Multi-scale feature mixed attention network for cloud and snow segmentation in remote sensing images

remote sensing

MDPI

Article

Multi-Scale Feature Mixed Attention Network for Cloud and Snow Segmentation in Remote Sensing Images

Liling Zhao [1,2,*], Junyu Chen [1,2], Zichen Liao [1,3] and Feng Shi [1,2]

[1] School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202212490018@nuist.edu.cn (J.C.); fj808642@student.reading.ac.uk (Z.L.); 202312490429@nuist.edu.cn (F.S.)
[2] Jiangsu Key Laboratory of Big Data Analysis Technology, B-DAT, Nanjing University of Information Science and Technology, Nanjing 210044, China
[3] Department of Computer Science, University of Reading, Whiteknights, Reading RG6 6DH, UK
* Correspondence: zhaoliling@nuist.edu.cn; Tel.: +86-025-58731272

Abstract: The coexistence of cloud and snow is very common in remote sensing images. It presents persistent challenges for automated interpretation systems, primarily due to their highly similar visible light spectral characteristic in optical remote sensing images. This intrinsic spectral ambiguity significantly impedes accurate cloud and snow segmentation tasks, particularly in delineating fine boundary features between cloud and snow regions. Much research on cloud and snow segmentation based on deep learning models has been conducted, but there are still deficiencies in the extraction of fine boundaries between cloud and snow regions. In addition, existing segmentation models often misjudge the body of clouds and snow with similar features. This work proposes a Multi-scale Feature Mixed Attention Network (MFMANet). The framework integrates three key components: (1) a Multi-scale Pooling Feature Perception Module to capture multi-level structural features, (2) a Bilateral Feature Mixed Attention Module that enhances boundary detection through spatial-channel attention, and (3) a Multi-scale Feature Convolution Fusion Module to reduce edge blurring. We opted to test the model using a high-resolution cloud and snow dataset based on WorldView2 (CSWV). This dataset contains high-resolution images of cloud and snow, which can meet the training and testing requirements of cloud and snow segmentation tasks. Based on this dataset, we compare MFMANet with other classical deep learning segmentation algorithms. The experimental results show that the MFMANet network has better segmentation accuracy and robustness. Specifically, the average MIoU of the MFMANet network is 89.17%, and the accuracy is about 0.9% higher than CSDNet and about 0.7% higher than UNet. Further verification on the HRC_WHU dataset shows that the MIoU of the proposed model can reach 91.03%, and the performance is also superior to other compared segmentation methods.

Keywords: remote sensing; segmentation; deep learning; attention mechanism

Academic Editor: Mohammad Awrangjeb

Received: 28 March 2025
Revised: 15 May 2025
Accepted: 23 May 2025
Published: 28 May 2025

Citation: Zhao, L.; Chen, J.; Liao, Z.; Shi, F. Multi-Scale Feature Mixed Attention Network for Cloud and Snow Segmentation in Remote Sensing Images. *Remote Sens.* **2025**, *17*, 1872. https://doi.org/10.3390/rs17111872

1. Introduction

The rapid development of remote sensing technology has significantly improved people's understanding of the Earth. In the large-scale snow depth estimation task, remote sensing images play a pivotal role. Currently, researchers widely use multi-source remote sensing data and auxiliary data for snow depth retrieval [1,2]. However, cloud and snow exhibit highly similar spectral reflectance and color characteristics in optical remote sensing images, leading to frequent misclassification and posing great challenges for accurate snow depth estimation. Efficient and precise cloud/snow detection and segmentation are

therefore critical for reducing misclassification errors and obtaining high-quality snow cover information.

Early cloud/snow segmentation methods primarily relied on spectral features and empirical rules. Threshold-based algorithms, such as the Normalized Difference Snow Index (NDSI) and Cloud Index (CI), were developed to leverage reflectance differences in the shortwave infrared (SWIR) and thermal infrared (TIR) bands [3–5]. The Fmask algorithm [6], a classical benchmark method, integrates multiple thresholds to identify cloud, snow, and shadows. However, although these methods are generally effective in ideal conditions, most of them struggle in complex environments. If these methods meet thin cloud and high-altitude snow-covered regions, they often lead to high misclassification rates due to spectral confusion. Furthermore, such algorithms are heavily dependent on specific sensor bands (e.g., Landsat's SWIR channel), limiting their adaptability to high-resolution imagery with limited spectral bands (e.g., WorldView-3). Additionally, the dynamic nature of cloud cover and the seasonal stability of snow have fostered the development of multi-temporal analysis methods, such as Tmask, which detects transient cloud cover based on time-series reflectance variations. The multi-temporal analysis method of remote sensing images can effectively detect dynamic changes in ground objects and improve the accuracy of target recognition [7,8]. However, these approaches require dense temporal image sequences, resulting in high data processing costs and poor adaptability to sudden landscape changes (e.g., post-wildfire surface reflectance variations).

The introduction of machine learning marks a shift toward intelligent cloud/snow segmentation. Researchers have employed Support Vector Machines (SVM) [9] and Random Forests (RF) [10] with manually designed texture and shape features (e.g., gray-level co-occurrence matrix contrast and entropy) to improve classification accuracy, achieving approximately 15% higher accuracy than threshold-based methods in complex terrains. The multi-granularity cascade forest (gcForest) [11] further enhances feature extraction, reducing processing time by 30% in HJ-1A/1B imagery compared to traditional methods. However, conventional machine learning approaches still remain limited by their dependence on handcrafted features and model capacity. These methods often perform inconsistently in complex scenarios, particularly in nonlinear spectral transitions at cloud/snow boundaries and low signal-to-noise ratio shadow regions. When cloud/snow coexistence exceeds 40%, or in heavily mixed cloud/snow regions (e.g., snowmelt transition zones), classification accuracy drops below 75%, sometimes performing even worse than the Fmask baseline, highlighting the limitations of traditional machine learning models.

Deep learning, as a data-driven approach, has the ability to learn complex nonlinear relationships embedded within datasets through neural networks. It has been widely applied in fields such as semantic segmentation [12,13] and change detection of optical remote sensing images [14,15], achieving high-precision semantic segmentation. The development of Convolutional Neural Networks (CNNs) has further boosted cloud/snow segmentation accuracy. For example, the Fully Convolutional Networks (FCN) achieved the first end-to-end pixel-level classification [16], and encoder-decoder architectures such as U-Net [17], through skip connections, enabled multi-scale feature fusion, improving the cloud/snow intersection-over-union (IoU) to 89% on Sentinel-2 imagery. CDNetV2, developed by the Guo team, used an encoder-decoder structure to extract cloud regions in satellite thumbnails but lacked adaptability for high-resolution data [18]. A multi-scale feature fusion network proposed by H Du effectively mitigated cloud/snow confusion, reducing misdetection rates to 8.7% on Landsat-8 data, but still not that robust under some complex circumstances [19]. Due to the similarity in many attributes of cloud and snow, cloud/snow detection in remote sensing images is inherently more difficult compared to other tasks. While existing methods have reduced the interaction between cloud and

snow to some extent, they still have limitations, especially in complex scenes, and cannot guarantee robustness. To address this, Li Y proposed an improved cloud/snow detection method based on the UNet3+ network, leveraging its advantage in feature fusion [20]. Fang Z developed a new deep learning model for high-resolution remote sensing images from different latitudes. This method first extracts the texture and spectral information of various objects, then processes these features, and finally generates the final cloud/snow mask image [21]. Furthermore, Xi Wu introduced geographical information to address the cloud/snow feature differences in different geographic locations, proposing a new neural network model to improve the adaptability of cloud/snow detection. This method demonstrated stronger robustness and broader applicability across various scenarios compared to existing techniques [22]. In recent years, the development of the Transformer has injected new momentum into semantic segmentation. The Vision Transformer (ViT) [23] utilized self-attention mechanisms to model global contextual dependencies, achieving an IoU greater than 91% in cloud/snow wide-distribution scenarios. However, pure Transformer models still face some limitations due to their enormous data requirements for pretraining (billions of samples) and high computational complexity, hindering practical applications. As a result, researchers have shifted towards exploring CNN–Transformer hybrid architectures. PVT, developed by Wang, integrated a pyramid structure with a Transformer, improving the mIoU to 44.8% on the ADE20K dataset [24]. Wu's Convolutional Vision Transformer (CVT) introduced local inductive biases while maintaining the advantages of attention mechanisms, achieving an F1-score of 0.893 in cloud detection tasks. However, the enormous parameter count (356M) limits practical application [25]. While these hybrid models can significantly improve segmentation performance in complex scenes, they still face issues with inflated parameter sizes and training costs.

Based on the above research, we find that the cloud-snow segmentation task has achieved good accuracy, but multiple states of cloud are similar to the texture and color attributes of snow, and it is challenging to achieve higher accuracy segmentation. Secondly, the cloud and snow boundary information is still not well restored, and the blurring of boundary information is the pain point of existing research, which has great limitations in fine application scenarios. In order to solve these problems, this paper proposes a Multi-scale Feature Mixed Attention Network (MFMANet). The main contributions are as follows:

- Design of the Multi-scale Pooling Feature Perception (MPFP): This module integrates multi-scale strip pooling operations with a self-attention mechanism to enhance global context modeling. By capturing dependencies and structural characteristics of cloud and snow across varying scales, it improve the identification accuracy and reduces misclassification.
- Proposal of the Bilateral Feature Mixed Attention Module (BFMA): This module combines spatial and channel attention to address the irregular morphology of cloud and snow. In order to preserve the edge features and avoid the loss of edge features of target classification caused by direct global pooling, Global Feature Channel Attention (GFCA) sets two branches to use global average pooling and global maximum pooling respectively, which can extract richer global features, and the Multi-Branch Spatial Aggregation (MBSA) refines boundary details through multi-kernel convolutions. This dual attention framework significantly enhances segmentation accuracy in regions with spectral confusion and complex spatial distributions.
- Develop the Multi-scale Feature Convolution Fusion Module (MFCF): To mitigate edge blurring caused by scale mismatches, this module employs directional strip convolutions (e.g., 1×7, 7×1) to fuse multi-scale features. By leveraging elongated kernels aligned with cloud/snow textures, it effectively restores fine-grained boundary

details and improves segmentation robustness, thus getting superior performance in recovering tortuous and irregular cloud/snow edges compared to conventional square convolutions.

