6, 2\}$
- ArgusNet-S: $C = \{96, 192, 384, 512\}$, layer numbers = $\{2, 2, 9, 2\}$
- ArgusNet-B: $C = \{128, 256, 512, 768\}$, layer numbers = $\{2, 2, 9, 2\}$

where C denotes the number of channels in each stage.

4. Experiments and Results

4.1. Datasets

The effectiveness of our proposed ArgusNet has been validated on two datasets: LoveDA Dataset [55] and Potsdam Dataset [56].

- (1) LoveDA Dataset: A high-resolution remote sensing dataset for domain-adaptive semantic segmentation, addressing urban–rural disparities. Collected by Wuhan University’s RSDIEA team, it includes 5987 images (0.3 m resolution) covering 536.15 km² across three Chinese cities, with 166,768 annotated objects and 7 land-cover categories. The dataset is split into urban (buildings, roads) and rural (farmland, water bodies) domains to study cross-domain adaptation. Challenges include multi-scale objects and complex backgrounds. In our study, we merged urban and rural data, trained/validated the model, and evaluated results via the official test platform.
- (2) Potsdam Dataset: Potsdam dataset contains 38 high-resolution images of 6000 × 6000 pixels over Potsdam City, Germany, and the ground sampling distance is 5 cm. The dataset is split into 24 images for training and 14 images for validation/testing. There are two modalities included in Potsdam dataset, i.e., true orthophoto (TOP) and digital surface model (DSM). While DSM consists of the near-infrared (NIR) band, TOP is corresponding to RGB image. In this work, we use TOP images from Potsdam and ignore DSM images. Following the experiment setup [57], we divide the dataset into 24 images for training and 14 images for testing. The testset of 14 images includes 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, and 7_13. This dataset contains five categories, namely impervious surface, building, tree, low vegetation, and car. Each image was divided into a series of 512 × 512 sub-images through an overlapping partitioning method with a step size of 256 × 256.

4.2. Implementation Details

- (1) Training Settings: To ensure a fair comparison, all competing methods are implemented within the MMSegmentation framework with a fixed random seed. All experiments are conducted on two NVIDIA RTX3090 GPUs (NVIDIA Corporation, Santa Clara, CA, USA) with a batch size of 16. For faster convergence, we adopt the AdamW optimizer with an initial learning rate of 3×10^{-4} and employ a PolyLR schedule to progressively decay the learning rate, without any fancy parameter tuning beyond that. For the competing methods, we use either the default configuration files provided by the framework or their officially claimed optimal configurations. The same data augmentation pipeline, including random scaling, random rotation, random cropping, random flipping, and random photometric distortion, was applied to all experiments.
- (2) Loss Function: We use cross-entropy loss (CE) as the loss function:

$$loss = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_i \log(\hat{y}_i) \quad (24)$$

where N represents the number of samples, C indicates the number of categories, y_i is the true label of pixel i , $\hat{y}_i$ is the predicted probability that pixel i belongs to the target class.

- (3) **Evaluation Metrics:** We Assess the model performance using metrics, such as overall mean F1(mF1) score and mean intersection over union (mIoU). They are defined as follows:

$$Recall = \frac{TP}{TP + FN} \quad (25)$$

$$Precision = \frac{TP}{TP + FP} \quad (26)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (27)$$

$$IoU = \frac{TP}{TP + FN + FP} \quad (28)$$

where TP , TN , FP , and FN denote the true positive, true negative, false positive, and false negative, respectively; mF1 denotes F1 score averaged over all the categories; and mIoU denotes the IoU averaged over all the categories.

4.3. Ablation Study

To evaluate the effectiveness of the proposed method, we conduct ablation experiments on the ArgusNet-Tiny and the more challenging LoveDA dataset and report the performance scores on the validation set.

- (1) **Effectiveness of Routing Strategies.** To evaluate the effectiveness of the routing strategy, this subsection conducts a controlled experiment. The baseline model is defined as a single shared vector with routing disabled (Table 1, last row). Multiple parameter settings are compared using a controlled-variable approach, and the performance is assessed on the validation set of the LoveDA dataset (see Table 1 for details). The results indicate that the baseline model achieves the lowest performance (mIoU = 50.12, mF1 = 65.37), whereas all routing-enabled models exhibit significant improvements. The best performance is obtained when $K_s = 1$, $K_e = 1$, and $T = 16$, where mIoU increases to 53.22 (+3.1 percentage points) and mF1 increases to 68.44 (+3.07 percentage points). This configuration demonstrates a balance between feature diversity and utilization. However, when T is increased from 16 to 32 under the setting $K_s = K_e = 1$, mIoU decreases by 0.39 and mF1 decreases by 0.33. Under the setting $K_s = 0$, $K_e = 2$, the reductions are more pronounced, with mIoU and mF1 decreasing by 0.53 and 0.42, respectively. This degradation arises because an excessively large T leads to sparse feature allocation. In contrast, when T is fixed, increasing the number of experts K_e results in only minor fluctuations in performance, suggesting that the current number of experts is already sufficient for the given task. Further increasing K_e introduces knowledge redundancy and has a limited impact on overall performance.

Table 1. Ablation Study on Different Configurations of K_s , K_e and T (Baseline: Shared Vector-only in Last Row, **Bold:** Best Result).

K_s	K_e	T	mIoU	mF1
0	1	16	52.72	68.1
0	1	32	52.17	67.67
0	2	16	52.73	68.11
0	2	32	52.2	67.69
1	1	16	53.22	68.44
1	1	32	52.83	68.11
1	2	16	53.2	68.43
1	2	32	52.79	68.07
1	0	0	50.12	65.37

To further elucidate the effectiveness of the dynamic routing mechanism in AWAA, we visualized the selection preferences of routing vectors under different local semantic conditions, as shown in Figure 6. Specifically, we selected a representative image with dense small building targets and visualized its routing selection preferences in the stage 2 block 2 and stage 3 block 3. From the figure, it can be observed that the routing mechanism selects the same vector for windows with similar features. In particular, in stage 2, there appears to be a preference for selecting vector 1 for dense buildings, while in stage 3, there seems to be a preference for selecting vector 4. This adaptive selection mechanism is a key reason for the effective performance of AWAA.

