International Journal of Geo-Information

MDPI

Article

Multi-Supervised Feature Fusion Attention Network for Clouds and Shadows Detection

Huiwen Ji [1], Min Xia [1,*], Dongsheng Zhang [1] and Haifeng Lin [2]

[1] Jiangsu Key Laboratory of Big Data Analysis Technology, CICAEET, Nanjing University of Information Science and Technology, Nanjing 210044, China; 20211222029@nuist.edu.cn (H.J.); 20201249164@nuist.edu (D.Z.)
[2] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China; haifeng.lin@njfu.edu.cn
* Correspondence: xiamin@nuist.edu.cn

Abstract: Cloud and cloud shadow detection are essential in remote sensing imagery applications. Few semantic segmentation models were designed specifically for clouds and their shadows. Based on the visual and distribution characteristics of clouds and their shadows in remote sensing imagery, this paper provides a multi-supervised feature fusion attention network. We design a multi-scale feature fusion block (FFB) for the problems caused by the complex distribution and irregular boundaries of clouds and shadows. The block consists of a fusion convolution block (FCB), a channel attention block (CAB), and a spatial attention block (SPA). By multi-scale convolution, FCB reduces excessive semantic differences between shallow and deep feature maps. CAB focuses on global and local features through multi-scale channel attention. Meanwhile, it fuses deep and shallow feature maps with non-linear weighting to optimize fusion performance. SPA focuses on task-relevant areas through spatial attention. With the three blocks above, FCB alleviates the difficulties of fusing multi-scale features. Additionally, it makes the network resistant to background interference while optimizing boundary detection. Our proposed model designs a class feature attention block (CFAB) to increase the robustness of cloud detection. The network achieves good performance on the self-made cloud and shadow dataset. This dataset is taken from Google Earth and contains remote sensing imagery from several satellites. The proposed model achieved a mean intersection over union (MIoU) of 94.10% on our dataset, which is 0.44% higher than the other models. Moreover, it shows high generalization capability due to its superior prediction results on HRC_WHU and SPARCS datasets.

Keywords: cloud and cloud shadow detection; convolutional neural networks; remote sensing; semantic segmentation

Citation: Ji, H.; Xia, M.; Zhang, D.; Lin, H. Multi-Supervised Feature Fusion Attention Network for Clouds and Shadows Detection. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 247. https://doi.org/10.3390/ijgi12060247

Academic Editors: Christos Chalkias, Marinos Kavouras, Margarita Kokla, Mara Nikolaidou and Wolfgang Kainz

Received: 17 April 2023
Revised: 5 June 2023
Accepted: 16 June 2023
Published: 18 June 2023

1. Introduction

Detection of clouds and their shadows is essential in preprocessing optical remote sensing images. That is because they may obscure the task-relevant objects, thus causing difficulties for various remote sensing applications, such as land cover detection [1], flood monitoring [2], and forest change analysis [3]. Therefore, it is crucial to segment clouds and their shadows before applying remote sensing imagery. Traditional algorithms are based on the spectral and geometric properties of clouds and shadows. A cloud has specific spectral properties in different bands so that they can be recognized by specific spectral thresholds [4,5]. Due to the spectral uncertainty of cloud shadows, geometric relationships are often utilized to locate the shadow corresponding to the cloud. Specifically, the relative position of the cloud and its shadow is calculated to recognize the corresponding shadow based on prior knowledge (the satellite sensor view angle, the satellite sensor altitude, the illumination angle, etc.) [6]. It helps to make up for the shortcomings of spectral analysis.

Zhu et al. suggested an algorithm termed function of mask (Fmask) for detecting clouds and their shadows in Landsat imagery [6]. The technique combines spectral thresh-

olds with cloud probability algorithms (including temperature, brightness, and spectral variability algorithms) to derive potential cloud layers. Then, it matches potential cloud shadow layers through geometric analysis to generate the final cloud and cloud shadow masks. After that, several improved versions were developed [7–10]. Li et al. offered an automatic multi-feature combined (MFC) approach to locate the cloud and its shadow in GF-1 WFV imagery. The method enhances the performance of cloud and shadow recognition based on their textural features [11]. In addition to the single-image method, the multi-temporal method was employed to recognize clouds and their shadows. Since they have significant spectral changes in time series, their characteristic information can be captured at various times to detect the cloud and its shadow [12–14].

The algorithms described above are computationally challenging and require a large amount of prior knowledge. Furthermore, they limit the range of applications to a large extent, as they are primarily designed for specific sensors. Increasing numbers of machine learning algorithms are employed in clouds and their shadow segmentation as a result of the advancement of machine learning. They can automatically learn the optimal model, saving time and reducing manpower costs. Traditional machine learning algorithms such as decision trees [15,16], random forests [17–19], and neural networks [20,21] were used to detect cloud and cloud shadow. However, these methods still rely on spectral analysis. They are scarcely applicable to images from muti-sensors since the multi-spectral bands of remote sensing imagery generally vary from satellite to satellite.

Deep learning, as a branch of machine learning, recently developed rapidly in the field of image processing. More and more semantic segmentation networks (such as SegNet [22], FCN [23], DeepLab [24], and UNet [25]) are used to locate clouds and shadows. These algorithms, based on convolutional neural networks (CNNs), can consider both spatial and spectral information. Furthermore, they have the advantages of simple data preprocessing, convenient operation, wide applicability, and more. Shendryk et al. applied CNN to detect clouds and their shadows in multi-sensor imagery, proving the potential of CNN [26]. Segal-Rozenhaimer et al. combined CNN and the domain adversarial neural network (DANN) to segment clouds and shadows in multi-sensor imagery [24]. They experimentally proved that DeepLab [27] had higher prediction accuracy than VGG-16 [28]. Wieland et al. proved the superior performance of UNet for segmenting clouds and their shadows on the SPARCS dataset [25].

To further improve segmentation performance, semantic segmentation networks, as a universal network model, can adjust the network structure according to the characteristics of the segmented objects. Some customized semantic segmentation networks were used in the field of remote sensing, such as change detection [29,30], land cover [31–33], and water body segmentation [34–36]. However, because most semantic segmentation networks are based on CNNs, they face challenges such as reduced feature resolution (caused by max pooling and down-sampling operations) and being easily trapped in local information (caused by convolutional kernels), making it difficult to segment boundaries and small-scale objects. Several improvement measures, such as dual-branch structure [37,38], multi-scale fusion [39], and attention mechanism [40–43], were proposed to address these challenges.

For now, few researchers designed semantic segmentation models specifically for clouds and cloud shadows based on their visual and distribution properties in the imagery. There are three characteristics of cloud and shadow images.

(1) Clouds and their shadows have complex and diverse boundaries, uneven distribution, and large-scale changes. As is mentioned above, although the general semantic segmentation models can extract rich semantic information, it is easy to overlook details and they are not friendly to the boundary and small-scale segmentation. Multi-level fusion mechanisms were used to overcome this shortcoming [44–46]. This enhances the performance of the model by fusing shallow feature maps with rich location information and deep feature maps with abundant semantic information.
(2) Cloud and cloud shadow images have complex and diverse backgrounds since they carry massive amounts of information about ground objects. The high-intraclass and

low-interclass variance make semantic segmentation more difficult. The problem of how to segment small-scale targets from interference deserves to be studied.

(3) Clouds have a unique visual appearance in imagery but can still be mistaken for other objects under some backgrounds. Cloud shadows are also difficult to identify under certain circumstances.

Based on the above characteristics of clouds and their shadows in remote sensing imagery, our model made the following three innovations:

(1) This paper designs a multi-scale feature fusion block. Unlike existing multi-scale feature fusion networks [23,47], which fuse features in a simple linear fashion, the block we built allows the network to automatically learn the parameters that aid in selecting appropriate feature representations from deep and shallow feature maps.

(2) This work develops a multi-scale channel attention mechanism to focus on local as well as global information, hence improving the expression of local features. It can assist us in more precisely identifying cloud shadow boundaries and detailed information. The background of remote sensing images is complex and diverse, and paying too much attention to local details may introduce interference. As a result, we also added a spatial attention mechanism to focus on task-related areas while limiting background interference.

(3) Some clouds and shadows are inconspicuous. Since they have unique properties, we provided a class feature attention mechanism to learn class features. It helps balance background interference and local representations.

2. Proposed Network

Based on the previous discussion, a semantic segmentation network is suggested to address the challenges of cloud and shadow detection. This section presents the architecture of the proposed network for the detection of clouds and shadows. Next, the sub-modules, including the fusion convolution blocks (FCB), channel attention blocks (CAB), spatial attention blocks (SPA), and category feature attention blocks (CFAB), will be discussed in detail.

2.1. Network Architecture

Our model is based on the encoder–decoder structure. The encoder conducts feature extraction, gradually reducing the resolution of features and enriching them with semantic information. We chose to use a residual network (ResNet) as the encoder since its residual blocks help resolve issues with gradient explosion and disappearance resulting from network deepening [48]. Many researchers improved ResNet with different versions [49–52]. Merely replacing ResNet with the improved versions in our model framework can improve the experimental results. The decoder gradually recovers lost spatial details from the encoder's features, which is critical for semantic segmentation. Our model focuses on designing the decoder to ensure that it recovers lost details and spatial information efficiently. As shown in Figure 1, the decoder in our model comprises four feature fusion blocks (FFBs) and one category feature attention block (CFAB). Each FFB consists of two fusion convolution blocks (FCBs), a channel attention block (CAB), and two spatial attention blocks (SPA). FFBs fuse low-level and high-level feature maps by extracting multi-scale features, thus recovering the lost details. FCBs reduce the significant semantics gap between shallow and deep feature maps. CAB learns the importance of each channel to achieve remarkable segmentation performance with a slight increase in computation. SPA reconstructs the spatial details to focus on the task-related areas. CFAB is introduced after the final FFB to enhance semantic information by learning class features. These blocks will be described in detail later.

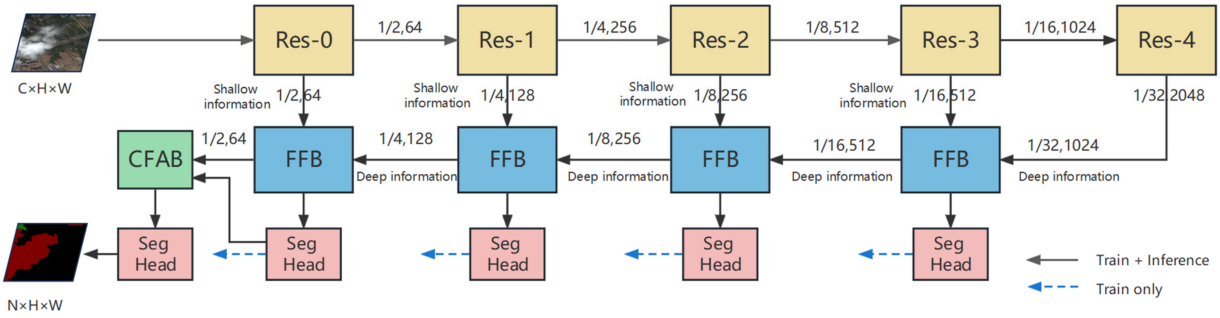

Figure 1. Framework of our proposed cloud and shadow detection network. Here, *C* represents the number of channels, N represents the number of classes, and *H*, W represent the height and width of the input images.