2. Methodology

The purpose of semantic segmentation is to assign a distinct class label to each pixel in an image. However, cloud and snow remote sensing images often show characteristics such as high background complexity, small inter-class variance, and large intra-class variance, posing great challenges for segmentation tasks. The primary difficulty in semantic segmentation of cloud and snow lies in their diverse and complex forms. The segmentation boundaries of cloud and snow are usually tortuous and irregular, presenting various shapes, colors, and textures, thereby making it hard for precise segmentation. Therefore, deep neural network models designed for cloud and snow segmentation should not only accurately identify these morphologically diverse subjects but also possess strong edge perception capabilities. This enhanced capability is quite essential for better handling of complex boundaries, preventing false positives or omissions, and ultimately achieving more accurate segmentation outcomes.

2.1. Framework

To address the challenges mentioned above, this chapter introduces a novel Multi-scale Feature Mixed Attention Network (MFMANet), as illustrated in Figure 1, which enhances the accuracy of semantic segmentation for cloud and snow remote sensing images by integrating the advantages of multiple network sub-modules.

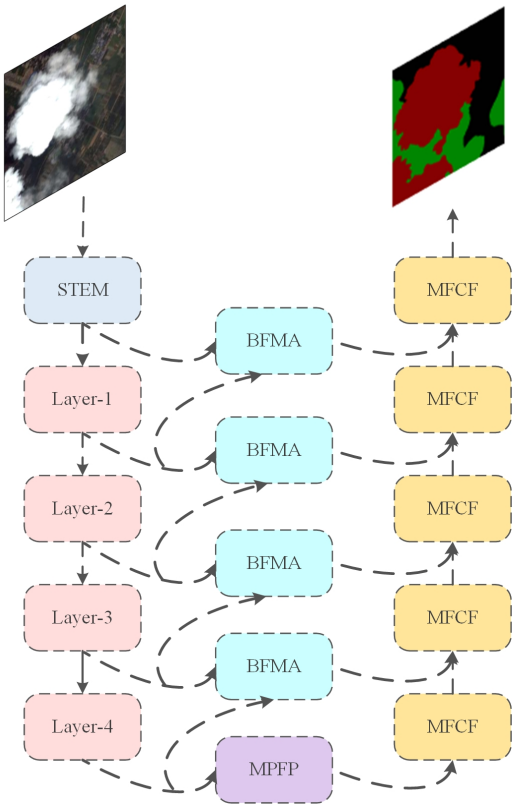

Figure 1. The overall structure of the Multi-scale Feature Mixed Attention Network (MFMANet).

Firstly, MFMANet adopts ResNet50 as the backbone network to extract multi-level deep features, enabling effective capture of the complex structures of cloud and snow across various scales. Secondly, to further enhance the model's capability in context

modeling, a Multi-scale Pooling Feature Perception (MPFP) is designed. By employing multi-scale pooling operations, MPFP aids the model in better understanding the structural characteristics of cloud and snow at different scales. Subsequently, a Bilateral Feature Mixed Attention (BFMA) is introduced. This module aggregates spatial and channel-wise feature information, allowing for a detailed capture of boundary characteristics of cloud and snow regions from multiple perspectives, thereby further improving segmentation accuracy. Lastly, to recover edge information of cloud and snow and enhance the detail representation of segmentation results, a Multi-scale Feature Convolution Fusion (MFCF) is designed. Through multi-scale convolution operations, MFCF fuses features from various levels, effectively mitigating problems related to edge blurring caused by scale differences.

By combining these network sub-modules, MFMANet can more accurately restore complex boundary information and precisely segment morphologically diverse cloud and snow subjects in segmentation tasks. The following sections will provide a detailed explanation of the design principles of each specific module within the MFMANet model, highlighting its advantages in the semantic segmentation task of cloud and snow.

2.2. Backbone

In our model, we have adopted ResNet50 as the backbone network. ResNet50 is widely used in the field of computer vision as a backbone network due to its numerous advantages. Firstly, compared to earlier networks like VGG, ResNet50 features a deeper architecture capable of extracting richer features. Moreover, it mitigates the vanishing gradient problem encountered during the training of deep networks through residual connections. Additionally, while offering lower computational costs than deeper architectures such as ResNet101 and ResNet152, ResNet50 strikes an excellent balance between computational efficiency and performance, making it particularly suitable for resource-constrained environments.

Furthermore, ResNet50 benefits from pretraining on large-scale datasets, making its pretrained weights with robust transfer learning capabilities. Many advanced object detection and semantic segmentation networks often choose ResNet50 as their foundational backbone network, making use of its efficient feature extraction capabilities. Therefore, ResNet50 not only ensures strong feature learning ability but also takes into account computational efficiency, feasibility, and transferability. Considering factors such as computational power and memory, we eventually opted for ResNet50 as our backbone network.

The expression for a residual block can be formulated as Equations (1) and (2):

$$x_i' = \mathrm{Conv}_{3\times3}(\sigma\mathrm{Conv}_{1\times1}(x_i)) \tag{1}$$

$$x_{i+1} = x_i + \mathrm{ReLU}\big(\mathrm{Conv}_{1\times1}\{\sigma(x_i')\}\big) \tag{2}$$

where x_i is the input to the residual block, x_i' is the output of the intermediate block, and x_{i+1} is the output of the residual block. Conv_k denotes convolution operations with a kernel size of k, and σ represents the BatchNorm normalization algorithm followed by the ReLU activation function.

We made two improvements on the backbone network: first, we introduced dilated convolutions in the L3 and L4 layers of the backbone network, which are designed to increase the receptive field for better capture of global semantic information, extract multi-scale features while retaining spatial resolution, and reduce computational cost to some extent. Second, we removed the last average pooling layer and fully connected layer of ResNet. The structure of the backbone network is shown in Table 1.

Table 1. Backbone network structure.

Layer	Structure	Feature Map Size
Stem	7×7 conv, stride 2	1/2
3×3 Max pool	stride 2	1/4
L1	$\begin{cases} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 3 \times 3, 64 \end{cases} \times 3$	1/4
L2	$\begin{cases} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 3 \times 3, 512 \end{cases} \times 4$	1/8
L3 (Dilated conv)	$\begin{cases} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 3 \times 3, 1024 \end{cases} \times 6$	1/16
L4 (Dilated conv)	$\begin{cases} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 3 \times 3, 2048 \end{cases} \times 3$	1/32

2.3. Multi-Scale Pooling Feature Perception Module (MPFP)

In the architecture of deep neural networks, shallower layers tend to extract local, low-level edge and texture features, whereas deeper layers are capable of capturing abstract and more discriminative semantic features. For tasks such as cloud and snow segmentation, effectively utilizing these deep semantic features is very important. Long-range dependencies within images play a significant role in accurate segmentation. Traditional convolutional neural networks struggle with learning correlations between long-range features due to their reliance on limited receptive fields.

To address this challenge, the self-attention mechanism has been introduced [26]. This mechanism exhibits position-invariance in the field of visual image processing, enabling the capture of dependencies between distant pixels, thus emphasizing the recognition of global features. Moreover, the global dependencies provided by the self-attention mechanism effectively expand the receptive field, aiding in the detection of thin cloud more efficiently.

Based on this principle, we propose Multi-scale Pooling Feature Perception (MPFP), and embed it in the deepest layer of the backbone network. The design of the MPFP module aims at effectively perceiving global features, thereby guiding the network to perform cloud and snow target segmentation more accurately. The structure is shown in Figure 2.

As shown in Figure 2, we use multiple groups of 1×N kernel-sized (e.g., N = 1, 3, 5, 7) horizontal and vertical stripe average pooling layers to process the input high-level semantic information. Horizontal stripe convolutions are dedicated to learning some horizontal features of cloud and snow in images, while vertical stripe convolutions focus on capturing longitudinal features of cloud and snow. The use of multiple groups of stripe average pooling facilitates better multi-scale global information correlation.

In addition to these, we employ a group of global average pooling (GAP) and global max pooling (GMP) to extract global features. After obtaining multiple sets of feature maps through pooling operations, they are concatenated and then passed through a 1×1 convolution layer to adjust the number of channels, which is subsequently divided into two branches. One branch serves as the self-attention branch aimed at fully exploring global contextual information, while the other branch undergoes a 3×3 convolution to enhance local feature information.

