Semantic segmentation of clouds and cloud shadows using state space models

Zhang, Z., Hu, Z. ORCID: https://orcid.org/0000-0001-9994-771X, Xia, M. ORCID: https://orcid.org/0000-0003-4681-9129, Yan, Y. ORCID: https://orcid.org/0000-0002-3609-0496, Zhang, R., Liu, S. and Li, T. ORCID: https://orcid.org/0000-0002-3418-275X (2025) Semantic segmentation of clouds and cloud shadows using state space models. Remote Sensing, 17 (17). 3120. ISSN 2072-4292 doi: 10.3390/rs17173120 Available at https://centaur.reading.ac.uk/124421/

remote sensing

MDPI

Article

Semantic Segmentation of Clouds and Cloud Shadows Using State Space Models

Zhixuan Zhang [1,2,3], Ziwei Hu [2], Min Xia [2,*], Ying Yan [2], Rui Zhang [3], Shengyan Liu [2,4] and Tao Li [2]

[1] School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; zhangzhixuan@nuist.edu.cn
[2] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202212490021@nuist.edu.cn (Z.H.); ying.yan@nuist.edu.cn (Y.Y.); yq835099@student.reading.ac.uk (S.L.); litaojia@nuist.edu.cn (T.L.)
[3] China Aero Geophysical Survey and Remote Sensing Center for Natural Resources, Beijing 100083, China; 2106040326@st.btbu.edu.cn
[4] Department of Computer Science, University of Reading, Whiteknights House, Reading RG6 6DH, UK
* Correspondence: xiamin@nuist.edu.cn

Abstract

In remote sensing image processing, cloud and cloud shadow detection is of great significance, which can solve the problems of cloud occlusion and image distortion, and provide support for multiple fields. However, the traditional convolutional or Transformer models and the existing studies combining the two have some shortcomings, such as insufficient feature fusion, high computational complexity, and difficulty in taking into account local and long-range dependent information extraction. In order to solve these problems, this paper proposes the MCloud model based on Mamba architecture is proposed, which takes advantage of its linear computational complexity to effectively model long-range dependencies and local features through the coordinated work of state space and convolutional support and the Mamba-convolutional fusion module. Experiments show that MCloud have the leading segmentation performance and generalization ability on multiple datasets, and provides more accurate and efficient solutions for cloud and cloud shadow detection.

Keywords: cloud and cloud shadow; semantic segmentation; remote sensing images; Mamba; state-space models; deep learning

Academic Editors: Xin Zhang, Xueyao Hu, Yang Li, Fernando José Aguilar and Muhammad Yasir

Received: 13 July 2025
Revised: 31 August 2025
Accepted: 4 September 2025
Published: 8 September 2025

Citation: Zhang, Z.; Hu, Z.; Xia, M.; Yan, Y.; Zhang, R.; Liu, S.; Li, T. Semantic Segmentation of Clouds and Cloud Shadows Using State Space Models. *Remote Sens.* **2025**, *17*, 3120. https://doi.org/10.3390/rs17173120

1. Introduction

In the field of remote sensing image processing, the recognition of clouds and cloud shadows is an important research direction. However, recognizing clouds and their shadows only represents one approach to mitigating the interference of cloud cover; another effective technical solution lies in cloud removal, which directly restores ground information obscured by clouds. Cloud cover has a significant impact on the quality and accuracy of remote sensing images, which changes the reflectance spectrum of ground objects, further increasing the difficulty of remote sensing data processing and analysis [1–3]. Especially in the fields of agriculture, urban planning, and natural disaster monitoring, its occlusion can lead to image brightness distortion and spectral characteristics change, which increases the difficulty of data processing and analysis [4,5]. In agricultural resource management, cloud occlusion can affect crop growth assessment, while in natural disaster monitoring, cloud cover can help assess the severity and extent of disasters [6,7]. The persistence of cloud cover over disaster-stricken areas can indicate prolonged adverse weather conditions,

aiding in assessing secondary disaster risks. Thus, understanding cloud properties is not only essential for mitigating occlusion impacts but also provides valuable insights for disaster assessment and response strategies. The wide application of cloud detection technology in the fields of earth resources research, agriculture and forestry research, and meteorological forecasting provides important information support for the study of global climate change [8–10]. In terms of earth resources, cloud detection of remote sensing images can help monitor and evaluate land use, distribution, and utilization of water resources, and changes in the ecological environment [11,12]. In agriculture and forestry, cloud detection of remote sensing images can help agricultural producers better assess the production status, planting status, and crop growth status of farmland, thereby improving the production efficiency of farmland [13]. At the same time, cloud detection technology also plays an important role in the monitoring of natural disasters, which can provide necessary data support for pre- and post-disaster assessment [14]. Therefore, the accurate identification of clouds and cloud shadows is not only an important topic in the field of remote sensing image processing, but also a necessary condition to meet the needs of remote sensing technology development and information acquisition on the Earth's surface. Before deep learning technology was widely used, the early cloud detection methods mainly included the following types: first, image analysis-based techniques [15,16], which used visual computing methods such as texture analysis and edge recognition to extract features, and classified pixels in the image with the help of classifiers. Secondly, the threshold-based method [17] distinguishes clouds from the surface by setting fixed thresholds in different spectral intervals according to the characteristics that the reflectance of clouds in the visible and near-infrared spectra is generally higher than that of most surface objects. Thirdly, techniques based on frequency domain analysis [18], which identify clouds and their shadows by studying the characteristics of remote sensing images in the frequency domain, such as spectral distribution and energy density, include Fourier analysis, wavelet analysis, and frequency domain filtering. Finally, techniques based on classical machine learning, such as support vector machines and random forests, learn to classify labeled training data to identify clouds and their shadows. With the rapid development of deep learning technology in the field of computer vision [19,20], how to build a deep learning network model to achieve more accurate segmentation of clouds and their shadows in remote sensing images has important application value for environmental monitoring, climate forecasting, and hydrological model construction [21].

In the early research on cloud and cloud shadow detection, the texture feature analysis method proposed by Haralick et al. in 1973 [22] provided a theoretical basis for subsequent cloud detection research. In addition, Hyeungu Choi et al. proposed a method that combines shadow matching techniques and normalized snow index threshold decision-making [23] to achieve cloud detection in polar ice sheet regions through the combination of texture analysis and shadow features. The threshold-based approach is one of the simplest and most widely used techniques in early cloud detection. Simpson et al. (2002) proposed a cloud detection method based on 1.6 micron band data [24] to classify Arctic sea ice, clouds, and water bodies by setting thresholds for spectral reflectance.

Traditional cloud and cloud shadow semantic segmentation methods play an important role in remote sensing image processing. The technique based on image analysis and the method based on threshold setting have been widely used in early research because of their simplicity and efficiency, but their adaptability in complex scenarios is poor. The technique based on frequency domain analysis can effectively extract the local features of the image, but the computational complexity is high, and the noise is sensitive. The classical machine learning-based method shows good performance when dealing with complex scenes, but requires a large number of training samples and feature engineering. Although

these traditional methods can meet the needs of cloud and cloud shadow detection to a certain extent, with the continuous improvement of the resolution of remote sensing data and the increasing complexity of application scenarios, the limitations are gradually revealed.

Significant progress has been made in cloud and cloud shadow detection methods based on deep learning. Early Fully Convolutional Networks (FCNs) and U-Net [25] achieved end-to-end pixel-level prediction through an encoder–decoder architecture, which solved the problem of traditional methods relying on manual features. Subsequently, the introduction of multi-scale contextual modeling, such as dilated convolution, spatial pyramid pooling, and attention mechanisms, has further improved the adaptability of the model to complex scenes, especially in the distinction between thin clouds and cloud shadows [26]. In addition, lightweight design, such as deep separable convolution, and multimodal data fusion have become research hotspots, which significantly improve the efficiency and generalization performance of the model [27,28]. However, the limitations of the Convolutional Neural Network (CNN) in long-range dependency modeling have prompted researchers to explore the Transformer architecture, while Transformers provide the ability to extract global semantic features, they also increase the number of parameters, so their high computational complexity still needs to be optimized. Transformer modules with a high number of parameters are susceptible to overfitting, particularly on small- to medium-sized datasets [29].

In order to overcome the shortcomings of the Transformer [30], recent research has turned to state-space models (SSMs), especially the Mamba architecture [31], which has shown potential in visual tasks due to its linear computational complexity and dynamic feature selection ability. MCloud also shows obvious advantages in segmentation results, especially in complex scenarios, not only against traditional deep learning frameworks but also among state-of-the-art Mamba-based models [32]. Models such as Vision Mamba [33] and VMamba have been successful in image classification and segmentation tasks, while Pan-Mamba [34] and RSMamba [35] have further adapted them to remote sensing image processing. However, the application of Mamba in cloud and cloud shadow detection has not been fully explored, which provides an important direction for future research. At present, the combination of CNN's local feature extraction and Mamba's global modeling capabilities may be the key breakthrough point to improve the detection accuracy and efficiency.

In the field of cloud and cloud shadow semantic segmentation, deep learning-based methods have become the mainstream research direction, but there are still some problems, such as the lack of generalization ability, easy loss of space and detailed information, and false positives. At present, although there are many attempts, such as using transformers to extend vision tasks and building dual-branch fusion networks, the segmentation accuracy and reliability in complex scenarios still need to be improved. In this context, this study is the first to explore the application of state-space model-based networks to cloud and cloud shadow semantic segmentation of remote sensing images, and MCloud is proposed. The encoder includes a state-space architecture branch and a convolutional architecture branch, which are used for long-range dependency and local feature learning modeling, respectively. At the same time, the MC module is designed to integrate the global context modeling ability of the Mamba architecture [36,37] and the local feature perception advantages of the convolutional network, realize the cross-scale feature interaction mechanism, and carry out multi-modal feature fusion through the adaptive weight mechanism, which significantly enhances the model's ability to analyze the features of complex ground objects, provides new research ideas and directions for the development of this field, and proves the feasibility of the Mamba architecture in this direction. It is worth noting that after

achieving accurate segmentation, the subsequent challenge of cloud and shadow removal from contaminated images, which is crucial for downstream applications, has also been explored, as seen in studies like [38,39].