ResNet comprises five encoder stages, depicted in Figure 1. The initial stage, known as Res-0, performs a 7×7 convolution with a stride of 2. Subsequently, the feature maps are subsampled by maximum pooling, resulting in a $1/4$ size reduction prior to entering the remaining four stages. The four subsequent stages are constructed using bottlenecks and exhibit comparable architectures. The decoder receives the output features from the five encoder stages as input. To reduce computation, the decoder compresses the output channels of the latter four encoder stages. A 1×1 convolution is applied to halve the number of channels. Auxiliary training can enhance the predictive performance of the model and hasten the network convergence. Our model is equipped with auxiliary heads to supervise each decoder stage. The segmentation head (seg head) generates prediction maps during auxiliary training. The seg head utilized for both main and auxiliary training is made up of convolution and dropout layers. The decoder is augmented with these heads at each FFB for auxiliary training purposes.

2.2. Feature Fusion Block (FFB)

Feature fusion of different scales can improve the performance of semantic segmentation. Shallow feature maps have a high resolution with more location and detailed information. They pass fewer convolutions with more noise and less semantic information. Deep feature maps have significant semantic information, but they have a low resolution and poor detail perception. Efficiently fusing both is crucial when it comes to enhancing segmentation accuracy. The feature fusion block implements this function. Usually, the common feature fusion is achieved through simple linear operations such as summation or concatenation. However, merely adding or concatenating features can result in significant inconsistencies in scale and semantics and may limit the model's performance. The fusion convolution block in this paper processes the shallow and deep feature maps before they fuse. It reduces excessive semantic differences. Another measure to overcome this difficulty is to learn the importance of fusion features and then fuse features again through nonlinear weight. Some previous researchers used these methods, such as SKNet and ResNeSt. They learn fusion weights through the global channel attention mechanism [50,52]. However, these methods are suitable for global information but not for small targets. Our objectives are to detect both massive cumulus clouds and small discrete clouds. Since detecting targets of different scales is vital, this work offers a multi-scale channel attention mechanism that focuses on global and local information.

FFB comprises FCB, CAB, and SPA. Figure 2 illustrates the structure of the feature fusion block. The shallow and deep features extracted by the backbone network are used as inputs. The deep feature maps are first up-sampled. Then, the deep and shallow feature maps pass through FCB, respectively, and are fused for the first time. After that, the fusion feature maps pass through CAB, which will be introduced in the next section, to obtain fusion weights. Then, the shallow and deep features are weighted on the channel dimension, respectively. SPA, which will also be introduced in the following section, is

added to the weighted feature maps to provide better spatial information. Finally, the processed shallow and deep features are fused again.

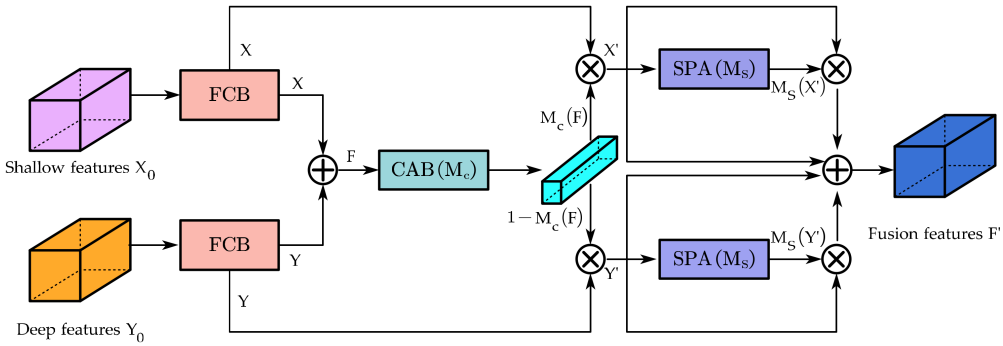

Figure 2. Feature fusion block. Here, FCB represents the fusion convolution block, CAB represents the channel attention block, and SPA represents the spatial attention block.

The shallow feature information $X_0 \in R^{C\times H\times W}$ and the deep feature information $Y_0 \in R^{C\times H\times W}$ are first processed by FCB and then element-wise added to obtain the fusion feature information $F \in R^{C\times H\times W}$, where C, H, and W represent the channel count, height, and width of the input maps. M_s and M_c refer to the spatial and channel attention operation, respectively. The operating procedure of FFB is as follows:

$$F = FCB(X_0) + FCB(Y_0) = X + Y \tag{1}$$

$$X' = M_c(F) \otimes X \tag{2}$$

$$Y' = (1 - M_c(F)) \otimes Y \tag{3}$$

$$F' = M_s(X') \otimes X' + M_s(Y') \otimes Y' + X' + Y'. \tag{4}$$

The symbols + and $\otimes$, respectively, denote the element-wise addition and multiplication in the given formulas. X_0 and Y_0 represent the original shallow and deep features, while X' and Y' denote the channel-wise weighted shallow and deep feature information. F' represents the final fusion feature information. The sigmoid activation function in M_c constrains the fusion weights $M_c(F)$ from 0 to 1. Subtracting the fusion weights from 1 aims to calculate the weighted average of X and Y, which is equivalent to a soft selection. The fusion weights are learned during training.

2.2.1. Fusion Convolution Block (FCB)

Large semantic differences between the two input feature maps significantly affect the fusion weights and limit the semantic segmentation performance. To address this issue, this paper proposes FCB. This block is made up of two multi-scale convolution blocks and is inspired by residual connections in the middle layer of Res2Net [49]. The features first pass through the first point-wise convolution layer and are split equally into six subsets along the channel dimension, as shown in Figure 3. The first point-wise convolution layer adjusts the channel number to multiples of 6 (C//6×6, where C represents the channel number). To reduce the number of model parameters, five of these sub-features undergo 1×3 and 3×1 convolution kernels instead of a single 3×3 convolution kernel. Then, the feature maps are reconstructed by concatenating the six subsets and passing a 1×1 convolution layer. Finally, the processed features and the initial features are element-wise added to form the output feature maps. This block achieves more fine-grained receptive fields at the convolution level.

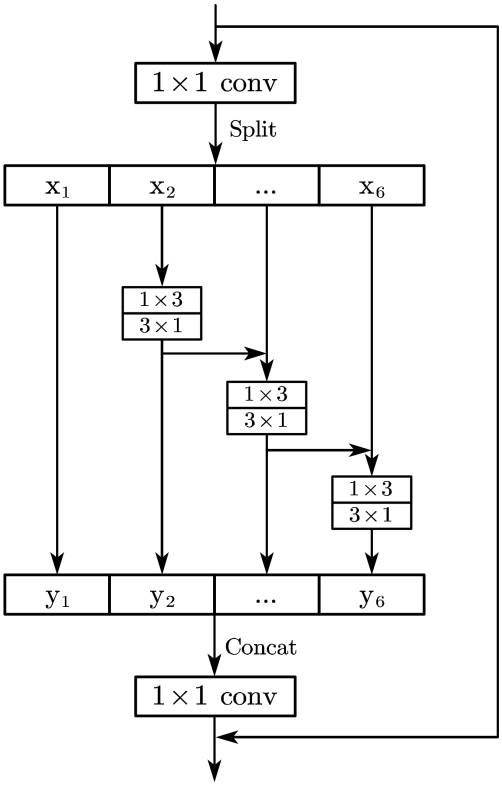

Figure 3. Multi-scale convolution block.

The feature subsets are denoted by x_i, where i ranges from 1 to 6. The number of channels per subset is one-sixth of the original. Except for x_1, each x_i goes through 1×3 and 3×1 convolution kernels, denoted by $C_i(\cdot)$. Thus, each output y_i can be written by the following formula:

$$y_i = \begin{cases} x_i & i = 1 \\ C_i(x_i) & i = 2 \\ C_i(x_i + y_{i-1}) & 2 < i \leqslant 6 \end{cases}. \tag{5}$$

2.2.2. Channel Attention Block (CAB)

CAB focuses on the information from the channel dimension. Each channel of the feature maps represents a feature, and CAB obtains the importance of each channel to assign a weight value. This enables CAB to enhance or suppress the corresponding feature by learning the importance of each channel. To obtain the importance of each channel, each feature map of size $H \times W$ is typically compressed into a scalar. However, this rough operation is only suitable for large targets, not for small ones. To address this issue, this paper proposes the improved channel attention block.

Figure 4 shows the structure of our CAB. The upper and lower branches, respectively, employ the global and local channel attention mechanisms. The dual-branch design mitigates issues associated with the spatial variability of clouds and their shadows, as it obtains channel attention at multiple scales. Subsequently, the outputs of the two branches are element-wise added and passed through a sigmoid nonlinear activation function to obtain channel weights.

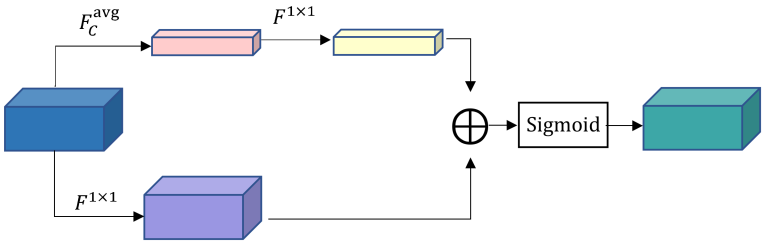

Figure 4. Channel attention block. Here, F_C^{avg} represents average pooling, and $F^{1\times1}$ represents a point-wise convolution layer.