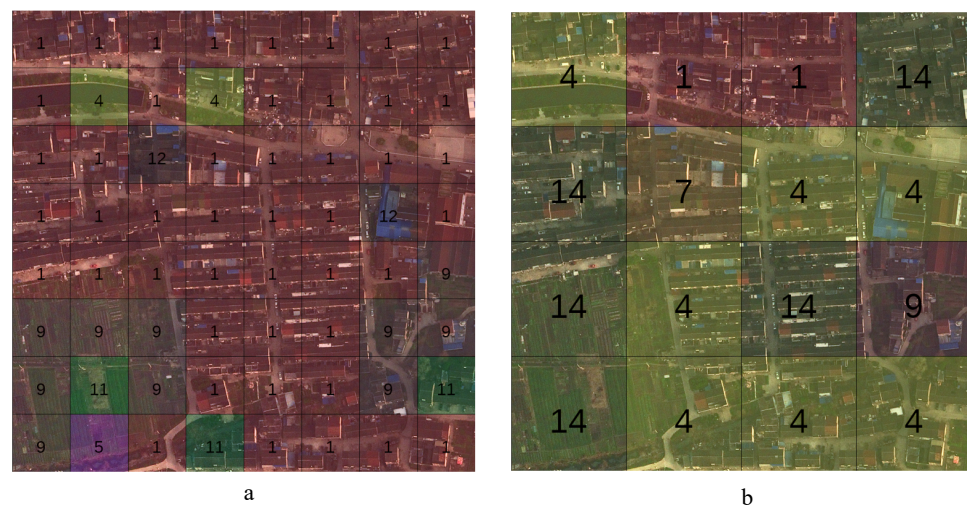


Figure 6. Visualization of the routable vector selection preferences for each window in stage 2 and stage 3 of ArgusNet. The figure illustrates that the routing mechanism assigns the same vector to windows with similar features, which is the key reason why AWAA improves performance. The numbers represent vector indices, while different colors denote different windows. (a) Visualization of selection preferences in stage 2. (b) Visualization of selection preferences in stage 3.

- (2) Effectiveness of HGLB and MGM. Within the HGLB, AWAA and VSS are responsible for local and global feature modeling, respectively. To verify their necessity, we adopt a switch-based controlled variable strategy, combined with the on/off configuration of MGM, and the complete results are summarized in Table 2. The experiments demonstrate that the best segmentation performance is achieved when all modules are enabled (last row of Table 2). When VSS is retained while AWAA is disabled (second row of Table 2), the network can still capture long-range dependencies, but the lack of local boundary characterization leads to a drop in mIoU from 53.22% to 49.77% and in mF1 to 65.66%.

In particular, the IoU of Building and Road decreases by 3.88 and 3.59 percentage points, respectively. Qualitative visualizations in Figure 7 further corroborate this observation. The red box in the first row shows that large-scale buildings exhibit incomplete segmentation with noticeable holes, and in the second row, the boundaries of small-scale water bodies become entangled. Conversely, when AWAA is retained but VSS is disabled (first row of Table 2), the absence of global contextual information causes the mIoU to drop further to 47.26% and mF1 to 64.2%. As shown in the first-row red box of Figure 7, large-scale buildings are almost entirely misclassified, while the second-row water boundaries remain intact. This indicates that large-scale and continuously distributed objects strongly rely on the global receptive field. Only when AWAA and VSS are simultaneously activated can local details and global semantics complement each other, leading to superior performance across all categories. Although certain building edges remain imperfectly

segmented, the overall results are substantially improved. Building on this configuration, disabling MGM (third row of Table 2) causes significant performance degradation, with mIoU reduced by 1.09 percentage points and mF1 reduced by 1.02 percentage points. The qualitative results also reveal incomplete segmentation of large-scale buildings and indistinct boundaries in small-scale water bodies, suggesting that the complementary use of global and local information is not sufficiently exploited.

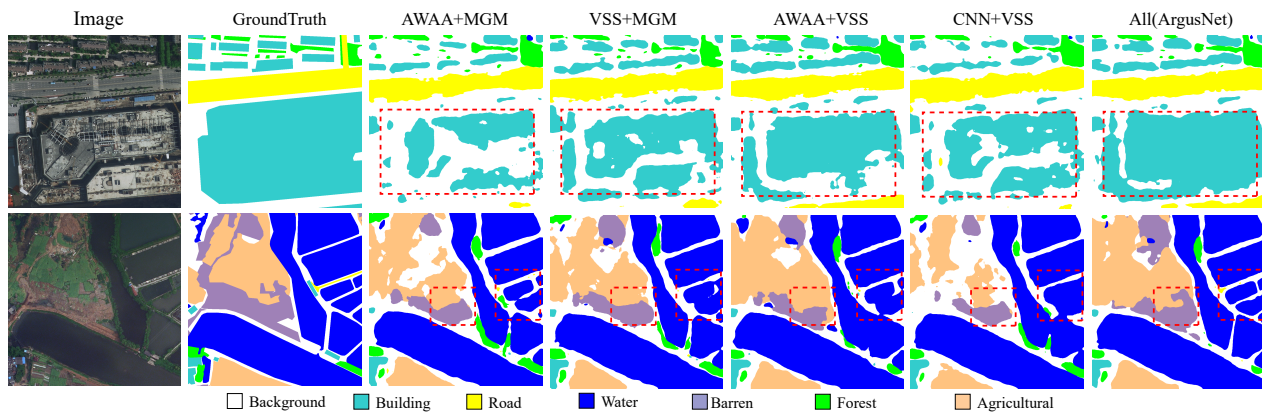


Figure 7. Visualization Results of Ablation Experiments on the LoveDA Val Dataset (Red box: Areas requiring particular comparison).

To intuitively demonstrate the efficacy of the MGM, we conducted a comparative visualization of the feature maps before and after processing (Figure 8). The results show that our module effectively suppresses noisy responses in background regions, resulting in cleaner feature maps with more focused activations on the target areas. Specifically, in the topographically complex transition zones between forests and farmland (as shown in the first and second columns), the boundaries between the two can be more clearly distinguished after processing with the MGM. Furthermore, MGM demonstrates significant advantages in challenging scenarios with shadows and occlusions (as shown in the third and fourth columns). In the third column, despite the occlusion caused by buildings, the features of the trees located between them are effectively enhanced, whereas they are mostly ignored without processing; in the fourth column, the noise in the building feature map is significantly suppressed after MGM processing, resulting in a cleaner and sharper feature representation.