Compared to traditional methods, such as the GGLCM threshold method, which have limitations such as insufficient robustness and the inability to reuse different sensor data, MCloud can operate reliably in a diverse range of sensor data environments. This ensures continuity and accuracy in cloud information collection during disaster monitoring. For methods with weak cross-sensor generalization capabilities, such as U-Net cloud, MCloud can effectively obtain consistent cloud distribution information of multi-source data in complex disaster scenarios with its strong adaptability, so as to comprehensively evaluate the disaster scope covered by the cloud.

2. Network Architecture

2.1. Backbone Architecture

The MCloud model proposed in this chapter is designed for efficient and accurate segmentation of clouds and cloud shadows in remote sensing images. The model adopts a dual-branch encoder–decoder architecture, as illustrated in Figure 1. The encoder consists of two parallel branches: one based on the state space model, referred to as the Variational State Space Model (VSSM) branch, and the other based on traditional convolutional architecture. These two branches are responsible for modeling the long-range dependencies and local detail features of clouds and cloud shadows in remote sensing images, respectively, enabling the model to capture and interpret complex information from different perspectives.

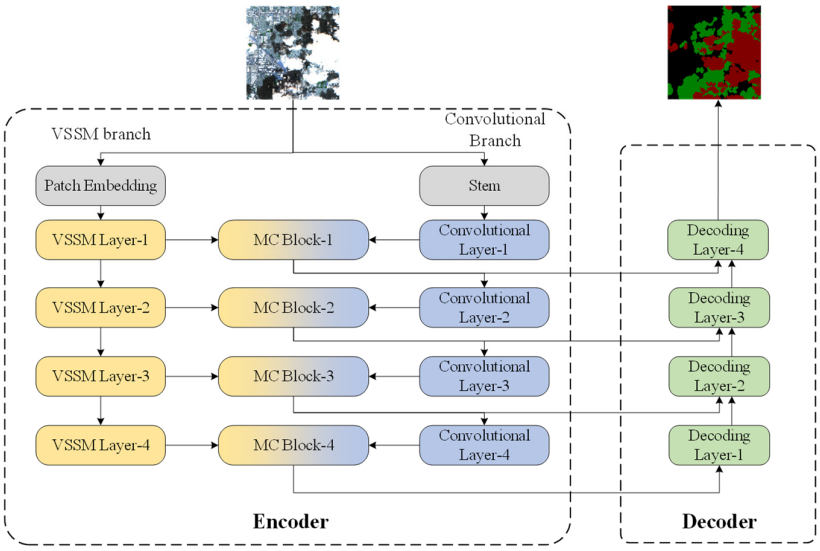

Figure 1. Overall structure diagram of MCloud.

The VSSM branch leverages the powerful capabilities of state space models to effectively capture the long-range dependencies of clouds and cloud shadows in images, which is crucial for handling their large-scale influence and interactions with other ground objects. On the other hand, the convolutional architecture branch focuses on extracting local detail features such as edges and textures, which are essential for accurately identifying the boundaries and shapes of clouds and cloud shadows. In this study, we aim to ensure model performance while minimizing computational complexity. Therefore, for the convolutional branch responsible for extracting local features, we selected a computationally efficient structure. After balancing computational complexity and performance, we adopted the ResNet-34 module to construct the convolutional branch.

For the decoder design, we employed a cascaded upsampling strategy. Specifically, this approach is similar to the U-Net architecture, where skip connections are used to concatenate features from the encoder with the upsampled output features from the previous decoder layer along the channel dimension. This is followed by convolutional operations to further process and fuse the features, generating more precise segmentation results. Through this design, we effectively reduce computational complexity while ensuring model performance to meet the objectives of this study.

During the feature extraction process, the features generated by the VSSM branch are fed into the corresponding scale of the MC module in the convolutional branch for feature fusion. This fusion mechanism allows the model to integrate global and local information across multiple scales, enhancing its ability to represent cloud and cloud shadow features. Specifically, after feature extraction and fusion at four different scales, we obtain rich and multi-scale feature representations. These multi-scale features are then passed to the corresponding decoder via skip connections. The skip connections are designed to directly transfer features from different levels of the encoder to the corresponding levels of the decoder, preserving more detailed information during the restoration of spatial resolution. The decoder utilizes these multi-scale features to gradually restore the spatial resolution of the image and optimize segmentation accuracy, ensuring that the final segmentation results are both accurate and detailed.

In summary, the MCloud model, through its unique dual-branch encoder–decoder architecture, combines the strengths of the VSSM branch and the convolutional branch. By incorporating mechanisms such as multi-scale feature fusion and skip connections, it achieves efficient and precise segmentation of clouds and cloud shadows in remote sensing images, providing a reliable foundation for subsequent image analysis and applications.

2.2. VSSM Module

The VSSM branch uses the VSSM block as its core building unit, and the specific structure of the VSSM block is shown in Figure 2. In this branch, the input features first enter the initial branch, where the feature channels are expanded through a linear layer. Subsequently, the expanded features undergo a series of operations, including depthwise convolution, nonlinear transformation via the SiLU activation function, and processing through SS2D and layer normalization. After these operations, the resulting features are aggregated with the input features mapped by another linear layer, further integrating information.

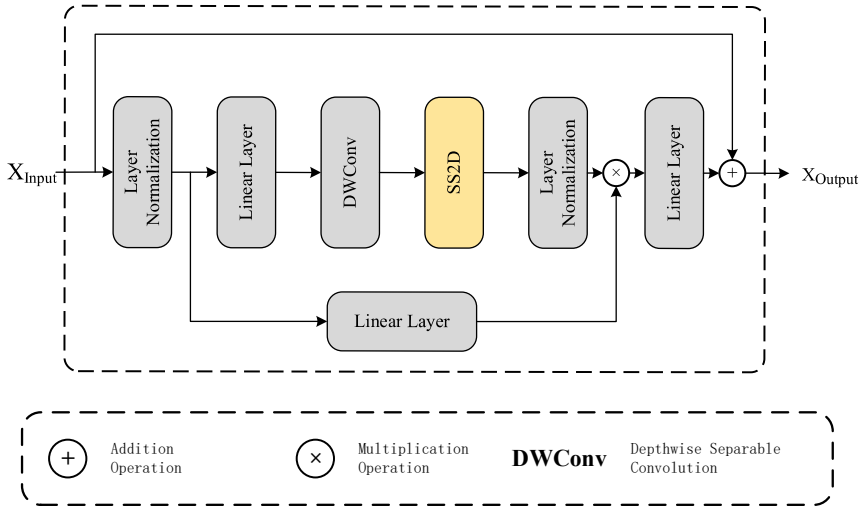

Figure 2. VSSM structure diagram.

To capture long-range dependencies while maintaining computational efficiency, the VSSM block employs the state space model (SSM). The state space equations governing the VSSM block are defined as follows:

$$h_{t+1} = \mathrm{A}h_t + Bx_t \tag{1}$$

$$y_t = Ch_t + Dx_t \tag{2}$$

$$h_k = \widetilde{A}h_{k-1} + \widetilde{B}x_k \tag{3}$$

$$y_k = Ch_k + Dx_k \tag{4}$$

Here, x_t and y_t represent the input and output of the system at time t, respectively. h_t represents the internal state of the system at time t. The matrices A, B, C, and D are the parameter matrices of the model. A captures information from previous states and constructs new states. B represents the degree of influence of the input on the system, C defines how the state is transformed into the output, and D acts as a direct signal from input to output, similar to a residual connection. These equations enable the VSSM block to effectively model long-range dependencies while preserving local feature details.

Finally, the aggregated features are combined with the original input features through a residual connection mechanism, enabling efficient feature propagation and stable network training. The output of this branch serves as the foundation for subsequent processing and analysis.

The SS2D (Selective Scan 2D) module is the core component of the VSSM module, and its structure is illustrated in Figure 3.

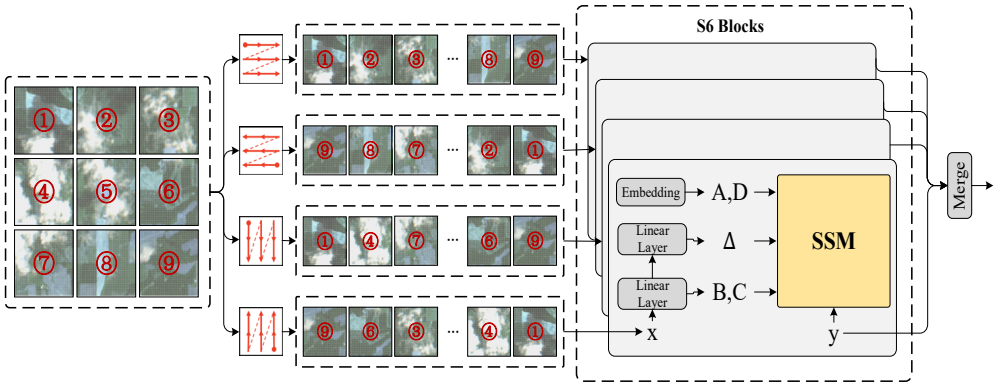

Figure 3. SS2D structure diagram.