The fusion feature information $F \in R^{C\times H\times W}$ is taken as the input of our CAB. The global attention branch includes average pooling F_C^{avg}, and a point-wise convolution layer $F^{1\times1}$. Pointwise convolution (1×1 convolution), rather than larger convolution kernels, is selected to reduce the computational complexity of our proposed module. The process of our proposed channel attention operation M_c is formulated as:

$$F_C^{avg}(F) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} F(i,j) \tag{6}$$

$$F^{1\times1}(F) = B(f_2^{1\times1}\left(Relu\left(B\left(f_1^{1\times1}(F)\right)\right)\right)) \tag{7}$$

$$M_c(F) = \sigma(F^{1\times1}\left(F_C^{avg}(F)\right) + F^{1\times1}(F)). \tag{8}$$

In the given formulas, $f^{1\times1}$ represents 1×1 convolution, B represents batch normalization, and Relu and σ represent the Relu and sigmoid activation functions; $f_1^{1\times1}$ reduces the number of input feature channels to $\frac{1}{r}$ of the original, while $f_2^{1\times1}$ restores the initial number of channels. Here, r represents the channel scaling ratio.

2.2.3. Spatial Attention Block (SPA)

Spatial information plays a vital role in image segmentation because even the slightest positional change can significantly impact the segmentation results. Therefore, it is necessary to incorporate spatial attention (SPA) to boost spatial information. Shallow features are rich with location information, while deep features contain effective semantic information. Passing both shallow and deep features through SPA enhances spatial information further. Average pooling is the most common way of aggregating spatial information. Alternatively, maximum pooling gathers unique spatial details from average pooling, enhancing the most representative spatial information. Previous research showed that combining both average and maximum pooling in spatial attention mechanisms leads to performance improvements [53].

As shown in Figure 5, the input feature maps $Z \in R^{C\times H\times W}$ undergo maximum and average pooling to obtain two one-channel features, $F_s^{avg}(Z) \in R^{1\times H\times W}$ and $F_s^{max}(Z) \in R^{1\times H\times W}$, respectively. Next, these two features are concatenated together. Finally, the stitched feature maps pass through a 7×7 convolution layer and a nonlinear activation function to obtain spatial weights. The process of the proposed spatial attention operation M_s can be represented as follows:

$$F_s^{avg}(Z) = \frac{1}{C}\sum_{k=1}^{C} Z(k) \tag{9}$$

$$F_s^{max}(Z) = \max(Z(k)),\ \text{for } k = 1,2,\ldots \tag{10}$$

$$M_s(Z) = \sigma\left(f^{7\times7}\left(concat\left(F_s^{avg}(Z), F_s^{max}(Z)\right)\right)\right). \tag{11}$$

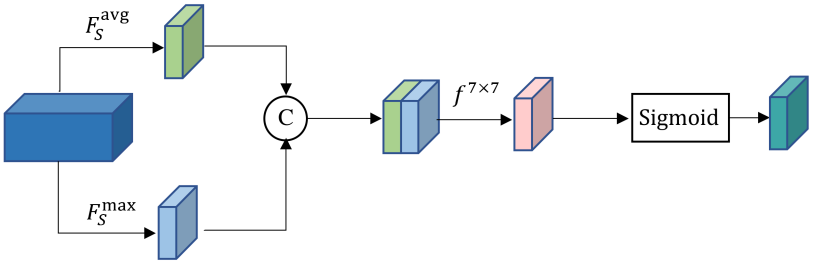

Figure 5. Spatial attention block. Here, F_s^{avg} represents average pooling, F_s^{max} represents maximum pooling, and $f^{7\times7}$ represents 7×7 convolution.

Here, $f^{7\times7}$ represents 7×7 convolution, *concat* represents concatenation along channels, and σ represents sigmoid activation function. Z could be X' or Y' as mentioned in the above section.

2.3. Category Feature Attention Block (CFAB)

Remote sensing images have complex geological backgrounds. The high-intraclass and low-interclass variance make cloud and cloud shadow segmentation difficult. Inspired by OCRNet [54], a category feature attention block (CFAB) is designed in the proposed model to learn class feature representations. Figure 6 illustrates the architecture of the CFAB, which accepts two inputs: the output feature maps from the last feature fusion block (FFB) in the decoder stage and the coarse semantic segmentation maps derived from it. By taking the product of these two inputs, the CFAB generates class feature maps with each column representing feature representations of a specific class. The coarse semantic segmentation maps are used directly as attention maps, and the class feature maps are multiplied with them to calculate class attention feature maps. Finally, these class attention feature maps are merged with the original feature maps as the final output.

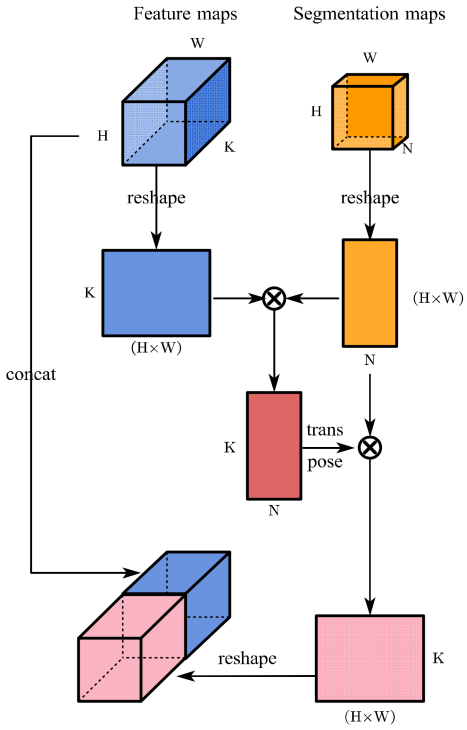

Figure 6. Category feature attention block. Here, N represents the number of classes, and K, H, and W represent the channel, height, and width of feature maps.

The feature maps $F \in R^{K\times H\times W}$, marked in blue in Figure 6, are reshaped into a two-dimensional vector $F \in R^{K\times H*W}$. Each column of F represents the K features of a given pixel. Similarly, the coarse semantic segmentation maps $S \in R^{N\times H\times W}$, marked in orange, are reshaped into a two-dimensional vector $S \in R^{H\times W\times N}$. Each row of S represents the class probability of a given pixel. The class feature $F_c \in R^{K\times N}$, marked in red, is calculated by the matrix multiplication formula $F_c = F \times S$. Each column of F_c represents the K features of a given class. Specifically, it consolidates the feature representations of all pixels that are classified as a given class. Pixels that have high probability values in S are more likely to be accurate, so they should have more significant contributions while computing class features. Therefore, the class attention feature $F_a \in R^{K\times H\times W}$, marked in pink in Figure 6, is applied. It is calculated by the matrix multiplication formula $F_a = F_c \times S^T$. In this way, the CFAB can create a mapping of features from pixels to categories and back to pixels, and it can extract the feature information of pixels that share the same category more effectively.

To confirm the efficacy of CFAB, the heatmaps of output features without and with CFAB are shown in Figure 7. The heatmaps denote high-weight areas in red and low-weight areas in blue. Figure 7 depicts that the network with CFAB exhibits higher weights at cloud and shadow boundaries. In particular, for thin clouds in the white box and cloud shadows that tend to be confused with the background in the red box, the CFAB can allocate more attention to them. Therefore, the model with CFAB is more accurate and stable in predicting clouds and cloud shadows.

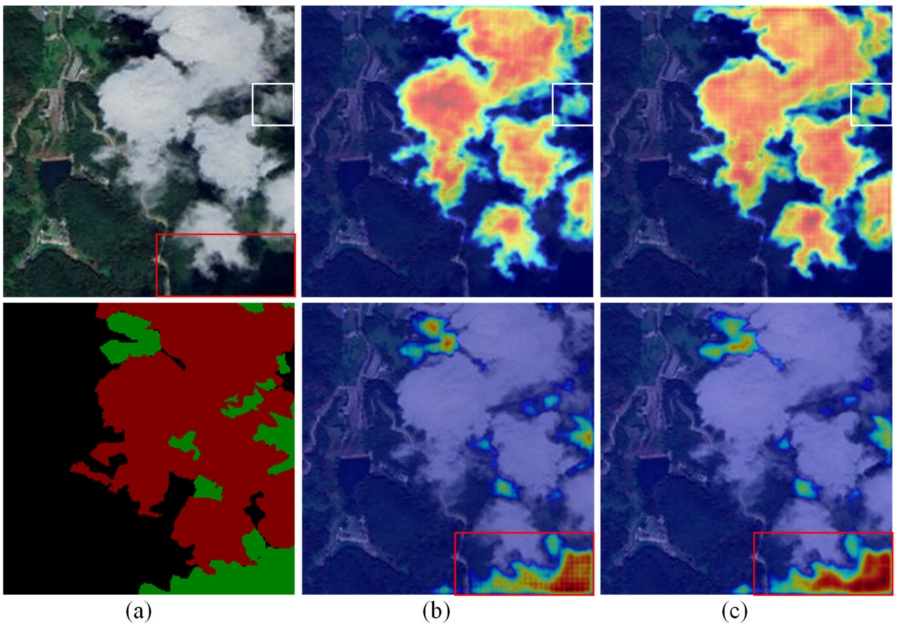

Figure 7. (**a**) Test image and its label; (**b**) heatmaps without CFAB; and (**c**) heatmaps with CFAB. In (**b**,**c**), the top row shows the heatmaps focusing on clouds, while the bottom row shows the heatmaps focusing on cloud shadows. The white and red rectangles highlight the regions where CFAB has made a considerable impact on the model's performance.

3. Experiments and Result Analysis

3.1. Experimental Datasets

The dataset used in this study for clouds and cloud shadows was derived from high-resolution remote sensing imagery available on Google Earth. The majority of the images in this dataset were obtained from QuickBird and WorldView-4 satellites. QuickBird satellite images contain a panchromatic band (450–900 nm), a blue band (450–520 nm), a green band (520–600 nm), a red band (630–690 nm), and a near-infrared band (760–900 nm). The WorldView-4 sensor captures the remote sensing images in specific bands, including a panchromatic band (450–800 nm), a blue band (450–510 nm), a green band (510–580 nm), a red band

(655–690 nm), and a near-infrared band (780–920 nm). To ensure credible experimental results, several high-resolution images with complex and changeable backgrounds were selected as the original image data. These images, distributed in the plains of China, have a resolution of 30 m. The chosen images contain various backgrounds, including urban areas, farmland areas, plant areas, water areas, and wasteland areas.

The original high-resolution cloud and shadow images were resized to 224 × 224 in size, considering the limitation of GPU memory. These images have three channels, RGB. A total of 1916 images were obtained by cropping and filtering, 1545 images among them were utilized for training, and 371 were used as the validation set. We manually labeled and classified each pixel of the cloud and shadow images into three categories: backgrounds, clouds, and shadows. The annotations were performed by a team of five trained annotators who are engaged in the field of remote sensing images. The annotators worked independently. In cases where there was disagreement, the annotators discussed the examples until a consensus was reached. The decision process was based on full agreement, meaning that all five annotators had to agree on the classification of a feature before it was included in the dataset. Figure 8 illustrates several representative images and their corresponding labels with various backgrounds, where clouds, shadows, and backgrounds are labeled in red, green, and black, respectively.