To further validate the effectiveness of the overall HGLB design, we also replace the internal AWAA with a convolutional layer of comparable parameter size and disable MGM (fourth row of Table 2, where “★” denotes the convolution layer). This technique is widely adopted in other methods to mitigate limitations associated with local information. The experimental results show that our AWAA remains the most competitive approach. In summary, AWAA, VSS, and MGM are mutually supportive, collectively forming the cornerstone of the proposed method’s effectiveness in remote sensing semantic segmentation tasks.

Table 2. Ablation Study of the Proposed Module on the LoveDA Val Dataset (**Bold:** Best Results, ✓: Enable this module, ✗: Disable this module, ★: Replace this module with a CNN).

HGLB			IoU Per Class (%)							mIoU	mF1
AWAA	VSS	MGM	Background	Building	Road	Water	Barren	Forest	Agricultural		
✓	✗	✓	49.83	54.44	47.95	64.74	27.36	38.52	48.01	47.26	64.2
✗	✓	✓	52.14	57.57	53.04	65.62	27.5	42.73	49.75	49.77	65.66
✓	✓	✗	53.29	60.17	55.81	70.48	28.8	38.28	58.06	52.13	67.42
★	✓	✗	52.83	58.72	54.69	68.02	27.31	38.11	56.14	50.83	65.92
✓	✓	✓	54.63	61.45	56.63	71.0	29.44	39.2	60.17	53.22	68.44

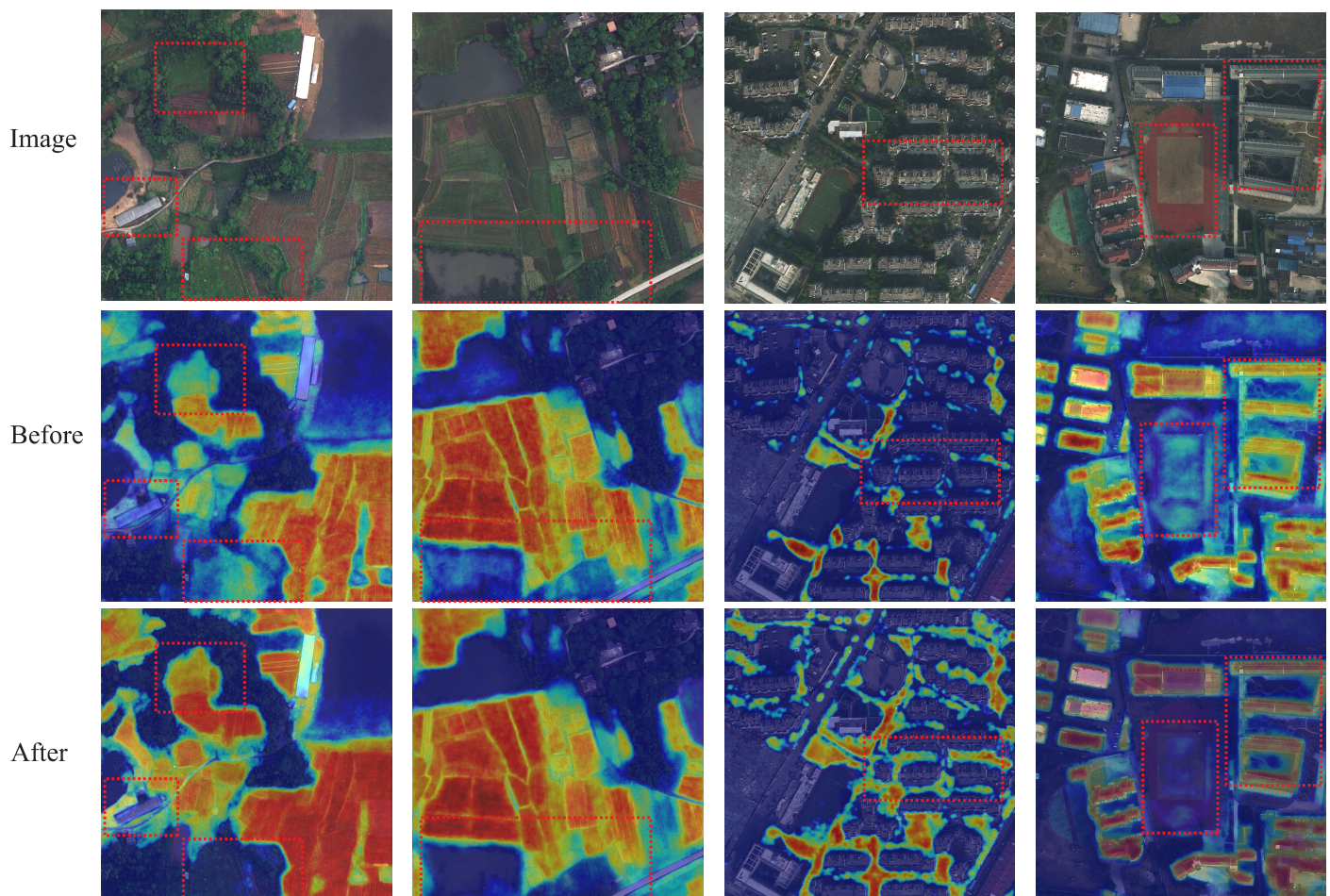


Figure 8. Comparison of class activation maps before and after MGM processing. Results demonstrate that the MGM module effectively suppresses background noise (Red box: Areas requiring particular comparison).

4.4. Comparison to the State-of-the-Art Methods

To comprehensively evaluate the effectiveness of ArgusNet, we designed and conducted a systematic comparative study covering a wide range of state-of-the-art semantic segmentation methods. Specifically, for CNN-based approaches, we selected DeepLabV3+ [44] and FCN [5] (both employing ResNet50 [58] as the backbone) as well as the recently proposed general-purpose backbone OverLoCK [59]. For Transformer-based and hybrid architectures, we included SwinTransformer [11], SegFormer [26], UNetFormer [25] (also with ResNet50 as the backbone), and AerialFormer [57]. In addition, RS3Mamba [35] and UNetMamba [60], which are based on the Mamba architecture, were incorporated as benchmarks. This comparative scheme spans traditional approaches to cutting-edge techniques, while taking into account the characteristics of remote sensing image segmentation tasks. Such a design ensures that the evaluation of ArgusNet's performance advantage is both objective and comprehensive. The parameter sizes and computational costs of all models are summarized in Table 3.