The SS2D module processes the image through four distinct scanning paths: from top-left to bottom-right, from bottom-right to top-left, from top-right to bottom-left, and from bottom-left to top-right. Each path unfolds the 2D image into a 1D sequence, allowing the model to capture contextual information from multiple directions. This multi-directional scanning mechanism ensures a comprehensive understanding of cloud and cloud shadow features in the image, regardless of their distribution direction. The sequences generated by each scanning path are processed by independent S6 blocks [40] for feature extraction. The S6 block possesses dynamic weight adjustment capabilities, enabling it to adaptively adjust model parameters based on the input data, thereby more effectively capturing complex patterns and features in the image. This dynamic characteristic allows the SS2D module to flexibly adjust its focus on image features when processing remote sensing images under different scenarios, enhancing the model's generalization ability and segmentation accuracy. Through this design, the SS2D module not only efficiently captures global and local features in the image but also achieves a good balance between computational complexity and

performance, providing robust technical support for the precise segmentation of clouds and cloud shadows in remote sensing images. The calculation formula for SS2D is as follows:

$$\overline{x_d} = S6(x_d) \tag{5}$$

$$y = merge(x_d) \tag{6}$$

Here, scan$(\cdot)$ and merge$(\cdot)$ represent processes similar to the multi-path scanning and scan merging in VMamba [40]. The S6 block is an improvement based on the state space model in the Mamba architecture, which can selectively retain or filter information based on the input content, thereby maintaining linear time complexity when processing long sequences. The state space model is a mathematical model derived from modern control theory, used to describe the dynamic behavior of systems. It represents the internal state of a system through a set of state variables and describes the evolution of state variables over continuous time, as well as the relationship between state variables and output variables, through state equations and output equations. SSM assumes that the state of a dynamic system can be predicted using the following two mathematical equations:

$$h(t+1) = \boldsymbol{A}h(t) + \boldsymbol{B}x(t) \tag{7}$$

$$y(t) = \boldsymbol{C}h(t) + \boldsymbol{D}x(t) \tag{8}$$

where $x(t)$ and $y(t)$ represent the input and output of the system at time, $h(t)$ represents the internal state of the system at time. A, B, C, D are the parameter matrices of the model. A used to capture information from previous states and construct new states. B represents the degree of influence of the input on the system, and C defines how the state is transformed into the output. D is similar to a residual connection, providing a direct signal from input to output.

However, in traditional SSMs, A, B, C the parameter matricesare static and do not change with the input content. This prevents the model from dynamically adjusting its information retention strategy based on the input, leading to a loss of contextual relevance. In the Mamba architecture, the S6 block associates the parameter matrices B, C and $\triangle$ with the input through linear layer projections, enabling the model to selectively retain or ignore contextual information and adjust the state update rate based on the input content. Additionally, the parameter matrix A is initialized as a HIPPO matrix [41]. Specifically, the initialization formula for matrix A is given by:

$$A = exp(\Delta A_c) \tag{9}$$

where ΔA_c represents the dynamic adjustment factor for long-range dependency modeling. This initialization ensures that the matrix A remains positive definite, which is crucial for maintaining numerical stability and ensuring that the model can adaptively capture long-range dependencies. Furthermore, the parameter matrix B is dynamically adjusted during training using the following equation:

$$B = \left(e^{\Delta A_c} - I\right) A_c^{-1} B_c \tag{10}$$

The exponential function ensures that the matrix A remains positive definite, which is crucial for maintaining numerical stability and ensuring that the model can adaptively capture long-range dependencies. These improvements enhance the model's content-awareness while maintaining low computational complexity. The parameter matrices B, C

and △ dynamic update formulas are shown in the Formulas (6)–(8), where SiLU is the SiLU activation function, and the projection and bias operations are denoted as follows:

$$\Delta = SiLU(W_\Delta x + b_\Delta) \tag{11}$$

$$\boldsymbol{B} = W_B x + b_B \tag{12}$$

$$\boldsymbol{C} = W_C x + b_C \tag{13}$$

$$h_{t+1} = \overline{\boldsymbol{A}} h_t + \overline{\boldsymbol{B}} x_t \tag{14}$$

$$y_t = \boldsymbol{C} h_t + \boldsymbol{D} x_t \tag{15}$$

Here, $\overline{A}$ and $\overline{B}$ are the parameter matrices are obtained by approximating the discretization of continuous equations using the zero-order hold method, as shown in Formulas (11) and (12). D represents the direct linear projection of the input to the output, similar to a residual connection.

$$\overline{A} = e^{\Delta \boldsymbol{A}} \tag{16}$$

$$\overline{\boldsymbol{B}} = \left(\Delta \boldsymbol{A}\right)^{-1} \left(e^{\Delta \boldsymbol{A}} - \boldsymbol{I}\right) \Delta \boldsymbol{B} \tag{17}$$

After processing by the S6 block, the sequences from the four scanning directions are reassembled. Through an inverse operation, the 1D sequences are restored into 2D image blocks, which contain contextual feature information integrated from four different directions. These features are then fused using the Hadamard product. The Hadamard product, as an element-wise multiplication operation, effectively integrates features from different directions, enhancing the model's understanding of global image information. This fusion method not only preserves the uniqueness of features from each direction but also highlights their common characteristics, thereby improving the model's adaptability to complex scenes.

Through this approach, the SS2D module can learn long-range dependency features while maintaining linear complexity, thereby enhancing the model's performance. This design enables the SS2D module to efficiently capture global and local features in remote sensing images, providing robust technical support for the precise segmentation of clouds and cloud shadows.

2.3. Mamba–Convolution Fusion Module

Most existing state space model networks overly emphasize the modeling of long-range dependencies while neglecting the importance of local feature information, which plays a crucial role in semantic segmentation. This leads to suboptimal performance in cloud and cloud shadow detection tasks. To address these issues, this study proposes the Mamba–Convolution Fusion Module, whose core design goal is to effectively integrate long-range dependency features from the VSSM branch and local feature information from the convolutional branch. This enables more comprehensive and richer feature representation, enhancing the model's performance on specific tasks. Its structure is illustrated in Figure 4.

For features extracted by the convolutional branch: Since these features are inherently obtained through convolution operations between convolutional kernels and local regions of the input image, they primarily contain local feature information. To more effectively integrate these local features into the global semantic understanding framework, a channel attention mechanism is first applied to optimize them. The channel attention mechanism evaluates and weights the importance of each channel feature, highlighting those channels that contribute more discriminative power to the current task, thereby enhancing the discriminability and expressiveness of the features. On this basis, a spatial attention mechanism is further introduced. This mechanism allocates attention weights to

different spatial locations in the feature map with relatively low computational complexity, capturing global feature information with significant semantic value. This achieves an effective combination of local and global features, improving the model's overall grasp and understanding of features. For features from the VSSM branch: Their advantage lies in being obtained through VSSM blocks with excellent long-range dependency learning capabilities, enabling the capture of long-range dependencies in images or data. This is crucial for semantic understanding and context modeling in many complex tasks. However, relying solely on long-range dependency features may result in insufficient description of local details. Therefore, convolutional operations at different scales are used to further learn detailed features of local regions. Multi-scale convolutional operations extract features from various perspectives and granularities, enriching the expressive power of VSSM branch features. This allows the features to retain long-range dependency information while also more accurately depicting local details, further enhancing the completeness and richness of the features.

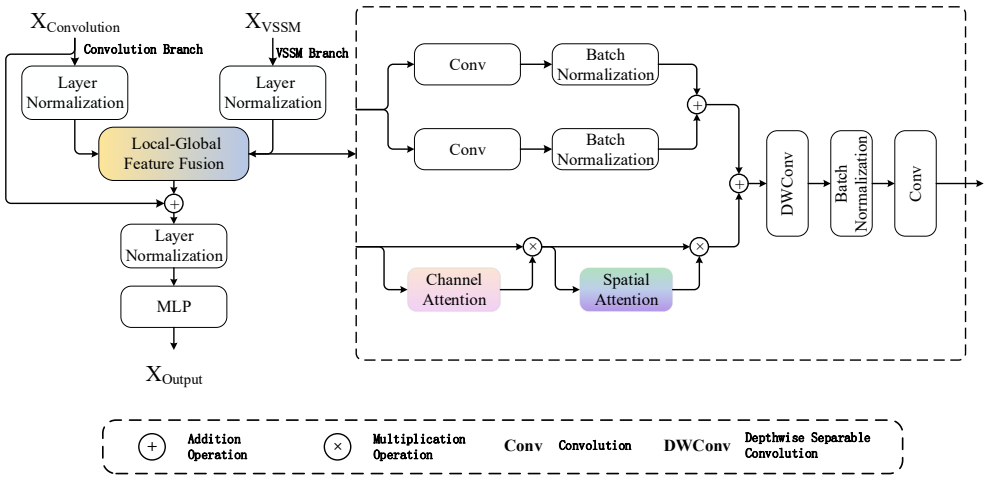

Figure 4. MC module structure diagram.

Finally, the processed features from the convolutional branch and the VSSM branch are fused and fed into the decoder through skip connections. This fusion method not only preserves the richness of the original features while maintaining low computational complexity but also transmits feature information at different levels and semantic hierarchies to the decoder via skip connections. Skip connections effectively prevent the loss and degradation of feature information during transmission, enabling the decoder to obtain more comprehensive and accurate feature representations. This significantly enhances the decoder's ability to reconstruct features and adapt to complex tasks, providing strong support for the model to achieve excellent performance in various application scenarios.

3. Experimental Results

3.1. Datasets

Three main datasets are used in this paper: CloudSEN-12, 38-Cloud, and SPARCS-Val.

(a) CloudSEN-12: This is a large-scale cloud semantic understanding dataset that covers multispectral imagery from the Sentinel-2 satellite, containing annotated information for multiple clouds and cloud shadows, and is widely distributed on all continents except Antarctica [42].The dataset has diverse band information and high-resolution image features, which provides a valuable resource for studying cloud and cloud shadow detection in complex scenes. And CloudSEN12 is a large dataset for cloud semantic understanding that consists of 9880 regions of interest (ROIs). Each ROI has

five 5090 m × 5090 m image patches (IPs) collected on different dates; we manually choose the images to guarantee that each IP inside an ROI matches one of the cloud cover groups.

(b) 38-Cloud: This dataset is focused on the cloud detection task of Landsat8 satellite images, which contains images of 38 scenes and their pixel-level annotations [43]. The dataset is characterized by its multispectral band configuration, which can effectively distinguish clouds from other highly reflective surface objects such as ice, snow, and buildings, thereby creating a more challenging data environment for model training. The dataset is binary and contains two classifications: cloud and background. The labeling process is performed manually by professionals, ensuring the high quality and accuracy of the labels.

(c) SPARCS-Val: This dataset was created by Oregon State University in the United States to validate the performance of cloud and cloud shadow removal algorithms [44]. The dataset not only contains a variety of feature types of annotations, but also covers complex scene combinations, each scene is equipped with manually annotated finely labeled images, and seven categories such as cloud shadow, cloud shadow on water, ice and snow, and cloud are annotated in detail, which further enriches the scene diversity of model validation and provides researchers with a rich data base.