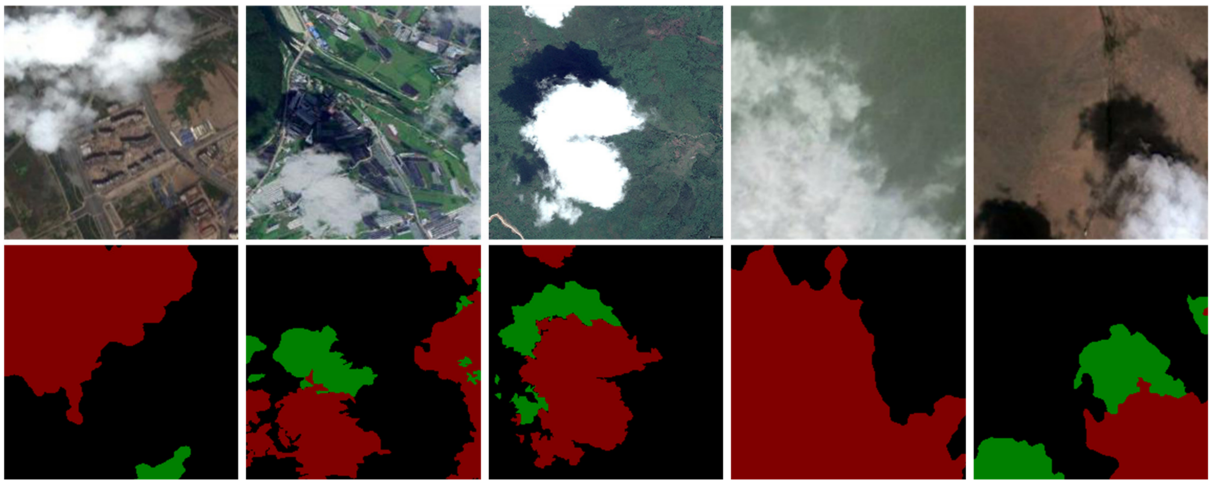

Figure 8. Some cloud images and their labels against different backgrounds. Clouds, cloud shadows, and backgrounds are marked red, green, and black, respectively. (**a**) urban areas; (**b**) farmland areas; (**c**) plant areas; (**d**) water areas; and (**e**) wasteland areas.

The generalization experiments were conducted on the public datasets, namely the HRC_WHU and SPARCS datasets, to prove the generalizability and validity of our model. The lab from Wuhan University offered the HRC WHU dataset [55]. The dataset contains 150 high-resolution images from around the world, with resolutions ranging from 0.5 to 15 m. The images come from Google Earth, and only clouds were labelled. The backgrounds of these images are divided into five main types of land cover (i.e., snow, urban, wasteland, plant, and water). To preserve GPU memory, 150 high-resolution images from HRC_WHU were cropped into 3488 images of the size 3 × 224 × 224. The training and validation sets were divided into a 4/1 ratio, with 2791 and 697 images, respectively.

The SPARCS dataset was produced by a team from Oregon State University. The dataset contains 80 images of the size 1000 × 1000 pixels with 10 wavebands [56]. The images were collected from the Landsat8 satellite in 2013 and 2014. Clouds, cloud shadows, snow, water, and backgrounds were labelled. Due to memory constraints, 80 images of size 1000 × 1000 from SPARCS were resized into 256 × 256, and the red, green, and blue (RGB) channels were selected as the image channels. A total of 1280 images were obtained. The model is vulnerable to overfitting due to the limited training data. Thus, the SPARCS

dataset needs to be expanded. This paper used random rotation and vertical and horizontal flips to augment the original 1280 images to 3168 images. The training and validation datasets were divided into 2534 and 634 images, respectively.

3.2. Execution Details

All works rely on the device equipped with NVIDIA GeForce RTX 3080. The operation platform is Windows, and the software platform is PyTorch 1.10.0.

3.2.1. Super Parameter Setting

The learning strategy affects the final training outcome, so choosing the right learning strategy is crucial. Our experiments set the initial learning rate, batch size, and iterations to 0.0001, 16, and 200, respectively. Considering that the Adam optimizer can converge quickly and stably, it was used as our optimizer, where β1 and β2 were set to 0.9 and 0.999. The exponential decay strategy was chosen as the learning rate decay strategy, and its formula is as follows:

$$\mathrm{lr} = \mathrm{lr}_{\mathrm{init}} \times \left(1 - \frac{\mathrm{epoch}}{\mathrm{num\_epoch}}\right)^{\mathrm{power}}. \tag{12}$$

Here, lr and $\mathrm{lr}_{\mathrm{init}}$ represent the current and initial learning rates, epoch represents the count of present iterations, and num_epoch represents the total count of iterations; power, which controls the shape of the curve, was set to 2.

3.2.2. Loss Function

In our experiments, the cross-entropy loss function, a typical loss function in semantic segmentation, was employed. This loss function examines each pixel one by one and compares the prediction vector with the true value vector. It is expressed as follows:

$$\mathrm{L} = -\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{c}=1}^{\mathrm{M}} \mathrm{y}_{\mathrm{ic}} \log(\mathrm{p}_{\mathrm{ic}}). \tag{13}$$

Here, N means the count of pixels; M means the count of categories; y_{ic} means the indicator variable (0 or 1), y_{ic} is 1 if the predicted class c of the pixel i is true, else 0; p_{ic} means the prediction probability of pixel i belonging to the c class.

In our proposed model, auxiliary loss was also set to supervise and optimize the learning process. The auxiliary segmentation heads are after four feature fusion blocks, as can be seen in Figure 1. They can help supervise each stage of the decoder without affecting the main loss. The final loss consists of one main loss and four auxiliary losses. The ratio of each loss was set to 1. To prevent overfitting and improve the generalization ability, L2 regularization is introduced in the loss function. The final form is as follows:

$$\mathrm{L_f}(\theta) = \mathrm{L_m}(\theta) + \sum_{\mathrm{i}=1}^{4} \mathrm{L_{a_i}}(\theta) + \frac{\lambda}{2} \parallel \theta \parallel^2. \tag{14}$$

Here, $\mathrm{L_f}$ denotes the final loss, $\mathrm{L_m}$ denotes the main loss, $\mathrm{L_{a_i}}$ denotes the ith auxiliary loss, and θ denotes the network parameters. The coefficient λ in L2 regularization was set to 1.

3.2.3. Evaluation Indicators

Some common evaluation indicators for semantic segmentation were used in this work to assess and compare the results of different segmentation models. The following is a brief introduction to the evaluation indicators used in our experiments.

The percentage of all correctly categorized pixels of all pixels is known as pixel accuracy (PA). Precision (P) is the percentage of pixels accurately categorized per category of those predicted for that category. Recall (R) is the ratio of the count of pixels accurately categorized

per class to the total count of pixels per class, mP is the cumulative mean of P for each class, and mR is the cumulative average of R for each class. Their formulas are as follows:

$$PA = \sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij}} \tag{15}$$

$$P_i = \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij}}, \ i = 1, \ \ldots \ , \ k \tag{16}$$

$$R_i = \frac{p_{ii}}{\sum_{j=1}^{k} p_{ji}}, \ i = 1, \ \ldots \ , \ k \tag{17}$$

$$mP = \frac{1}{k} \sum_{i=1}^{k} P_i \tag{18}$$

$$mR = \frac{1}{k} \sum_{i=1}^{k} R_i. \tag{19}$$

Here, k denotes the count of categories, p_{ii} denotes the count of pixels predicted correctly in class i, p_{ij} denotes the count of pixels predicted as class j in class i, and p_{ji} denotes the count of pixels predicted as class i in class j. F1 score is the harmonic average of P and R. If the two are quite out of balance, where one is particularly high and the other is particularly low, the resulting F1 score will be particularly low. The F1 score will only be high when both are very high. Its formula is as follows:

$$F1 = \frac{1}{k} \sum_{i=1}^{k} \frac{2 \times P_i \times R_i}{P_i + R_i}. \tag{20}$$

Here, P_i denotes the precision of class i, and R_i denotes R of class i. IoU is the proportion of the intersection and union of the predicted and true value for each category. MIoU is the cumulative mean of IoU for each category. The purpose of FWIoU is to weigh IoU according to the frequency of each class and then sum them. Their formulas are as follows:

$$IoU_i = \frac{pii}{\sum_{j=1}^{k} p_{ij} + \sum_{j=1}^{k} p_{ji} - p_{ii}} \tag{21}$$

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

$$FWIoU = \frac{1}{\sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij}} \sum_{i=1}^{k} (IoU_i \sum_{j=1}^{k} p_{ij}). \tag{23}$$

3.3. Ablation Study on Cloud and Cloud Shadow Dataset

Ablation experiments were performed on four modules of our proposed model, namely CAB, FCB, SPA, and CFAB. The details of them were described in the second section. Among them, CAB, FCB, and SPA belong to the feature fusion block. Table 1 compares the evaluation indicators, the quantity of parameters, and the volume of calculations for each ablation module. ResNet50 was used as the backbone in our ablation experiments. It can save time and memory with relatively high accuracy. The reference model uses ResNet50 as the encoder and performs simply additive fusion in the decoder stage. As seen in Table 1, the MIoU of it is 92.18% and PA is 96.58%.

Table 1. Comparison of each ablation model (the best model is in bold).

Models	Parameters (M)	Flops (G)	PA (%)	MIoU (%)
ResNet50	27.00	4.95	96.58	92.18
ResNet50 + CAB	27.53	5.17	96.93	92.79
ResNet50 + FCB + CAB	43.53	7.94	97.25	93.60
ResNet50 + FCB + CAB + SPA	43.53	7.95	97.36	93.82
ResNet50 + FCB + CAB + SPA + CFAB	**44.06**	**8.52**	**97.39**	**93.99**

Addition testing for CAB: CAB weighs and averages the channel weights from the fusion features to the shallow and deep features, then fuses them again to improve the fusion effect. It increases PA from 96.58% to 96.93% and MIoU from 92.18% to 92.79%.

Addition testing for FCB: FCB includes two multi-scale convolution blocks (MSC). It is added before the feature maps are fused. MSC can enlarge the receptive field by multi-scale convolution at the convolution level. It reduces the excessive semantic differences between deep and shallow features, thus improving the fusion performance. FCB raises PA from 96.93% to 97.25% and MIoU from 92.79% to 93.60%.

