*remote sensing*

MDPI

*Article*

# FEMNet: A Feature-Enriched Mamba Network for Cloud Detection in Remote Sensing Imagery

**Weixing Liu [1], Bin Luo [1], Jun Liu [1,*], Han Nie [1] and Xin Su [2]**

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; weix_liu@whu.edu.cn (W.L.); luob@whu.edu.cn (B.L.); niehan@whu.edu.cn (H.N.)
[2] School of Artificial Intelligence, Wuhan University, Wuhan 430079, China; xinsu.rs@whu.edu.cn
* Correspondence: liujunand@whu.edu.cn

**Abstract**

Accurate and efficient cloud detection is critical for maintaining the usability of optical remote sensing imagery, particularly in large-scale Earth observation systems. In this study, we propose FEMNet, a lightweight dual-branch network that combines state space modeling with convolutional encoding for multi-class cloud segmentation. The Mamba-based encoder captures long-range semantic dependencies with linear complexity, while a parallel CNN path preserves spatial detail. To address the semantic inconsistency across feature hierarchies and limited context perception in decoding, we introduce the following two targeted modules: a cross-stage semantic enhancement (CSSE) block that adaptively aligns low- and high-level features, and a multi-scale context aggregation (MSCA) block that integrates contextual cues at multiple resolutions. Extensive experiments on five benchmark datasets demonstrate that FEMNet achieves state-of-the-art performance across both binary and multi-class settings, while requiring only 4.4M parameters and 1.3G multiply–accumulate operations. These results highlight FEMNet's suitability for resource-efficient deployment in real-world remote sensing applications.

**Keywords:** cloud detection; remote sensing image; Mamba; deep learning

Academic Editor: Silvia Liberata Ullo

Received: 11 June 2025
Revised: 12 July 2025
Accepted: 21 July 2025
Published: 30 July 2025

**Citation:** Liu, W.; Luo, B.; Liu, J.; Nie, H.; Su, X. FEMNet: A Feature-Enriched Mamba Network for Cloud Detection in Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 2639. https://doi.org/10.3390/rs17152639

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Cloud contamination represents one of the most significant challenges in optical remote sensing, with studies indicating that up to 66% of satellite images may be partially or completely affected by cloud cover [1,2]. This widespread occlusion severely limits the effective utilization of Earth observation data by masking critical land surface features and introducing radiometric uncertainties that compromise downstream analysis. Consequently, accurate and efficient cloud detection has become an indispensable preprocessing step for numerous remote sensing applications, including land cover classification, environmental monitoring, and change detection [3,4].

The complexity of cloud detection stems from the highly variable nature of cloud formations—ranging from thin, semi-transparent cirrus to thick stratocumulus—which often exhibit spectral signatures similar to other bright surface features such as snow and ice [5,6]. Additionally, cloud boundaries are frequently irregular and diffuse [7], while cloud shadows introduce further complexity, as they often share low reflectance characteristics with other dark surfaces [8] such as water bodies and terrain shadows. These spectral ambiguities increase the risk of misclassification, rendering robust pixel-wise discrimination particularly challenging [2].

Early cloud detection methods primarily relied on rule-based algorithms and spectral thresholding techniques such as FMask [9]. While computationally efficient, these methods exhibit limited generalization capability across diverse atmospheric conditions and sensor characteristics. The advent of deep learning revolutionized cloud detection methodologies, with CNNs emerging as the dominant paradigm. Representative models like CDNet [10], LPMSNet [11], and CRSNet [12] employ encoder–decoder architectures enhanced with attention mechanisms and multi-scale feature fusion. However, CNNs face inherent limitations in capturing long-range spatial dependencies due to their localized receptive fields [13,14]. This makes it difficult to exploit the contextual relationship between spatially disconnected but semantically related regions, such as clouds and their corresponding shadows [15], especially in large-scale scenes.

To address these limitations, Vision Transformers have been increasingly adopted in remote sensing applications [16,17], leveraging self-attention mechanisms to capture global dependencies. Many hybrid approaches [18–21] combine Transformers with CNNs, seeking to balance global semantic understanding with fine-grained spatial detail preservation. Despite their effectiveness, Transformer-based architectures introduce significant computational overhead due to quadratic attention complexity, limiting their scalability for high-resolution imagery and real-time applications [22]. Recently, State Space Models (SSMs) have emerged as a compelling alternative. The Mamba architecture [23] introduces selective state space modeling with linear time complexity, while VMamba [24] adapts this to 2D spatial modeling. Remote sensing applications including ChangeMamba [25], RSCaMa [26], and LCCDMamba [27] have demonstrated Mamba's effectiveness for scalable representation learning.

While Mamba-based architectures show promise for efficient global modeling, they do not fully address specific structural challenges in encoder–decoder segmentation networks. Two critical limitations persist: first, conventional skip connections often introduce semantic misalignment between spatially detailed shallow features and semantically rich deep features, particularly problematic for thin cloud boundaries; second, standard decoders lack explicit multi-scale context integration, reducing their ability to distinguish between small fragmented clouds and large homogeneous formations. To address these challenges, we propose FEMNet, a dual-stream architecture combining Mamba-based long-range modeling with lightweight CNN-based spatial encoding. Our approach incorporates two specialized modules as follows: the cross-stage semantic enhancement (CSSE) module addresses semantic misalignment through adaptive gating, while the multi-scale context aggregation (MSCA) module enriches multi-scale understanding through resolution-aware pooling and fusion.

We evaluate FEMNet on five public datasets covering both binary and multi-class cloud segmentation scenarios. Our results demonstrate consistent improvements over state-of-the-art methods, including SCTNet and HRCloudNet, while maintaining computational efficiency. Notably, FEMNet achieves a 3.67% improvement in mIoU on the L8 Biome dataset with only 4.4 million parameters and 1.3G MACs, confirming its practical value for operational deployment.

Our main contributions are as follows:

- We propose FEMNet, a novel dual-stream architecture that effectively combines Mamba-based global modeling with CNN-based spatial detail encoding for efficient and accurate cloud segmentation.
- We design the cross-stage semantic enhancement (CSSE) module to resolve semantic misalignment between encoder and decoder features through adaptive gating guided by high-level contextual information.

- We introduce the multi-scale context aggregation (MSCA) module to enhance decoder-scale awareness through lightweight multi-resolution pooling and fusion strategies.
- We demonstrate superior segmentation accuracy and computational efficiency across five diverse datasets, validating FEMNet's effectiveness.

## 2. Related Work

### *2.1. CNN-Based Cloud Segmentation*

Convolutional Neural Networks established the foundation for modern cloud detection through learnable feature representations. CDNet [10] pioneered encoder–decoder structures with skip connections, while DBNet [28] and CDUNet [29] introduced dual-branch architectures for enhanced contextual fusion. Recent advances include LPM-SNet [11] with location-aware pooling and multi-scale attention and CRSNet [12] targeting small cloud detection through strip pyramid attention. However, CNNs face fundamental limitations in modeling long-range dependencies due to their local receptive fields, particularly problematic when distinguishing clouds from spectrally similar surface features across large spatial extents.

### *2.2. Transformer and Hybrid Architectures*

Vision Transformers address CNN limitations through self-attention mechanisms capable of capturing global contextual relationships. SCTNet [21] demonstrated token-wise self-attention for global context integration, while hybrid approaches like MAFNet [20], and MCANet [18] combine CNN spatial processing with Transformer global modeling. These methods seek to balance fine-grained detail preservation with semantic abstraction. Despite promising results, Transformer-based approaches introduce significant computational overhead due to quadratic attention complexity, limiting their scalability for high-resolution imagery and real-time applications.

### *2.3. Mamba and State Space Models*

State Space Models offer a promising alternative with linear complexity for long-range modeling. Mamba [23] introduces selective state space mechanisms with hardware-aware implementations, while VMamba [24] adapts this to 2D spatial modeling via SS2D modules. In remote sensing, recent works, including Samba, ChangeMamba [25], RSCaMa [26], and LCCDMamba [27], demonstrate Mamba's effectiveness across diverse applications from semantic segmentation to change detection. These approaches confirm the potential of SSM-based designs for scalable and globally aware modeling in geospatial tasks.

While each architectural paradigm addresses specific limitations of its predecessors, many recent works have adopted dual-branch encoder designs to improve semantic representation. DDRNet [30] builds two deep resolution-specific branches with repeated bilateral fusion, enhancing boundary detail and global context in urban scenes. DE-UNet [31] integrates two parallel CNN encoders into a U-Net framework, enabling the simultaneous capture of fine-grained textures and holistic structure. ST-UNet [32] combines CNNs with Swin Transformers through a relational aggregation module to integrate local and global cues hierarchically. Similarly, RS3Mamba [33] constructs a Mamba-based auxiliary branch alongside a convolutional backbone, using a collaborative fusion mechanism to adaptively refine dual-stream features. Inspired by these dual-stream principles, our proposed FEMNet employs a lightweight parallel encoder composed of a Mamba branch for global modeling and a CNN branch for spatial detail preservation.

## 3. Methodology

FEMNet is a lightweight yet effective semantic segmentation framework, developed for multi-class cloud detection in remote sensing imagery. It is specifically designed to tackle the challenges posed by spectral ambiguity—such as the confusion between clouds, snow, and bright land surfaces—and the need for fine-grained structural delineation across diverse terrestrial backgrounds. The architecture simultaneously captures high-resolution spatial textures and long-range semantic dependencies, while maintaining computational efficiency suitable for large-scale satellite data processing.