3.2. Experiments Setup

Experiments were performed on NVIDIA RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA) using PyTorch (v2.6.0), and since most of the models in this experiment converged after 250 iterations, the epoch number was fixed at 300 and the batch size was 16. In this study, the cross-entropy loss function was used, and the AdamW optimizer was used, and the weight attenuation coefficient was 0.001. The Poly learning rate strategy is used during training. The initial learning rate is set to 0.001, the PolyPower is set to 2, and the learning rate LR of each round of training is described as follows:

$$\mathrm{LR} = 0.001 \times \left(1 - \frac{\mathrm{epoch}}{300}\right)^2 \tag{18}$$

3.3. Ablation Experiments

Ablation experiments were conducted on the CloudSEN-12 dataset to evaluate the contribution of different components in the model to the final segmentation performance. The following are the different combinations and their corresponding MIoU metrics:

$$I_i = \sum_{x,y}[Pred(x,y) = i \cap GT(x,y) = i] \tag{19}$$

$$U_i = \sum_{x,y}[Pred(x,y) = i \cup GT(x,y) = i] \tag{20}$$

$$IoU_i = \frac{I_i}{U_i} \tag{21}$$

$$MIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i \tag{22}$$

$$PA = \frac{\sum_{x,y}[Pred(x,y) = GT(x,y)]}{\sum_{x,y} 1} \tag{23}$$

$$PA_i = \frac{\sum_{x,y}[Pred(x,y) = i \cap GT(x,y) = i]}{\sum_{x,y}[GT(x,y) = i]} \tag{24}$$

$$MPA = \frac{1}{N}\sum_{i=1}^{N} PA_i \tag{25}$$

where $Pred(x, y) = i$ represents that at the image coordinate (x, y), the predicted class by the model is class i. In other words, the model predicts that the pixel at this position belongs to class i. $GT(x, y) = i$ is the abbreviation of "Ground Truth", means that at the image coordinate (x, y), the actual (true) class is class i, that is, the real class that the pixel belongs to. I_i counts the number of pixels where both the predicted class and the ground truth class are class i. U_i counts the number of pixels where either the predicted class is class i or the ground-truth class is class i. MIoU is the average of the intersection over Union IoU_i for all N classes. Mean Pixel Accuracy (MPA) is an index obtained by calculating the Pixel Accuracy (PA) of each category and then averaging the accuracy of all categories, which is used to evaluate the pixel-level prediction accuracy of the image segmentation model on all categories, which can more evenly reflect the segmentation ability of the model for different categories, especially small categories. lt is a simple and intuitive indicator to reflect the overall correct prediction degree. The results are shown in Table 1.

Table 1. Ablation experiments of different modules in the network.

Methods	MIoU(%)
Convolutional Branch	73.22
Convolutional Branch + VSSM Branch	76.3 (3.08↑)
Convolutional Branch + VSSM Branch + MC Module	78.19 (1.89↑)

(a) Convolutional Branch: The baseline model uses only the convolutional branch, achieving an MIoU of 73.22% and an MPA (Mean Pixel Accuracy). While the convolutional branch effectively extracts local features, it struggles with capturing long-range dependencies, this limitation is not only reflected in the moderate MIoU but also in the relatively low MPA, especially for small-scale cloud regions. The low MPA indicates that the baseline model frequently misclassifies these small cloud regions as non-cloud areas, as it cannot integrate global contextual information to distinguish them from similar-textured ground objects.

(b) Convolutional Branch + VSSM Branch: After adding the VSSM branch to the baseline model, the MIoU increased to 76.30%, an improvement of 3.08%, and the MPA has increased. The VSSM branch captures long-range dependencies through the state space model, significantly enhancing the model's ability to perceive global information. The larger improvement in MPA confirms that the VSSM branch effectively addresses the baseline model's weakness in classifying small or scattered cloud categories, which are more sensitive to MPA metrics.

(c) Convolutional Branch + VSSM Branch + MC Module: With the further addition of the MC module, the MIoU increased to 78.19%, an improvement of 1.89% and MPA increased. The MC module integrates the global features from the VSSM branch and the local features from the convolutional branch, enabling cross-scale feature interaction. This further enhances feature representation and strengthens the model's ability to interpret complex cloud and cloud shadow features.

3.4. Comparative Experiments

3.4.1. Generalization Experiments on the CloudSEN-12 Dataset

In this section, the proposed MCloud network is compared with state-of-the-art models, which are categorized into three main types based on their architectures: Convolution-based models, such as FCN, DeepLab, and OCRNet; Transformer-based models, such as SETR, PVT, and SwinUNet; Convolution-Transformer hybrid models, such as CVT, MPViT, and DBNet; Mamba-based models, such as CCViM, VM-UNet, and RS3Mamba.

The experiment designed for thick and thin cloud scenarios on the CloudSEN-12 dataset needs to combine the physical characteristics of the two types of clouds, with high reflectivity, clear boundaries, and continuous spatial distribution. Thin clouds have low reflectivity, blurred boundaries, and are distributed in discrete filaments.

Tables 2 and 3 present the evaluation metrics of different models on the CloudSEN-12 dataset. In terms of the overall ranking based on the MIoU metric, the proposed MCloud model achieves the best performance, outperforming CNN-based, Transformer-based, hybrid, and Mamba-based networks in MIoU, PA, and MPA metrics, with scores of 78.19%, 90.13%, and 88.85%, respectively. These results demonstrate that MCloud has a significant advantage in the task of cloud and cloud shadow semantic segmentation.

Table 2. Comparison of overall evaluation metrics of different models on the CloudSEN-12 dataset.

Architecture	Model	MIoU(%)	PA(%)	MPA(%)
CNN	FCN-32s	71.23	86.98	84.8
	DANet	71.79	87.02	84.14
	BiSeNetV2	74.19	88.12	85.47
	PAN	74.83	88.48	86.48
	CGNet	74.98	88.59	86.23
	LinkNet	75.19	88.63	86.64
	DenseASPP	75.32	88.77	86.61
	DeeplabV3	75.33	88.71	86.79
	HRNet	76.45	89.25	87.03
	OCRNet	76.74	89.5	87.66
	SegNet	77.01	89.62	87.56
Transformer	SETR	73.9	87.78	85.38
	PVT	76.62	89.03	86.59
	SwinUNet	77.53	89.78	87.61
CNN-Transformer Hybrid	CVT	73.93	87.93	85.16
	MPViT	77.22	88.89	87.37
	DBNet	77.37	89.71	87.4
Mamba	CCViM	74.5	88.1	85.3
	VM-UNet	77.13	89.53	86.29
	RS3Mamba	77.91	90.05	86.34
	MCloud	78.19	90.13	88.85

In comparison with models based on other architectures, MCloud not only outperforms most CNN and Transformer-based models but also significantly surpasses models with CNN-Transformer hybrid architectures. For instance, compared to SegNet, which performs well in the CNN architecture (MIoU of 77.01%), and SwinUNet, which performs well in the Transformer architecture (MIoU of 77.53%), MCloud's MIoU is higher by 1.18% and 0.66%, respectively. Moreover, MCloud also outperforms models in CNN-Transformer hybrid architectures, such as DBNet (MIoU of 77.37%), which was proposed in a previous study on cloud and cloud shadow semantic segmentation using attention mechanism-based multi-scale feature extraction (MIoU of 77.85%). This indicates that MCloud demonstrates stronger feature extraction and fusion capabilities. In comparisons within models based on the Mamba architecture, MCloud also stands out. Compared to CCViM (MIoU of 74.5%), VM-Unet (MIoU of 77.13%), and RS3Mamba (MIoU of 77.91%), MCloud's MIoU exceeds theirs by 3.69%, 1.06%, and 0.28%, respectively. This suggests that the MCloud, based on the Mamba architecture and paired with the network structure and MC module designed in this study, has superior feature extraction and fusion capabilities, enabling it to more effectively handle complex cloud and cloud shadow segmentation tasks.

Table 3. Comparison of classification evaluation metrics of different models on the CloudSEN-12 dataset.

Architecture	Model	Cloud			Cloud Shadow		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
CNN	FCN-32s	87.55	89.85	88.68	78.7	61.31	68.92
	DANet	88.59	88.44	88.51	75.51	66.02	70.44
	BiSeNetV2	90.08	89.81	89.94	77.26	71.03	74.01
	PAN	91.43	88.67	90.02	79.97	70.48	74.92
	CGNet	90.27	90.39	90.32	79	71.04	74.8
	LinkNet	91.96	88.38	90.13	80	71.5	75.51
	DenseASPP	91.44	89.23	90.32	79.71	71.36	75.3
	DeeplabV3	89.79	90.66	90.22	81.03	70.98	75.67
	HRNet	91.35	90.04	90.69	80	74.27	77.02
	OCRNet	91.44	90.32	90.87	81.91	72.8	77.08
	SegNet	91.22	90.63	90.92	81.5	73.14	77.09
Transformer	SETR	88.92	89.96	89.43	78.07	71.29	74.52
	PVT	89.67	90.43	90.04	78.98	72.34	75.51
	SwinUNet	91.16	91.19	91.17	80.88	75.96	78.34
CNN-Transformer Hybrid	CVT	89.62	89.44	89.52	76.55	71.44	73.9
	MPViT	91.66	89.67	90.65	78.28	73.79	75.96
	DBNet	91.7	90.51	91.1	80.03	76.17	78.05
Mamba	CCViM	89.5	88.2	88.8	75.4	74.1	74.7
	VM-UNet	90.08	91.72	90.9	76.1	80.13	78.12
	RS3Mamba	90.93	91.76	91.34	74.63	83.02	78.83
	MCloud	92.15	92.50	92.32	83.00	82.50	82.75

From the perspective of classification indicators, MCloud performs well in the segmentation task of both cloud and cloud shadow. In the cloud segmentation, MCloud's P, R, and F1 reach 92.15%, 92.50%, and 92.32%, respectively, which is significantly better than other models. In the segmentation of cloud shadow, MCloud's P, R, and F1 reach 83.00%, 82.50%, and 82.75%, respectively, which is also better than other models. This shows that MCloud can not only accurately detect the target area in the segmentation task of cloud and cloud shadow, but also retain rich boundary details to reduce false positives and missed judgments.

