JARS Journal of Applied Remote Sensing

RESEARCH PAPER

DB-Net: dual-branch deep learning network for cloud detection utilizing attention mechanism and cascaded multiscale feature fusion

Ruyan Zhou,[a] Chenhao Zhang,[a] Haiyan Pan,[a,*] Haiyang He,[b] Yun Zhang,[a]
Yanling Han,[a] Jing Wang,[a] Shuhu Yang,[a] and Zhonghua Hong[a]
[a]Shanghai Ocean University, School of Information, Remote Sensing Navigation Laboratory, Shanghai, China
[b]Sense Time Intelligent Technology Co., Ltd., Shanghai, China

ABSTRACT. Remote sensing images often contain a significant amount of clouds, which can result in substantial resource costs during transmission and storage. Cloud detection can reduce these costs. Although current cloud detection methods perform well in extracting large and thick clouds, there are still some issues, such as missed detection of small and thin clouds and false detection in non-cloud areas. Therefore, we propose a deep learning framework called DB-Net. It consists of three main modules: feature extraction module (FEM), cascaded feature enhancement module (CFEM), and feature fusion module (FFM). In the FEM, we leverage the advantages of both convolutional neural network and Transformer by utilizing two branches to reduce the loss of semantic information. To enhance the acquisition capability of multi-scale semantic information, in the CFEM, regular convolutions are replaced with deformable convolutions to adaptively capture cloud features of various sizes, and a cascaded structure is designed to enhance the interaction of information among different scales. Furthermore, to focus on small and thin cloud information and suppress non-cloud background information, we designed the FFM using attention mechanisms to enhance the target information in the features extracted by FEM and CFEM. Extensive experiments were conducted on the GF1-WHU dataset, and comparisons were made with mainstream cloud detection networks. The experimental results indicate that the proposed DB-Net method reduces cloud information omission, effectively focuses on thin clouds and small clouds, and improves overall cloud detection performance.

[DOI: 10.1117/1.JRS.18.034524]

Keywords: cloud detection; dual branch; transformer; convolutional neural network; attention mechanisms; remote sensing

Paper 240266G received May 4, 2024; revised Aug. 16, 2024; accepted Aug. 28, 2024; published Sep. 27, 2024.

1 Introduction

With the advancement of technology, a great deal of artificial satellites are being launched into space, leading to a rapid growth in the quantity of remote-sensing images. These data are widely applied in the fields of surface observation,[1] natural disaster analysis,[2] land use and land cover change detection,[3] and three-dimensional terrain reconstruction.[4] However, according to the International Satellite Cloud Climatology Project,[5] more than 66% of the earth's surface is regularly covered by clouds. Most of the data covered by the cloud layer are redundant and invalid and bring burden and resource overhead to the analysis,[6] storage, and transmission of remote

*Address all correspondence to Haiyan Pan, hy-pan@shou.edu.cn sensing data, especially true for space-borne satellites. Therefore, accurate cloud detection plays a crucial role in improving transmission efficiency and saving memory storage overhead.

In the field of deep learning, cloud detection is considered an end-to-end encoding–decoding model that establishes complex nonlinear mappings between inputs and outputs, enabling automatic extraction and classification of cloud features. Xie et al.[7] designed a dual-branch deep convolutional neural network (CNN) that effectively predicts thick clouds, thin clouds, or non-clouds. Shao et al.[8] proposed a network model named multiscale features CNN (MF-CNN). The model employs four different pooling operations to capture multi-scale features of clouds. Chai et al.[9] introduced the MSegNet network, which effectively enriched the semantic information of clouds using full-spectrum images. The aforementioned cloud detection methods based on CNNs achieve decent results, but they suffer from information loss during the feature extraction process and lack the capability to effectively extract foreground information (clouds). Attention mechanisms[10–13] can reduce information loss and enhance the foreground information. Zhang et al.[14] designed a cloud pixel detection model for GF1-WFV images. The model employs a U-shaped structure and utilizes spatial attention modules to capture information at different scales. Yu et al.[15] proposed a novel CNN architecture called MFG-Net for cloud detection in GF-5 images. They incorporated channel attention and spatial attention into the pyramid pooling module to capture cloud channel and spatial information. In addition, some researchers enhance the acquisition of foreground information by designing special convolutional kernels or employing information augmentation techniques. He et al.[16] designed a framework called DAB-Net. The framework consists of a feature extraction backbone and a deformable context feature pyramid module to capture features at different scales, reducing the computational complexity of the model while extracting multi-scale cloud information. Wu et al.[17] proposed BoundaryNet. It extracts cloud features at different scales and then leverages a differential boundary network for integrating multi-scale features and edge information. Guowei et al.[18] proposed BABFNet, which is specifically targeted at challenging areas in remote sensing image cloud detection, such as cloud boundaries and thin clouds. By introducing a boundary prediction branch and a bilateral fusion module, the performance of cloud detection was significantly improved. Zhang et al.[19] enhanced the extraction and fusion of global and multi-scale contextual information by designing the Resblock-cloud and two context information fusion modules. Zhang et al.[20] introduced a multi-branch residual context semantic module, a multi-scale convolution sub-channel attention module, and a feature fusion upsampling module to enhance feature extraction and edge information.