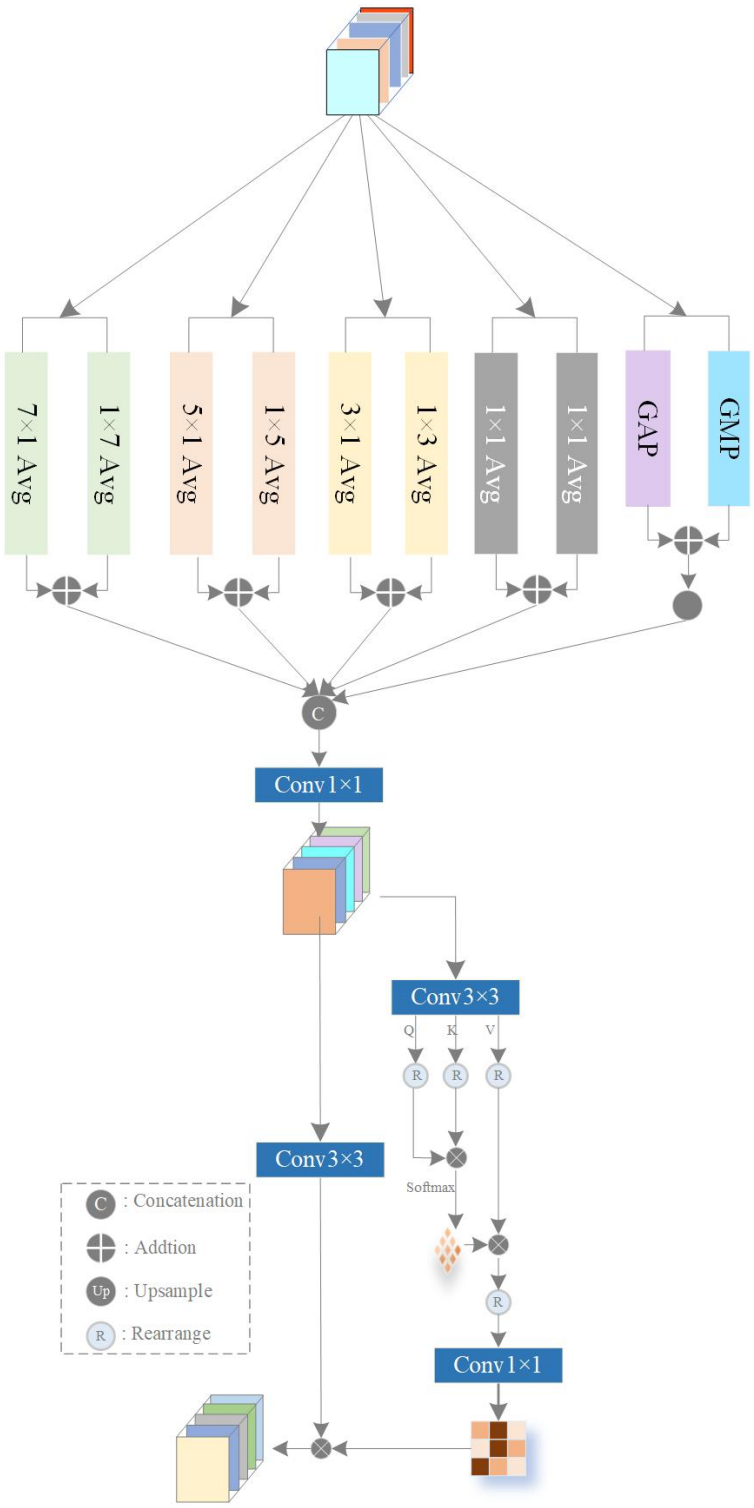

Figure 2. The structure of the Multi-scale Pooling Feature Perception (MPFP).

In the self-attention branch, depthwise separable convolutions with a 3 × 3 kernel size are used to map the input feature maps into three distinct representation vectors (query vector Q, key vector K, value vector V). Following this, Q and K are rearranged, and their dot product interaction generates a transposed attention map. This attention map is then dotted with the rearranged V to form a weight map for each pixel. Finally, the weight map from the attention branch is multiplied with the feature map from the other branch to produce the output of the MPFP module.

2.4. Bilateral Feature Mixed Attention Module (BFMA)

Within the realm of computer vision, attention mechanisms have garnered increasingly widespread application. Channel attention and spatial attention are two critical branches of attention mechanisms. The core idea is to dynamically adjust the model's focus on various parts of the input data, thereby enhancing its feature extraction capability.

Channel attention primarily captures dependencies between channels through global pooling, enhancing the model's sensitivity to significant channels. A prominent example is the Squeeze-and-Excitation Network (SENet) [27], which boosts model performance by compressing channel dimensions and reallocating weights. Channel attention has shown remarkable effectiveness in image classification and object recognition tasks, aiding models in better capturing features relevant to target categories.

Spatial attention focuses on the spatial information within feature maps, typically using local pooling to capture the importance rate of different regions within an image. This mechanism enhances target regions while suppressing non-target areas, thus improving the model's sensitivity to classification features. In semantic segmentation tasks, introducing spatial attention helps models more accurately locate target classifications and significantly enhance boundary features of objects, thereby increasing segmentation accuracy. Recently, scholars have increasingly combined spatial and channel attention, creating hybrid mechanisms that leverage the strengths of both for feature extraction, effectively capturing both global and local information from feature maps. A notable example is the Convolutional Block Attention Module (CBAM) [28], which uses convolution to acquire channel and spatial information from feature maps to improve object recognition capabilities. However, CBAM's reliance on convolution limits its ability to fully understand global contextual features due to the inherently local nature of convolution operations.

Given the distinct physical properties and uncertain morphologies of cloud and snow, cloud and snow segmentation networks must not only correctly identify cloud and snow bodies but also precisely recover boundary information. To address these challenges, we designed the Bilateral Feature Mixed Attention (BFMA), comprising the Global Feature Channel Attention (GFCA) and the Multi-Branch Spatial Aggregation (MBSA), as illustrated in Figure 3.

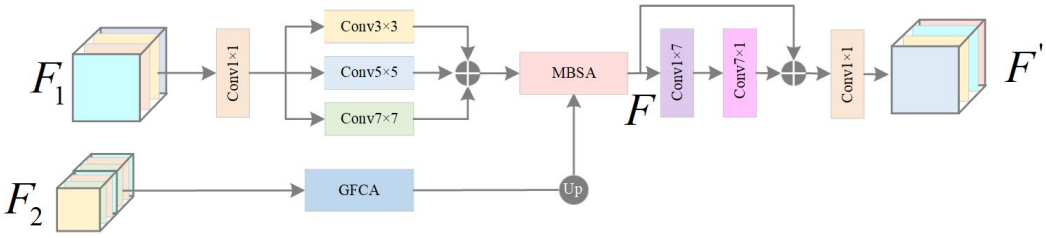

Figure 3. The structure of the Bilateral Feature Mixed Attention (BFMA).

In Figure 3, the inputs to the BFMA module are high-level features F_2 and low-level features F_1. For the low-level feature branch, we enhance the features using a set of convolution kernels with sizes 1×1, 3×3, 5×5, and 7×7. These enhanced features are then fed into the MBSA module as one of its inputs. Unlike traditional single-size convolution kernels, employing a diverse set of kernel sizes avoids information loss caused by a single convolution kernel and effectively extracts multi-scale features. For the high-level feature branch, given its complex channel information, these features are processed through the GFCA module. This process enhances significant channels while suppressing others, reducing the interference from noisy features. The output is then upsampled on the channel axis to match the scale required for the MBSA module's input. Through the computations performed in the MBSA module, the features in the target regions of

the feature maps are amplified, while irrelevant region features are suppressed, thereby enhancing the model's capability to identify cloud and snow bodies. Finally, to further refine the boundary information of cloud and snow, a set of stripe convolutions with sizes 1×7 and 7×1 are used to extract boundary features, improving the model's ability to recognize edge details. The resulting features undergo channel adjustment via a 1×1 convolution operation to produce the final output feature map F'.

2.4.1. GFCA Module

Inspired by channel attention mechanisms such as SENet, FCANet [29], and ECANet [30], we designed the GFCA within the BFMA module, as shown in Figure 4. To preserve edge features and avoid the loss of boundary features caused by direct global pooling, the GFCA module is structured with two branches. These branches respectively use global average pooling and global max pooling to extract global features from the feature maps, enabling the capture of richer global information.

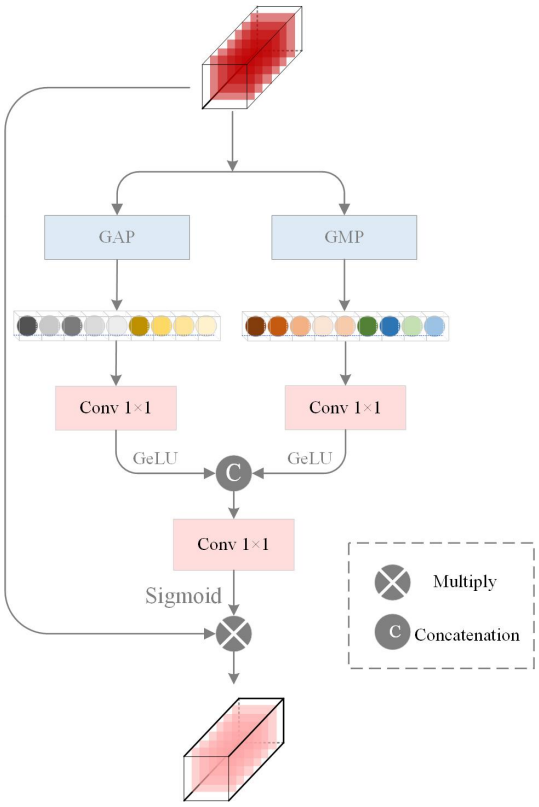

Figure 4. The structure of the Global Feature Channel Attention (GFCA) module.