- (1) Comparisons on the LoveDA Dataset: The comparative results on the LoveDA dataset are summarized in Table 4. Across three model scales (T, S, B), ArgusNet consistently achieves segmentation performance that is comparable to or better than existing state-of-the-art methods. Overall, ArgusNet-B attains an mIoU of 54.89%, the highest among all compared models, exceeding OverLoCK-B with a similar parameter scale

by 1.32 percentage points. This demonstrates the effectiveness of the proposed synergistic design of HGLB and MGM. At the medium scale, ArgusNet-S achieves an mIoU of 53.36%, outperforming SegFormer-B4 (52.94%) and Swin-T (51.90%), indicating that the proposed approach offers superior cross-scale feature integration while maintaining lower computational cost. The lightweight ArgusNet-T also achieves competitive results, with an mIoU of 52.7%, surpassing SegFormer-B2 (52.35%) and UNetMamba (51.20%), highlighting its favorable parameter–efficiency trade-off. From the perspective of class-wise performance, the ArgusNet series delivers pronounced advantages in categories with complex textures and high inter-class similarity, such as Forest and Agricultural. For instance, ArgusNet-B achieves 65.61% IoU on the Agricultural category, exceeding the second-best method by 2.42 percentage points, thereby verifying that AWAA significantly enhances detail preservation in boundary-complex regions. Moreover, ArgusNet also maintains competitive performance on structured categories such as Building, reflecting the general adaptability of local–global feature modeling across diverse land-cover classes.

Table 3. Information of Different Models, where the Computational Complexity is Measured on 512×512 Images.

Method	Year	Type	Params (M)	Flops (G)
DeepLabV3+ (R50)	2018	CNN	41.2	177
SegFormer-B2	2021	Attention	24.7	25
UNetFormer (R50)	2021	Hybrid	32	66
UNetMamba	2024	SSM	14	38
RS3Mamba	2024	SSM	43.4	40
ArgusNet-T(Ours)	–	Hybrid	38.5	67
FCN (R50)	2015	CNN	47.1	198
SegFormer-B4	2021	Attention	61.4	59
SwinTransformer-T	2021	Attention	58.9	236
ArgusNet-S(Ours)	–	Hybrid	60.6	111
SwinTransformer-B	2021	Attention	120	298
AerialFormer-B	2024	Hybrid	114	133
OverLoCK-B	2025	CNN	127	297
ArgusNet-B(Ours)	–	Hybrid	113	197

Table 4. Comparison of Segmentation Results on the LoveDA Dataset (**Bold:** Best Overall Results, Underline: Best Among Similar-Sized Models).

Method	IoU Per Class (%)							mIoU
	Background	Building	Road	Water	Barren	Forest	Agricultural	
DeepLabV3+ R50	41.41	55.88	54.25	79.62	22.56	42.69	62.47	51.7
SegFormer B2	43.56	56.13	56.11	79.52	22.38	46.54	62.18	52.35
UNetFormer R50	41.95	55.95	53.65	77.97	16.8	44.68	61.85	50.4
UNetMamba	42.55	54.8	54.91	78.58	<u>24.14</u>	45.35	58.11	51.2
RS3Mamba	43.69	55.17	54.88	77.63	16.72	47.23	62.76	51.15
ArgusNet-T(Ours)	<u>44.4</u>	<u>56.2</u>	<u>56.91</u>	<u>79.73</u>	21.35	<u>47.45</u>	<u>62.88</u>	<u>52.7</u>
FCN R50	44.83	57.3	55.86	79.17	25.08	45.36	62.48	52.87
SegFormer B4	44.89	<u>57.41</u>	55.12	79.58	<u>25.23</u>	45.69	62.67	52.94
Swin-T	44.27	54.44	56.46	78.42	23.99	44.12	61.58	51.9
ArgusNet-S(Ours)	<u>45.47</u>	56.64	<u>56.95</u>	<u>79.94</u>	23.26	<u>47.49</u>	<u>63.8</u>	<u>53.36</u>
Swin-B	44.62	58.11	58.18	79.18	23.5	46.53	58.42	52.65
AerialFormer-B	46.12	59.15	58.12	80.57	17.47	46.47	63.14	53.0
OverLoCK-B	46.15	58.62	55.45	80.12	24.81	46.68	63.19	53.57
ArgusNet-B(Ours)	45.85 ± 0.12	$\mathbf{59.27} \pm 0.17$	$\mathbf{59.3} \pm 0.04$	$\mathbf{80.67} \pm 0.09$	$\mathbf{25.7} \pm 0.23$	$\mathbf{47.85} \pm 0.13$	$\mathbf{65.61} \pm 0.2$	$\mathbf{54.89} \pm 0.04$

A qualitative comparison is presented in Figure 9. In the first row (red box), ArgusNet delineates water boundaries more precisely, effectively preventing background noise erosion. In the second row, under scenarios with highly complex spatial structures and high inter-class similarity, ArgusNet achieves more accurate building segmentation, markedly reducing boundary blurring and misclassification. In the third row (red box), ArgusNet demonstrates robust discriminability between highly similar Forest and Agricultural textures, not only preserving the integrity of forest regions but also correctly identifying the Agricultural area in the lower-right corner. In the same row, the central red box further illustrates ArgusNet's ability to generate clearer and more distinguishable boundaries in dense small-object scenarios. These observations validate the effectiveness of the collaborative mechanism of AWAA, VSS, and MGM in handling complex remote sensing scenes.