The aforementioned methods are all based on convolution operations, which can only capture local information. Although it is possible to obtain a larger receptive field and integrate a wider range of information through downsampling, this process may lead to a loss of cloud contextual semantic information. To overcome the limitations of CNNs, some researchers have recently explored the use of Transformers[21–23] to obtain global information. Zhao et al.[24] introduced the MMANet, harnessing the synergy of multi-scale patch embedding, multi-path pyramid vision transformers, and strip convolution to adeptly synthesize comprehensive and detailed feature representations. Zhang et al.[25] proposed Cloudformer V2, which introduces Transformers into cloud detection, addressing the issues of low accuracy in cloud detection. Tan et al.[26] proposed SwinUnet, which utilized the Swin Transformer to detect cloud and cloud shadow. Ma et al.[27] proposed a hybrid CNN-Transformer network called CNN-TransNet for cloud detection. It combines the advantages of Transformers and CNNs to enhance finer details and establish long-term dependencies.

As observed from the above analysis, to enhance the detection capability of thin clouds and small clouds, the network needs to capture both local and global features while possessing the ability to differentiate similar features. Therefore, the objective of this study is to preserve multi-scale and global information while enhancing local information, aiming to minimize the loss of information for small clouds and thin clouds. In addition, the attention mechanism is employed to focus on cloud-related foreground information and suppress interference from features similar to clouds in the scene. The contributions are as follows:

1. This paper proposes a feature extraction module (FEM) by combining CNN and Swin Transformer. In the FEM, a U-shaped CNN branch is designed to focus on extracting local

texture details, whereas the Swin branch utilizes the Swin Transformer to extract global information. The information from both branches complements each other, enhancing the ability to extract and preserve cloud information. The FEM will be introduced in Sec. 3.1.

2. This paper proposes a cascaded feature enhancement module (CFEM) based on atrous spatial pyramid pooling (ASPP).[28] It enhances the interaction of information across different scales by concatenating feature maps from different scales, thereby reducing the loss of multi-scale information. The CFEM will be introduced in Sec. 3.2.
3. This paper adopts a feature fusion module (FFM) based on an attention mechanism that combines the global information from the Swin branch with the local detail information from the CNN branch. It adaptively focuses on the semantic information of small and thin clouds and ignores the semantic information of features similar to clouds in the scene that are non-clouds, thereby improving the performance of cloud detection. The FFM will be introduced in Sec. 3.3.

2 Dataset

The GF1-WHU[29] dataset consists of 108 GF1 wide-field-of-view (WFV) 2A-grade images captured by the GF1 wide-angle camera, along with their corresponding reference clouds and cloud masks. The cloud masks for the GF1-WHU dataset were manually drawn by experienced experts. They visually inspected the boundaries of clouds and their shadows and manually delineated them. Pixel values 0, 1, 128, and 255 represent missing values, background, cloud shadows, and clouds, respectively. In this study, the focus was primarily on the detection of clouds. Therefore, irrelevant information was removed, and missing values, cloud shadows, and background were unified as the background class, whereas clouds were classified as a separate class. The background is represented by 0, and the clouds are represented by 255.

In this study, we selected 12 scenes of satellite imagery for training and three scenes for testing. The testing set includes three types of clouds: thin, thick, and small clouds. All images were cropped to a size of 384 × 384 pixels. The final training set consists of 14,698 cropped images, whereas the testing set contains a total of 5544 images, including 1890 images with thick clouds, 1890 images with small clouds, and 1764 images with thin clouds. Figure 1 displays a portion of the training data, whereas Fig. 2 shows a subset of the three scenes of testing data.