As shown in Figure 1, FEMNet comprises the following four main components: a dual-branch encoder, a multi-scale context aggregator (MSCA), a cross-scale semantic enhancement (CSSE) module, and a lightweight decoder. The encoder integrates convolutional blocks for spatial detail extraction with a Mamba-based sequence encoder for capturing semantic context. The MSCA module refines deep features by integrating contextual cues across scales. In parallel, the CSSE module modulates early-stage features using semantic information from deeper layers, promoting alignment between low- and high-level representations. The decoder then progressively fuses features from multiple levels to produce the final segmentation map.

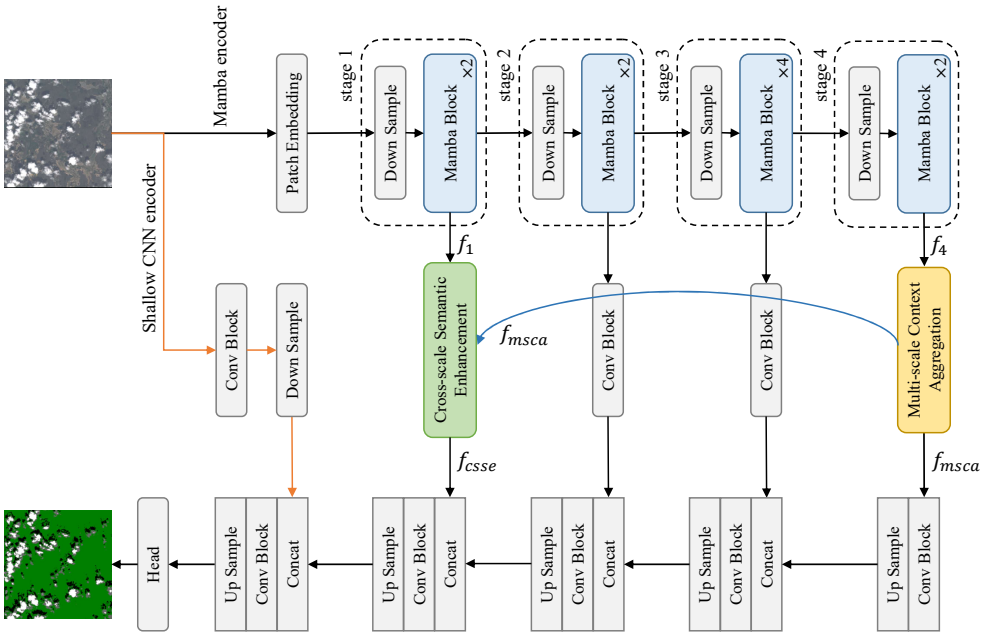


**Figure 1.** Overview of the proposed FEMNet architecture. The network adopts a dual-branch encoder design as follows: a Mamba-based encoder extracts hierarchical semantic features $f_1$ to $f_4$, while a shallow CNN encoder retains high-resolution spatial details. The MSCA module processes $f_4$ using average pooling at multiple scales and produces a context-enriched feature $f_{\text{msca}}$. The CSSE module receives low-level feature $f_1$ and high-level semantic feature $f_{\text{msca}}$, and it outputs an aligned feature $f_{\text{csse}}$ through gated fusion. The decoder progressively upsamples and refines features to recover full-resolution predictions. See Section 3 for the internal structure of each module.

This section is organized as follows: Section 3.1 introduces the overall architecture. Section 3.2 explains the Mamba-based multi-scale encoder. Sections 3.3 and 3.4 describe the MSCA and CSSE modules. Section 3.5 details the loss function and evaluation metrics.

### *3.1. Network Architecture Overview*

FEMNet processes the input image using two parallel encoding branches. The first pathway is a shallow convolutional encoder that extracts high-resolution spatial details, which are essential for recovering thin cloud edges and subtle shadow boundaries. In par-

allel, the second pathway utilizes a Mamba-based SegMAN encoder that operates hierarchically to extract multi-level semantic representations. This dual-stream design ensures that both local textures and global context are preserved throughout the network.

Formally, given an input image $x \in \mathbb{R}^{3\times H\times W}$, the CNN branch generates a low-level feature map $f_{\text{cnn}} \in \mathbb{R}^{16\times \frac{H}{2}\times \frac{W}{2}}$ through a sequence of convolutional and downsampling blocks. Simultaneously, the SegMAN encoder produces a set of hierarchical semantic features $\{f_1, f_2, f_3, f_4\}$ with progressively decreasing spatial resolutions and increasing channel dimensions. Specifically, $f_1 \in \mathbb{R}^{32\times \frac{H}{4}\times \frac{W}{4}}$, $f_2 \in \mathbb{R}^{64\times \frac{H}{8}\times \frac{W}{8}}$, $f_3 \in \mathbb{R}^{144\times \frac{H}{16}\times \frac{W}{16}}$, and $f_4 \in \mathbb{R}^{192\times \frac{H}{32}\times \frac{W}{32}}$, capturing increasingly abstract semantic representations.

The deepest feature $f_4$ is refined by the MSCA module, while the lowest-level feature $f_1$ is modulated by the CSSE block using high-level priors from $f_4$. The decoder integrates these representations through progressive upsampling and convolution, with skip connections enabling multi-scale fusion. At the final stage, the decoder output is concatenated with the shallow CNN feature (projected to a compatible dimension), and it is fused through a convolutional block. This design ensures both semantic completeness and spatial precision in the final prediction.

To qualitatively assess the contributions of the dual-stream encoder design, we visualize the intermediate feature maps from the CNN and Mamba branches, along with the fused representation. As shown in Figure 2, the CNN branch captures fine-grained texture details such as thin clouds and sharp edges, but it tends to produce noisy or fragmented responses. In contrast, the Mamba encoder focuses on global semantic consistency, leading to smoother but spatially coarser activation maps. The fused features successfully combine the strengths of both branches, which aligns well with our design motivation for combining local and global representations in FEMNet.

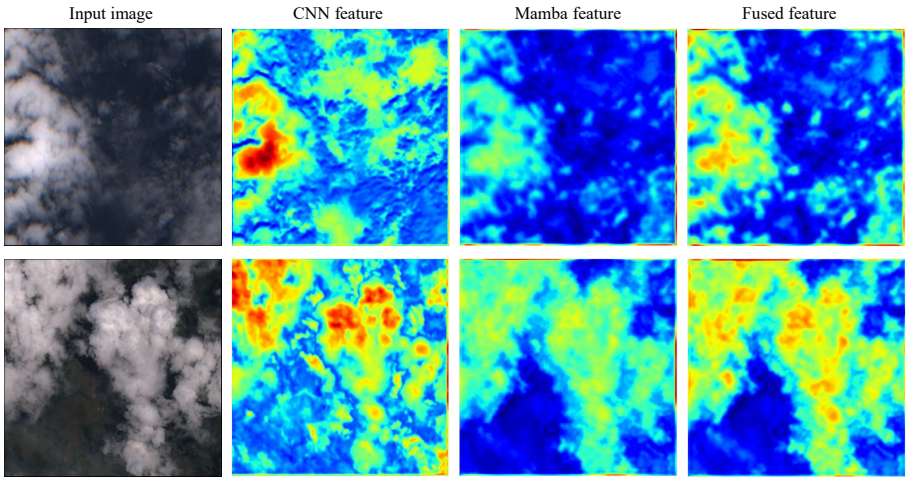


**Figure 2.** Feature visualization of FEMNet's dual-branch encoder. The first column shows the input remote sensing images. The second and third columns visualize features extracted from the shallow CNN branch and the Mamba branch, respectively. The last column presents the fused features after dual-branch integration.

### *3.2. Mamba-Based Multi-Scale Encoder*

Traditional convolutional encoders, while efficient for local feature extraction, struggle to model long-range dependencies, which are critical for capturing large-scale and spatially disconnected cloud patterns. Transformer-based models address this limitation through global self-attention but suffer from quadratic computational cost, which hinders their scalability for high-resolution remote sensing imagery. To balance efficiency and modeling capacity, we adopt Mamba—a state space model (SSM) that offers linear-time sequence modeling with dynamic parameterization.

Mamba formulates sequence modeling as a continuous-time dynamical system as follows:

$$\frac{dh(t)}{dt} = Ah(t) + Bx(t), \quad y(t) = Ch(t) + Dx(t), \tag{1}$$

where $h(t)$ is the latent state vector and $A, B, C, D$ are learnable matrices. Discretization via zero-order hold yields the following:

$$\bar{A} = \exp(\Delta A), \quad \bar{B} = A^{-1}(\bar{A} - I)B\Delta, \tag{2}$$

$$h_k = \bar{A}h_{k-1} + \bar{B}x_k, \quad y_k = Ch_k + Dx_k. \tag{3}$$

To enable adaptive modeling, Mamba [23,24] makes the parameters $B, C$ and step size $\Delta$ input-dependent. As illustrated in Figure 3, each SS2D block consists of the following three key components: cross-scan, selective scanning with S6 blocks, and cross-merge. Cross-scan unfolds the 2D feature map into four sequences along different spatial traversal directions (e.g., left-to-right, top-to-bottom, and their reverses), enabling each pixel to participate in multiple directional contexts. Selective scanning applies a dedicated S6 block to each directional sequence independently, allowing efficient 1D modeling with selective information flow. Finally, cross-merge reshapes and aggregates the four directional outputs to reconstruct the 2D feature map, typically by summing corresponding pixel responses across directions.