Addition testing for SPA: SPA focuses on the spatial information of cloud and shadow. Through increasing the almost negligible amount of model parameters and calculations, it improves PA from 97.25% to 97.36% and MIoU from 93.60% to 93.82%.

Addition testing for CFAB: Cloud and cloud shadow have obvious category characteristics. CFAB is added after the final part of the decoder to learn category feature information. It increases PA from 97.36% to 97.39% and MIoU from 93.82% to 93.99%.

3.4. Analysis of Comparative Experiments

To compare with our suggested model, SegNet, U-Net, FCN8s, DeepLabV3+, PSPNet, HRNet, and OCRNet were used in this work. SegNet is a semantic segmentation model with an encoder–decoder architecture [57]. The encoder of it uses VGG16 as the backbone and is symmetrical to the decoder. U-Net is one of the most common and simple semantic segmentation models and was designed originally for segmenting medical imagery [58]. Its structure is U-shaped. FCN is the most basic network in image segmentation, substituting a convolutional layer for the final fully connected layer [59]. FCN8s in this paper used ResNet50 as the backbone with down-sampling 8x. The latter two down-sampling operations of ResNet50 were replaced with dilated convolution.

DeepLabV3+ is the improved model of the DeepLab series, proposed for image segmentation by the Google Team [60]. Based on the dilated convolution and ASSP module, it fuses multi-scale information through the encoder–decoder architecture. DeepLabV3+ in this paper used ResNet50 as the backbone with 8x down-sampling. PSPNet is the scene analysis network proposed for the disadvantages of FCN [61]. It collects contextual information more effectively through the pyramidal aggregation mechanism. PSPNet in this paper used ResNet50 as the backbone with 8x down-sampling. HRNet is a key point detection network proposed for human pose estimation [62]. High-resolution and low-resolution features in HRNet are connected in parallel and exchange information constantly. OCRNet is an advanced segmentation model that transforms pixel-level classification issues into object-level classification issues [54]. It uses HRNet as the backbone and explicitly enhances object-related information through the OCR concept.

Table 2 shows the comparative results of the above segmentation algorithms on our self-made dataset. The backbone of each model is shown in the table too. All models were pretrained in our experiments. The results show that our proposed model is superior to other algorithms for segmenting clouds and their shadows. In order to facilitate comparison, ResNet50, which was also used in the majority of other models, was used as the backbone of our proposed model. The MIoU of our model was 93.99%. It is 1.14% higher than PSPNet using the same backbone, and 0.33% higher than advanced OCRNet. Our proposed model can improve semantic segmentation results by changing the improved backbone. The MIoU of our model based on Res2net50 was 94.10%, which is 0.44% higher than OCRNet.

Considering that our model parameters and calculations are smaller than OCRNet, it is a good result. In addition, Table 3 provides a detailed list of the accuracy improvement of our proposed model under different surface types. Our model shows varying degrees of improvement in segmentation performance across five different backgrounds (water, farmland, plant, wasteland, and urban) in our self-made dataset.

Table 2. Comparative cloud and shadow detection results of various algorithms (the best model is in bold).

Methods	Backbone	PA (%)	mP (%)	mR (%)	F1 (%)	MIoU (%)	FwioU (%)
SegNet [57]	VGG16	95.58	94.32	94.79	94.55	89.74	91.60
U-Net [58]	VGG16	96.46	95.51	95.84	95.67	91.74	93.20
FCN8s [59]	Resnet50	96.55	95.76	95.74	95.75	91.89	93.36
DeepLabv3+ [60]	Resnet50	96.85	96.09	96.18	96.14	92.59	93.91
PSPNet [61]	Resnet50	96.96	96.25	96.30	96.28	92.85	94.12
HRNet [62]	HRNet-W48	97.09	96.16	96.74	96.45	93.17	94.38
OCRNet [54]	HRNet-W48	97.30	96.63	96.80	96.71	93.66	94.76
Ours	Resnet50	97.44	96.95	96.84	96.89	93.99	95.02
Ours	**Res2net50**	**97.48**	**97.03**	**96.88**	**96.95**	**94.10**	**95.10**

Table 3. The MIoU of our model under different backgrounds (the best model is in bold).

Methods	Urban (%)	Farmland (%)	Plant (%)	Water (%)	Wasteland (%)
SegNet	85.12	81.76	83.96	84.50	87.60
U-Net	85.84	82.66	84.75	85.21	88.73
FCN8s	87.07	83.92	87.40	86.53	90.56
DeepLabv3+	88.01	85.27	87.88	87.21	91.07
PSPNet	87.60	85.28	88.13	87.29	91.30
HRNet	88.61	85.11	89.64	88.12	92.25
OCRNet	89.22	85.40	90.05	88.30	93.77
Ours	**90.38**	**85.98**	**90.74**	**88.94**	**93.90**

The predicted pictures of different segmentation models are shown in Figures 9–12, which intuitively reflect the superior performance of our proposed model. Our model can obtain good recognition results for small-scale targets. It benefits from multi-scale fusion and category feature learning in our designed network. As is shown in Figure 9, our proposed model recognized the small-scale thin cloud marked by the white boxes better than other algorithms.

The proposed model can better identify the boundary between cloud, cloud shadow, and background. As shown in Figure 10, FCN8s, DeepLabv3+, PSPNet, and HRNet failed to recognize the holes of the cloud shadow very well in the area marked by the white boxes. This is due to the insufficient perception of local feature information. In the area marked by the blue boxes, OCRNet had identification errors. Our designed model enhances the expression of global and local information through attention mechanisms. Therefore, it has a higher boundary recognition ability and a lower error rate.

Our model can identify details better than other models. As shown in the area marked by the white boxes in Figure 11, FCN8s, DeepLabv3+, PSPNet, and HRNet only recognized the rough outline of the cloud, losing too much detailed information. On the other hand, our network performs several effective multi-scale feature fusion operations to improve the ability to perceive details. Even compared to OCRNet, which has high semantic segmentation performance, our model still identified more abundant and accurate details.

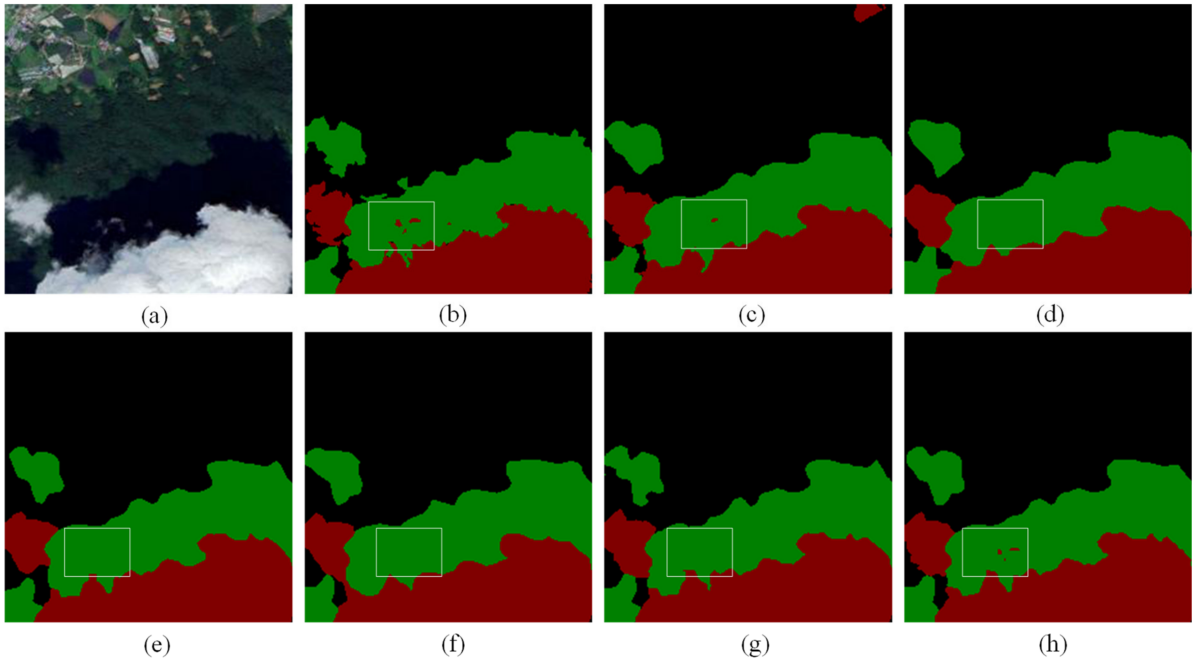

Figure 9. The predicted images of different segmentation models under the plant or farmland area. Clouds, cloud shadows, and backgrounds are marked red, green, and black, respectively. The white rectangles highlight the areas where our model has shown significant improvements compared to other models. (**a**) The original picture, (**b**) the label, (**c**) FCN8s, (**d**) DeepLabv3+, (**e**) PSPNet, (**f**) HRNet, (**g**) OCRNet, and (**h**) ours.

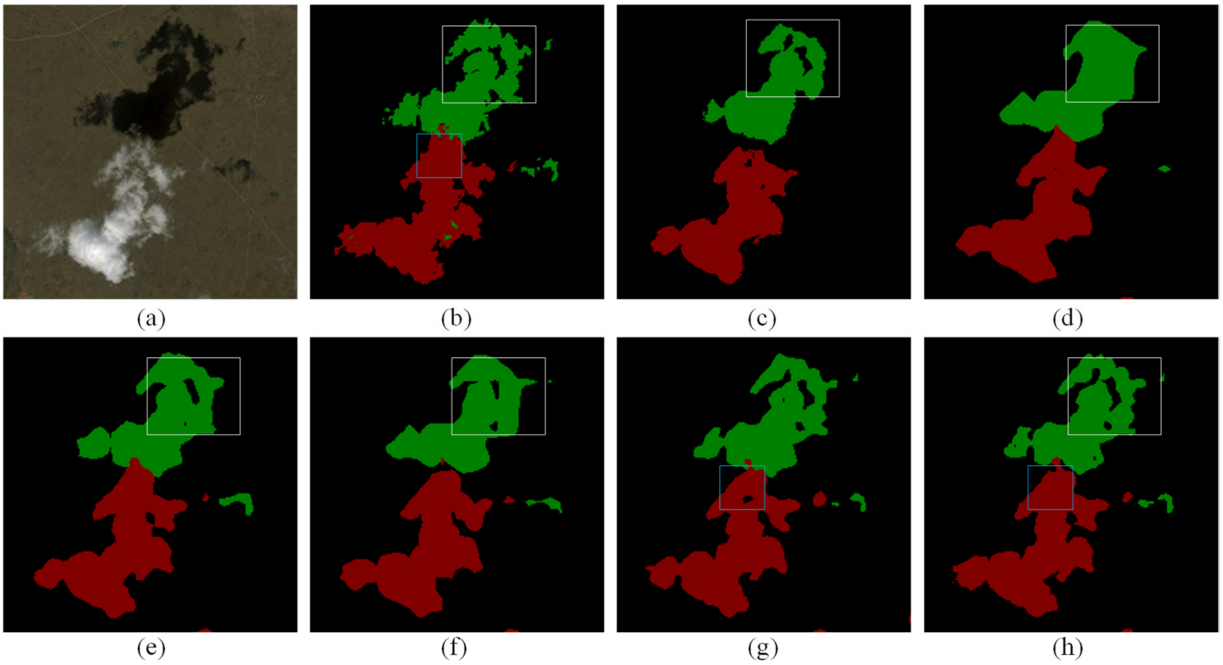

Figure 10. The predicted images of different segmentation models under the wasteland area. Clouds, cloud shadows, and backgrounds are marked red, green, and black, respectively. The white and blue rectangles highlight the areas where our model has shown significant improvements compared to other models. (**a**) The original picture, (**b**) the label, (**c**) FCN8s, (**d**) DeepLabv3+, (**e**) PSPNet, (**f**) HRNet, (**g**) OCRNet, and (**h**) ours.