In order to further verify the performance of MCloud, this study randomly selected five images in different scenarios such as urban, rural, open space, and water, and compared the segmentation results using several models at the top of the MIoU index, as shown in Figure 5. From the visualization results, MCloud performs best in the segmentation task of cloud and cloud shadow. The segmentation results are better than other models in terms of edge accuracy and local detail, and can accurately identify small areas of clouds and cloud shadows, reducing misjudgment.

In summary, MCloud significantly outperforms models from other Mamba architectures on the CloudSEN-12 dataset, as well as most models with CNN, Transformer, and CNN-Transformer hybrid architectures. This shows that MCloud has significant advantages in feature extraction, fusion and multi-scale information processing, and can effectively improve the accuracy and robustness of cloud and cloud shadow segmentation.

3.4.2. Generalization Experiments on the 38-Cloud Dataset

In order to evaluate the segmentation performance and generalization ability of our proposed MCloud network, generalization experiments were carried out on the 38-Cloud dataset. Tables 4 and 5 show how our network compares to the current state-of-the-art model on the 38-Cloud dataset.

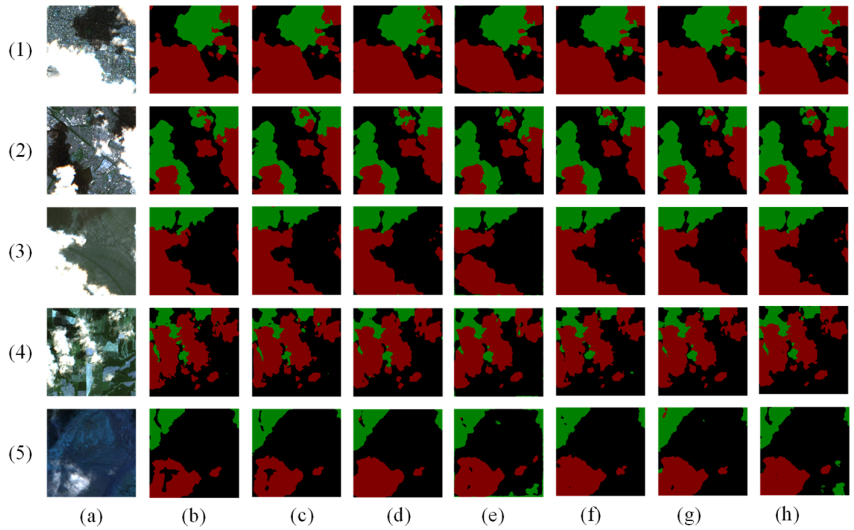

Figure 5. Comparison of segmentation results of different networks in several scenarios. (**a**) Test Image; (**b**) Label; (**c**) MCloud; (**d**) RS3Mamba; (**e**) VM-Unet; (**f**) SwinUNet; (**g**) DBNet; (**h**) SegNet.

Table 4. Comparison of overall evaluation metrics of different models on the 38-Cloud dataset.

Architecture	Model	MIoU(%)	PA(%)	MPA(%)
CNN	DANet	87.69	93.44	93.45
	FCN-32s	88.67	94	93.99
	BiSeNetV2	91.28	95.44	95.45
	LinkNet	91.48	95.55	95.55
	DenseASPP	91.62	95.62	95.63
	PAN	91.69	95.66	95.66
	DeeplabV3	91.86	95.75	95.77
	CGNet	92.24	95.96	95.98
	PSPNet	92.34	96.02	96.01
	SegNet	92.58	96.14	96.16
	HRNet	92.63	96.17	96.17
	CDUNet	92.64	96.18	96.19
	OCRNet	92.69	96.2	96.21
Transformer	SETR	82.65	90.5	90.51
	SwinUNet	93.1	96.42	96.42
CNN-Transformer Hybrid	CVT	87.92	93.57	93.56
	MPViT	92.86	95.96	95.97
	DBNet	93.27	96.52	96.51
Mamba	RS3Mamba	93	96.38	96.38
	VM-UNet	93.5	96.85	96.88
	CCViM	94.1	97.15	97.2
	MCloud	94.6	97.58	97.62

From the perspective of network architecture, the model based on the Mamba architecture shows significant advantages in cloud detection tasks, and its performance exceeds that of traditional CNN, Transformer, and hybrid architecture models. In terms of comprehensive performance, traditional CNN models, such as DANet and FCN-32s, and pure transformer models, which include SETR, performed the weakest, with their MIoU values falling below 90%. Hybrid architecture models, such as DBNet, outperform single-architecture models, but they are still inferior to Mamba models. Specifically, the Mamba architecture's MCloud topped the list with 94.60% MIoU, 97.58% PA, and 97.62% MPA, significantly ahead of other models. The Mamba-based algorithm shows significant advantages in cloud segmentation tasks, and its comprehensive performance surpasses that of traditional convolutional, Transformer, and hybrid architecture models. Traditional

convolutional networks, such as DANet and BiSeNetV2, and pure transformer models, such as SETR, exhibit insufficient segmentation accuracy in complex scenarios due to their limited feature modeling capabilities. Although hybrid architecture models improve performance by integrating multiple types of features, they are still limited by computational complexity and local-global information interaction efficiency. In contrast, the Mamba architecture achieves a breakthrough balance between long-range dependency modeling and computational efficiency through the co-design of state-space models and convolution.

Table 5. Comparison of classification evaluation metrics of different models on the 38-Cloud dataset.

Architecture	Model	Cloud			Background		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
CNN	DANet	93.69	93.08	93.38	93.2	93.79	93.5
	FCN-32s	93.78	94.17	93.98	94.21	93.82	94.02
	BiSeNetV2	94.68	96.24	95.46	96.22	94.65	95.43
	LinkNet	95.63	95.41	95.52	95.47	95.68	95.58
	DenseASPP	95.35	95.87	95.61	95.9	95.38	95.64
	PAN	95.33	95.99	95.66	96	95.35	95.67
	DeeplabV3	94.75	96.83	95.79	96.79	94.69	95.74
	CGNet	94.88	96.12	95.49	97.08	94.81	95.95
	PSPNet	95.44	96.61	96.02	96.61	95.43	96.02
	SegNet	96.22	96.02	96.12	96.07	96.27	96.17
	HRNet	96.02	96.29	96.16	96.32	96.06	96.19
	CDUNet	95.52	96.86	96.19	96.85	95.5	96.18
	OCRNet	96.48	95.87	96.17	95.94	96.54	96.24
Transformer	SETR	89.86	91.2	90.53	91.16	89.82	90.49
	SwinUNet	96.4	96.41	96.41	96.45	96.44	96.44
CNN-Transformer Hybrid	CVT	93.51	93.56	93.54	93.62	93.58	93.6
	MPViT	95.97	95.93	95.94	95.23	95.74	95.48
	DBNet	96.82	96.16	96.49	96.22	96.67	96.44
Mamba	RS3Mamba	95.99	96.75	96.37	96.76	96.01	96.38
	VM-UNet	96.82	96.5	96.66	96.92	96.8	96.86
	CCViM	97.1	96.95	97.03	97.3	97.1	97.20
	MCloud	97.5	97.2	97.35	97.8	97.3	97.55

The time complexity of Transformer is $O(n^2 \times d)$; and the time complexity of VSSM is $O(n \times d)$ or $O(n \times d \times logn)$, respectively. The $O(n \times d)$ complexity arises when no pre-computation is used, relying solely on direct sampling or approximation. While simple to implement, this approach becomes highly inefficient for large filters due to explicit computation costs. In contrast, the $O(n \times d \times \log n)$ complexity leverages acceleration structures (e.g., MIPMAP, SAT) to enable efficient large-filter operations. These methods trade pre-computation overhead for runtime efficiency by employing statistical approximations and hierarchical optimizations.

For the MCloud network, when evaluating its segmentation performance and generalization on the 38-Cloud dataset, with Tables 4 and 5 comparing to state-of-the-art models, the Mamba-based architecture stands out. Traditional CNNs, such as DANet and FCN-32s, and pure Transformers, such as SETR, are constrained by limitations in feature modeling, yielding weak performance (MIoU < 90%) and often incurring high complexities like $O(n^2 \times d)$. Hybrid models (e.g., DBNet), though better than single architectures, still face bottlenecks from computational complexity and suboptimal information interaction, likely adhering to costlier complexity patterns. In contrast, our MCloud leverages Mamba's state-space model and convolution co-design. This not only delivers superior performance, topping metrics with 94.60% MIoU, 97.58% PA, and 97.62% MPA, but also achieves a breakthrough balance: it avoids the high quadratic complexity of traditional models, align-

ing more closely with the lower-order, efficient complexities $O(n \times d)$ or $O(n \times d \times logn)$. Thus, in both performance and computational efficiency, the Mamba-based method outpaces competitors, boasting the smallest complexity while delivering state-of-the-art cloud segmentation results.

Our proposed MCloud network achieves a balance between global context perception and local detail extraction through the collaborative design of state-space branches and convolutional branches, while discarding the dependence on multiband input to reduce the computational complexity, and achieving the leading performance with visible light data alone. Compared with DBNet, the MIoU is improved by 1.33%, and the parameter volume is reduced by 43%, providing an efficient and reliable solution for real-time remote sensing image processing. This result validates the feasibility and potential of the Mamba architecture in cloud and cloud shadow remote sensing tasks.

Figure 6 shows the segmentation results of MCloud, DBNet, and SwinUNet in complex background scenarios such as cloudless and multi-cloud. From the visualization results, it can be seen that the segmentation results of MCloud are significantly better than other models in terms of edge continuity and detail restoration ability. In the prediction results of DBNet and CDUNet, obvious jagged fractures appear at the cloud boundary, particularly in thin cloud areas, where local misjudgment is prone to occur, as shown in Figure 2. Although OCRNet improves the detection accuracy of cloud subjects through multi-scale feature extraction, it is not adaptable enough to the changes in the internal texture of clouds, resulting in the over-smoothing of thick cloud segmentation. SwinUNet's Transformer-based global modeling capability improves the coherence of the cloud contour, but there are still missing detections in small-scale cloud block detection.