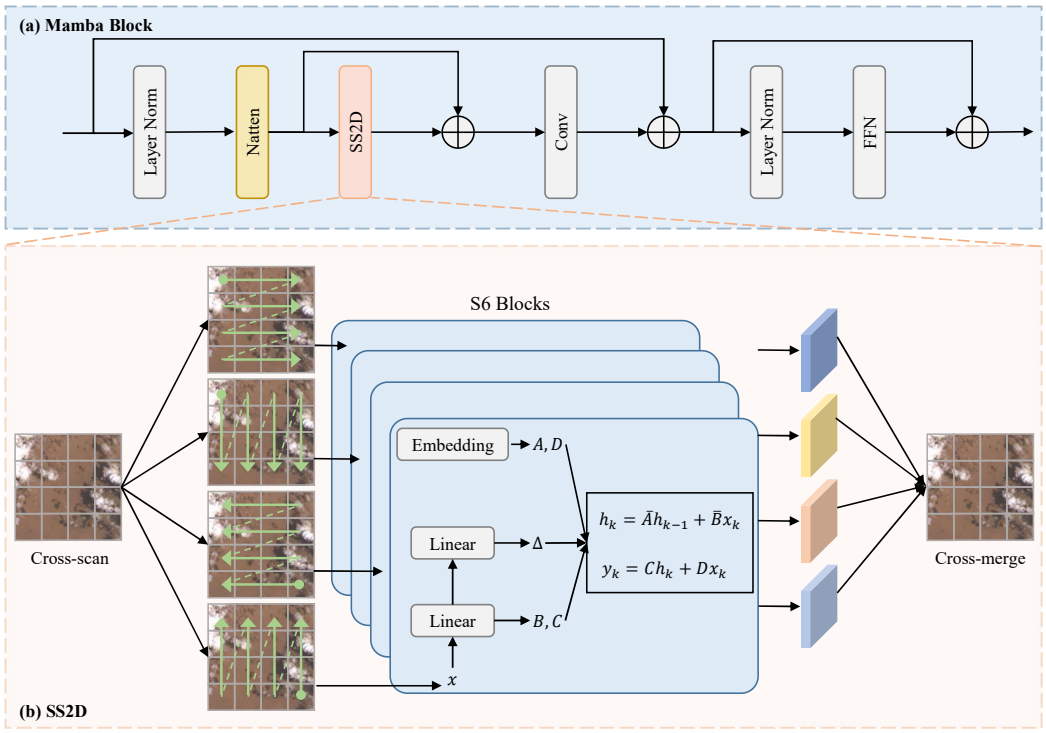


**Figure 3.** Internal structure of the SegMAN encoder with Mamba. (**a**) Each stage consists of a residual backbone composed of Layer Normalization, Neighborhood Attention (NAttn), 2D Selective-Scan (SS2D) module, and convolution, followed by a Feed-Forward Network (FFN). (**b**) The SS2D module unfolds the input feature map along four spatial directions via cross-scan (left-to-right, right-to-left, top-to-bottom, bottom-to-top), transforming the 2D input into directional sequences. Each sequence is processed independently by a selective state space model (S6 block), which dynamically generates SSM parameters using learnable linear projections (e.g., for $B, C, \Delta$). These outputs are then reshaped and aggregated via Cross-Merge to reconstruct a globally contextualized 2D representation.

This SS2D design empowers each pixel to aggregate contextual information from multiple orientations, thereby constructing a global receptive field in a computation-friendly

manner. Compared to traditional 2D convolutions or attention, SS2D maintains linear complexity while enhancing spatial awareness. Integrated into a four-stage encoder with neighborhood attention and residual connections, this module enables FEMNet to capture both local detail and long-range semantic structure in complex cloud scenes.

### *3.3. Multi-Scale Context Aggregation*

Contextual modeling plays a crucial role in cloud detection, particularly when segmenting large or diffuse cloud regions where local features alone may be insufficient. However, deep layers in encoder–decoder networks typically suffer from a loss of spatial resolution, which limits their ability to retain multi-scale semantic cues. To address this limitation, we propose the Multi-Scale Context Aggregation (MSCA) module to enhance the semantic depth of high-level features while recovering spatial context.

As illustrated in Figure 4a, the Multi-Scale Context Aggregation (MSCA) module enhances the deepest semantic feature $f_4 \in \mathbb{R}^{192\times\frac{H}{32}\times\frac{W}{32}}$ by aggregating contextual cues from multiple spatial scales.

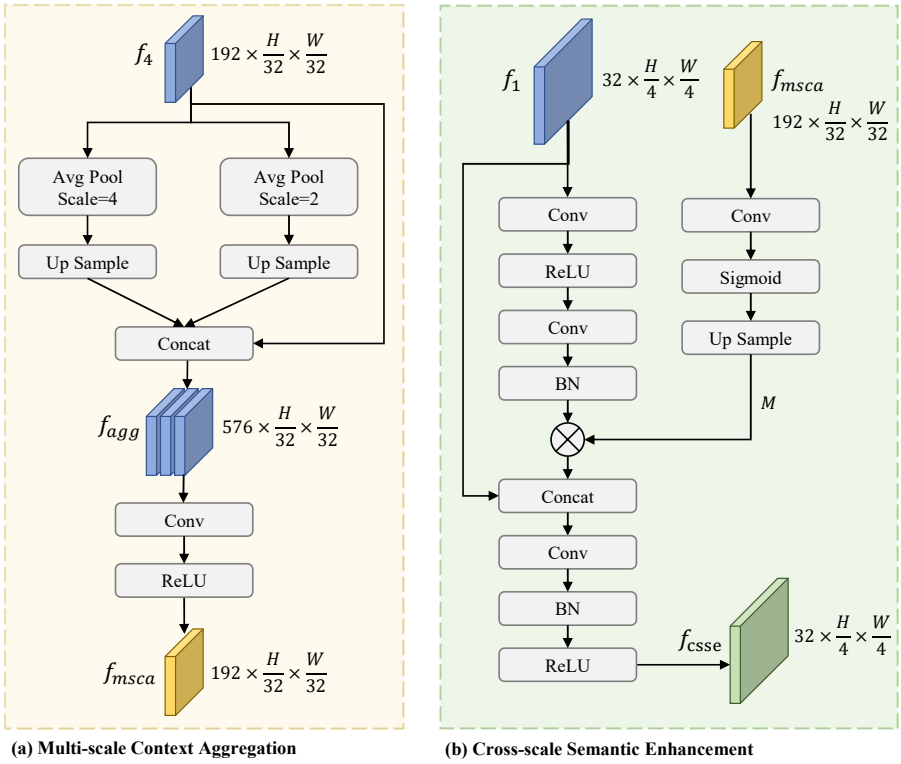


**Figure 4.** Auxiliary modules. (**a**) Multi-Scale Context Aggregator (MSCA): The input $f_4$ is the deepest semantic feature map, which typically exhibits sparse and fragmented activations due to resolution loss. MSCA enhances this feature by applying average pooling at multiple scales (2, 4), upsampling, and concatenation, followed by a convolutional refinement to produce $f_{\text{msca}}$, which contains more coherent and context-aware activations. (**b**) Cross-Scale Semantic Enhancement (CSSE): The module takes the low-level feature $f_1$ and semantic feature $f_{\text{msca}}$ as input. The latter is transformed into a spatial attention map $M$ via a $1 \times 1$ convolution and sigmoid activation. $M$ modulates $f_1$ through element-wise multiplication, and the result is concatenated with the original $f_1$, followed by a convolutional block to produce the enhanced feature $f_{\text{csse}}$. This improves semantic consistency and boundary localization in the decoding path.

Specifically, two parallel branches apply average pooling with kernel sizes of 2 and 4, respectively, followed by bilinear upsampling to restore the original resolution. The resulting pooled features are concatenated with the original $f_4$, yielding an aggregated tensor of size $\mathbb{R}^{576\times\frac{H}{32}\times\frac{W}{32}}$:

$$f_{\text{agg}} = \text{Concat}(f_4, \text{Up}(\text{AvgPool}_2(f_4)), \text{Up}(\text{AvgPool}_4(f_4))). \tag{4}$$

The fused representation $f_{\text{agg}}$ is then passed through a convolutional layer and a ReLU activation to produce the output feature $f_{\text{msca}} \in \mathbb{R}^{192 \times \frac{H}{32} \times \frac{W}{32}}$:

$$f_{\text{msca}} = \text{ReLU}(\text{Conv}(f_{\text{agg}})). \tag{5}$$

This design enables FEMNet to capture richer contextual information across multiple receptive fields without significantly increasing model complexity. As illustrated in Figure 5, the MSCA module transforms fragmented semantic features into more regionally coherent activations, reinforcing large-area consistency and improving the network's ability to delineate spatially extensive or spectrally ambiguous cloud structures.

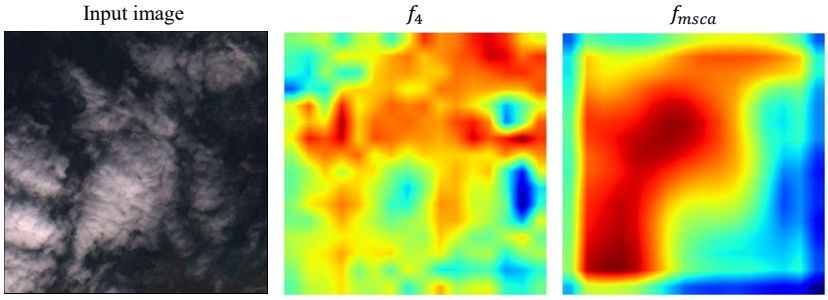


**Figure 5.** Visualization of feature maps before and after the Multi-Scale Context Aggregation (MSCA) module.

### *3.4. Cross-Stage Semantic Enhancement*

Encoder–decoder architectures often suffer from semantic inconsistencies when merging low-level features (rich in spatial detail) with high-level features (rich in semantics). This is especially detrimental in cloud segmentation, where boundary precision and semantic coherence are critical. To alleviate this, we propose the Cross-Stage Semantic Enhancement (CSSE) module.