Then each branch processes the features through a 1×1 convolution to alter the channel dimensions, enhancing interactions between different channels. The results from the two branches are then activated using the Gaussian Error Linear Unit (GeLU) activation function before being concatenated. Next, a 1×1 convolution is applied to adaptively focus on the target classification regions, capturing cloud and snow feature information while suppressing noise from non-target areas.

Finally, the output is weighted by applying the Sigmoid activation function to the original features, producing the final output features. The computation process can be summarized by the following formulas:

$$x_{\max} = \mu(\mathrm{Conv}(g_{\max}(x_i))) \tag{3}$$

$$x_{\text{avg}} = \mu(\text{Conv}(g_{\text{avg}}(x_i))) \tag{4}$$

$$x_{i+1} = x_i + \sigma(x_{\max} + x_{\text{avg}}) \tag{5}$$

where $\mathcal{G}_{\max}$ and $\mathcal{G}_{\text{avg}}$ respectively represent the global max pooling operation and the global average pooling operation. Conv denotes the 1×1 convolution operation. Cat represents the concatenation operation, and σ indicates the non-linear activation function Sigmoid.

2.4.2. MBSA Module

The detection and recognition of cloud and snow bodies are very important in cloud and snow segmentation. The introduction of spatial attention can better achieve target body discrimination, reducing missed and incorrect predictions. As shown in Figure 5, we designed a MBSA within the BFMA, aiming to better learn and identify the main features of cloud and snow.

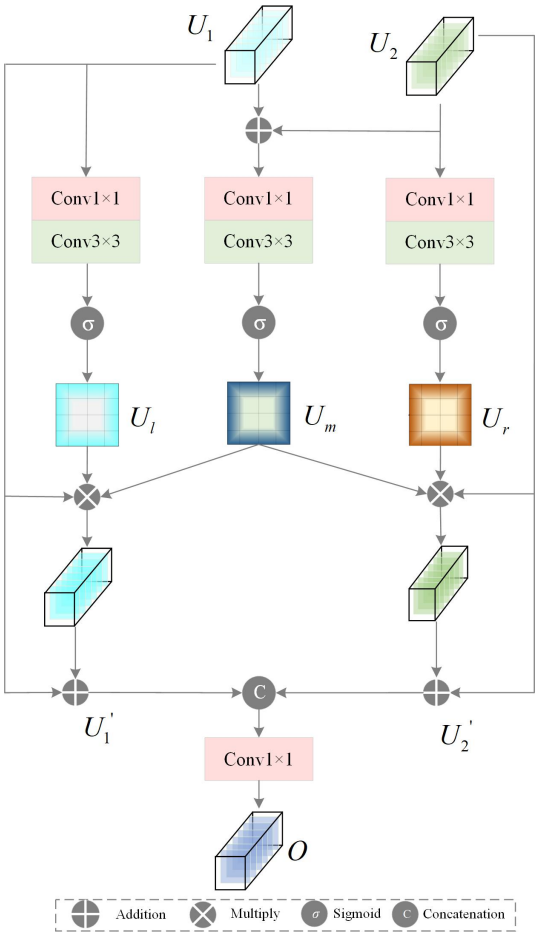

Figure 5. The structure of the Multi-Branch Spatial Aggregation (MBSA) module.

The MBSA takes as input the upsampled features U_1 from the channel attention module and the enhanced low-level features U_2. The MBSA is roughly divided into three branches: two input features each form one branch, and the sum of the two input features forms the intermediate branch. Each branch adaptively enhances the spatial features of the feature maps through a 1×1 convolution, followed by a 3×3 convolution to extract contextual information, capturing the main features of cloud and snow. This enhances the network's ability to recognize cloud and snow, effectively reducing misclassification and missed detection rates.

The enhanced features from each branch are passed through the Sigmoid activation function to obtain three spatial feature weights: U_l, U_m, and U_r. The spatial feature weight U_m of the intermediate branch is then pixel-wise multiplied with the spatial feature weights U_l and U_r of the two input branches. The resulting weight matrices are element-wise multiplied with the original input features to produce feature maps U_1' and U_2' with cloud and snow target attention attributes.

Inspired by residual connections, we concatenate the two feature maps and add them to the original feature map. The result is adjusted via a 1×1 convolution to align the channels, producing the output O of the MBSA module. The mathematical expressions involved in this module are as follows:

$$U_l = \sigma(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(U_1))) \tag{6}$$

$$U_r = \sigma(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(U_2))) \tag{7}$$

$$U_m = \sigma(\text{Conv}_{3\times3}(\text{Conv}_{1\times1}(U_3))) \tag{8}$$

$$U_f = U_l + U_r + U_m \tag{9}$$

$$U_{\text{out}} = \text{ReLU}(\text{Conv}_{1\times1}(U_f)) \tag{10}$$

$$X_{i+1} = X_i + U_{\text{out}} \tag{11}$$

where σ represents the non-linear activation function sigmoid, Cat denotes the concatenation operation, and $\times$ represents element-wise multiplication. $\text{Conv}_{1\times1}$ denotes the convolution operation with a 1×1 kernel, and $\text{Conv}_{3\times3}$ denotes the convolution operation with a 3×3 kernel.

2.5. Multi-Scale Feature Convolution Fusion Module (MFCF)

After extracting the aforementioned multi-scale features, addressing the fusion of features across different scales is crucial for achieving accurate cloud and snow segmentation. Simple methods such as broadcasting mechanisms or straightforward concatenation and combination of features at different scales struggle to recover the edge details of cloud and snow. Therefore, to effectively integrate information from different scales, we propose the MFCF.

As shown in Figure 6, the inputs to the module are two feature maps at different scales, U_1 and U_2. For the U_1 branch, after upsampling, the feature map is split into two sub-branches for convolution operations. One sub-branch consists of stripe convolutions with kernel sizes 1×3 and 3×1, while the other sub-branch consists of stripe convolutions with kernel sizes 1×5 and 5×1. The results from these two sub-branches are then added element-wise.

The U_2 branch is similar, but since the input feature map has a larger scale, the kernel sizes of the stripe convolutions are set to 5 and 7, respectively, producing U_1' and U_2'. After upsampling U_1', it is concatenated with U_2'. The concatenated features are then split and passed through a 1×1 convolution followed by the Sigmoid activation function. The outputs are multiplied element-wise, and after a series of convolution operations, the final output feature map O is obtained.

To prevent the addition of this module from significantly increasing the network's parameters, we adopt depthwise separable convolutions within the module, following the approach of Chollet [31], to reduce the number of parameters.

It is worth mentioning that, considering the characteristics of the cloud and snow segmentation task, in addition to achieving precise segmentation and classification of the two main recognition targets (cloud and snow), accurately identifying the edges of cloud

and snow is also crucial. Better recognition and learning of edge features could significantly contribute to the accuracy of cloud and snow segmentation. Through observations of large cloud and snow datasets, we found that the edges of both cloud and snow tend to be narrow and elongated, with texture distributions with certain directional patterns.

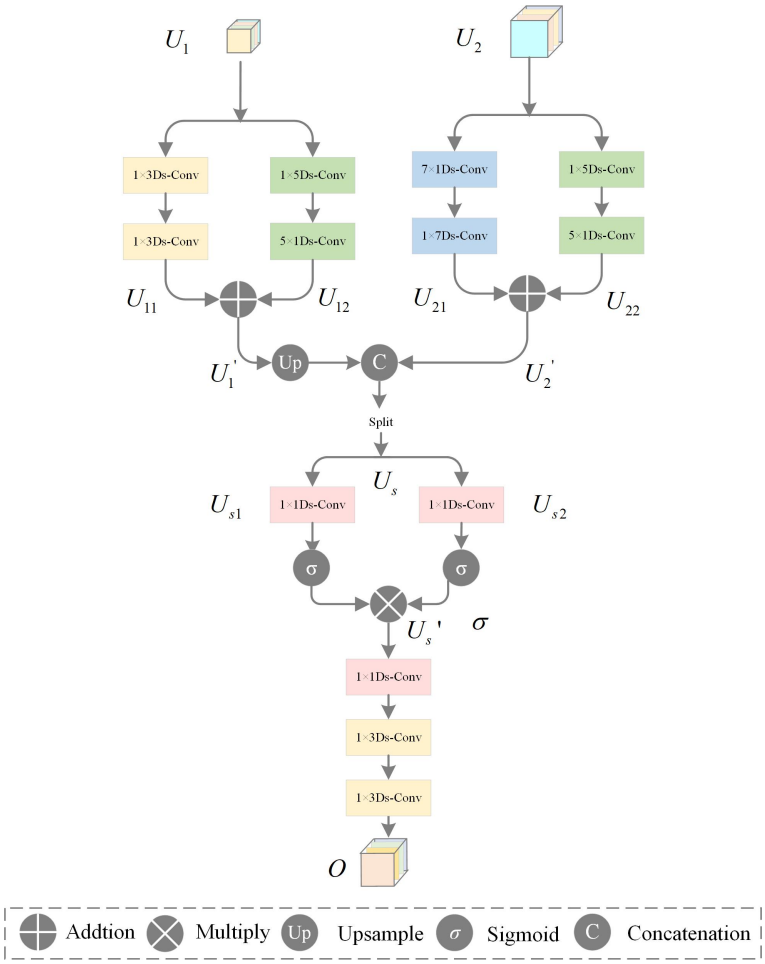

Figure 6. The structure of the Multi-scale Feature Convolution Fusion (MFCF).