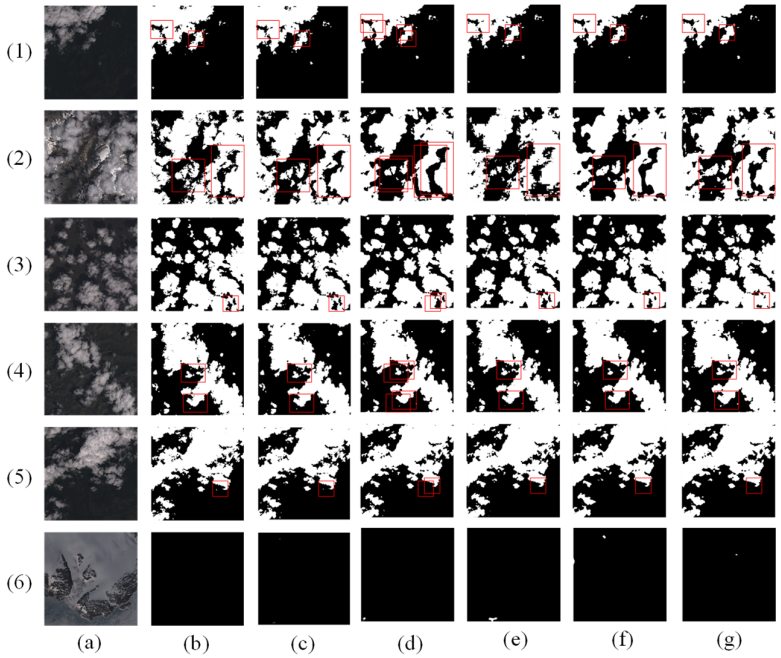

Figure 6. Comparison of different models on the 38-Cloud dataset. (**a**) Test Image; (**b**) Label Image. (**c**) MCloud; (**d**) DBNet; (**e**) SwinUNet; (**f**) OCRNet; (**g**) CDUNet.

The segmentation results of MCloud proposed by us show remarkable robustness and accuracy, and it effectively captures the continuity characteristics of cloud distribution and avoids the edge fracture problem through the long-range dependence of state-space branching modeling. In the dense cloudy area, the local texture information extracted from the convolution branch and the global semantic guidance of the state space branch synergize, and the fine distinction of the thick and thin areas in the cloud layer is realized.

However, in the snow area with high reflection background interference, as shown in Figure 6, MCloud significantly suppresses the false detection noise by dynamically filtering the cross-scale context features, while completely preserving the boundary details of clouds and ground objects. It is worth noting that MCloud only relies on visible band input, which confirms the potential of state-space architecture in complex feature modeling.

In summary, MCloud achieves an efficient balance between global context perception and local feature resolution through the deep collaboration between the state space model and the convolution module, and its output results reach the advanced level in terms of edge accuracy, noise suppression and scene adaptability, which provides a new solution for the semantic segmentation task of remote sensing images.

3.4.3. Generalization Experiments on the SPARCS-Val Dataset

In order to further evaluate the segmentation performance and generalization ability of MCloud networks, comparative experiments were also carried out on the SPARCS-Val dataset with more classifications and more scenarios. The experimental results are shown in Tables 6 and 7, where Table 6 show the overall indicators and Table 7 show the pixel accuracy of different models for each category.

Table 6. Comparison of overall evaluation metrics of different models on the SPARCS-Val dataset.

Model	Overall Data				
	MIoU(%)	**PA(%)**	**MPA(%)**	**R(%)**	**F1(%)**
DANet	55.61	85.04	70.28	67.12	66.76
FCN-32s	61.38	88.03	75.41	71.2	72.4
BiSeNetV2	64.38	88.57	80.33	73.26	75.8
SegNet	65.86	89.3	80.74	75.18	77.53
CGNet	66.82	89.93	80	76.37	77.31
PSPNet	67.23	89.92	82.5	75.23	77.81
DenseASPP	67.73	89.81	82.42	76.21	78.63
DeeplabV3	68.26	90.06	82.94	76.9	79.05
LinkNet	68.62	90.84	83.38	76.8	79.24
HRNet	69.74	90.98	84.61	77.3	80.51
OCRNet	69.91	90.94	86.21	77.15	80.04
SETR	63.59	87.73	79.58	72.89	75.38
PVT	68.54	89.28	83.02	77.57	80.2
SwinUNet	73.0	91.86	86.44	80.44	83.1
CVT	62.68	87.03	78.3	72.42	74.6
MPViT	72.98	90.02	85.26	79.49	82.27
DBNet	74.04	92.54	87.26	81.01	83.65
VM-UNet	74.9	92.14	88.24	81.95	84.59
RS3Mamba	75.36	93.14	86.86	83.52	84.98
CCViM	76.46	93.02	**89.1**	83.17	85.61
MCloud	77.47	93.77	88.23	85.06	86.5

Table 7. Comparison of classification evaluation metrics of different models on the SPARCS-Val dataset.

Model	Class Pixel Accuracy (%)						
	CS	**CSOW**	**W**	**I/S**	**L**	**C**	**F**
DANet	55.44	30.37	89.21	87.63	91.07	76.89	61.36
FCN-32s	64.63	37.51	89.57	89.56	92.36	83.1	71.17
BiSeNetV2	73.24	57.35	90.2	92.17	91.65	83.18	74.52
SegNet	76.98	55.74	86.82	92.49	92.66	83.5	77.02
CGNet	72.47	50.32	93	90.53	94.02	84.17	75.49
PSPNet	76.72	56.59	93.06	92.59	92.75	83.81	82.02

Table 7. *Cont.*

Model	Class Pixel Accuracy (%)						
	CS	**CSOW**	**W**	**I/S**	**L**	**C**	**F**
DenseASPP	77.13	57.26	94.13	93.23	93.46	81.2	80.58
DeeplabV3	79.87	60.99	91.28	91.56	93.71	81.97	81.21
LinkNet	79.3	60.06	87.36	91.46	93.2	88.7	83.58
HRNet	82.77	65.31	89.36	93.22	93.23	85.91	82.52
OCRNet	81.64	68.4	94.27	93.53	92.46	87.4	85.78
SETR	71.16	55.83	89.85	92.31	91.89	78.9	77.14
PVT	76.34	62.57	90.42	93.23	92.46	84.6	81.55
SwinUNet	80.14	70.09	92.34	94.17	94.15	88.27	85.94
CVT	67.11	52.85	88.78	93.11	91.64	77.79	76.85
MPViT	80.42	68.95	91.03	93.77	92.96	84.56	85.18
DBNet	81.74	70.21	93.3	94.15	94.44	89.38	87.62
VM-UNet	82.43	75.21	94.15	94.88	95.04	85.22	90.76
RS3Mamba	85.66	70.69	93.5	94.41	95.25	89.84	78.71
CCViM	83.39	77.55	94.27	93.95	95.23	89.01	90.3
MCloud	81.95	74.38	95.34	94.31	95.86	92.06	83.74

CS refers to cloud shadow classification, CSOW refers to cloud shadow classification on water, W refers to water classification, I/S refers to ice and snow classification, L refers to land classification, C refers to cloud classification, and F refers to flood classification.

Experimental results show that MCloud performs better on the SPARCS-Val dataset than most models with other architectures. Specifically, MCloud's MIoU, PA, MPA, R, and F1 indicators reached 77.47%, 93.77%, 88.23%, 85.06%, and 86.5%, respectively, and performed well among all models.

This indicates that MCloud has strong generalization ability when dealing with complex datasets. From the perspective of category pixel accuracy, MCloud performed well in cloud (C) and land (L) classification, with pixel accuracy of 92.06% and 95.86%, respectively. MCloud also achieved good results in the classification of cloud shadow (CS) and cloud shadow over water (CSOW), with 81.95% and 74.38%, respectively. In addition, MCloud's pixel accuracy of 95.34% and 94.31%, respectively, for water (W) and ice and snow (I/S) classifications is equally excellent. Specifically, MCloud's high pixel accuracy on cloud (C) and land (L) classifications indicates that it has high accuracy in distinguishing between these two common feature categories. Although the accuracy of cloud shadow (CS) and cloud shadow over water (CSOW) classification is relatively low, it still shows good recognition ability, which may be related to the complexity and diversity of cloud shadow. In the classification of water (W) and ice and snow (I/S), MCloud's high pixel accuracy further proves its effectiveness when dealing with these features with different spectral and spatial characteristics. MCloud has a well-balanced segmentation performance in different categories, and can effectively handle the segmentation tasks of various complex features. This shows that MCloud not only performs well in terms of overall performance, but also has high accuracy and stability in the face of different types of feature classification.

Figure 7 shows the segmentation results of multiple models in different scenarios of the SPARCS-Val dataset. As can be seen from the figure, the segmentation results of the traditional convolutional structure network DeepLab V3 have the problems of rough edges and false detections, especially in the classification of water areas. Although the SwinUNet of the Transformer structure and the DBNet of the hybrid structure have relatively good segmentation results, there are still a certain range of false detections. In contrast, MCloud's segmentation results are outstanding. In the segmentation task of cloud and cloud shadow, MCloud can accurately segment the boundary between cloud and cloud shadow, retain rich boundary details, and reduce the occurrence of false detection. This is mainly due to the

introduction of the Mamba architecture based on the state space model in MCloud, which enables the effective modeling of long-range dependencies and local features through the collaborative work of state space architecture branches and convolutional architecture branches. In addition, the MC module designed by MCloud further enhances the model's ability to resolve complex features, so that the model can segment clouds and cloud shadows more accurately.