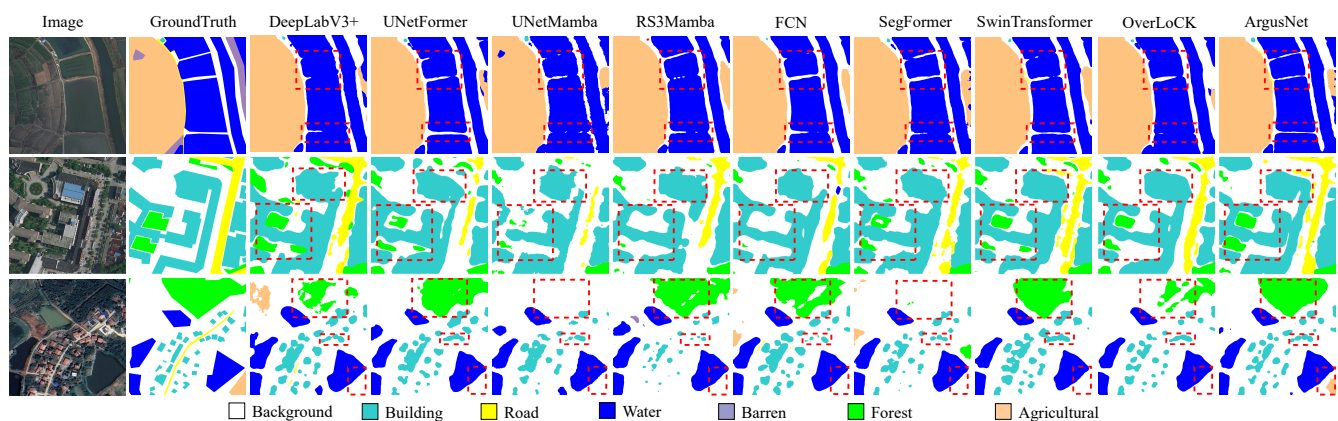


Figure 9. Comparison of the proposed ArgusNet with other methods on the LoveDA dataset (Red box: Areas requiring particular comparison).

- (2) Comparisons on the Potsdam Dataset: The comparative results on the Potsdam dataset are summarized in Table 5. ArgusNet consistently exhibits superior segmentation performance across three model scales. In particular, ArgusNet-B achieves the highest overall performance, with 89.31% in mIoU and 94.55% in mF1, and also obtains the best results in all categories. Compared with OverLoCK-B of similar parameter size, ArgusNet-B improves mF1 by 0.56 percentage points, indicating its advantage in fine-grained segmentation of high-resolution urban scenes. The medium-scale ArgusNet-S attains 88.87% in mIoU and 94.13% in mF1, outperforming SegFormer-B4 (88.50% and 93.92%) and Swin-T (88.23% and 93.61%). It also achieves leading performance in all categories, demonstrating the efficiency and generalization of the joint modeling of AWAA and VSS. The lightweight ArgusNet-T achieves 88.74% in mIoU, which surpasses UNetMamba (86.90%) and RS3Mamba (87.72%). Category-wise analysis shows that the ArgusNet series achieves accuracy comparable to the best competing methods in well-structured classes such as Impervious Surface and Building, while delivering the highest accuracy in small-object categories such as Car. Notably, ArgusNet-B reaches 94.77% in the Car category, highlighting its advantage in handling small objects and capturing precise boundaries. Furthermore, ArgusNet also performs favorably in the Tree category, indicating that the proposed global–local feature fusion strategy enables stable performance in vegetation classes with large texture variations.

Qualitative comparisons are presented in Figure 10. Methods such as UNetMamba produce incorrect predictions due to occlusions and illumination variations, which can be attributed to insufficient exploitation of global and local contextual information. In contrast, ArgusNet yields more accurate predictions. For instance, the gap between two buildings in

the third row is often misclassified under shadow and poor illumination, whereas ArgusNet alleviates this issue and produces clearer boundaries. Overall, the experiments on Potsdam further confirm that ArgusNet effectively balances fine-grained detail representation with large-scale contextual understanding in high-resolution remote sensing imagery.

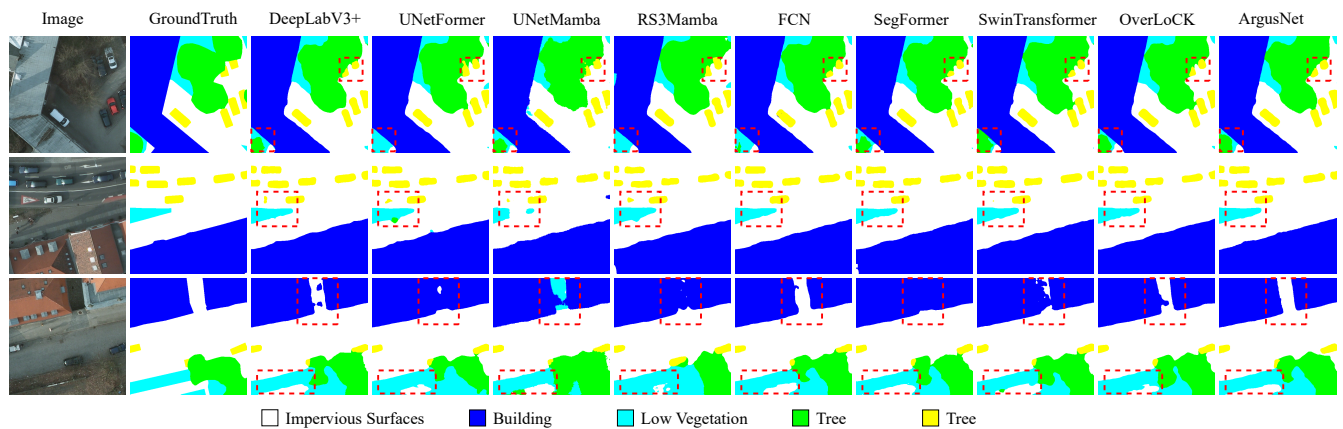


Figure 10. Comparison of the proposed ArgusNet with other methods on the Potsdam dataset (Red box: Areas requiring particular comparison).

Table 5. Comparison of Segmentation Results on the Potsdam Dataset (**Bold:** Best Overall Results, Underline: Best Among Similar-Sized Models).