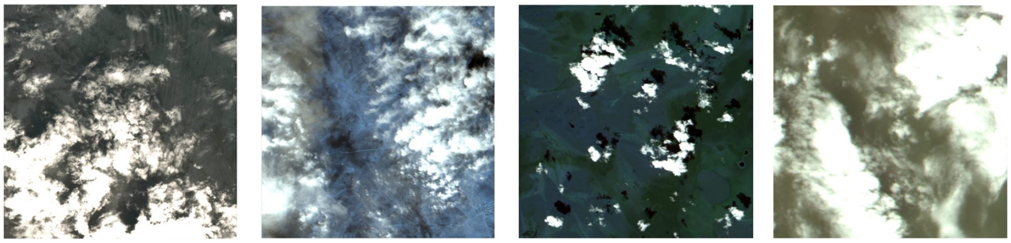

Fig. 1 Display of some training data from the GF1-WHU dataset.

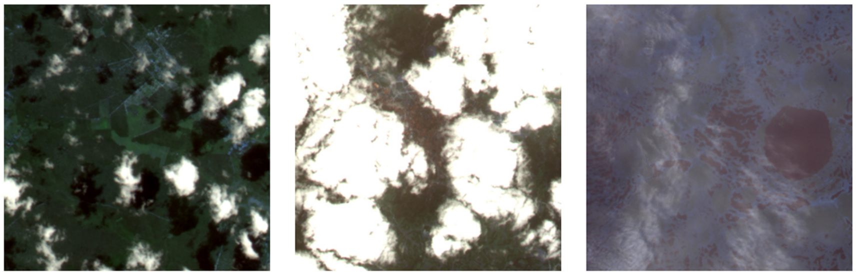

Fig. 2 Display of some testing data from the GF1-WHU dataset.

3 Proposed Method

DB-Net is an encoder–decoder-based pixel-level segmentation model that classifies each pixel as either cloud or non-cloud. The specific structure of DB-Net is shown in Fig. 3. It consists of three components: the FEM, CFEM, and FFM.

FEM has two branches dedicated to global and local cloud information extraction. CFEM is added to the Fusion layer of the Swin branch and in the middle of the CNN branch to achieve multi-scale cloud information. FFM combines information from both branches and emphasizes the extraction of cloud information while minimizing the loss of information from small and thin clouds. In addition, concatenation conv and upsample (CCU) is to fuse the input features and restore the original number of input channels. The detailed introduction of the three modules can be found in Secs. 3.1–3.3.

3.1 Feature Extraction Module by the Combination of Swin Transformer and CNN

Traditional convolutional neural networks excel at extracting local features but are limited by the fixed size of their convolutional kernels, making it challenging to effectively capture global contextual information. This limitation can impact the ability to extract features of clouds and consequently affect the accuracy of cloud detection.

In contrast to CNNs, Swin Transformer employs a hierarchical attention mechanism to efficiently handle global long-range dependencies in images. It decomposes the image into a series of small patches and utilizes multiple layers of transformer modules for information propagation and feature extraction. Within each local window, the features are encoded as vector representations and interact through a self-attention mechanism to capture global dependencies.

To fully utilize the strengths of Swin Transformer and CNNs, we employ the FEM in the encoding stage of the model. The Swin branch captures the global contextual information of images, whereas the CNN branch captures the local texture and detail information. In the FFM, the global and local information is fused.

3.1.1 *CNN branch*

The CNN branch is designed to capture local texture details of clouds, and it consists of two parts: encoding and decoding. The encoding part consists of two elements, conv block (CB) and downsample block (DB). CB is composed of a convolution, BatchNorm, rectified linear unit (ReLU) activation function, and basic block. The structure of the basic block is consistent with that of ResNet.[30] DB is composed of a downsampling operation and multiple basic blocks. In CB, the convolutional kernel has a size of 3 with a stride of 2 and a padding of 1. In DB1 to DB4, the number of basic blocks is 3, 4, 6, and 3, respectively. The CFEM is incorporated in the middle of the encoding–decoding process to perform multiscale feature extraction. In the decoding part, the features from the encoding part are concatenated and upsampled, gradually restoring them to their original sizes. In UP, the convolutional kernel has a size of 3 with a stride of 1 and a