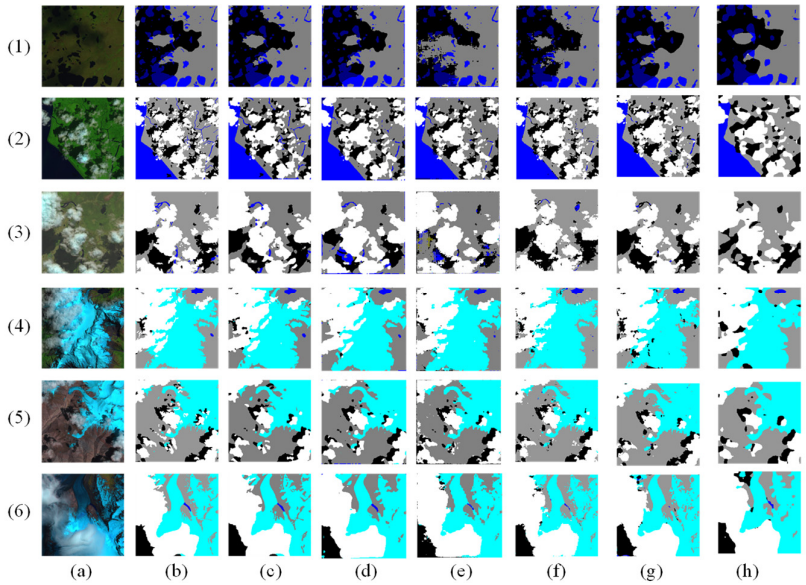

Figure 7. Comparison of different models on the SPARCS-Val dataset.(**a**) Test Image; (**b**) Label; (**c**) MCloud; (**d**) RS3Mamba; (**e**) VM-UNet; (**f**) SwinUNet; (**g**) DBNet; (**h**) DeepLab V3.

In complex scenes, such as ice and snow noise interference, MCloud can still maintain good segmentation performance. This is mainly due to the fact that MCloud's Mamba architecture is able to effectively capture global context information, while the convolutional architecture branch is able to extract local feature details. By synergistically integrating these two features, the MC module realizes a cross-scale feature interaction mechanism, thereby improving the robustness and adaptability of the model in complex scenarios.

Compared with other networks based on the Mamba architecture, such as RS3Mamba and VM-UNet, MCloud also shows significant advantages in segmentation results. Although RS3Mamba also has a good performance in the segmentation of cloud and cloud shadow, there are still some false detections when dealing with complex scenes. VM-UNet also has similar problems in the segmentation results, especially in the case of ice and snow noise interference, the false detection phenomenon is obvious. In contrast, MCloud further enhances the model's ability to resolve complex feature features by introducing the MC module, so as to show excellent segmentation performance in different scenarios.

In summary, the performance of MCloud on the SPARCS-Val dataset proves its generalization ability and robustness in cloud and cloud shadow semantic segmentation tasks. This is mainly due to the introduction of the Mamba architecture based on the state space model in MCloud, which enables the effective modeling of long-range dependencies and local features through the collaborative work of the state space architecture branch and the convolutional architecture branch. At the same time, the design of the MC module further enhances the model's ability to analyze complex ground features, so that MCloud can maintain excellent segmentation performance on different datasets. The results are shown in Figure 8.

In summary, MCloud's accurate segmentation in complex surface, low-contrast and broken cloud scenes can capture local details such as cloud shadow edges and thin cloud

textures through feature visualization depth analysis, shallow volume branching, and this hierarchical feature division and module collaboration clearly express MCloud's decision-making logic from local details to global semantics, so that the advantages and disadvantages of segmentation results can be explained.

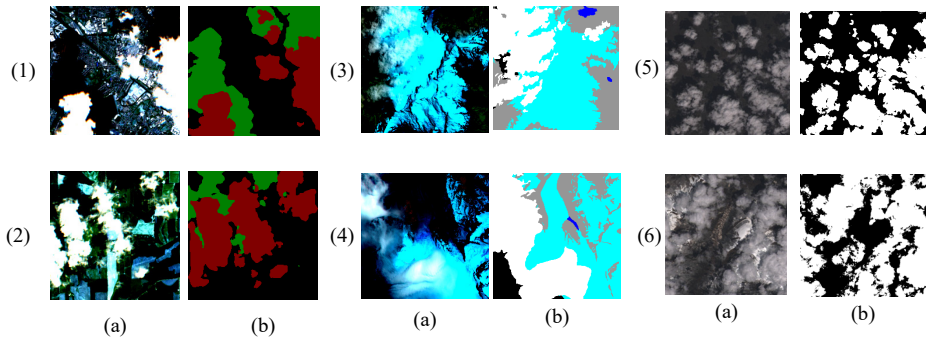

Figure 8. The comparison of the MCloud model on different datasets. (**a**) Test Image; (**b**) MCloud.

MCloud has achieved significant results in cloud and cloud shadow semantic segmentation tasks, but there are still some areas that can be improved. In future work, we will further optimize the model to improve its inference speed. This study proves the feasibility of Mamba architecture in cloud and cloud shadow semantic segmentation in remote sensing images, which is of great significance for promoting the development of cloud and cloud shadow semantic segmentation.

4. Performance Analysis

The original motivation of this study was to explore the possibility of introducing the Mamba architecture based on the state space model to the cloud and cloud shadow detection tasks in order to solve the high computational complexity of the Transformer architecture. In order to comprehensively evaluate the balance between the computational overhead and segmentation accuracy of the MCloud model proposed in this study, the parameters and computational complexity of MCloud and other advanced models were compared with the MIoU indicators on the CloudSEN-12, 38-Cloud and SPARCS-Val datasets, and the results are shown in Figures 9–11.

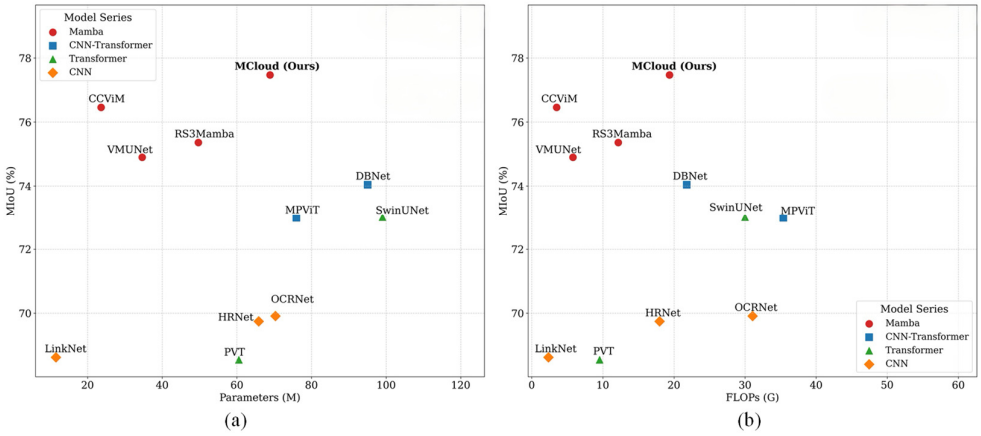

Figure 9. Performance analysis of models on the SPARCS-Val dataset. (**a**) Comparison of parameter count and MIoU metric; (**b**) comparison of computational complexity and MIoU Metric.

Through comparative analysis, it is found that in the models based on Mamba architecture, although the number of parameters is slightly higher than that of RS3Mamba, CCViM and VM-UNet, the segmentation performance of MCloud is significantly better

than that of these models. This shows that MCloud can effectively improve the feature expression ability through reasonable network design, especially the introduction of MC module, so that the model can capture the morphological features of clouds and cloud shadows more accurately. Compared with CCViM and VM-UNet, which have lower computational complexity, the performance improvement of MCloud is significantly higher than the increase in computational complexity, reflecting a good balance between efficiency and performance.

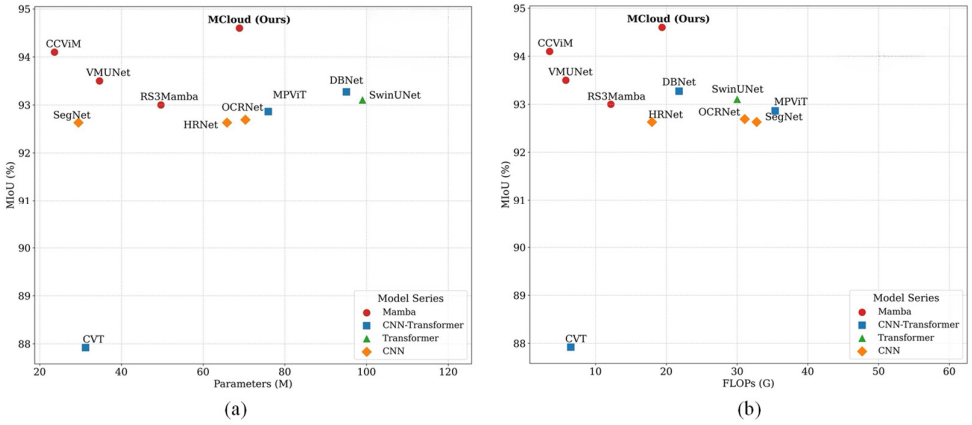

Figure 10. Performance analysis of models on the 38-Cloud Dataset. (**a**) Comparison of parameter count and MIoU metric; (**b**) comparison of computational complexity and MIoU Metric.

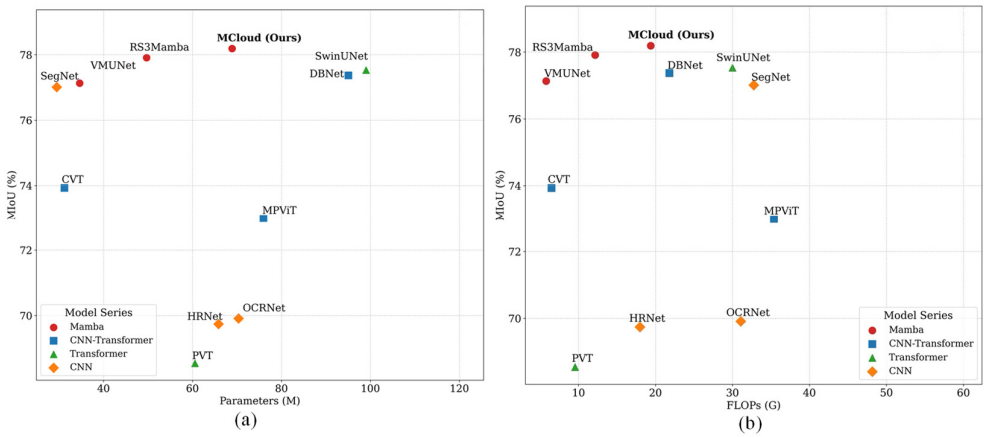

Figure 11. Performance analysis of models on the CloudSEN-12 dataset. (**a**) Comparison of parameter count and MIoU Metric; (**b**) comparison of computational complexity and MIoU Metric.