As illustrated in Figure 4b, the CSSE module takes a low-level spatial feature $f_1 \in \mathbb{R}^{32 \times \frac{H}{4} \times \frac{W}{4}}$ and a high-level semantic feature $f_{\text{msca}} \in \mathbb{R}^{192 \times \frac{H}{32} \times \frac{W}{32}}$ as inputs.

First, $f_{\text{msca}}$ is projected using a $1 \times 1$ convolution and activated by a sigmoid function to generate a spatial attention map $M$, which is then upsampled to match the resolution of $f_1$:

$$M = \sigma(\text{Up}(\text{Conv}(f_{\text{msca}}))), \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid activation.

Meanwhile, $f_1$ is refined through two convolutional layers with interleaved ReLU and batch normalization as follows:

$$f_1' = \text{BN}(\text{Conv}(\text{ReLU}(\text{Conv}(f_1)))). \tag{7}$$

Then, the attention map $M$ is applied to the refined feature $f_1'$ via element-wise multiplication, and the result is concatenated with the original $f_1$ as follows:

$$f_{\text{cat}} = \text{Concat}(f_1, f_1' \odot M), \tag{8}$$

where $\odot$ denotes element-wise multiplication.

Finally, the concatenated feature is passed through another convolution–BN–ReLU sequence to produce the output feature $f_{\text{csse}}$ as follows:

$$f_{\text{csse}} = \text{ReLU}(\text{BN}(\text{Conv}(f_{\text{cat}}))) \in \mathbb{R}^{32\times\frac{H}{4}\times\frac{W}{4}}. \tag{9}$$

As shown in Figure 6, the CSSE module suppresses redundant textures in shallow features and emphasizes cloud boundaries, promoting more semantically consistent decoding. By selectively preserving spatially relevant information, it improves the network's ability to recover fine cloud structures and mitigate false positives in complex backgrounds.

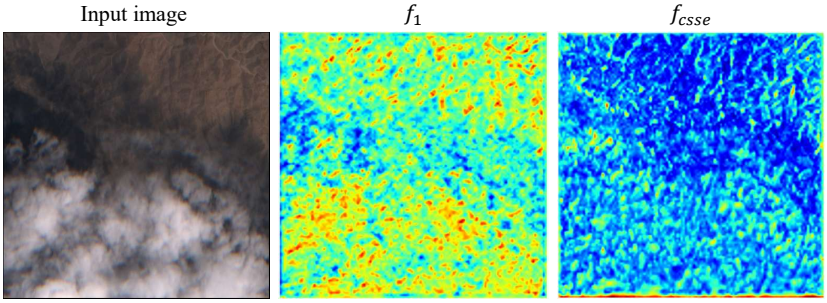


**Figure 6.** Visualization of feature maps before and after the Cross-Stage Semantic Enhancement (CSSE) module.

### *3.5. Loss Function and Evaluation Metrics*

To supervise the multi-class cloud segmentation task, we adopt the standard pixel-wise cross-entropy loss, formulated as follows:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^{(n)} \log \hat{y}_k^{(n)}, \tag{10}$$

where $N$ denotes the number of training pixels (or samples), $K$ is the number of cloud categories (e.g., clear sky, cloud shadow, thin cloud, and thick cloud), $y_k^{(n)} \in \{0,1\}$ is the ground truth label for the $k$-th class of the $n$-th sample, and $\hat{y}_k^{(n)} \in [0,1]$ is the predicted probability for that class. This formulation penalizes class-wise divergence between predictions and ground truth distributions.

To evaluate segmentation performance, we adopt four commonly used metrics computed at the pixel level: overall accuracy (aAcc), mean accuracy (mAcc), mean Intersection over Union (mIoU), and mean Dice coefficient (mDice). These metrics are defined as follows:

Overall Accuracy (aAcc) measures the proportion of correctly classified pixels over the entire dataset as follows:

$$\text{aAcc} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{11}$$

where $TP$, $TN$, $FP$, and $FN$ denote the total number of true positives, true negatives, false positives, and false negatives across all classes.

Mean Accuracy (mAcc) computes the average per-class accuracy, helping mitigate class imbalance as follows:

$$\text{mAcc} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i}, \tag{12}$$

where $TP_i$ and $FN_i$ refer to the number of true positives and false negatives for class $i$, respectively.

Mean Intersection over Union (mIoU) evaluates the average region overlap between prediction and ground truth for each class as follows:

$$\text{mIoU} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i + FN_i}. \tag{13}$$

Mean Dice Coefficient (mDice) provides a harmonic measure of precision and recall, particularly sensitive to boundary and small-object segmentation as follows:

$$\text{mDice} = \frac{1}{C}\sum_{i=1}^{C}\frac{2TP_i}{2TP_i + FP_i + FN_i}. \tag{14}$$

These metrics jointly reflect the model's performance in both region-level accuracy and class-specific consistency, providing a comprehensive evaluation for multi-class cloud segmentation tasks.

## 4. Experiments

In this section, we evaluate FEMNet's performance on both binary and multi-class cloud detection tasks using multiple public remote sensing datasets. We detail the datasets, implementation setup, baseline methods, and quantitative and qualitative comparisons with state-of-the-art approaches.

### *4.1. Experimental Setup*

#### 4.1.1. Datasets

For binary cloud segmentation, we conduct experiments on the following three publicy available binary cloud detection datasets: HRC_WHU [34], GF1MS-WHU [35], and GF2MS-WHU [35]. All images in these datasets are uniformly padded and cropped into non-overlapping patches of size 256 × 256 for training and evaluation.

- HRC_WHU is a high-resolution dataset collected from Google Earth, with spatial resolutions ranging from 0.5 to 15 m. It covers diverse geographic regions across the globe and consists of 120 training images and 30 test images.
- GF1MS-WHU is derived from the GF-1 satellite with an 8 m resolution. It includes 6344 training and 4086 test images, collected over various regions in China.
- GF2MS-WHU originates from the GF-2 satellite, offering a 4 m resolution. This dataset contains 14,357 training and 7560 test images.

For multi-class cloud segmentation, we evaluate our method on the following two multi-class datasets: L8 Biome [36] and CloudSEN12 [37]. The multi-class cloud segmentation task involves the following four semantic categories: clear sky, thin cloud, thick cloud, and cloud shadow. For both datasets, images are preprocessed into 512 × 512 patches to facilitate efficient training.

- L8 Biome consists of imagery from the Landsat-8 OLI sensor. Scenes are categorized into eight biome types as follows: Urban, Barren, Forest, Shrubland, Grass/Cropland, Snow/Ice, Wetlands, and Water. Thin clouds are defined as semi-transparent regions with moderate reflectance and soft boundaries, whereas thick clouds are optically dense, high-albedo regions with distinct shapes. Cloud shadows are annotated by visual inspection based on spatial alignment with overlying clouds and their low reflectance signatures. The dataset includes 7931 training and 2643 test images.
- CloudSEN12 is constructed from Sentinel-2 satellite imagery and provides both Level-1C (L1C) and Level-2A (L2A) data. Each version contains 8490 training and 975 test images. Thin clouds exhibit weaker opacity and are often partially transparent, while

thick clouds are optically saturated. Shadows are inferred based on cloud position and solar geometry but may exhibit labeling uncertainty in complex terrain.

#### 4.1.2. Implementation Details

All experiments are conducted on a single NVIDIA RTX 2080Ti GPU (11 GB). The models are trained using the RMSprop optimizer with a learning rate of $1 \times 10^{-4}$ and a batch size of 4. The maximum number of training iterations is set to 40,000. The Mamba encoder is initialized using SegMAN-Tiny [38]. Standard data augmentations including flipping, scaling, and color jittering are applied.

We evaluate and compare our method against a wide range of existing cloud detection approaches, including CDNetv1 [10], CDNetv2 [39], KappaMask [40], DBNet [28], UNetMobv2 [37], SCNN [41], MCDNet [42], HRCloudNet [43], SCTNet [21], and SegMAN [38]. To better understand the characteristics of existing methods, we categorize them based on their architectural design. CDNetv1, CDNetv2, and KappaMask are representative single-stream CNN models that follow the encoder–decoder paradigm. UNetMobv2 and SCNN are lightweight CNN-based methods, where the former employs depthwise separable convolutions for efficiency, and the latter uses a minimal three-layer structure without any pooling operations to preserve fine details. DBNet and SCTNet are hybrid models that integrate CNN and Transformer branches to balance local and global features; among them, SCTNet is specifically optimized for lightweight deployment. MCDNet introduces a dual-input mechanism to better distinguish thin clouds, while HRCloudNet leverages multi-resolution representations for enhanced cloud structure awareness. SegMAN employs a fully Mamba-based encoder–decoder architecture, enhanced with Neighborhood Attention to mitigate detail loss by capturing local dependencies.

### *4.2. Comparisons with State-of-the-Art Methods*

#### 4.2.1. Results of Binary Cloud Detection

We report binary cloud segmentation results on the HRC_WHU, GF1MS-WHU, and GF2MS-WHU datasets. Quantitative comparisons are presented in Tables 1–3, while representative visualizations are provided in Figure 7, where cloud regions are displayed in white. Across all three datasets, FEMNet achieves leading performance, with mIoU values of 89.68%, 93.32%, and 88.24% on HRC_WHU, GF1MS-WHU, and GF2MS-WHU, respectively. These consistent gains highlight FEMNet's ability to adapt across diverse spatial resolutions and scene complexities.