Considering the features mentioned above, in the MFCF module, we extensively utilize stripe convolutions. These stripe convolutions can extract features along the directional patterns of these shapes, effectively recovering texture and edge details of cloud and snow in specific directions. Compared to traditional square convolution kernels, which may introduce irrelevant information that interferes with cloud and snow segmentation, stripe convolutions are more targeted in extracting boundary features. This approach focuses more on extracting and fusing cloud and snow features, thereby enhancing segmentation accuracy.

Furthermore, stripe convolutions can process features at multiple scales by combining kernels of varying lengths and orientations, enabling effective fusion of multi-scale features. The corresponding expressions are as follows:

$$U_{11} = \mathrm{Conv}_{3\times1}(\mathrm{Conv}_{1\times3}(U_1)) \tag{12}$$

$$U_{12} = \mathrm{Conv}_{5\times1}(\mathrm{Conv}_{1\times5}(U_1)) \tag{13}$$

$$U_{13} = \mathrm{Conv}_{7\times1}(\mathrm{Conv}_{1\times7}(U_1)) \tag{14}$$

$$U_{1f} = U_{11} + U_{12} + U_{13} \tag{15}$$

$$U_{1\text{out}} = \text{ReLU}(\text{Conv}_{1\times1}(U_{1f})) \tag{16}$$

$$X_{i+1} = X_i + U_{1\text{out}} \tag{17}$$

Among them, U_1 represents high-level features, U_2 represents low-level features, *Conv* denotes depthwise separable convolution operations, and the subscripts indicate the size of the convolution kernels. Up denotes upsampling operations, σ represents the sigmoid activation function, and U_s' is the element-wise product result after the convolution operations of branches U_{s1} and U_{s2}.

3. Experiments

3.1. Dataset

Deep learning is a data-driven approach that relies on training with large datasets. The dataset forms the foundation for completing deep learning experiments. In this study of cloud and snow segmentation, we experiment on remote sensing images from the CSWV and HRC_WHU datasets.

3.1.1. The CSWV Dataset

The CSWV dataset was constructed from WorldView-2 satellite imagery acquired between 2014–2016 over the North American Cordillera region. It offers a benchmark for cloud-snow segmentation research. Comprising 27 multispectral images spanning resolutions from 0.5 to 10 m, the dataset includes diverse environments such as glaciers, forests, towns, and deserts. After preprocessing (cropping and resizing to 256 × 256 pixels), it provided 9594 samples addressing challenges posed by spectral ambiguities between cloud (e.g., cumulus, cirrus) and snow (permanent, unstable, discontinuous), as well as melting-phase surface conditions. The dataset is also divided into two subsets—CSWV_S6 (6 sub-meter resolution images) and CSWV_M21 (21-m interpolated images)—to study resolution-dependent feature identifiability. The pixel-level annotations are used to classify pixels into three categories: cloud (purple), snow (white), and background (black, encompassing vegetation and bare ground). Some samples are shown in Figure 7.

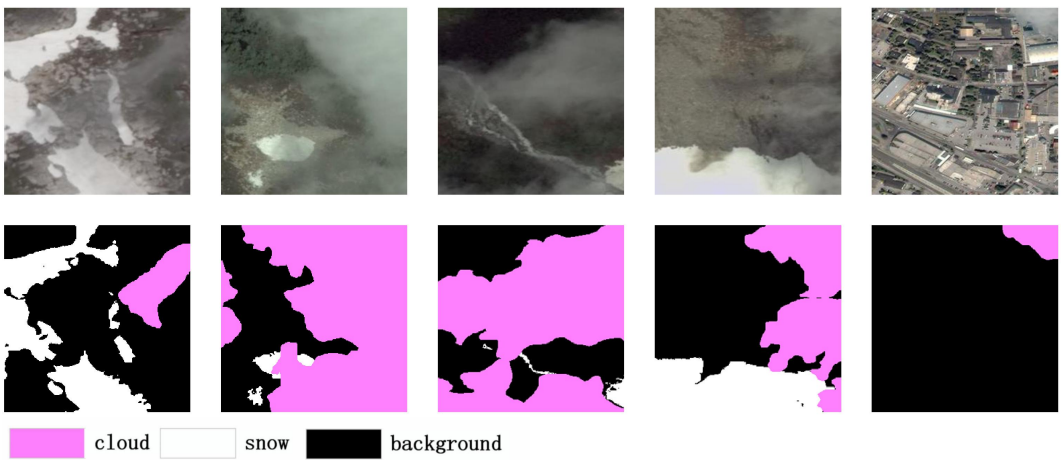

Figure 7. Sample images and their labels from the CSWV dataset.

3.1.2. The HRC_WHU Dataset

The HRC_WHU dataset has made great contributions to cloud detection research and the academic community, serving as a high-resolution public cloud detection dataset

that meets the training and testing needs of deep learning models. It includes 150 high-resolution images with RGB channels, featuring resolutions ranging from 0.5 to 15 m, covering different regions globally. These images are sourced from Google Earth, integrating satellite imagery, aerial photography, and Geographic Information System (GIS) data. The associated reference cloud masks were digitized by experts in remote sensing image interpretation at Wuhan University. Using the HRC_WHU dataset could facilitate performance benchmarking for deep learning models in image classification tasks. In Figure 8, we display some images from the dataset. The first row shows the original color images, while the second row presents the corresponding classification label maps for the remote sensing images. In these label maps, white areas represent snow cover, and black areas denote the background, as shown in Figure 8.

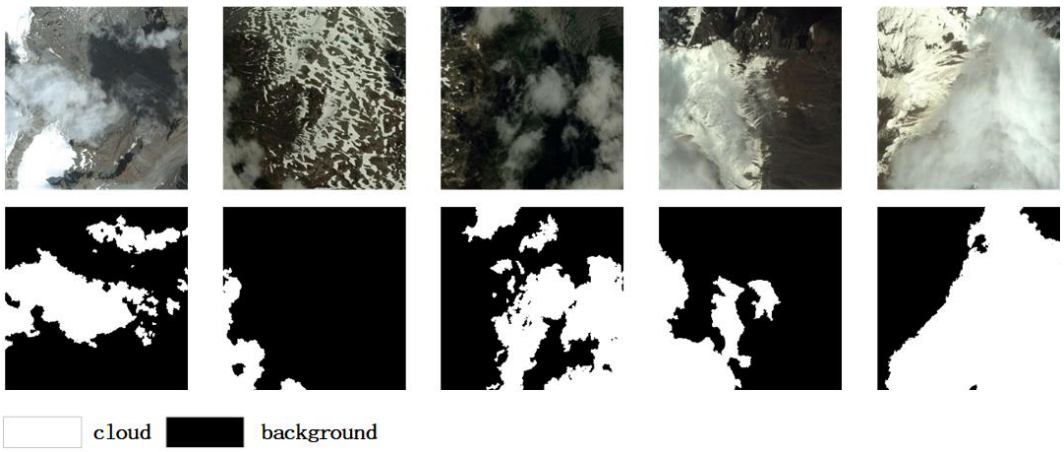

Figure 8. Sample images and their labels from the HRC_WHU dataset.

3.2. Experimental Parameter Setting

All experiments in this paper were conducted using the PyTorch 2.3. The computing environment was set up on a computer with the Windows 10 operating system, equipped with an Intel Core i5-12400F CPU and an Nvidia GeForce RTX 4070 Ti GPU. The optimizer we use is Adaptive Moment Estimation (Adam), and the learning rate scheduling strategy follows the "ploy" method, which is defined by the following formula:

$$lr = lr_{base} \times \left(1 - \frac{e}{e_m}\right)^p \tag{18}$$

In this equation, lr is the updated learning rate; lr_{base} is the base learning rate; e is the current iteration count; e_m is the maximum number of iterations; and p controls the shape of the curve. In all experiments in this work, lr_{base} was set to 0.001, e_m was set to 250, and p was set to 0.9. This is the optimal combination obtained after we try a variety of hyper-parameter combinations such as lr of 0.0005 and 0.0015, em of 200 and 300. During the training process, we did not use pre-trained parameters. Due to computational power and memory size limitations, in the cloud and snow semantic segmentation task, the batch size was set to 16. To prevent model overfitting, we adopted multiple methods including data augmentation, dropout, and normalization. The loss function used in the cloud and snow semantic segmentation study was Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss).

3.3. Metrics

Our experiments adopted a $K \times K$ confusion matrix to evaluate pixel-level classification performance in semantic segmentation tasks. As shown in Table 2, in the cloud and

snow semantic segmentation task, TP (True Positive) represents samples where both the true class and the model's prediction are cloud (or snow), FN (False Negative) represents samples where the true class is cloud (or snow) but the model's prediction is incorrect, FP (False Positive) represents samples where the true class is not cloud (or snow) but the model's prediction is cloud (or snow), and TN (True Negative) represents samples where both the true class and the model's prediction are not cloud (or snow).

Table 2. Confusion matrix structure.

Confusion Matrix		**Predicted Value**	
		Positive	**Negative**
True Value	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

We use Precision (P), Recall (R), F1 Score, Pixel Accuracy (PA), Mean Pixel Accuracy (MPA), Intersection over Union (IoU), and Mean Intersection over Union ($MIoU$) as the evaluation metrics for this experiment. Their mathematical formulas are as follows:

$$P = \frac{TP}{TP + FP} \tag{19}$$

$$R = \frac{TP}{TP + FN} \tag{20}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{21}$$

$$PA = \frac{\sum_{i=0}^{k} \rho_{i,j}}{\sum_{i=0}^{k} \sum_{j=0}^{k} \rho_{i,j}} \tag{22}$$

$$MPA = \frac{1}{k} \sum_{i=0}^{k} \frac{\rho_{i,j}}{\sum_{j=0}^{k} \rho_{i,j}} \tag{23}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{24}$$

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\rho_{i,j}}{\sum_{j=0}^{k} \rho_{i,j} + \sum_{j=0}^{k} \rho_{j,i} - \rho_{i,i}} \tag{25}$$

Here, k represents the object segmentation categories (excluding background), $\rho_{i,i}$ represents the true class, and $\rho_{i,j}$ represents the number of pixels belonging to class i but predicted as class j.