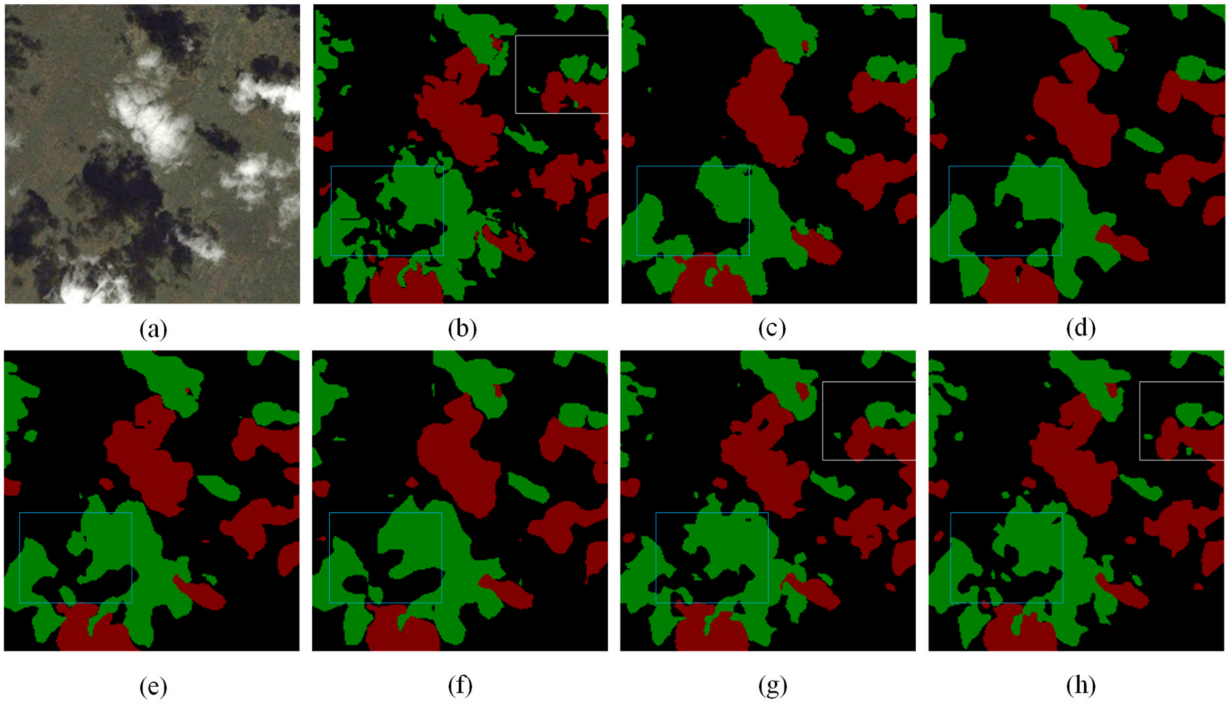

Figure 11. The predicted images of different segmentation models under the water areas. Clouds, cloud shadows, and backgrounds are marked red, green, and black, respectively. The white and blue rectangles highlight the areas where our model has shown significant improvements compared to other models. (**a**) The original picture, (**b**) the label, (**c**) FCN8s, (**d**) DeepLabv3+, (**e**) PSPNet, (**f**) HRNet, (**g**) OCRNet, and (**h**) ours.

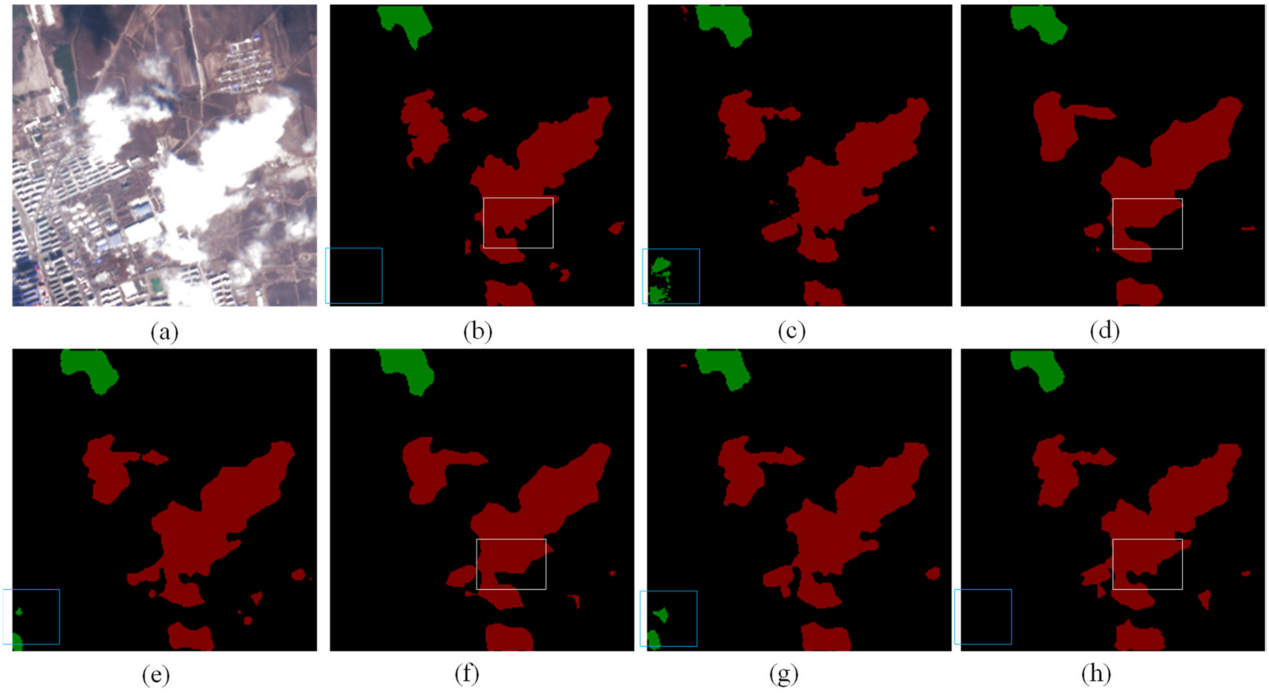

Figure 12. The predicted images of different segmentation models under the urban area. Clouds, cloud shadows, and backgrounds are marked red, green, and black, respectively. The white and blue rectangles highlight the areas where our model has shown significant improvements compared to other models. (**a**) The original picture, (**b**) the label, (**c**) FCN8s, (**d**) DeepLabv3+, (**e**) PSPNet, (**f**) HRNet, (**g**) OCRNet, and (**h**) ours.

Our model can maintain good prediction performance even under complex backgrounds. To prove this, we selected an image with an urban background for testing. The background contains rich ground object information, making it challenging to identify clouds and shadows accurately. As seen in the areas marked by the blue boxes in Figure 12, DeepLabv3+, HRNet, and OCRNet falsely predicted other objects as cloud shadows. The reason for these false predictions is that the background of this area is similar to cloud shadow. On the other hand, FCN8s and PSPNet had correct predictions, not because of their superior semantic segmentation performance, but because they focus on global information and ignore local information. As shown in the areas marked by the white boxes, these models did not predict the details of the cloud boundary very well. In contrast, our model adds spatial attention mechanisms, focusing on task-related areas. Thus, only our model can balance local information and noise interference effectively.

3.5. Generalization Performance Analysis

Comparative studies were carried out on the HRC_WHU and SPARCS datasets to further demonstrate the viability of our suggested model. The experiments on HRC_WHU help analyze the performance of cloud recognition. Table 4 lists evaluation indicators of various algorithms on HRC_WHU. From the table, the proposed model has superior performance for cloud recognition compared to other algorithms. The MioU of our model was 95.26%, which is 0.73% higher than the best model (i.e., PSPNet) of the other algorithms. In addition to our model, the predicted images of the three other models that achieved the best results among these algorithms (i.e., DeepLabv3+, OCRNet, and PSPNet) are shown in Figure 13. In order to be more convincing, the test images under different backgrounds are specially selected. It is evident that the visual images produced by our model have more cloud details and better segmentation effects compared to others.

Table 4. Comparative results of various algorithms on HRC_WHU (the best results are in bold).

Methods	PA (%)	mP (%)	mR (%)	F1 (%)	MIoU (%)	FwioU (%)
SegNet	94.73	94.63	94.54	94.58	89.74	89.99
UNet	95.94	95.86	95.81	95.84	92.01	92.20
HRNet	95.97	95.87	95.86	95.87	92.07	92.27
FCN8s	96.87	96.75	96.84	96.80	93.80	93.94
DeepLabv3+	97.04	96.91	97.02	96.97	94.11	94.26
OCRNet	97.24	97.14	97.19	97.16	94.49	94.63
PSPNet	97.25	97.16	97.21	97.18	94.53	94.65
Ours	**97.63**	**97.55**	**97.59**	**97.57**	**95.26**	**95.37**

The experiments on the SPARCS dataset can analyze our model performance in the face of complex segmentation problems. The segmentation models need to identify all pixels of the test images into five classes, including clouds, cloud shadows, snow, water, and backgrounds. Table 5 shows evaluation indicators of different segmentation algorithms on the SPARCS dataset. From the table, our proposed model maintains a good segmentation performance for multi-class classification compared to other algorithms. The MIoU of our model was 91.31%, which is 0.34% higher than the best model (i.e., OCRNet) among the other algorithms. The performance of these models for various classes was assessed by the evaluation indicator called IoU. Table 6 shows the comparative results of each class. As seen from the table, our model has better results for each class than the other models. Our model maintains a relatively high recognition accuracy in the face of indistinguishable clouds and snow.