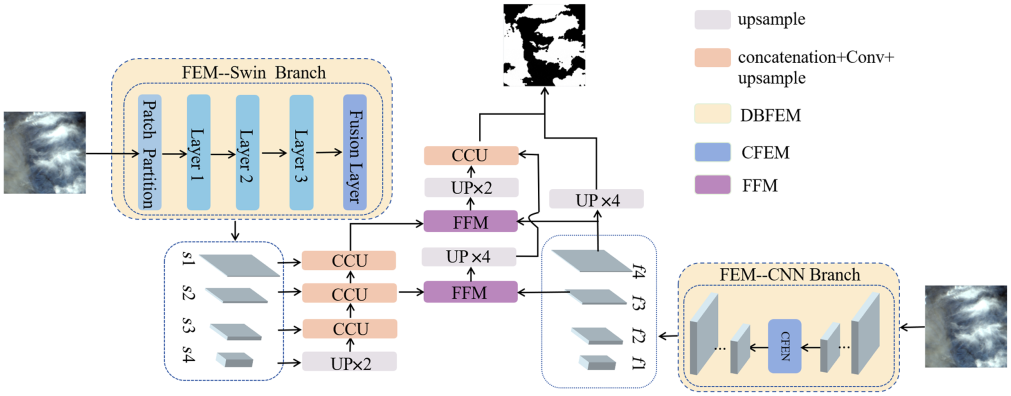

Fig. 3 Overall architecture of the proposed DB-Net.

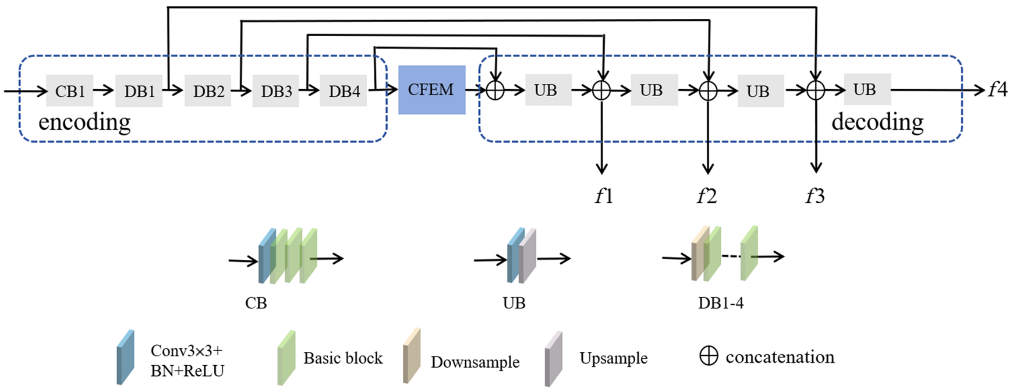

Fig. 4 Architecture of the CNN branch.

padding of 1. The feature sizes of f_1,f_2, f_3, and f_4 are 1/32, 1/16, 1/8, and 1/4 of the original size, respectively. The specific structure of the CNN branch is shown in Fig. 4.

3.1.2 *Swin branch*

The Swin branch is used to extract the global semantic information of clouds. From Fig. 5, it can be seen that the Swin branch mainly consists of patch partition, patch embedding, patch merging, Swin Transformer block, and CFEM. The Swin branch consists of four layers, each of which extracts features corresponding to different scales to capture semantic information at different levels. In our implementation, the number of Swin Transformer blocks in each layer is set to 2, 2, 6, and 2. The feature sizes of S_1, S_2, S_3, and S_4 are 1/4, 1/8, 1/16, and 1/32 of the original size.