3.4. Ablation Studies

To quantify the contribution of each module in the model, this section conducts ablation studies by integrating each of the modules designed in this chapter (MPFP, BFMA, MFCF) into the backbone network one by one, with experiments performed on the CSWV dataset. As shown in Table 3, we use the MIoU value as the evaluation metric to assess the effectiveness of each module. According to the ablation study results, it is evident that all the proposed modules are effective and contribute to achieving precise cloud and snow segmentation.

Table 3. Ablation study of MFMANet modules.

Method	MIoU (%)
Backbone	87.69
Backbone + MPFP	88.32 (0.63 ↑)
Backbone + MPFP + BFMA	88.78 (0.46 ↑)
Backbone + MPFP + BFMA + MFCF	**89.17** (0.39 ↑)

Bold indicates the best result, ↑ indicates that the accuracy improved.

In ablation studies, heatmaps are often introduced to visualize the effects of integrating various modules. We selected two high-resolution remote sensing images from the CSWV dataset and applied our network to cloud and snow recognition under both thick cloud and thin cloud meteorological conditions, then utilized the best parameters obtained from model training to generate heatmaps for different module combinations. As shown in Figure 9, each remote sensing image corresponds to two rows of heatmaps: the first row represents the attention heatmap for cloud, and the second row represents the attention heatmap for snow. The heatmaps visually display feature importance, with regions of higher model attention appearing in red, followed by yellow-green, and then blue. The intensity of the red color indicates the level of attention given to key areas. It is evident from the figures that as modules are progressively integrated, the performance of our cloud and snow classification network improves. After each module is added, the red areas in the heatmaps become darker, the yellow-green areas gradually turn blue, and the boundaries between cloud, snow, and other surface features become clearer. This verifies that the modules we designed enhance the network's focus on cloud and snow, achieving better cloud and snow segmentation.

Notably, when only the MFPF module is added, the main bodies of large-area cloud and snow can be well identified and focused on, but there are still issues of misclassification and missed detection. With the addition of BFMA, which introduces spatial attention and channel attention, it is clear that the model's ability to recognize the main bodies of cloud and snow significantly improves, and the boundaries of cloud and snow become more distinct. After integrating the MFCF module into the model, the boundaries of cloud and snow become even clearer. Meanwhile, from the heatmaps, we can identify the shortcomings of our model in cloud and snow segmentation. Under thin cloud meteorological conditions, there are still some inaccuracies in determining cloud and snow boundaries, and similar regions are prone to confusion.

- MPFP Module: According to the experimental results, after integrating the MPFP module into the deepest layer of the model, the segmentation metric MIoU improved by 0.63%. The experimental results demonstrate that the effective extraction of semantic information from the deepest layer guided the network in making a preliminary judgment on cloud and snow, as can be seen from the heatmap (C). However, as the deep semantic information is mixed with a significant amount of noise interference, there are some misclassifications and missed detections of cloud and snow. More modules need to be introduced into the network to improve the discrimination of details;
- BFMA Module: We introduced Multi-Branch Spatial Aggregation (MBSA) and Global Feature Channel Attention (GFCA) through the BFMA module. After integrating the BFMA module, the MIoU improved by 0.46%. From the heatmaps, it is evident that from (C) to (B), the misclassifications and missed detections of the main cloud and snow areas were significantly reduced, and the interference from noisy elements was suppressed. This allows for better detection of cloud and snow, effectively verifying that this module can fuse bilateral features and focus better on cloud and snow targets in the spatial domain;

- MFCF Module: The experimental results intuitively show that after integrating the MFCF module into the network, the model's MIoU improved by 0.39%. It can be seen that compared to simple feature concatenation, MFCF can effectively fuse multi-scale features. From the heatmaps, it is also evident that from (B) to (A), with the addition of the MFCF module, the boundaries of cloud and snow became clearer and more distinct. This effectively restores the edge details of cloud and snow, verifying that strip convolution is highly sensitive to linear and edge features and can effectively extract the boundary features of cloud and snow, resulting in more accurate segmentation. Therefore, adding the MFCF module is effective for the cloud and snow segmentation task. It not only fuses multi-scale features effectively but also restores the edge information of classification targets well.

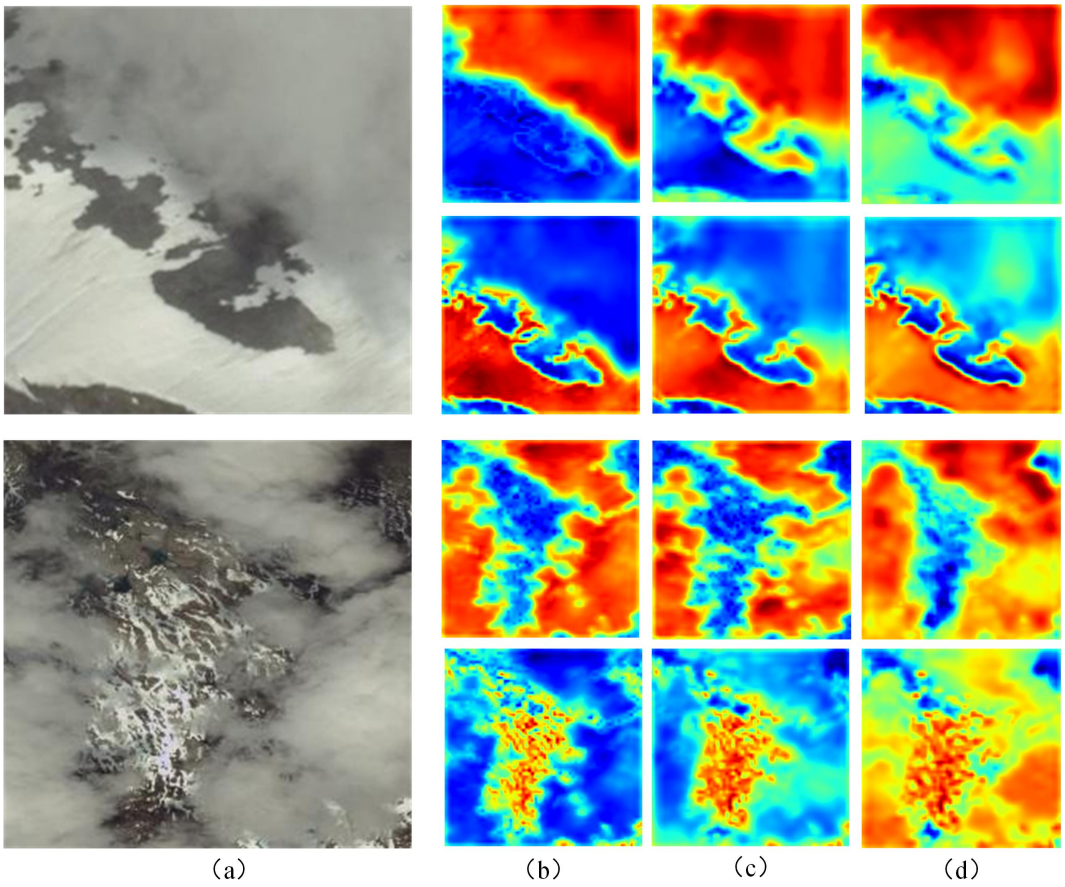

Figure 9. Heatmaps of cloud and snow under different module combinations. (**a**) Test image; (**b**) MPFP + BFMA + MFCF; (**c**) MPFP + BFMA; (**d**) MPFP.

3.5. Comparative Experiments

Through ablation experiments, we have verified the effectiveness of each module in our model. Next, maintaining consistency in experimental settings, we selected several classical semantic segmentation deep learning models for comparative experiments on the CSWV dataset. The evaluation metrics used include Pixel Accuracy (PA), F1 Score, Mean Pixel Accuracy (MPA), and Mean Intersection over Union (MIoU). The compared models include ACFNet [32], BiSeNetV2 [33], CvT, CSDNet [34], SegNet [35], HRNet [36], DeepLabV3Plus [37], DABNet [38], DFN [39], FCN8s, U-Net, PSPNet [40], PAN [41], among others. We evaluated the performance of these models using both visual segmentation results and quantitative metrics, accordingly gave the parameters (Params), computational complexity(FLOPs) to consider our model.

Table 4 shows the segmentation metrics for all models on the CSWV dataset. Our MFMANet achieved the best results across all four evaluation metrics. To visually compare the segmentation performance, we selected seven high-resolution cloud and snow images from the test set. These images represent various forms of cloud and snow. Figure 10 presents the segmentation results of each model on these seven test images. Overall, our network accurately identifies the main bodies of cloud and snow with low misclassification rates. In detail, our model effectively recovers the boundaries of cloud and snow. Compared to other models, which often misclassify shallow snow and cloud and are less sensitive to boundary information, our MFMANet demonstrates superior performance, confirming its effectiveness and broad application prospects in cloud and snow segmentation.

Table 4. Evaluation results of different models on CSWV dataset.