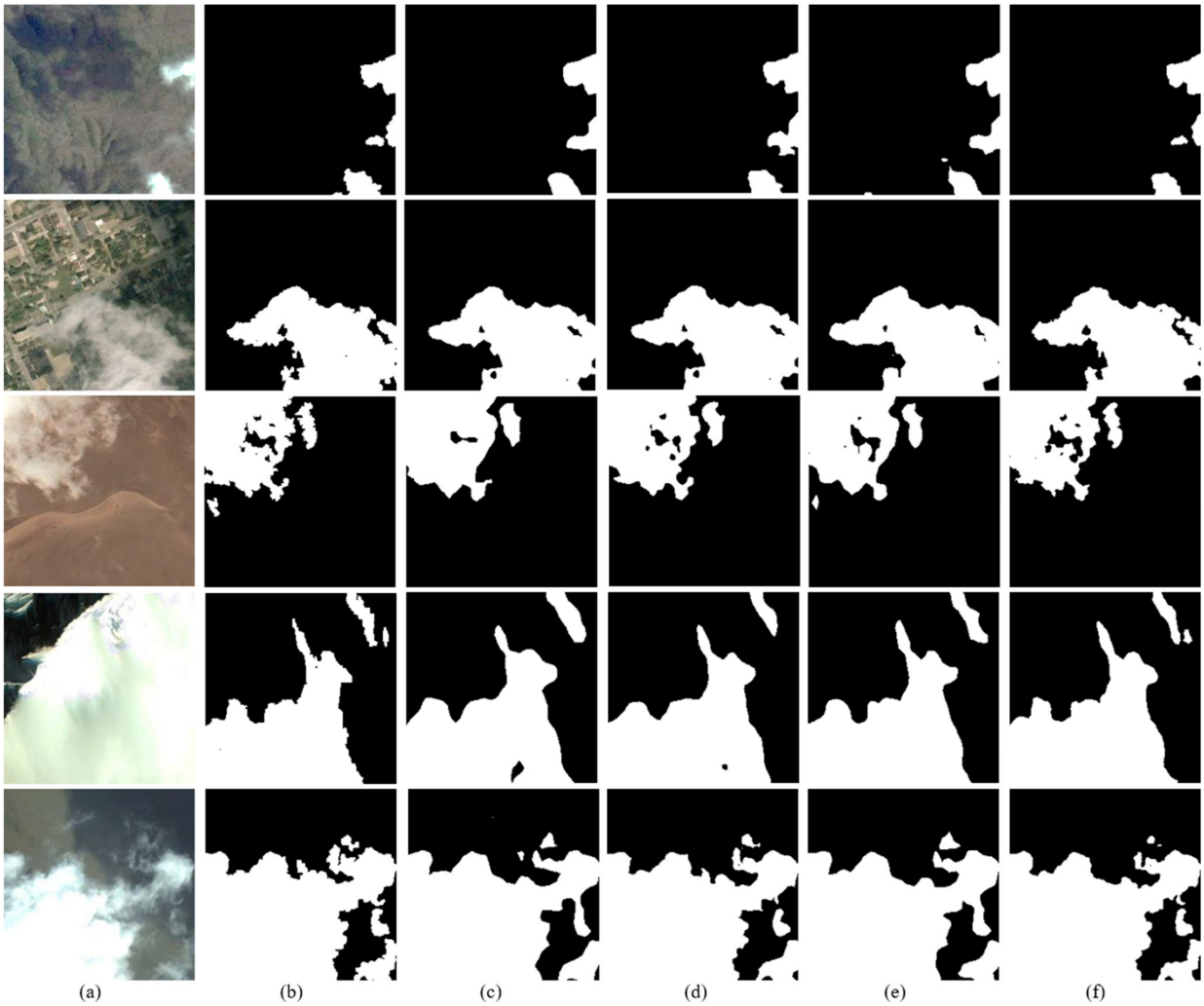

Figure 13. The predicted images of different segmentation models under the background of vegetation, urban, barren, snow, and water. (**a**) The original pictures, (**b**) the labels, (**c**) DeepLabv3+, (**d**) OCRNet, (**e**) PSPNet, and (**f**) ours.

Table 5. Comparative results of different algorithms on SPARCS (the best results are in bold).

Methods	PA (%)	mP (%)	mR (%)	F1 (%)	MIoU (%)	FwioU (%)
SegNet	92.19	89.46	87.64	88.48	80.07	85.80
PSPNet	95.56	94.02	92.81	93.39	87.84	91.61
UNet	95.48	93.92	92.92	93.40	87.91	91.47
FCN8s	95.60	94.27	92.68	93.44	87.94	91.69
DeepLabv3+	96.43	95.21	94.16	94.67	90.05	93.18
HRNet	96.60	95.15	94.90	95.03	90.67	93.51
OCRNet	96.79	95.66	94.76	95.20	90.97	93.85
Ours	**96.90**	**95.76**	**95.03**	**95.39**	**91.31**	**94.05**

In addition to our model, Figure 14 shows the predicted images of the other three models that have the best results (i.e., DeepLabv3+, HRNet, and OCRNet). Clouds, cloud shadows, snow, water, and backgrounds are marked white, black, cyan, blue, and grey, respectively. As seen from the picture, our model performs better for boundary recognition than the others.

Table 6. Comparative results (IoU) of each class obtained by different algorithms on SPARCS (the best model is in bold).

Methods	Clouds (%)	Cloud Shadows (%)	Snow (%)	Water (%)	Backgrounds (%)	Average (%)
SegNet	81.71	59.67	89.28	78.92	90.75	80.07
PSPNet	89.42	75.77	93.33	85.98	94.72	87.84
UNet	88.57	74.22	93.48	88.61	94.68	87.91
FCN8s	89.59	75.97	93.41	85.95	94.75	87.94
DeepLabv3+	91.45	80.31	94.50	88.27	95.72	90.05
HRNet	91.57	81.24	94.87	89.73	95.93	90.67
OCRNet	92.39	82.46	95.05	88.78	96.15	90.97
Ours	**92.41**	**82.63**	**95.26**	**89.90**	**96.32**	**91.31**

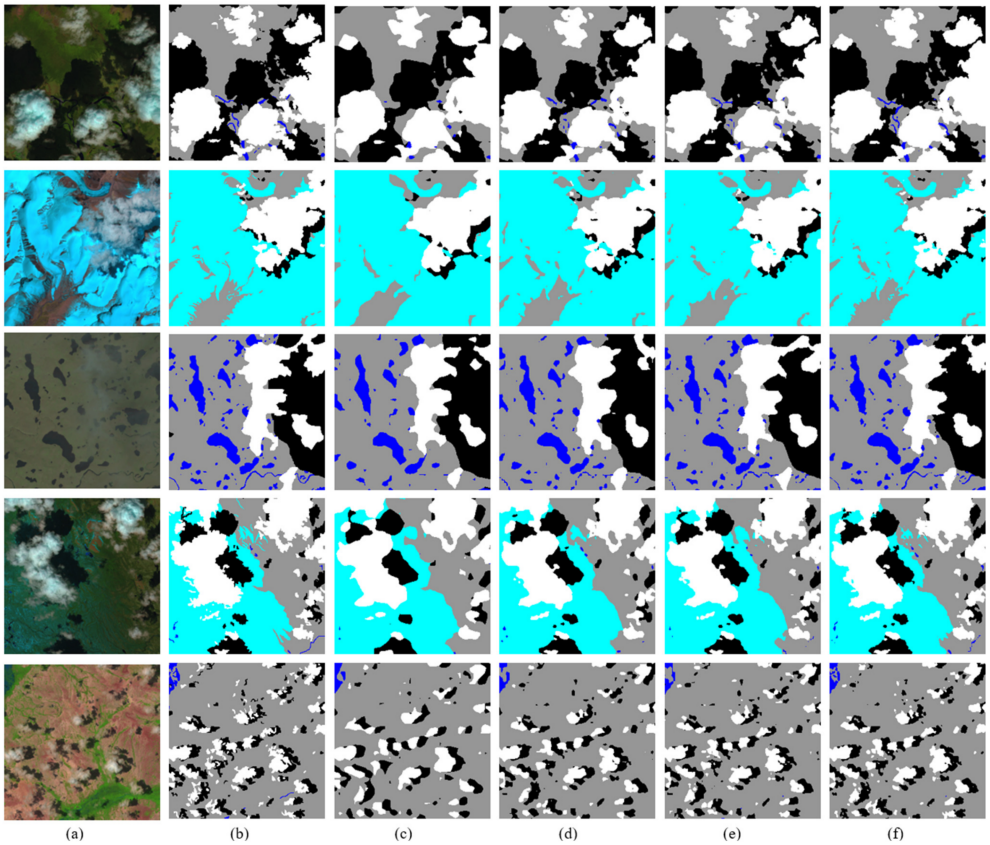

Figure 14. The predicted images of different segmentation algorithms. Clouds, cloud shadows, snow, water, and backgrounds are marked white, black, cyan, blue, and grey, respectively. (**a**) The original pictures, (**b**) the labels, (**c**) DeepLabv3+, (**d**) HRNet, (**e**) OCRNet, (**f**) ours.

4. Conclusions

Cloud and cloud shadow segmentation is an important research direction in the field of remote sensing. This paper proposes a multi-supervised feature fusion attention semantic segmentation network. The network chooses the ResNet series as the encoder. The decoder includes the feature fusion block (FFB) and category feature attention block (CFAB), which restore the spatial information lost during the encoder stage. FFB consists of FCB, CAB, and SPA, which effectively fuse the multi-scale features from the encoder. Since cloud and cloud shadow have obvious feature information, this paper designs the CFAB to learn category feature information before outputting the final segmentation maps. Compared to previous segmentation models, our algorithm can better capture the details and boundaries of clouds and their shadows and reduce interference under complex backgrounds. Our model performs well on the homemade dataset, and experiments on the HRC_WHU and SPARCS

datasets show the good generalization performance of our model. However, our network still has some shortcomings: (1) Our model can have a relatively accurate segmentation effect for clouds and cloud shadows under complex backgrounds, but it may have errors identifying nearby similar objects as clouds or shadows. (2) It is worth investigating how to reduce the number of parameters and calculations in our model without significantly reducing recognition accuracy.