The Swin branch divides the image into a set of patches using a patch partition. Each patch is then linearly embedded into the Swin Transformer block by patch embedding. Patch merging involves dividing the features equally and placing them in different channels, similar to a pooling operation. The CFEM is utilized for multiscale feature extraction, similar to the CNN branch. The Swin Transformer block consists of four main components: window-based multi-head self-attention, shifted window-based multi-head self-attention, multi-layer perceptron (MLP), and layer normalization. The equations are as follows:

$$\overset{\Lambda}{Y}^{l} = W - \mathrm{MSA}(\mathrm{LN}(Y^{l-1})) + Y^{l-1}, \tag{1}$$

$$Y^{l} = \mathrm{MLP}(\mathrm{LN}(\overset{\Lambda}{Y}^{l})) + \overset{\Lambda}{Y}^{l}, \tag{2}$$

$$\overset{\Lambda}{Y}^{l+1} = \mathrm{SW} - \mathrm{MSA}(\mathrm{LN}(Y^{l})) + Y^{l}, \tag{3}$$

$$Y^{l+1} = \mathrm{MLP}(\mathrm{LN}(\overset{\Lambda}{Y}^{l+1})) + \overset{\Lambda}{Y}^{l+1}. \tag{4}$$

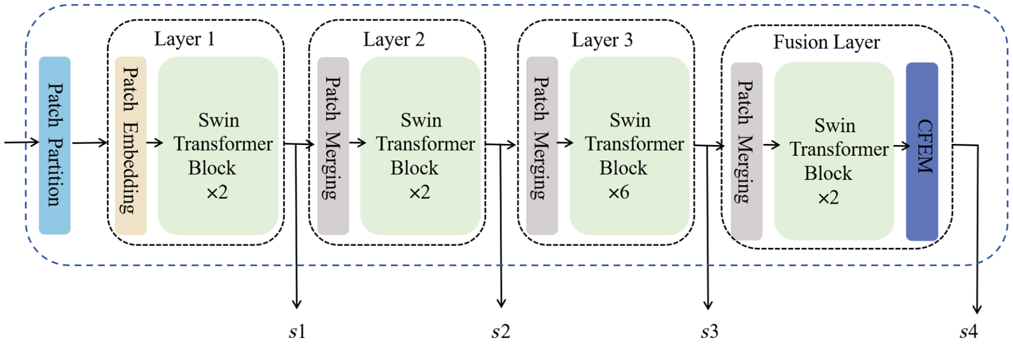

Fig. 5 Architecture of the Swin branch.

3.2 Cascaded Feature Enhancement Module Based on ASPP

Multiscale information plays a crucial role in improving the detection accuracy of clouds. General multiscale feature extraction architectures often lack effective information interaction among different scales, which can negatively impact detection performance. To address these challenges and effectively extract multiscale information, we propose the CFEM. It consists of two components: the multi-scale feature extraction module (MFC) and channel weight selection (CWS). The overall CFEM is given by Eq. (5)

$$\mathrm{CFEM} = \mathrm{CWS}(\mathrm{MFC}(x)), \tag{5}$$

and the specific structure of the CFEM is shown in Fig. 6.

The MFC component includes a pooling block, a 1 × 1 convolution, and three parallel deformable convolution (DCN)[31] modules. The pooling block takes the input features and applies both max pooling and average pooling operations. The results of these pooling operations are then element-wise added together. The resulting feature map is then upsampled using linear interpolation to restore it to the original size. DCNs use different dilation rates to capture cloud information at different scales. Compared with regular convolutions, the DCN is better suited to adapt to the shape of clouds and extract their semantic information. The regular convolution has a dilation rate of 1 and padding set to 0, whereas DCN has dilation rates and padding set to 6, 12, and 18. The kernel size of the regular convolution is 1 × 1, with a stride of 1. All DCNs have a convolution kernel size of 3 × 3 and a stride of 1. To enhance the cloud semantic information, we respectively concatenate f_1, f_2, and $f_3$with the input features x. The MFC module is defined as shown in Eq. (6)

$$\mathrm{MFC} = \begin{cases} f_1 = \mathrm{conv}_{1\times 1}(x) \\ f_2 = \mathrm{DCN}_{3\times 3}(\mathrm{concat}(x, f_1)) \\ f_3 = \mathrm{DCN}_{3\times 3}(\mathrm{concat}(x, f_2)) \\ f_4 = \mathrm{DCN}_{3\times 3}(\mathrm{concat}(x, f_3)) \\ f_5 = \mathrm{upsample}(\mathrm{add}(\mathrm{maxpool}(x), \mathrm{averagepool}(x))) \end{cases}, \tag{6}$$

where the DCN is an extension of the regular convolution that introduces additional offset terms ΔX_n, which are computed through another convolution operation. The DCN is defined as shown in Eq. (7)