Method	PA (%)	MPA (%)	F1 (%)	MIoU (%)	Params (M)	FLOPs (G)
CvT	89.92	88.47	88.27	78.87	0.82	5.73
DeepLabV3Plus	91.47	90.36	89.93	81.93	7.83	6.03
HRNet	91.63	90.58	90.51	82.82	65.85	23.31
SegNet	91.76	91.41	90.85	83.32	29.48	42.52
BiSeNetV2	91.88	91.53	90.93	84.83	3.62	3.20
DABNet	91.91	91.68	90.77	85.47	0.752	1.27
FCN8s	92.07	91.72	91.55	86.55	18.64	20.06
ACFNet	92.54	92.27	91.73	86.53	89.97	99.32
PSPNet	92.71	92.57	91.86	87.46	49.07	46.07
PAN	92.82	93.93	92.67	87.67	23.65	5.37
DFN	93.52	93.86	92.79	87.49	42.53	10.51
CSDNet	93.66	93.33	93.28	88.28	8.66	21.90
UNet	93.98	94.35	93.61	88.44	13.41	30.94
MFMANet	95.12	94.69	94.34	89.17	25.39	34.78

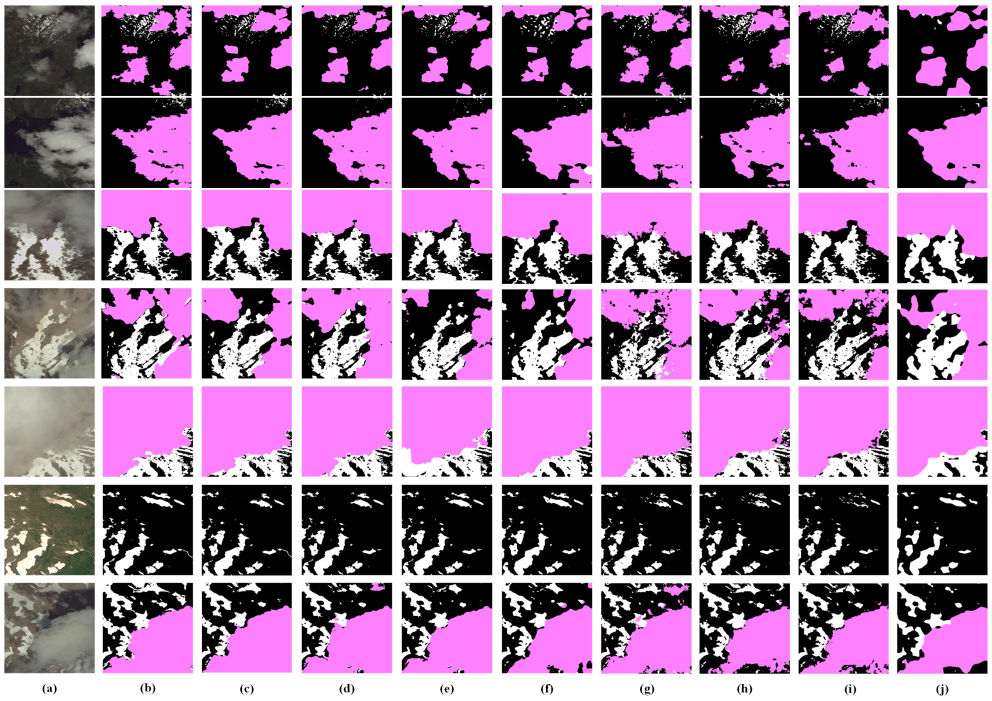

Figure 10. Visualization of segmentation performance on the CSWV dataset. (**a**) Test image; (**b**) Ground truth; (**c**) MFMANet; (**d**) U-Net; (**e**) CSDNet; (**f**) DFNNet; (**g**) PAN; (**h**) PSPNet; (**i**) ACFNet; (**j**) DeepLabV3+. In this figure, pink represents cloud, black represents background, and white represents snow.

Our model has achieved very good experimental results on the CSWV dataset, especially with good detail segmentation ability for cloud/snow images. In order to demonstrate the advantages of our algorithm more clearly, as shown in the Figure 11, it is evident that our model exhibits strong perception ability for cloud boundaries and can effectively segment both in areas where cloud and snow coexist.

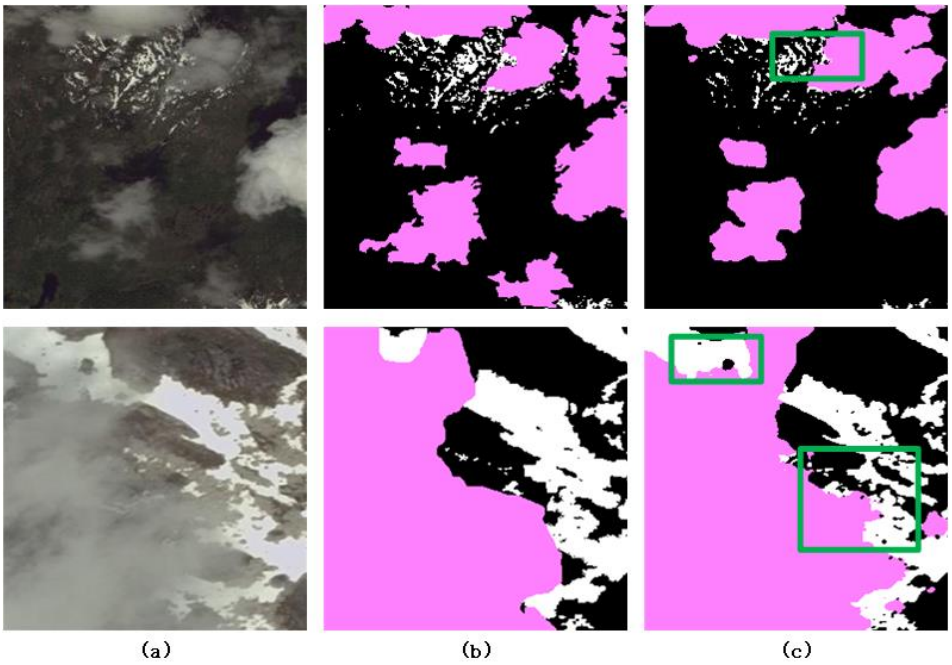

Figure 11. Detailed experimental results. (**a**) Test image; (**b**) Ground truth; (**c**) MFMANet; The area in the green rectangle represents that MFMANet has good segmentation ability for cloud/snow details and overlapping areas. In this figure, pink represents cloud, black represents background, and white represents snow.

3.6. Generalization Experiments

The generalization experiment aims to evaluate the ability of our proposed MFMANet to generalize to unseen datasets, assessing the model's robustness and transferability while also exploring its application potential. In this section, we conducted experiments using the publicly available HRC_WHU dataset from Wuhan University, validating the superiority of our network in cloud and snow segmentation through both visual comparisons of segmentation results and evaluation metrics. During the experiment, we adhered to the principle of controlling variables, maintaining consistent experimental environments and parameter settings. As shown in Table 5 our model achieved 95.22%, 94.88%, 94.27%, and 91.03% on the four key metrics: PA, MPA, F1 Score, and MIoU, respectively. All these results represent the best performance among all compared models.

Additionally, we know that the HRC_WHU dataset contains various backgrounds, including snow, buildings, water bodies, and vegetation. Considering the specific focus of our cloud and snow segmentation task, we selected five cloud images with snow as the background from the test dataset for comparative experiments. As shown in Figure 12, our model demonstrated wonderful segmentation performance on the test images. In contrast, visualization results from other networks, such as DeepLabV3+ and DFN, exhibited noticeable misclassifications for some pixels. On the other hand, our MFMANet achieved low misclassification rates, accurately identifying cloud layers and effectively recovering boundaries. Therefore, we conclude that MFMANet demonstrates strong robustness in cloud and snow segmentation tasks and possesses certain transferability.

Table 5. Evaluation results of different models on HRC_WHU dataset.

Method	PA (%)	MPA (%)	F1 (%)	MIoU (%)
CvT	93.50	93.36	92.27	87.47
FCN8s	94.27	94.21	93.53	89.83
HRNet	94.29	94.93	93.02	88.50
UNet	94.34	94.32	93.47	88.96
PAN	94.37	94.23	93.12	88.72
BiSeNetV2	94.45	94.29	93.27	89.03
DABNet	94.47	94.31	93.36	89.17
DeepLabV3Plus	94.50	94.68	93.45	89.51
ACFNet	94.52	94.70	93.52	89.53
PSPNet	94.55	94.77	93.57	89.46
SegNet	94.59	94.93	93.59	89.67
DFN	94.72	94.86	93.63	90.19
CSDNet	94.86	95.03	93.80	90.54
ACFNet	95.09	94.91	93.76	90.59
MFMANet	95.22	94.88	94.27	91.03