**Table 1.** Evaluation results on the HRC_WHU dataset. Bold highlights the best result.

| Method | Venue | aAcc | mIoU | mAcc | mDice |
|---|---|---|---|---|---|
| CDNetv1 [10] | TGRS 2019 | 89.88 | 77.79 | 89.93 | 87.20 |
| CDNetv2 [39] | TGRS 2020 | 89.71 | 76.75 | 87.46 | 86.46 |
| KappaMask [40] | RS 2021 | 84.73 | 67.48 | 80.30 | 79.74 |
| DBNet [28] | TGRS 2022 | 90.11 | 77.78 | 88.80 | 87.17 |
| UNetMobv2 [37] | Sci. Data 2022 | 92.13 | 79.91 | 85.61 | 88.45 |
| SCNN [41] | ISPRS 2024 | 74.51 | 57.22 | 81.27 | 72.31 |
| MCDNet [42] | JAG 2024 | 75.14 | 53.50 | 68.91 | 67.96 |
| HRCloudNet [43] | JEI 2024 | 92.93 | 83.44 | 92.39 | 90.79 |
| SCTNet [21] | GRSL 2025 | 95.75 | 89.22 | 93.90 | 94.22 |
| Segman [38] | CVPR 2025 | 94.64 | 86.22 | 90.79 | 92.45 |
| Ours | RS 2025 | **95.86** | **89.68** | **95.17** | **94.49** |

**Table 2.** Evaluation results on the GF1MS-WHU dataset. Bold highlights the best result.

| Method | Venue | aAcc | mIoU | mAcc | mDice |
|---|---|---|---|---|---|
| CDNetv1 [10] | TGRS 2019 | 96.81 | 81.82 | 92.75 | 89.27 |
| CDNetv2 [39] | TGRS 2020 | 97.55 | 84.93 | 92.96 | 91.36 |
| KappaMask [40] | RS 2021 | 98.91 | 92.42 | 95.05 | 95.94 |
| DBNet [28] | TGRS 2022 | 98.73 | 91.36 | 95.19 | 95.33 |
| UNetMobv2 [37] | Sci. Data 2022 | 98.82 | 91.71 | 93.99 | 95.53 |
| SCNN [41] | ISPRS 2024 | 97.18 | 81.68 | 87.21 | 89.13 |
| MCDNet [42] | JAG 2024 | 97.55 | 85.16 | 93.97 | 91.51 |
| HRCloudNet [43] | JEI 2024 | 98.80 | 91.86 | 95.82 | 95.62 |
| SCTNet [21] | GRSL 2025 | 99.05 | 93.22 | **96.69** | 96.40 |
| Segman [38] | CVPR 2025 | 98.85 | 91.78 | 95.32 | 95.58 |
| Ours | RS 2025 | **99.07** | **93.32** | 96.65 | **96.46** |

**Table 3.** Evaluation results on the GF2MS-WHU dataset. Bold highlights the best result.

| Method | Venue | aAcc | mIoU | mAcc | mDice |
|---|---|---|---|---|---|
| CDNetv1 [10] | TGRS 2019 | 92.42 | 78.20 | 83.07 | 87.17 |
| CDNetv2 [39] | TGRS 2020 | 92.63 | 78.84 | 83.67 | 87.61 |
| KappaMask [40] | RS 2021 | 90.30 | 72.00 | 77.64 | 82.57 |
| DBNet [28] | TGRS 2022 | 92.61 | 78.68 | 83.39 | 87.50 |
| UNetMobv2 [37] | Sci. Data 2022 | 93.22 | 80.44 | 84.86 | 88.70 |
| SCNN [41] | ISPRS 2024 | 91.99 | 76.99 | 82.06 | 86.32 |
| MCDNet [42] | JAG 2024 | 92.30 | 78.36 | 83.95 | 87.31 |
| HRCloudNet [43] | JEI 2024 | 91.46 | 75.57 | 80.95 | 85.29 |
| SCTNet [21] | GRSL 2025 | 95.51 | 86.99 | 91.63 | 92.86 |
| Segman [38] | CVPR 2025 | 94.76 | 84.51 | 88.66 | 91.33 |
| Ours | RS 2025 | **96.00** | **88.24** | **92.17** | **93.61** |

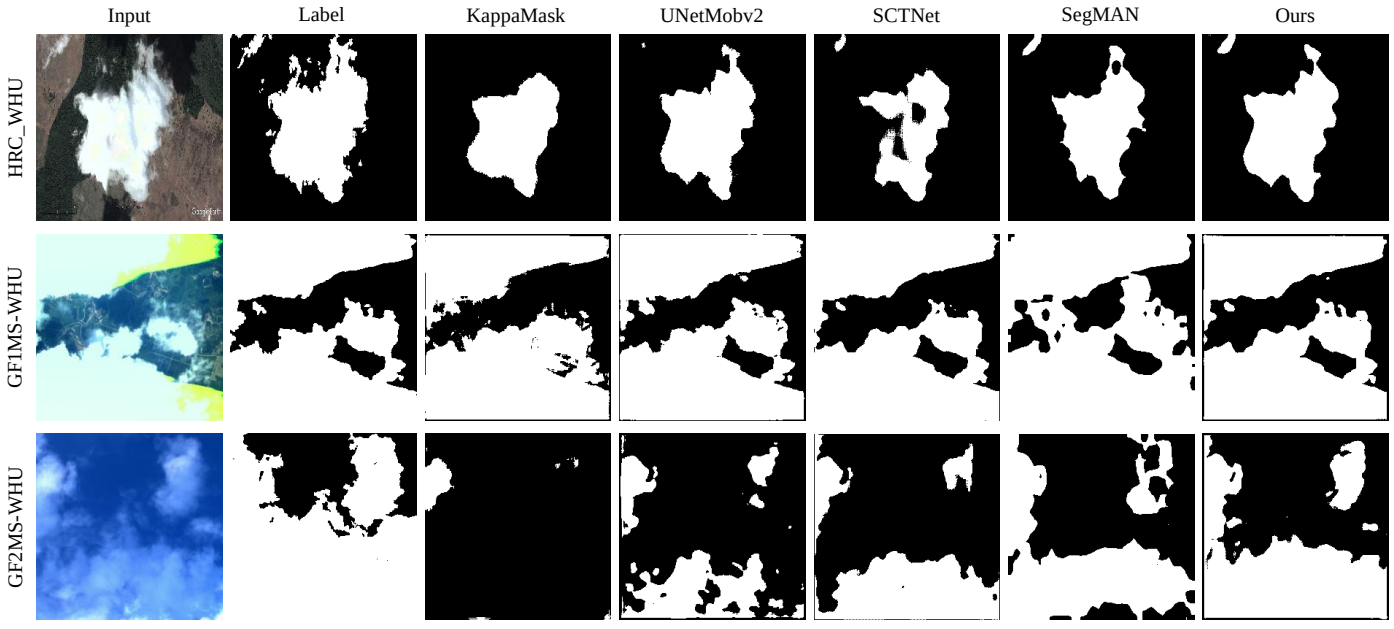


**Figure 7.** Qualitative results on the (**top**) HRC_WHU, (**middle**) GF1MS-WHU, and (**bottom**) GF2MS-WHU datasets.

On the HRC_WHU dataset, FEMNet exceeds SCTNet by 0.46% in mIoU. HRC_WHU features heterogeneous acquisition conditions and relatively sparse training data, making generalization especially challenging. Qualitatively, FEMNet produces cloud masks that

better conform to annotated shapes, avoiding both under-segmentation of thin structures and overextension into clear-sky areas. In contrast, models such as UNetMobv2 and SegMAN occasionally exhibit over-smoothing or omission of minor cloud components, particularly near ambiguous edges.

For GF1MS-WHU, where overall model performance is high due to abundant samples and consistent data quality, FEMNet still achieves a modest gain over SCTNet (93.32% vs. 93.22%). Compared to SegMAN, FEMNet produces comparable interior region segmentation but demonstrates improved boundary adherence, especially where annotations include fine protrusions or disconnected fragments.

On the GF2MS-WHU dataset, which contains higher variability in cloud thickness and complex land-atmosphere backgrounds, FEMNet yields the largest performance margin—a 1.25% mIoU gain over SCTNet and 3.73% over SegMAN. Visual comparisons show that FEMNet preserves fragmented or fuzzy cloud components more faithfully, particularly in low-contrast conditions. Unlike KappaMask or MCDNet, which frequently exhibit missing regions or false activations, FEMNet maintains spatial coherence without introducing artificial artifacts.

Real-world clouds, especially thin cirrus or complex formations, often possess irregular, fractal-like perimeters. Over-smoothing such boundaries can lead to the underrepresentation of valid cloud regions. FEMNet addresses this by adaptively balancing structure preservation with semantic consistency, avoiding excessive regularization while reducing noise. Taken together, these results affirm the benefit of FEMNet's dual-branch design and decoder enhancements. Its strong performance across both quantitative metrics and qualitative visual fidelity illustrates robustness across cloud morphologies, acquisition domains, and resolution settings.

#### 4.2.2. Results of Multi-Class Cloud Segmentation

Table 4 summarizes the quantitative results on the CloudSEN12 High L1C dataset, and the corresponding visual comparisons are shown in Figure 8. Compared with binary classification, this setting introduces additional ambiguity, particularly between visually similar classes such as thin clouds and haze, or between shadows and water bodies. To further validate the generalizability of FEMNet, we include comparison with DDRNet [30], a widely used real-time semantic segmentation model originally developed for traffic scene understanding. Its inclusion serves as a baseline for evaluating FEMNet against general purpose lightweight architectures.