Method	IoU Per Class (%)					mIoU	mF1
	Impervious Surface	Building	Low Vegetation	Tree	Car		
DeepLabV3Plus R50	90.01	95.13	79.26	80.08	93.4	87.58	92.92
SegFormer B2	90.03	95.28	80.1	80.15	93.24	87.76	92.81
UNetFormer R50	89.21	94.82	78.98	78.96	93.03	87.0	92.9
UNetMamba	89.37	94.52	79.41	79.32	91.92	86.9	92.88
RS3Mamba	90.09	95.48	80.12	79.79	93.13	87.72	93.33
ArgusNet-T(Ours)	<u>90.85</u>	<u>95.84</u>	<u>81.51</u>	<u>80.95</u>	<u>94.58</u>	<u>88.74</u>	<u>93.81</u>
FCN R50	89.99	95.0	79.94	80.9	94.42	88.05	93.52
SegFormer B4	90.61	95.38	81.09	81.05	94.38	88.5	93.92
Swin T	90.25	95.47	81.14	80.16	94.11	88.23	93.61
ArgusNet-S(Ours)	<u>90.98</u>	<u>95.89</u>	<u>81.63</u>	<u>81.11</u>	<u>94.73</u>	<u>88.87</u>	<u>94.13</u>
Swin-B	91.02	96.07	81.59	81.38	94.4	88.89	94.0
AerialFormer-B	90.55	95.78	80.55	80.07	94.1	88.21	93.6
OverLoCK-B	90.91	96.13	81.36	81.34	94.55	88.86	93.99
ArgusNet-B(Ours)	91.56 $\pm$ 0.03	96.81 $\pm$ 0.02	81.69 $\pm$ 0.04	81.72 $\pm$ 0.04	94.77 $\pm$ 0.05	89.31 $\pm$ 0.0004	94.55 $\pm$ 0.03

- (3) Complexity Comparison: The complexity of each model was measured in terms of the number of parameters (M) and floating-point operations (FLOPs). All results were obtained on a single NVIDIA 3090 GPU with an input size of 512×512 , as reported in Table 3. Methods based on CNNs, such as DeepLabV3+ and FCN, require substantially higher computation due to dilated convolutions in the ResNet backbone. Under comparable parameter sizes, their FLOPs are more than 2.6 times those of ArgusNet-T, while producing lower segmentation accuracy. ArgusNet-S and ArgusNet-B achieve competitive accuracy with only half the computational cost of Swin Transformer, indicating that ArgusNet provides an effective balance between accuracy and efficiency.

Although ArgusNet adopts linear-complexity building blocks, its overall computational cost is higher than that of lightweight Transformer-based models such as SegFormer-B4. Specifically, ArgusNet-S exhibits approximately twice the FLOPs of SegFormer-B4 while achieving a moderate mIoU improvement. This gap primarily stems from the

parallel global–local design in HGLB, where AWAA and SS2D are executed concurrently to explicitly preserve fine-grained local structures while modeling long-range dependencies.

We note that ArgusNet is not designed to maximize accuracy-per-FLOP under strict efficiency constraints but rather to target high-resolution remote sensing scenarios where dense small objects, low inter-class contrast, and boundary ambiguity are dominant challenges. In such cases, the additional computational overhead enables more reliable local detail preservation and semantic consistency, which is reflected in improved qualitative results and boundary integrity.

Moreover, the use of linear-complexity attention and state-space modeling avoids the quadratic memory growth inherent to standard self-attention, making ArgusNet more scalable to ultra-high-resolution imagery despite higher constant factors in FLOPs. Therefore, ArgusNet represents a deliberate accuracy–efficiency trade-off, prioritizing robustness and segmentation quality in complex remote sensing scenes rather than minimal computational cost.

- (4) **Generalization Capability Test:** In order to evaluate the generalization ability of ArgusNet under extreme data conditions, we conducted an additional domain generalization experiment on the LoveDA dataset. The model was trained on the Urban domain and tested on the Rural domain, following the same training process as previously described. We compared ArgusNet-T with the second and third best-performing methods of similar model size from Table 4, and the comparison results are shown in Table 6. Due to severe data imbalance and significant style differences, all three methods, including ArgusNet, experienced varying degrees of performance degradation. Nevertheless, our method still achieved the best results, which is consistent with the observations from previous experiments. The experimental results further confirm the strong potential of ArgusNet.

Table 6. Experimental results of Urban-to-Rural domain generalization on the LoveDA dataset. (**Bold:** Best Overall Results).

Method	IoU Per Class (%)							mIoU
	Background	Building	Road	Water	Barren	Forest	Agricultural	
DeepLabV3+ R50	57.41	35.31	36.99	53.1	11.0	7.91	52.59	36.33
SegFormer B2	49.64	41.74	38.1	55.24	6.83	26.52	52.82	38.69
ArgusNet-T(Ours)	55.89 ± 0.77	46.25 ± 0.56	39.14 ± 0.79	53.87 ± 1.24	7.92 ± 0.55	26.68 ± 0.69	53.8 ± 0.68	40.51 ± 0.2

5. Discussion

In this study, we proposed ArgusNet, a hybrid global–local representation learning framework that integrates AWAA, SS2D, and the MGM to address the key challenges in remote sensing semantic segmentation. The experiments conducted on the LoveDA and Potsdam datasets demonstrate that ArgusNet achieves consistent performance improvements over representative CNN-, Transformer-, and Mamba-based models. These results confirm that the proposed architectural components effectively enhance both global contextual modeling and fine-grained local detail extraction.

From the perspective of feature representation, AWAA plays a central role in improving the model’s discriminative capability. By extending additive attention into a window-based structure and incorporating a dynamic routing mechanism, AWAA strengthens local feature modeling while maintaining efficient long-range perception. Its synergy with the VSS module enables stable coordination between global dependencies and local structures, which is crucial for handling high-similarity land-cover types and complex spatial layouts. The improvements observed in categories such as Agricultural, Forest, and Building indi-

cate that ArgusNet is capable of resolving class ambiguity and preserving subtle texture differences that traditional approaches often fail to capture.

The decoder optimization strategy further contributes to these gains. The three-stage feature fusion enhances multi-scale consistency, and the MGM effectively narrows the semantic gap during feature aggregation by leveraging high-level contextual guidance. The qualitative results illustrate that ArgusNet produces smoother object boundaries, mitigates background interference in Water regions, and achieves more precise delineation in densely built areas. These observations highlight the importance of harmonizing hierarchical features when dealing with remote sensing images characterized by large intra-class variability and intricate structural patterns.