Author Contributions: Conceptualization, Huiwen Ji, Min Xia and Dongsheng Zhang; methodology, Min Xia and Huiwen Ji; software, Huiwen Ji; validation, Huiwen Ji and Min Xia; formal analysis, Huiwen Ji and Haifeng Lin; investigation, Huiwen Ji and Min Xia; resources, Min Xia; data curation, Min Xia; writing—original draft preparation, Huiwen Ji; writing—review and editing, Huiwen Ji and Min Xia; visualization, Huiwen Ji; supervision, Min Xia; project administration, Min Xia; funding acquisition, Min Xia. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of the People's Republic of China (Grant No. 42075130).

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

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

References

1. Zhu, Z.; Woodcock, C.E. Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change. *Remote Sens. Environ.* **2014**, *152*, 217–234. [CrossRef]
2. Li, S.; Sun, D.; Yu, Y. Automatic cloud-shadow removal from flood/standing water maps using MSG/SEVIRI imagery. *Int. J. Remote Sens.* **2013**, *34*, 5487–5502. [CrossRef]
3. Huang, C.; Thomas, N.; Goward, S.N.; Masek, J.G.; Zhu, Z.; Townshend, J.R.G.; Vogelmann, J.E. Automated masking of cloud and cloud shadow for forest change analysis using Landsat images. *Int. J. Remote Sens.* **2010**, *31*, 5449–5464. [CrossRef]
4. Oishi, Y.; Ishida, H.; Nakamura, R. A new Landsat 8 cloud discrimination algorithm using thresholding tests. *Int. J. Remote Sens.* **2018**, *39*, 9113–9133. [CrossRef]
5. Irish, R.R.; Barker, J.L.; Goward, S.N.; Arvidson, T. Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 1179–1188. [CrossRef]
6. Zhu, Z.; Woodcock, C.E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **2012**, *118*, 83–94. [CrossRef]
7. 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]
8. Qiu, S.; He, B.; Zhu, Z.; Liao, Z.; Quan, X. Improving Fmask cloud and cloud shadow detection in mountainous area for Landsats 4–8 images. *Remote Sens. Environ.* **2017**, *199*, 107–119. [CrossRef]
9. Frantz, D.; Haß, E.; Uhl, A.; Stoffels, J.; Hill, J. Improvement of the Fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects. *Remote Sens. Environ.* **2018**, *215*, 471–481. [CrossRef]
10. Qiu, S.; Zhu, Z.; He, B. Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. *Remote Sens. Environ.* **2019**, *231*, 111205. [CrossRef]
11. 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]
12. Zhu, X.; Helmer, E.H. An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions. *Remote Sens. Environ.* **2018**, *214*, 135–153. [CrossRef]
13. Candra, D.S.; Phinn, S.; Scarth, P. Cloud and cloud shadow removal of landsat 8 images using Multitemporal Cloud Removal method. In Proceedings of the 2017 6th International Conference on Agro-Geoinformatics, Fairfax, VA, USA, 7–10 August 2017; pp. 1–5.
14. Lin, J.; Huang, T.Z.; Zhao, X.L.; Ding, M.; Chen, Y.; Jiang, T.X. A Blind Cloud/Shadow Removal Strategy for Multi-Temporal Remote Sensing Images. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 4656–4659.
15. Racoviteanu, A.; Williams, M.W. Decision tree and texture analysis for mapping debris-covered glaciers in the Kangchenjunga area, Eastern Himalaya. *Remote Sens.* **2012**, *4*, 3078–3109. [CrossRef]
16. Hollstein, A.; Segl, K.; Guanter, L.; Brell, M.; Enesco, M. Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in Sentinel-2 MSI images. *Remote Sens.* **2016**, *8*, 666. [CrossRef]
17. Lu, F.; Gong, Z. Construction of cloud-shadow-water mask based on Random Forests algorithm. *Remote Sens. Land Resour.* **2016**, *28*, 73–79.

18. Ghasemian, N.; Akhoondzadeh, M. Introducing two Random Forest based methods for cloud detection in remote sensing images. *Adv. Space Res.* **2018**, *62*, 288–303. [CrossRef]
19. Wei, J.; Huang, W.; Li, Z.; Sun, L.; Zhu, X.; Yuan, Q.; Liu, L.; Cribb, M. Cloud detection for Landsat imagery by combining the random forest and superpixels extracted via energy-driven sampling segmentation approaches. *Remote Sens. Environ.* **2020**, *248*, 112005. [CrossRef]
20. 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*, 5609519. [CrossRef]
21. 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]
22. Chai, D.; Newsam, S.; Zhang, H.K.; Qiu, Y.; Huang, J. Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks. *Remote Sens. Environ.* **2019**, *225*, 307–316. [CrossRef]
23. Mohajerani, S.; Saeedi, P. 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.
24. Segal-Rozenhaimer, M.; Li, A.; Das, K.; Chirayath, V. Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN). *Remote Sens. Environ.* **2020**, *237*, 111446. [CrossRef]
25. Wieland, M.; Li, Y.; Martinis, S. Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network. *Remote Sens. Environ.* **2019**, *230*, 111203. [CrossRef]
26. Shendryk, Y.; Rist, Y.; Ticehurst, C.; Thorburn, P. Deep learning for multi-modal classification of cloud, shadow and land cover scenes in PlanetScope and Sentinel-2 imagery. *ISPRS J. Photogramm. Remote Sens.* **2019**, *157*, 124–136. [CrossRef]
27. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
28. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
29. Wang, D.; Weng, L.; Xia, M.; Lin, H. MBCNet: Multi-Branch Collaborative Change-Detection Network Based on Siamese Structure. *Remote Sens.* **2023**, *15*, 2237. [CrossRef]
30. 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*, 32–43. [CrossRef]
31. Ma, Z.; Xia, M.; Lin, H.; Qian, M.; Zhang, Y. FENet: Feature enhancement network for land cover classification. *Int. J. Remote Sens.* **2023**, *44*, 1702–1725. [CrossRef]
32. Gao, J.; Weng, L.; Xia, M.; Lin, H. MLNet: Multichannel feature fusion lozenge network for land segmentation. *J. Appl. Remote Sens.* **2022**, *16*, 016513. [CrossRef]
33. Chen, B.; Xia, M.; Qian, M.; Huang, J. MANet: A multi-level aggregation network for semantic segmentation of high-resolution remote sensing images. *Int. J. Remote Sens.* **2022**, *43*, 5874–5894. [CrossRef]
34. Ma, Z.; Xia, M.; Weng, L.; Lin, H. Local Feature Search Network for Building and Water Segmentation of Remote Sensing Image. *Sustainability* **2023**, *15*, 3034. [CrossRef]
35. Hu, K.; Li, M.; Xia, M.; Lin, H. Multi-scale feature aggregation network for water area segmentation. *Remote Sens.* **2022**, *14*, 206. [CrossRef]
36. Chen, J.; Xia, M.; Wang, D.; Lin, H. Double Branch Parallel Network for Segmentation of Buildings and Waters in Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1536. [CrossRef]
37. Hu, K.; Zhang, E.; Xia, M.; Weng, L.; Lin, H. Mcanet: A multi-branch network for cloud/snow segmentation in high-resolution remote sensing images. *Remote Sens.* **2023**, *15*, 1055. [CrossRef]
38. Lu, C.; Xia, M.; Qian, M.; Chen, B. Dual-branch network for cloud and cloud shadow segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5410012. [CrossRef]
39. Miao, S.; Xia, M.; Qian, M.; Zhang, Y.; Liu, J.; Lin, H. Cloud/shadow segmentation based on multi-level feature enhanced network for remote sensing imagery. *Int. J. Remote Sens.* **2022**, *43*, 5940–5960. [CrossRef]
40. Chen, Y.; Weng, Q.; Tang, L.; Wang, L.; Xing, H.; Liu, Q. Developing an intelligent cloud attention network to support global urban green spaces mapping. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 197–209. [CrossRef]
41. Chen, Y.; Tang, L.; Huang, W.; Guo, J.; Yang, G. A Novel Spectral Indices-Driven Spectral-Spatial-Context Attention Network for Automatic Cloud Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3092–3103. [CrossRef]
42. Zhang, C.; Weng, L.; Ding, L.; Xia, M.; Lin, H. CRSNet: Cloud and Cloud Shadow Refinement Segmentation Networks for Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 1664. [CrossRef]
43. Qu, Y.; Xia, M.; Zhang, Y. Strip pooling channel spatial attention network for the segmentation of cloud and cloud shadow. *Comput. Geosci.* **2021**, *157*, 104940. [CrossRef]
44. Hu, K.; Zhang, D.; Xia, M. Cdunet: Cloud detection unet for remote sensing imagery. *Remote Sens.* **2021**, *13*, 4533. [CrossRef]
45. Lu, C.; Xia, M.; Lin, H. Multi-scale strip pooling feature aggregation network for cloud and cloud shadow segmentation. *Neural Comput. Appl.* **2022**, *34*, 6149–6162. [CrossRef]
46. Xia, M.; Wang, T.; Zhang, Y.; Liu, J.; Xu, Y. Cloud/shadow segmentation based on global attention feature fusion residual network for remote sensing imagery. *Int. J. Remote Sens.* **2021**, *42*, 2022–2045. [CrossRef]

47. Yan, Z.; Yan, M.; Sun, H.; Fu, K.; Hong, J.; Sun, J.; Zhang, Y.; Sun, X. Cloud and cloud shadow detection using multilevel feature fused segmentation network. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1600–1604. [CrossRef]
48. Wang, Z.; Xia, M.; Lu, M.; Pan, L.; Liu, J. Parameter Identification in Power Transmission Systems Based on Graph Convolution Network. *IEEE Trans. Power Deliv.* **2022**, *37*, 3155–3163. [CrossRef]
49. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef]
50. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R. Resnest: Split-attention networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 2736–2746.
51. Zhang, S.; Weng, L. STPGTN–A Multi-Branch Parameters Identification Method Considering Spatial Constraints and Transient Measurement Data. *Comput. Model. Eng. Sci.* **2023**, *136*, 2635–2654. [CrossRef]
52. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
53. 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.
54. Yuan, Y.; Chen, X.; Chen, X.; Wang, J. Segmentation transformer: Object-contextual representations for semantic segmentation. *arXiv* **2019**, arXiv:1909.11065.
55. Li, Z.; Shen, H.; Liu, Y. *HRC_WHU: High-Resolution Cloud Cover Validation Data*; Wuhan University: Wuhan, China, 2019.
56. Hughes, M.J.; Kennedy, R. High-Quality Cloud Masking of Landsat 8 Imagery Using Convolutional Neural Networks. *Remote Sens.* **2019**, *11*, 2591. [CrossRef]
57. 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]
58. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
59. 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.
60. 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.
61. 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.
62. 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.