**Table 4.** Evaluation results on the CloudSEN12 High L1C dataset. Bold highlights the best result.

| Method | Venue | aAcc | mIoU | mAcc | mDice |
|---|---|---|---|---|---|
| CDNetv1 [10] | TGRS 2019 | 83.48 | 60.35 | 72.46 | 73.30 |
| CDNetv2 [39] | TGRS 2020 | 86.68 | 65.60 | 75.65 | 77.83 |
| KappaMask [40] | RS 2021 | 76.27 | 41.27 | 50.76 | 49.33 |
| DBNet [28] | TGRS 2022 | 86.83 | 65.52 | 75.15 | 77.68 |
| UNetMobv2 [37] | Sci. Data 2022 | 89.52 | 71.65 | 81.05 | 82.47 |
| DDRNet [30] | TITS 2022 | 86.56 | 65.64 | 76.20 | 77.94 |
| SCNN [41] | ISPRS 2024 | 60.19 | 22.75 | 33.69 | 30.88 |
| MCDNet [42] | JAG 2024 | 72.68 | 44.80 | 59.99 | 58.08 |
| HRCloudNet [43] | JEI 2024 | 87.86 | 68.26 | 78.13 | 79.88 |
| SCTNet [21] | GRSL 2025 | 89.04 | 71.21 | **81.82** | 82.24 |
| SegMAN [38] | CVPR 2025 | 88.97 | 70.80 | 80.45 | 81.83 |
| Ours | RS 2025 | **89.62** | **71.78** | 80.90 | **82.56** |

FEMNet achieves the highest mIoU score of 71.78% with consistently strong performance across other metrics on the L1C dataset. Compared to SegMAN and SCTNet, FEMNet offers enhanced segmentation precision in spatially heterogeneous scenes. In Figure 8, FEMNet captures fine-scale structures of thin clouds more accurately, especially along coastlines and vegetated areas, where SegMAN tends to over-smooth or merge thin cloud boundaries with the background. For cloud shadows, which often overlap dark terrain or water, FEMNet demonstrates improved discriminative ability, reducing over-segmentation and better aligning with ground truth. These observations confirm that our model not only excels numerically but also produces more interpretable predictions in complex atmospheric contexts.

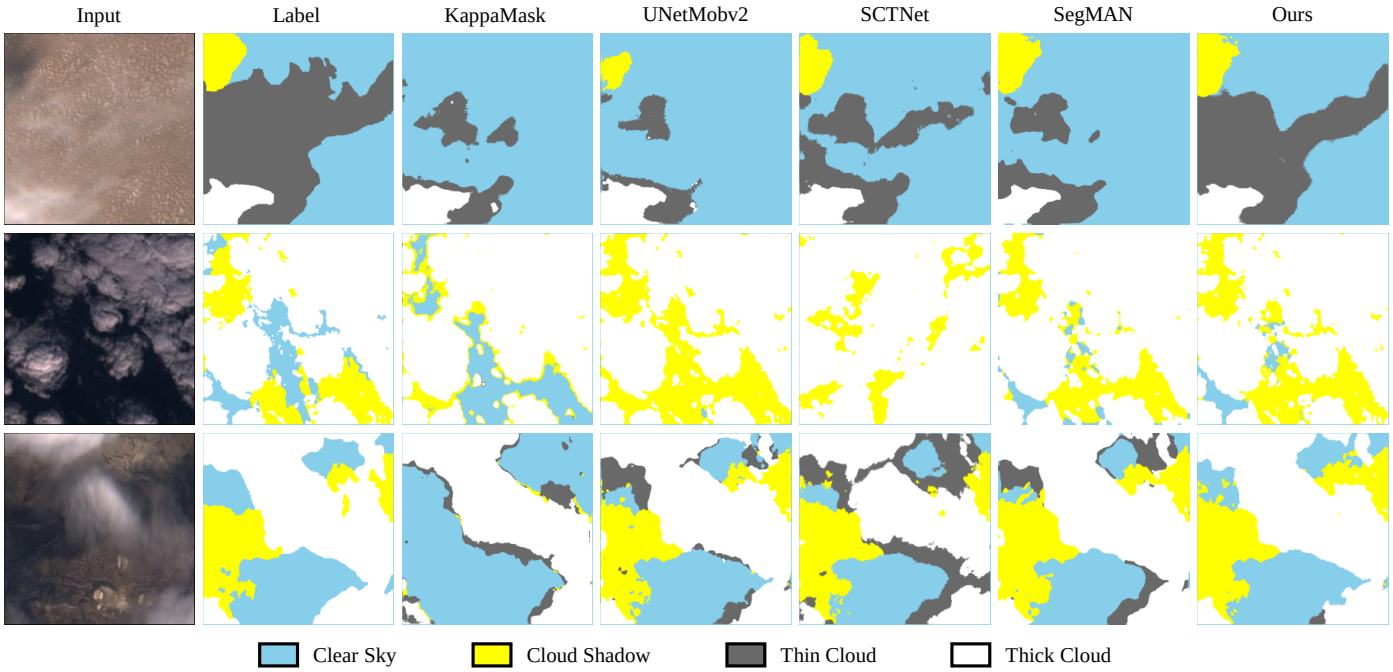


**Figure 8.** Qualitative comparison on the CloudSEN12 High L1C dataset.

On the CloudSEN12 High L2A dataset, FEMNet again achieves the best overall performance with a mIoU of 72.53%, aAcc of 89.64%, and mDice of 83.15%, outperforming the next-best model SCTNet by margins of 1.83% in mIoU and 1.27% in mDice (Table 5). Visual comparisons in Figure 9 show that FEMNet maintains semantic consistency across thin clouds and shadows, even when atmospheric correction increases spectral distortion. SegMAN and other baselines often misclassify thin clouds as clear sky or conflate cloud shadows with terrain shadows. FEMNet, by contrast, yields smoother and more distinct boundaries across all four semantic classes, confirming its robustness to radiometric variation and ambiguous boundaries.

To evaluate geographic generalization, we further conduct experiments on the L8 Biome dataset, which includes diverse land cover types such as Forest, Barren, Snow/Ice, and Wetlands. As reported in Table 6, FEMNet achieves the highest mIoU of 69.70% and mDice of 80.49%, outperforming SegMAN by 0.93% in mIoU and 1.08% in mDice, while SegMAN attains a slightly higher aAcc, FEMNet offers better inter-class discrimination, especially in spectrally complex regions.

**Table 5.** Evaluation results on the CloudSEN12 High L2A dataset. Bold highlights the best result.

| Method | Venue | aAcc | mIoU | mAcc | mDice |
|---|---|---|---|---|---|
| CDNetv1 [10] | TGRS 2019 | 84.74 | 62.39 | 73.50 | 75.15 |
| CDNetv2 [39] | TGRS 2020 | 86.88 | 66.05 | 75.91 | 78.19 |
| KappaMask [40] | RS 2021 | 79.60 | 45.28 | 55.47 | 53.56 |
| DBNet [28] | TGRS 2022 | 86.82 | 65.65 | 75.52 | 77.81 |
| UNetMobv2 [37] | Sci. Data 2022 | 88.96 | 70.36 | 79.57 | 81.45 |
| DDRNet [30] | TITS 2022 | 85.87 | 64.27 | 75.23 | 76.72 |
| SCNN [41] | ISPRS 2024 | 67.14 | 28.76 | 40.50 | 36.48 |
| MCDNet [42] | JAG 2024 | 75.67 | 46.52 | 58.54 | 59.59 |
| HRCloudNet [43] | JEI 2024 | 88.35 | 68.35 | 77.35 | 79.85 |
| SCTNet [21] | GRSL 2025 | 88.96 | 70.80 | 81.09 | 81.88 |
| SegMAN [38] | CVPR 2025 | 88.38 | 69.05 | 78.51 | 80.45 |
| Ours | RS 2025 | **89.64** | **72.53** | **83.14** | **83.15** |