We note that for elongated and thin structures such as roads, the visual superiority of ArgusNet over competing methods may appear less immediately perceptible in qualitative comparisons, as illustrated by the LoveDA visualizations in Figure 9. This phenomenon should not be interpreted as a degradation in segmentation performance. Instead, it reflects the intrinsic characteristics of road-like objects, which often extend across wide spatial ranges and are frequently affected by shadows, vegetation occlusion, and illumination variations. These factors make the balance between local texture cues and global structural information inherently difficult to perceive in a visually consistent manner, particularly along long and continuous boundaries. As a result, even when quantitative metrics such as IoU indicate improved segmentation accuracy, the corresponding visual differences may remain subtle. From this perspective, the observed behavior is better understood as a perceptual boundary rather than a fundamental methodological limitation. Nevertheless, the analysis also reveals concrete directions in which the current design could be further strengthened for elongated, slender structures. In particular, segmentation of road-like objects may benefit from explicitly encoding structural continuity and orientation consistency, for example, by incorporating structure-aware priors that emphasize connectivity and topology along long axes or by introducing directional context modeling that aligns feature aggregation with dominant geometric directions. In addition, geometry-constrained feature refinement—such as boundary-aware modulation or continuity-preserving regularization—could help reduce local fragmentation and improve visual coherence along extended boundaries. These strategies are complementary to the existing global–local fusion in ArgusNet and are expected to be especially effective in scenarios where quantitative improvements are driven by enhanced long-range consistency that is not immediately apparent in qualitative visualizations.

Another potential direction for improvement lies in the integration of multimodal data. In the Potsdam experiments, only TOP images were used, whereas DSM data could provide valuable elevation information for differentiating categories with height variations, such as Buildings and Trees. Extending ArgusNet to handle multimodal fusion or cross-view consistency may enhance robustness in broader remote sensing applications.

Overall, the findings in this work reinforce the significance of jointly optimizing global dependency modeling, local discriminative feature extraction, and multi-scale semantic alignment. ArgusNet provides an effective and balanced solution for remote sensing semantic segmentation, and the identified limitations offer meaningful insights for future research on structure-aware and scale-adaptive segmentation models.

6. Conclusions

In this work, we proposed ArgusNet, a hybrid global–local semantic segmentation framework for high-resolution remote sensing imagery. By integrating AWAA with SS2D-based state-space modeling in a unified HGLB, the proposed method effectively balances fine-grained local feature extraction and long-range global context modeling under linear

computational complexity. In addition, the MGM further enhances semantic consistency during feature fusion by dynamically guiding low-level representations with high-level semantic information.

Extensive experiments conducted on the LoveDA and Potsdam datasets demonstrate that ArgusNet achieves competitive and consistently strong segmentation performance compared with representative CNN-, Transformer-, and Mamba-based methods. Moreover, the cross-domain generalization experiments indicate that the proposed framework is able to maintain robust performance when transferred across datasets with distinct spatial resolutions and scene characteristics, highlighting its generalization capability within the evaluated experimental scope.

It is worth noting that, while the current study focuses on RGB-based remote sensing segmentation, the modular design of ArgusNet provides a flexible foundation for future extensions. In particular, exploring cross-domain generalization under more diverse acquisition conditions and validating the framework in multimodal settings (e.g., RGB–DSM fusion) remain promising directions for further investigation. These aspects may be valuable directions for future research.

Author Contributions: Conceptualization, H.C. and Y.F.; methodology, Y.F.; software, Y.F. and K.W.; validation, M.L.; formal analysis, Y.Z.; investigation, T.X. and J.J.; resources, H.Z.; data curation, C.W.; writing—original draft preparation, Y.F.; writing—review and editing, H.C.; visualization, K.W.; supervision, C.W.; project administration, C.W.; funding acquisition, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Industrial Technology and Development Project of the Development and Reform Commission of Jilin Province, grant number 2023C030-3.

Data Availability Statement: In this study, we utilized the Potsdam and LoveDA datasets. The Potsdam dataset can be accessed at <https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/default.aspx>, accessed on 9 July 2025. The LoveDA dataset is publicly available at <https://zenodo.org/records/5706578>, accessed on 9 July 2025. Please refer to the provided links for further details.

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

References

1. Maus, V.; Câmara, G.; Cartaxo, R.; Sanchez, A.; Ramos, F.M.; De Queiroz, G.R. A time-weighted dynamic time warping method for land-use and land-cover mapping. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 3729–3739. [CrossRef]
2. Sahar, L.; Muthukumar, S.; French, S.P. Using aerial imagery and GIS in automated building footprint extraction and shape recognition for earthquake risk assessment of urban inventories. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 3511–3520. [CrossRef]
3. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]
4. Chen, J.; Xia, M.; Wang, D.; Lin, H. Double branch parallel network for segmentation of buildings and waters in remote sensing images. *Remote Sens.* **2023**, *15*, 1536. [CrossRef]
5. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
6. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
7. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587. [CrossRef]
8. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual attention network. *Comput. Vis. Media* **2023**, *9*, 733–752. [CrossRef]
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
10. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

11. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
12. Wu, C.; Wu, F.; Qi, T.; Huang, Y.; Xie, X. Fastformer: Additive attention can be all you need. *arXiv* **2021**, arXiv:2108.09084. [[CrossRef](#)]
13. Shaker, A.; Maaz, M.; Rasheed, H.; Khan, S.; Yang, M.H.; Khan, F.S. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 17425–17436.
14. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752. [[CrossRef](#)]
15. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024; pp. 62429–62442.
16. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063.
17. Li, H.; Pan, H.; Liu, X.; Ren, J.; Du, Z.; Cao, J. GLVMamba: A Global-Local Visual State Space Model for Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4412115. [[CrossRef](#)]
18. Zhao, S.; Chen, H.; Zhang, X.; Xiao, P.; Bai, L.; Ouyang, W. Rs-mamba for large remote sensing image dense prediction. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5633314. [[CrossRef](#)]
19. Luo, H.; Chen, C.; Fang, L.; Zhu, X.; Lu, L. High-resolution aerial images semantic segmentation using deep fully convolutional network with channel attention mechanism. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3492–3507. [[CrossRef](#)]
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241.
21. Li, X.; Yang, X.; Li, X.; Lu, S.; Ye, Y.; Ban, Y. GCDB-UNet: A novel robust cloud detection approach for remote sensing images. *Knowl.-Based Syst.* **2022**, *238*, 107890. [[CrossRef](#)]
22. Gao, Y.; Cao, H.; Cai, W.; Zhou, G. Pixel-level road crack detection in UAV remote sensing images based on ARD-Unet. *Measurement* **2023**, *219*, 113252. [[CrossRef](#)]
23. Wang, Z.; Wang, J.; Yang, K.; Wang, L.; Su, F.; Chen, X. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+. *Comput. Geosci.* **2022**, *158*, 104969. [[CrossRef](#)]
24. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. [[CrossRef](#)]
25. 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](#)]
26. 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.
27. Hatamizadeh, A.; Yin, H.; Heinrich, G.; Kautz, J.; Molchanov, P. Global context vision transformers. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 12633–12646.
28. Maaz, M.; Shaker, A.; Cholakkal, H.; Khan, S.; Zamir, S.W.; Anwer, R.M.; Shahbaz Khan, F. Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 3–20.
29. Kitaev, N.; Kaiser, L.; Levskaya, A. Reformer: The efficient transformer. *arXiv* **2020**, arXiv:2001.04451. [[CrossRef](#)]
30. 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.
31. Hamilton, M.; Zhang, Z.; Hariharan, B.; Snavely, N.; Freeman, W.T. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv* **2022**, arXiv:2203.08414. [[CrossRef](#)]
32. Wang, X.; Girdhar, R.; Yu, S.X.; Misra, I. Cut and learn for unsupervised object detection and instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3124–3134.
33. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4015–4026.
34. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. Dinov2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
35. Ma, X.; Zhang, X.; Pun, M.O. Rs 3 mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6011405. [[CrossRef](#)]
36. Ding, H.; Xia, B.; Liu, W.; Zhang, Z.; Zhang, J.; Wang, X.; Xu, S. A novel mamba architecture with a semantic transformer for efficient real-time remote sensing semantic segmentation. *Remote Sens.* **2024**, *16*, 2620. [[CrossRef](#)]