Compared with the model based on the convolution-Transformer hybrid architecture, MCloud significantly reduces the computational complexity and parameter quantity while maintaining higher segmentation accuracy. This huge efficiency improvement is due to the linear computational complexity of the Mamba architecture, which proves the feasibility of introducing the Mamba architecture into remote sensing image cloud and cloud shadow semantic segmentation tasks.

From a broader perspective, different architectural models present different balances between efficiency and performance. Convolutional models such as LinkNet have the highest computational efficiency but limited performance; Transformer models such as SwinUNet have better performance but heavier computational burden. In particular, the MCloud proposed in this study achieves a segmentation performance that is close to or even surpasses that of most convolutional-Transformer hybrid architecture models when the number of parameters and computational complexity are only slightly higher than those of some convolutional models.

Overall, MCloud strikes a good balance between performance and computational complexity, providing an efficient and practical solution for cloud and cloud shadow semantic segmentation tasks. Compared to previous studies, MCloud significantly improves computing efficiency while maintaining near-top-tier segmentation performance, making it more suitable for real-world deployment applications. At the same time, MCloud's exploration of state-space models and Mamba architecture in the field of remote sensing image processing shows that this direction has broad research prospects and application potential.

5. Conclusions

In this chapter, MCloud, a state-space model-based cloud and cloud shadow semantic segmentation network, is proposed, which is the first time to introduce the state-space model-based Mamba architecture into the cloud and cloud shadow semantic segmentation task of remote sensing images. MCloud enables effective modeling of long-range dependencies and local features through the collaborative work of state-space architecture branches and convolutional architecture branches. The MC module designed in this study further enhances the model's ability to parse complex features, and realizes a cross-scale feature interaction mechanism by integrating the global context modeling capabilities of the Mamba architecture and the local feature perception advantages of the convolutional network. Recent years have witnessed significant advancements in computer vision technology, with deep learning providing promising solutions to change detection problems. Experimental results show that MCloud exhibits excellent segmentation performance and generalization ability on multiple datasets. Compared with traditional CNN and Transformer architecture models, MCloud shows greater robustness and adaptability when dealing with complex scenarios. Compared with other networks based on Mamba architecture, MCloud also shows obvious advantages in segmentation results, especially in complex scenarios, further solidifying its competitiveness not only against traditional deep learning frameworks but also among state-of-the-art Mamba-based models. However, it still relies on large-scale manual annotation data, which limits its application in scarce scenarios. Future research can further make up for the lack of information in thick clouds and low visibility scenarios, and further enhance the application of the model in business scenarios such as meteorological early warning and agricultural resource survey.

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

Funding: This work is supported in part by the Key Laboratory of Airborne Geophysics and Remote Sensing Geology Ministry of Nature Resources (2023YFL36), the National Key Research and Development Program of China "Cooperation research and demonstration application of monitoring technologies for the snow, glaciers and geohazards in High Mountain Asia and Arctic"(2021YFE0116800) and the National Natural Science Foundation of PR China (42075130).

Data Availability Statement: The data and the code of this study are available from the corresponding author upon request.

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

References

1. King, M.D.; Platnick, S.; Menzel, W.P.; Ackerman, S.A.; Hubanks, P.A. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 3826–3852. [CrossRef]
2. Papageorgiou, G.; Petrakis, C.; Ioannou, N.; Zagarelou, D. Effective business planning for sustainable urban development: The case of active mobility. In Proceedings of the ECIE 2019 14th European Conference on Innovation and Entrepreneurship (2 vols), Kalamata, Greece, 19–20 September 2019; p. 759.
3. McNally, A.P.; Watts, P.D. A cloud detection algorithm for high-spectral-resolution infrared sounders. *Q. J. R. Meteorol. Soc.* **2003**, *129*, 3411–3423. [CrossRef]
4. Tapakis, R.; Charalambides, A.G. Equipment and methodologies for cloud detection and classification: A review. *Sol. Energy* **2013**, *95*, 392–430. [CrossRef]
5. Goodman, A.H.; Henderson-Sellers, A. Cloud detection and analysis: A review of recent progress. *Atmos. Res.* **1988**, *21*, 203–228. [CrossRef]
6. Kazantzidis, A.; Tzoumanikas, P.; Bais, A.; Fotopoulos, S.; Economou, G. Cloud detection and classification with the use of whole-sky ground-based images. *Atmos. Res.* **2012**, *113*, 80–88. [CrossRef]
7. Zi, Y.; Xie, F.; Jiang, Z. A cloud detection method for Landsat 8 images based on PCANet. *Remote Sens.* **2018**, *10*, 877. [CrossRef]
8. Tian, M. A method for building a cadastral database of villages and towns based on ArcGIS. *Beijing Surv. Mapp.* **2015**, *6*, 94–98.
9. Huang, Q.; Zheng, X.J.; Liu, C. Non meteorlogic applications of meteorlogical satellite data in China. *China Aerosp.* **1997**, *7*, 14–17.
10. Xiang, D.X. Research on Drought Remote Sensing Monitoring Model Based on Cloud Parameter Method. Master's Thesis, Wuhan University, Wuhan, China, 2011.
11. Xu, M.; Wang, S.H.; Guo, R.Z.; Jia, X.; Jia, S. Review of Cloud Detection and Removal Methods for Remote Sensing lmages. *J. Comput. Res. Dev.* **2024**, *61*, 1585–1607.
12. 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]
13. Zhang, Q.; Xiao, C. Cloud detection of RGB color aerial photographs by progressive refinement scheme. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 7264–7275. [CrossRef]
14. Wang, L.; Li, X.; Bao, Y.X.; Shao, Y. Research progress of remote sensing application on transportation meteorological disasters. *Remote Sens. Land Resour.* **2018**, *30*, 1–7.
15. Chassery, J.M.; Garbay, C. An iterative segmentation method based on a contextual color and shape criterion. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *PAMI-6*, 794–800. [CrossRef] [PubMed]
16. Li, H.; Wang, Y.; Liu, K.J.R.; Lo, S.-C.B.; Freedman, M.T. Computerized radiographic mass detection. I. Lesion site selection by morphological enhancement and contextual segmentation. *IEEE Trans. Med. Imaging* **2001**, *20*, 289–301. [CrossRef] [PubMed]
17. Wang, H.X.; Jin, H.J.; Wang, J.L.; Jiang, W.S. Optimization Approach for Multi-scale Segmentation of Remotely SensedImagery under k-means Clustering Guidance. *Cehui Xuebao* **2015**, *44*, 526.
18. Huang, Z.K.; Chau, K.W. A new image thresholding method based on Gaussian mixture model. *Appl. Math. Comput.* **2008**, *205*, 899–907. [CrossRef]
19. 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.
20. Weng, L.; Pang, K.; Xia, M.; Lin, H.; Qian, M.; Zhu, C. Sgformer: A local and global features coupling network for semantic segmentation of land cover. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 6812–6824. [CrossRef]
21. Dong, Z.; Yang, D.; Reindl, T.; Walsh, W.M. Short-term solar irradiance forecasting using exponential smoothing state space model. *Energy* **2013**, *55*, 1104–1113. [CrossRef]
22. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
23. Choi, H.; Bindschadler, R. Cloud detection in Landsat imagery of ice sheets using shadow matching technique and automatic normalized difference snow index threshold value decision. *Remote Sens. Environ.* **2004**, *91*, 237–242. [CrossRef]
24. McIntire, T.J.; Simpson, J.J. Arctic sea ice, cloud, water, and lead classification using neural networks and 1.6-/spl mu/m data. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 1956–1972. [CrossRef]
25. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; part III 18; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
26. 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 June–12 June 2015; pp. 3431–3440.
27. Hong, D.; Zhang, B.; Li, H.; Li, Y.; Yao, J.; Li, C.; Werner, M.; Chanussot, J.; Zipf, A.; Zhu, X.X. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sens. Environ.* **2023**, *299*, 113856. [CrossRef]

28. Lu, H.Y. Research on Remote Sensing Image Water Body Extraction Based on CNN-Transformer and Semi-Supervised Adversarial Methods. Master's Thesis, Nanjing University of Information Science & Technology, Nanjing, China, 2024.
29. Cheng, P.; Xia, M.; Wang, D.; Lin, H.; Zhao, Z. Transformer Self-Attention Change Detection Network with Frozen Parameters. *Appl. Sci.* **2025**, *15*, 3349. [CrossRef]
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
31. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752. [CrossRef]
32. Ren, W.; Wang, Z.; Xia, M.; Lin, H. MFlNet: Multi-scale feature interaction network for change detection of high-resolution remote sensing images. *Remote Sens.* **2024**, *16*, 1269. [CrossRef]
33. 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 Forty-first International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
34. He, X.; Cao, K.; Zhang, J.; Yan, K.; Wang, Y.; Li, R.; Xie, C.; Hong, D.; Zhou, M. Pan-mamba: Effective pan-sharpening with state space model. *Inf. Fusion* **2025**, *115*, 102779. [CrossRef]
35. Chen, K.; Chen, B.; Liu, C.; Li, W.; Zou, Z.; Shi, Z. Rsmamba: Remote sensing image classification with state space model. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
36. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587. [CrossRef]
37. 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.
38. Gu, P.Z.; Liu, W.C.; Feng, S.Y.; Wei, T.Y.; Wang, J.; Chen, H. Hpn-cr: Heterogeneous parallel network for sar-optical data fusion cloud removal. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. [CrossRef]
39. He, Z.C.; Wang, P.; Zou, Y.K.; Huang, B.; Zhu, D.Y.; Harry, F.L.; Henry, L. DADIGAN: A dual attention blocks-based disentangled iterative Generative Adversarial Network for cloud and shadow removal on SAR and optical images. *Inf. Fusion* **2025**, *125*, 103487. [CrossRef]
40. 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.
41. Gu, A.; Dao, T.; Ermon, S.; Rudra, A.; Re, C. Hippo: Recurrent memory with optimal polynomial projections. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1474–1487.
42. 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]
43. SMohajerani, P. Saeedi, Cloud-net: An end-to-end cloud detection algorithm for landsat 8 imagery. In Proceedings of the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1029–1032.
44. Hughes, M.J.; Hayes, D.J. Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing. *Remote Sens.* **2014**, *6*, 4907–4926. [CrossRef]