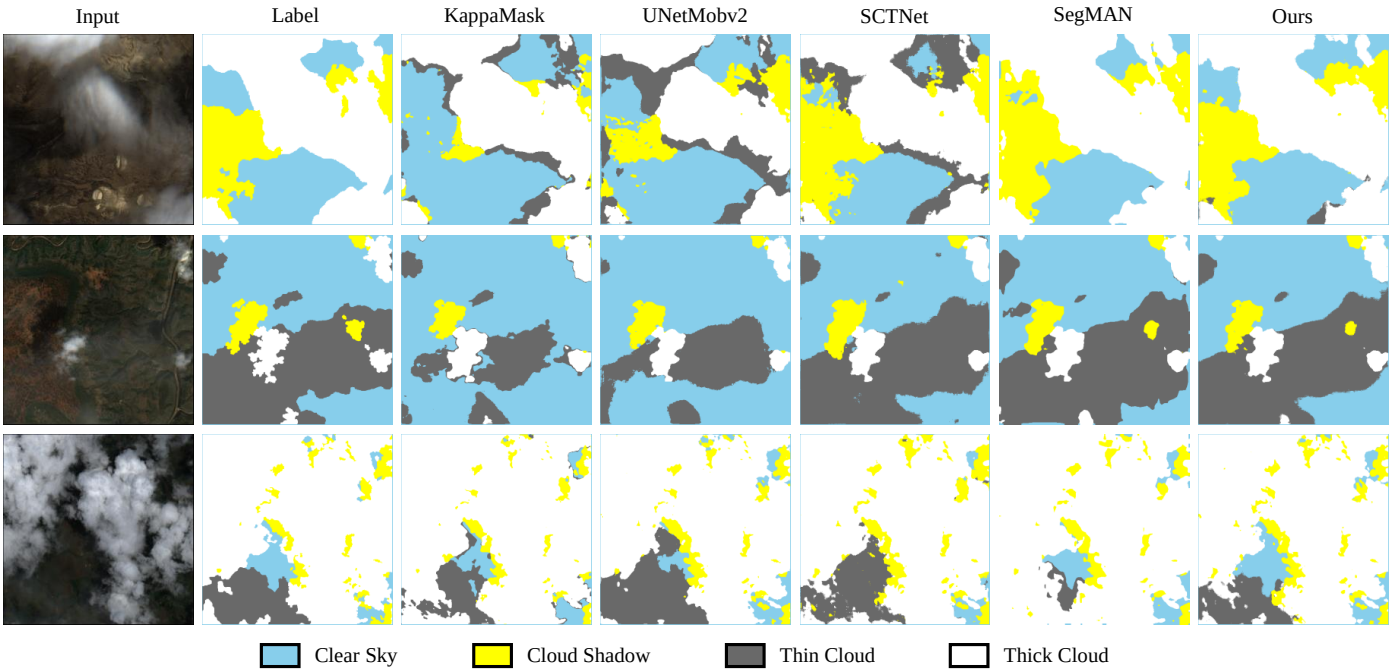


**Figure 9.** Qualitative comparison on the CloudSEN12 High L2A dataset.

**Table 6.** Evaluation results on the L8 Biome dataset. Bold highlights the best result.

| Method | Venue | aAcc | mIoU | mAcc | mDice |
|---|---|---|---|---|---|
| CDNetv1 [10] | TGRS 2019 | 68.16 | 34.58 | 45.59 | 45.80 |
| CDNetv2 [39] | TGRS 2020 | 78.56 | 43.63 | 52.98 | 52.75 |
| KappaMask [40] | RS 2021 | 76.63 | 42.12 | 51.73 | 51.35 |
| DBNet [28] | TGRS 2022 | 83.62 | 51.41 | 59.70 | 60.99 |
| UNetMobv2 [37] | Sci. Data 2022 | 82.00 | 47.76 | 56.29 | 56.91 |
| DDRNet [30] | TITS 2022 | 85.13 | 55.29 | 64.07 | 65.78 |
| SCNN [41] | ISPRS 2024 | 71.22 | 32.38 | 44.06 | 39.30 |
| MCDNet [42] | JAG 2024 | 69.75 | 33.85 | 44.82 | 42.76 |
| HRCloudNet [43] | JEI 2024 | 77.04 | 43.51 | 53.77 | 53.52 |
| SCTNet [21] | GRSL 2025 | 88.52 | 66.03 | 75.85 | 77.26 |
| SegMAN [38] | CVPR 2025 | **90.09** | 68.77 | 78.03 | 79.41 |
| Ours | RS 2025 | 89.78 | **69.70** | **82.98** | **80.49** |

In Figure 10, the visual analysis shows that FEMNet excels in separating thin clouds from snow surfaces, which is a frequent source of confusion for other methods. Moreover, cloud shadows are correctly distinguished from water bodies, which are often misclassified by SegMAN and KappaMask. These results confirm that FEMNet generalizes well across varying terrain and atmospheric conditions, maintaining both high accuracy and semantic coherence across different biomes.

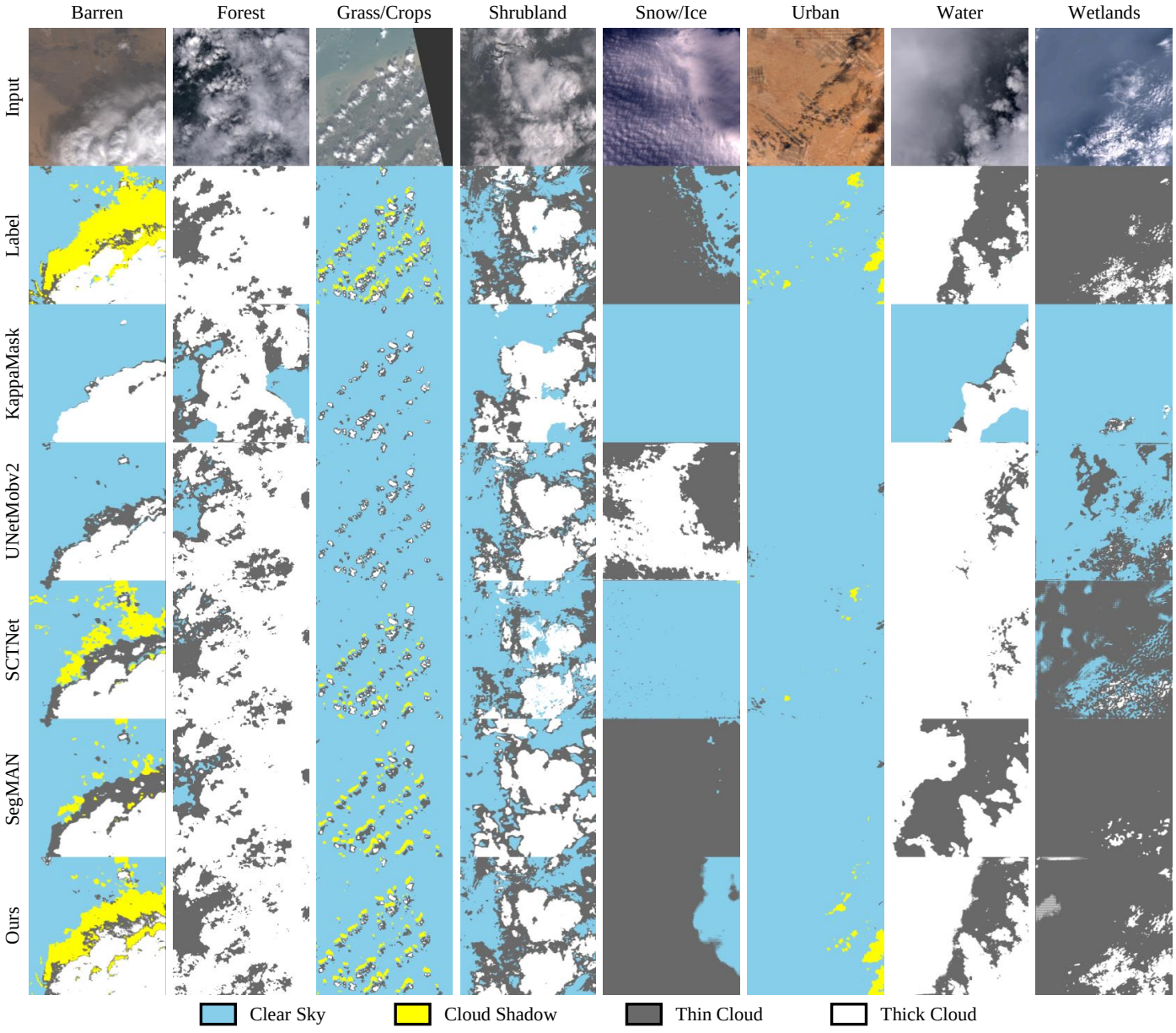


**Figure 10.** Qualitative comparison on the L8 Biome dataset.

## 5. Discussion

### *5.1. Ablation Study*

To assess the contribution of each proposed component, we conducted an ablation study on the L8 Biome dataset, as shown in Table 7. The baseline model comprises a Mamba-based encoder and a UNet-style decoder.

Among the variants, removing the Dual-Stream Encoder (DSE) causes the most notable performance degradation, with mIoU dropping from 69.70% to 66.67% and mDice from 80.49% to 78.00%. This emphasizes the essential role of combining convolutional and Mamba-based features for effective semantic segmentation. Similarly, excluding the

CSSE module results in a substantial performance drop, particularly in mIoU (69.01%) and mDice (79.79%), underlining its efficacy in modeling long-range contextual dependencies. Interestingly, removing the MSCA module yields the highest aAcc (90.19%), yet it significantly lowers mAcc (76.33%), suggesting that although overall pixel-level accuracy improves, class-wise balance deteriorates. Collectively, these results confirm that the integration of DSE, CSSE, and MSCA contributes complementary strengths, leading to optimal overall performance.

**Table 7.** Ablation results on the L8 Biome dataset. Bold highlights the best result.

| Method | aAcc | mIoU | mAcc | mDice |
|---|---|---|---|---|
| w/o DSE | 88.66 | 66.67 | 79.44 | 78.00 |
| w/o CSSE | 89.79 | 69.01 | 80.08 | 79.79 |
| w/o MSCA | **90.19** | 68.41 | 76.33 | 79.05 |
| Ours | 89.78 | **69.70** | **82.98** | **80.49** |

### *5.2. Model Efficiency*

We evaluate FEMNet's efficiency in terms of parameter count, computation, inference speed, model size, and memory usage. As shown in Table 8, FEMNet contains only 4.4 million parameters and 1.3 billion MACs, which is substantially lower than heavy models like DBNet and KappaMask. While SCTNet is even smaller (0.7M parameters), FEMNet achieves better segmentation performance with only a slight increase in computation.

On an NVIDIA RTX 2080Ti (batch size 64), FEMNet processes 331 images per second, offering a good trade-off between speed and accuracy. Although SCTNet reaches a higher throughput (1047 img/s), its segmentation quality is notably lower.

FEMNet also demonstrates compactness in storage and memory usage. The model file is only 17.9 MB, and inference with 256 $\times$ 256 inputs (batch size 1) requires 548 MB of GPU memory as measured by gpustat. This enables efficient deployment on memory-limited platforms such as edge devices and satellite systems.