37. Wan, Z.; Zhang, P.; Wang, Y.; Yong, S.; Stepputtis, S.; Sycara, K.; Xie, Y. Sigma: Siamese mamba network for multi-modal semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2025; pp. 1734–1744.
38. Ruan, J.; Li, J.; Xiang, S. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv* **2024**, arXiv:2402.02491. [\[CrossRef\]](#)
39. Zhou, W.; Kamata, S.I.; Wang, H.; Wong, M.S.; Hou, H.C. Mamba-in-mamba: Centralized mamba-cross-scan in tokenized mamba model for hyperspectral image classification. *Neurocomputing* **2025**, *613*, 128751. [\[CrossRef\]](#)
40. Dong, J.; Yin, H.; Li, H.; Li, W.; Zhang, Y.; Khan, S.; Khan, F.S. Dual hyperspectral mamba for efficient spectral compressive imaging. *arXiv* **2024**, arXiv:2406.00449. [\[CrossRef\]](#)
41. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
42. Wang, X.; Wang, S.; Ning, C.; Zhou, H. Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7918–7932. [\[CrossRef\]](#)
43. Gu, Y.; Wang, B.; Xu, B. A FPN-based framework for vehicle detection in aerial images. In *Proceedings of the 2018 2nd International Conference on Video and Image Processing*, Hong Kong, 29–31 December 2018; pp. 60–64.
44. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 8–14 September 2018; pp. 801–818.
45. Liu, W.; Shu, Y.; Tang, X.; Liu, J. Remote sensing image segmentation using dual attention mechanism Deeplabv3+ algorithm. *Trop. Geogr.* **2020**, *40*, 303–313.
46. Li, Y.; Chen, W.; Huang, X.; Gao, Z.; Li, S.; He, T.; Zhang, Y. MFVNet: A deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation. *Sci. China Inf. Sci.* **2023**, *66*, 140305. [\[CrossRef\]](#)
47. Ma, P.; Ren, J.; Sun, G.; Zhao, H.; Jia, X.; Yan, Y.; Zabalza, J. Multiscale superpixelwise prophet model for noise-robust feature extraction in hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5508912. [\[CrossRef\]](#)
48. Lobry, S.; Marcos, D.; Murray, J.; Tuia, D. RSVQA: Visual question answering for remote sensing data. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8555–8566. [\[CrossRef\]](#)
49. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual attention inception network for remote sensing visual question answering. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5606514. [\[CrossRef\]](#)
50. Zi, X.; Xiao, J.; Shi, Y.; Tao, X.; Li, J.; Braytee, A.; Prasad, M. RSVLM-QA: A Benchmark Dataset for Remote Sensing Vision Language Model-based Question Answering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, Dublin, Ireland, 27–31 October 2025; pp. 12905–12911.
51. Wang, J.; Xuan, W.; Qi, H.; Liu, Z.; Liu, K.; Wu, Y.; Chen, H.; Song, J.; Xia, J.; Zheng, Z.; et al. DisasterM3: A Remote Sensing Vision-Language Dataset for Disaster Damage Assessment and Response. *arXiv* **2025**, arXiv:2505.21089.
52. Wang, Z.; Prabha, R.; Huang, T.; Wu, J.; Rajagopal, R. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 5805–5813. [\[CrossRef\]](#)
53. Xiao, A.; Xuan, W.; Wang, J.; Huang, J.; Tao, D.; Lu, S.; Yokoya, N. Foundation models for remote sensing and earth observation: A survey. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 297–324. [\[CrossRef\]](#)
54. Lu, S.; Guo, J.; Zimmer-Dauphinee, J.R.; Nieusma, J.M.; Wang, X.; Wernke, S.A.; Huo, Y. Vision foundation models in remote sensing: A survey. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 190–215. [\[CrossRef\]](#)
55. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.
56. 2D Semantic Labeling Contest—Potsdam. ISPRS 2D Semantic Labeling Contest. Available online: <https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx> (accessed on 2 February 2026).
57. Hanyu, T.; Yamazaki, K.; Tran, M.; McCann, R.A.; Liao, H.; Rainwater, C.; Adkins, M.; Cothren, J.; Le, N. AerialFormer: Multi-resolution transformer for aerial image segmentation. *Remote Sens.* **2024**, *16*, 2930. [\[CrossRef\]](#)
58. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
59. Lou, M.; Yu, Y. OverLoCK: An Overview-first-Look-Closely-next ConvNet with Context-Mixing Dynamic Kernels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, Nashville, TN, USA, 11–15 June 2025; pp. 128–138.
60. Zhu, E.; Chen, Z.; Wang, D.; Shi, H.; Liu, X.; Wang, L. Unetmamba: An efficient unet-like mamba for semantic segmentation of high-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *22*, 6001205. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.