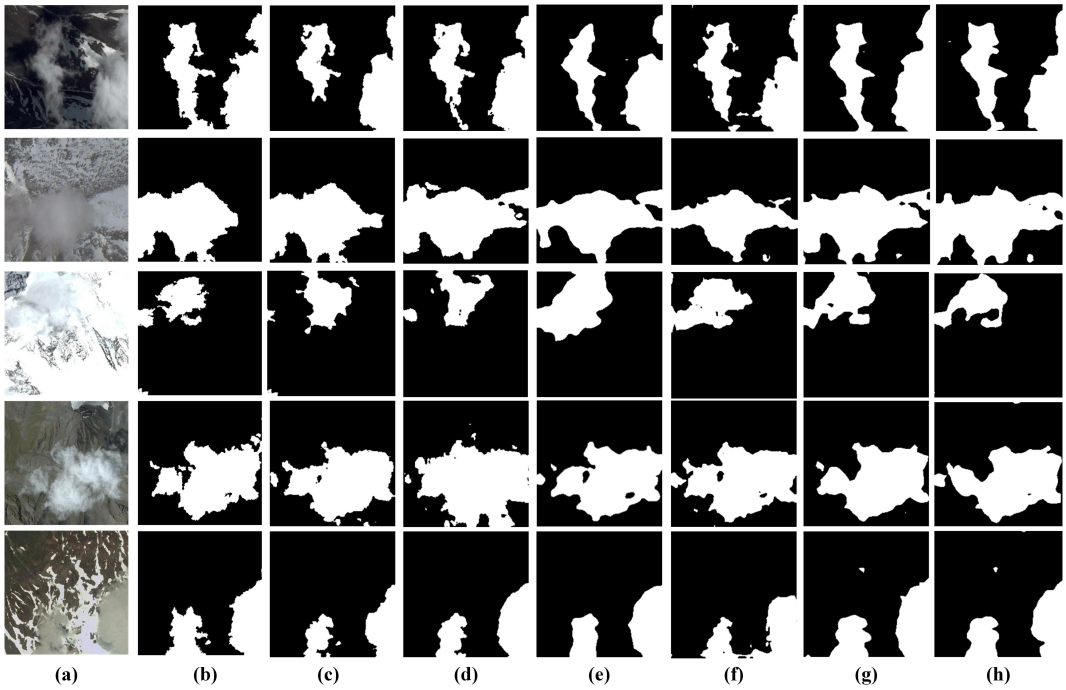

Figure 12. Comparison of segmentation results on the HRC_WHU dataset. (**a**) Test image; (**b**) Ground truth; (**c**) MFMANet; (**d**) OCRNet; (**e**) ACFNet; (**f**) PSPNet; (**g**) DeepLabV3+; (**h**) DFN. In this figure, black represents background, and white represents cloud.

4. Discussion

In addition to the above experiments, we have conducted additional validation using more complex and diverse imagery scenarios, including mountainous terrain, urban areas, bare soil, and vegetation. Shown as Figure 13, our experimental results demonstrate that the proposed algorithm maintains consistent performance in distinguishing cloud and snow, even when they exhibit similar spectral characteristics with urban environments.

Our work has achieved promising experimental results in the cloud and snow segmentation task. However, we must acknowledge there are still some limitations. First, cloud and snow exhibit highly similar morphologies and possess extremely complex and challenging boundary characteristics. Although our proposed MFMANet can effectively

distinguish large-scale cloud and snow regions, further improvements are needed to resolve small-scale cloud and thin snow layers.

Looking ahead, with the continuous advancement of radar remote sensing and optical remote sensing technologies, our method holds significant potential for future applications. In terms of data selection, given the spatial heterogeneity of snow, we can explore the use of remote sensing satellites or drone data with higher revisit frequencies and spatial resolutions to acquire more precise cloud and snow images. Additionally, we can integrate multi-source auxiliary data, such as temperature and elevation information. Regarding model development, we can experiment with more mainstream deep learning architectures or attention mechanisms, continuing to explore network structures tailored for cloud and snow segmentation, enabling more accurate segmentation under more complex meteorological conditions.

Figure 13. Cloud and snow segmentation validation using more complex and diverse imagery scenarios. In this figure, pink represents cloud, black represents background, and white represents snow.

5. Conclusions

This paper addresses the segmentation problem of coexisting cloud and snow targets in remote sensing images using deep learning approach. We propose a Multi-scale Feature Mixed Attention Network (MFMANet) for cloud/snow segmentation tasks, which could not only accurately identify the main bodies of cloud and snow but also effectively captures their edge details, achieving an MIoU of 89.17% on the CSWV dataset. The various network sub-modules proposed in this paper, such as MPFP, BFMA, and MFCF, could enable more targeted attention to cloud and snow regions, facilitating precise delineation of snow-covered areas. The method proposed in this paper can also provide cloud removal solutions for preparing large-scale, high-resolution image datasets for snow depth estimation in high-latitude cold regions around the world. This capability inspires further research into snow depth estimation, which is very crucial for water resource management and sustainable development in the future.

Author Contributions: Conceptualization, L.Z. and J.C.; Data curation, J.C and F.S.; Formal analysis, L.Z.; Funding acquisition, L.Z.; Investigation, J.C.; Methodology, L.Z. and J.C.; Software, J.C. and Z.L.; Supervision, L.Z.; Validation, L.Z. and J.C.; Writing—original draft, J.C. and Z.L.; Writing—review and editing, L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Shanghai Typhoon Research Foundation from Shanghai Typhoon Institute of China Meteorological Administration: TFJJ202208; National Natural Science Foundation of China: 42075130.

Data Availability Statement: The data are available from the corresponding author upon request.

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

References

1. Lu, X.; Wang, X.; Xie, G.; Cui, C.; Wang, J. Snow depth retrieval based on passive microwave remote sensing. *J. Arid Land Resour. Environ.* **2012**, *26*, 108–112. (In Chinese)
2. Li, H.; Xiao, P.; Feng, X.; Lin, J.; Wang, Z.; Man, W. Snow depth inversion method based on repeat-track InSAR. *J. Glaciol. Geocryol.* **2014**, *36*, 517–526. (In Chinese)
3. Braaten, J.D.; Cohen, W.B.; Yang, Z. Automated cloud and cloud shadow identification in Landsat MSS imagery for temperate ecosystems. *Remote Sens. Environ.* **2015**, *169*, 128–138. [CrossRef]
4. Li, Z.; Shen, H.; Li, H.; Xia, G.; Gamba, P.; Zhang, L. Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery. *Remote Sens. Environ.* **2017**, *191*, 342–358. [CrossRef]
5. Tapakis, R.; Charalambides, A.G. Equipment and methodologies for cloud detection and classification: A review. *Sol. Energy* **2013**, *95*, 392–430. [CrossRef]
6. Zhu, Z.; Wang, S.; Woodcock, C.E. Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel-2 images. *Remote Sens. Environ.* **2015**, *159*, 269–277. [CrossRef]
7. Jiang, B.; Li, X.; Chong, H.; Wu, Y.; Li, Y.; Jia, J.; Wang, S.; Wang, J.; Chen, X. A deep-learning reconstruction method for remote sensing images with large thick cloud cover. *Int. J. Appl. Earth Obs. Geoinf.* 2022, **115**, 103079. [CrossRef]
8. Zou, X.C.; Li, K.; Xing, J.L.; Zhang, Y.; Wang, S.Y.; Jin, L.; Tao, P. DiffCR: A fast conditional diffusion framework for cloud removal from optical satellite images. *IEEE Trans. Geosci. Remote Sens.* **2024**, **62**, 5612014. [CrossRef]
9. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [CrossRef] [PubMed]
10. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
11. Zhou, Z.H.; Feng, J. Deep forest. *Nat. Sci. Rev.* **2019**, *6*, 74–86. [CrossRef] [PubMed]
12. Dai, X.; Xia, M.; Weng, L.; Hu, K.; Lin, H.; Qian, M. Multiscale location attention network for building and water segmentation of remote sensing image. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–19. [CrossRef]
13. Chen, K.; Xia, M.; Lin, H.; Qian, M. Multiscale attention feature aggregation network for cloud and cloud shadow segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
14. Song, L.; Xia, M.; Weng, L.; Lin, H.; Qian, M.; Chen, B. Axial cross attention meets CNN: Bibranch fusion network for change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *16*, 21–32. [CrossRef]
15. Yin, H.; Weng, L.; Li, Y.; Xia, M.; Hu, K.; Lin, H.; Qian, M. Attention-guided siamese networks for change detection in high resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103206. [CrossRef]
16. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA; 7–12 June 2015, pp. 3431–3440.
17. 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, Proceedings, Part III 18*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
18. 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]
19. Du, H.; Li, K.; Guo, J.; Zhang, J.; Yang, J. Cloud and snow detection from remote sensing imagery based on convolutional neural network. In *Optoelectronic Imaging and Multimedia Technology VI*; SPIE: Bellingham, WA, USA, 2019; Volume 11187, pp. 260–266.
20. Li, Y.; Chen, W.; Zhang, Y.; Tao, C.; Xiao, R.; Tan, Y. Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning. *Remote Sens. Environ.* **2020**, *250*, 112045. [CrossRef]
21. Fang, Z.; Ji, W.; Wang, X.; Li, L.; Li, Y. Automatic cloud and snow detection for GF-1 and PRSS-1 remote sensing images. *J. Appl. Remote Sens.* **2021**, *15*, 024516. [CrossRef]
22. Wu, X.; Shi, Z.; Zou, Z. A geographic information-driven method and a new large scale dataset for remote sensing cloud/snow detection. *ISPRS J. Photogramm. Remote Sens.* **2021**, *174*, 87–104. [CrossRef]
23. 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.
24. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.

25. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. CVT: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 22–31.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
29. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 783–792.
30. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
31. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
32. Zhang, F.; Chen, Y.; Li, Z.; Hong, Z.; Liu, J.; Ma, F.; Han, J.; Ding, E. Acfnet: Attentional class feature network for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6798–6807.
33. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
34. Zhang, G.; Gao, X.; Yang, Y.; Wang, M.; Ran, S. Controllably deep supervision and multi-scale feature fusion network for cloud and snow detection based on medium-and high-resolution imagery dataset. *Remote Sens.* **2021**, *13*, 4805. [CrossRef]
35. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
36. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
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. Li, G.; Yun, I.; Kim, J.; Kim, J. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. *arXiv* **2019**, arXiv:1907.11357.
39. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1857–1866.
40. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
41. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.