**Table 8.** Computational complexity analysis (input size: 256 $\times$ 256). Bold highlights the best result.

| Method | Params (M) | MACs (G) | Throughput (img/s) | Model Size (MB) | Memory (MB) |
|---|---|---|---|---|---|
| KappaMask [40] | 31.0 | 54.7 | 137 | 124.2 | 734 |
| DBNet [28] | 95.1 | 28.5 | 127 | 381.8 | 1660 |
| UNetMobv2 [37] | 6.6 | 3.4 | 945 | 26.8 | 552 |
| SCTNet [21] | **0.7** | **1.0** | **1047** | **3.1** | **534** |
| SegMAN [38] | 7.4 | 1.3 | 307 | 29.8 | 674 |
| FEMNet | 4.4 | 1.3 | 331 | 17.9 | 548 |

## 6. Conclusions

This paper presents FEMNet, a lightweight and feature-enriched architecture for cloud detection in optical remote sensing imagery. By combining a Mamba-based state space encoder with a parallel CNN stream, FEMNet effectively captures both long-range semantic dependencies and fine-grained spatial details. To mitigate semantic inconsistency and enhance contextual awareness, which are two key limitations in conventional encoder–decoder frameworks, we introduce the following two targeted modules: the cross-stage semantic enhancement block for semantic alignment across feature hierarchies, and the multi-scale context aggregation module for efficient context fusion across spatial resolutions. Extensive experiments across five benchmark datasets validate FEMNet's superior performance over existing CNN-, Transformer-, and hybrid-based methods in both binary and multi-class segmentation scenarios. In future work, we aim to extend

FEMNet with domain-adaptive training strategies to improve its generalization across varying seasonal patterns, geographic regions, and sensor modalities.

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

**Funding:** This work was supported by the National Natural Science Foundation of China (grant number 61571332 and 42230108), and the Key Program for Basic Research of China (grant number JCKY2023206B026).

**Data Availability Statement:** Our research data are available at https://huggingface.co/datasets/XavierJiezou/cloudseg-datasets (accessed on 20 July 2025)

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

## References

1. Zhang, J.; Wu, J.; Wang, H.; Wang, Y.; Li, Y. Cloud detection method using CNN based on cascaded feature attention and channel attention. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–17. [CrossRef]
2. Wang, Z.; Zhao, L.; Meng, J.; Han, Y.; Li, X.; Jiang, R.; Chen, J.; Li, H. Deep learning-based cloud detection for optical remote sensing images: A Survey. *Remote Sens.* **2024**, *16*, 4583. [CrossRef]
3. Ghassemi, S.; Magli, E. Convolutional neural networks for on-board cloud screening. *Remote Sens.* **2019**, *11*, 1417. [CrossRef]
4. Li, Z.; Shen, H.; Weng, Q.; Zhang, Y.; Dou, P.; Zhang, L. Cloud and cloud shadow detection for optical satellite imagery: Features, algorithms, validation, and prospects. *ISPRS J. Photogramm. Remote Sens.* **2022**, *188*, 89–108. [CrossRef]
5. Mohajerani, S.; Krammer, T.A.; Saeedi, P. Cloud detection algorithm for remote sensing images using fully convolutional neural networks. *arXiv* **2018**, arXiv:1810.05782. [CrossRef]
6. Francis, A.; Sidiropoulos, P.; Muller, J.P. CloudFCN: Accurate and robust cloud detection for satellite imagery with deep learning. *Remote Sens.* **2019**, *11*, 2312. [CrossRef]
7. Wu, K.; Xu, Z.; Lyu, X.; Ren, P. Cloud detection with boundary nets. *ISPRS J. Photogramm. Remote Sens.* **2022**, *186*, 218–231. [CrossRef]
8. Zhu, Z.; Woodcock, C.E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **2012**, *118*, 83–94. [CrossRef]
9. Qiu, S.; Zhu, Z.; He, B. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. *Remote Sens. Environ.* **2019**, *231*, 111205. [CrossRef]
10. Yang, J.; Guo, J.; Yue, H.; Liu, Z.; Hu, H.; Li, K. CDnet: CNN-based cloud detection for remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6195–6211. [CrossRef]
11. Dai, X.; Chen, K.; Xia, M.; Weng, L.; Lin, H. LPMSNet: Location pooling multi-scale network for cloud and cloud shadow segmentation. *Remote Sens.* **2023**, *15*, 4005. [CrossRef]
12. Zhang, C.; Weng, L.; Ding, L.; Xia, M.; Lin, H. CRSNet: Cloud and cloud shadow refinement segmentation networks for remote sensing imagery. *Remote Sens.* **2023**, *15*, 1664. [CrossRef]
13. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. In Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016.
14. Chen, X.; Li, Z.; Jiang, J.; Han, Z.; Deng, S.; Li, Z.; Fang, T.; Huo, H.; Li, Q.; Liu, M. Adaptive effective receptive field convolution for semantic segmentation of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 3532–3546. [CrossRef]
15. Li, J.; Wang, Q. CSDFormer: A cloud and shadow detection method for landsat images based on transformer. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103799. [CrossRef]
16. 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 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
17. 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.
18. Hu, K.; Zhang, E.; Xia, M.; Weng, L.; Lin, H. Mcanet: A multi-branch network for cloud/snow segmentation in high-resolution remote sensing images. *Remote Sens.* **2023**, *15*, 1055. [CrossRef]

19. Ge, W.; Yang, X.; Jiang, R.; Shao, W.; Zhang, L. CD-CTFM: A lightweight CNN-transformer network for remote sensing cloud detection fusing multiscale features. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2024**, *17*, 4538–4551. [CrossRef]
20. Gu, H.; Gu, G.; Liu, Y.; Lin, H.; Xu, Y. Multi-Branch Attention Fusion Network for Cloud and Cloud Shadow Segmentation. *Remote Sens.* **2024**, *16*, 2308. [CrossRef]
21. Liu, W.; Luo, B.; Liu, J.; Nie, H.; Su, X. SCTNet: A Shallow CNN-Transformer Network with Statistics-Driven Modules for Cloud Detection. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1–5. [CrossRef]
22. Pang, Y.; Yao, L.; Luo, Y.; Dong, C.; Kong, Q.; Chen, B. RepSViT: An efficient vision transformer based on spiking neural networks for object recognition in satellite on-orbit remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [CrossRef]
23. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752. [CrossRef]
24. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Proc. Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063.
25. Chen, H.; Song, J.; Han, C.; Xia, J.; Yokoya, N. Changemamba: Remote sensing change detection with spatio-temporal state space model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–20. [CrossRef]
26. Liu, C.; Chen, K.; Chen, B.; Zhang, H.; Zou, Z.; Shi, Z. Rscama: Remote sensing image change captioning with state space model. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
27. Huang, J.; Yuan, X.; Lam, C.T.; Wang, Y.; Xia, M. LCCDMamba: Visual State Space Model for Land Cover Change Detection of VHR Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2025**, *18*, 5765–5781. [CrossRef]
28. Lu, C.; Xia, M.; Qian, M.; Chen, B. Dual-branch network for cloud and cloud shadow segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
29. Hu, K.; Zhang, D.; Xia, M. CDUNet: Cloud detection UNet for remote sensing imagery. *Remote Sens.* **2021**, *13*, 4533. [CrossRef]
30. Pan, H.; Hong, Y.; Sun, W.; Jia, Y. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Trans. Intell. Transp. Syst.* **2022**, *24*, 3448–3460. [CrossRef]
31. Liu, Y.; Song, S.; Wang, M.; Gao, H.; Liu, J. DE-Unet: Dual-Encoder U-Net for Ultra-High Resolution Remote Sensing Image Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 12290–12302. [CrossRef]
32. 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*, 1–15. [CrossRef]
33. Ma, X.; Zhang, X.; Pun, M.O. RS3Mamba: Visual State Space Model for Remote Sensing Image Semantic Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
34. Li, Z.; Shen, H.; Cheng, Q.; Liu, Y.; You, S.; He, Z. Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 197–212. [CrossRef]
35. Zhu, S.; Li, Z.; Shen, H. Transferring Deep Models for Cloud Detection in Multisensor Images via Weakly Supervised Learning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–18. [CrossRef]
36. Foga, S.; Scaramuzza, P.L.; Guo, S.; Zhu, Z.; Dilley, R.D., Jr.; Beckmann, T.; Schmidt, G.L.; Dwyer, J.L.; Hughes, M.J.; Laue, B. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sens. Environ.* **2017**, *194*, 379–390. [CrossRef]
37. Aybar, C.; Ysuhuaylas, L.; Loja, J.; Gonzales, K.; Herrera, F.; Bautista, L.; Yali, R.; Flores, A.; Diaz, L.; Cuenca, N.; et al. CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2. *Sci. Data* **2022**, *9*, 782. [CrossRef] [PubMed]
38. Fu, Y.; Lou, M.; Yu, Y. SegMAN: Omni-scale Context Modeling with State Space Models and Local Attention for Semantic Segmentation. In Proceedings of the Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 19077–19087.
39. Guo, J.; Yang, J.; Yue, H.; Tan, H.; Hou, C.; Li, K. CDnetV2: CNN-based cloud detection for remote sensing imagery with cloud-snow coexistence. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 700–713. [CrossRef]
40. Domnich, M.; Sünter, I.; Trofimov, H.; Wold, O.; Harun, F.; Kostiukhin, A.; Järveoja, M.; Veske, M.; Tamm, T.; Voormansik, K.; et al. KappaMask: AI-based cloudmask processor for Sentinel-2. *Remote Sens.* **2021**, *13*, 4100. [CrossRef]
41. Chai, D.; Huang, J.; Wu, M.; Yang, X.; Wang, R. Remote sensing image cloud detection using a shallow convolutional neural network. *ISPRS J. Photogramm. Remote Sens.* **2024**, *209*, 66–84. [CrossRef]
42. Dong, J.; Wang, Y.; Yang, Y.; Yang, M.; Chen, J. MCDNet: Multilevel cloud detection network for remote sensing images based on dual-perspective change-guided and multi-scale feature fusion. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103820. [CrossRef]
43. Li, J.; Xue, T.; Zhao, J.; Ge, J.; Min, Y.; Su, W.; Zhan, K. High-resolution cloud detection network. *J. Electron. Imaging* **2024**, *33*, 043027. [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.