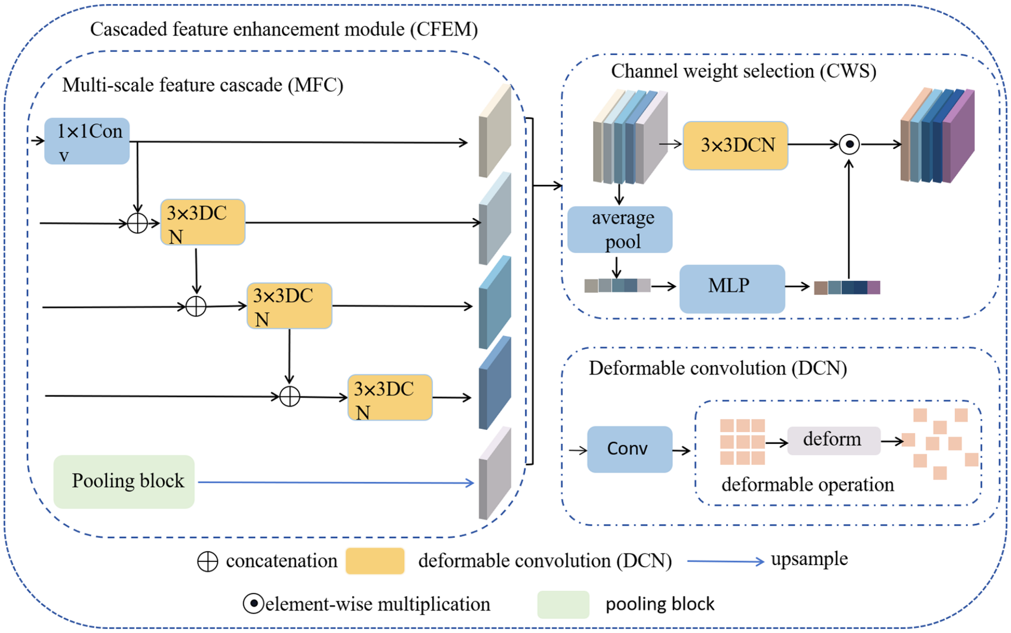
Cascaded feature enhancement module (CFEM)
Multi-scale feature cascade (MFC)
1×1Con
v
3×3DC
N
3×3DC
N
3×3DC
N
Pooling block
Channel weight selection (CWS)
3×3DCN
average
pool
MLP
Deformable convolution (DCN)
Conv
deform
deformable operation
⊕ concatenation
deformable convolution (DCN)
upsample
⊙element-wise multiplication
pooling block

Fig. 6 Architecture of CFEM.

$$\mathrm{DCN}(x_0) = \sum_{x_n \in R} w(x_n) \cdot y(x_0 + x_n + \Delta x_n), \tag{7}$$

where y represents the pixel matrix at the corresponding position of the convolution kernel, x_n represents the offset for each point on the convolutional output receptive field, x_0 represents the coordinates of the convolutional kernel center, w represents the weights corresponding to the sampling points, and R represents the set of offset coordinates.

CWS component includes a 3 × 3 DCN, average pool, and MLP. The channel concatenation in the MFC module unavoidably leads to redundant information, which can negatively impact detection performance. CWS is utilized to concentrate on the significant channel information related to clouds and to diminish information redundancy. The specific equation of CWS is shown in Eq. (8)

$$\mathrm{CWS}(x) = \mathrm{MLP}(\mathrm{averagepool}(x)) * \mathrm{DCN}_{3\times 3}(x), \tag{8}$$

where x represents the result of concatenating f_1,f_2, f_3,f_4, andf_5.

3.3 Feature Fusion Module Based on Attention Mechanism

In Secs. 3.1 and 3.2, the global contextual semantic information, local texture detail information, and multiscale information from cloud images are extracted by FEM and CFEM. However, the information contains some irrelevant noise and background information. We designed the FFM to fuse the aforementioned information and focus on foreground (cloud) information. The FFM consists of the spatial attention module (SAM) and multi-scale channel attention module (MCAM). SAM and MCAM operate on the outputs of the CNN and Swin branches, respectively. SAM is responsible for focusing on the spatial information and local texture details of clouds in the input information. MCAM extracts multiscale information from the intermediate layers and focuses on important channel information and spatial information relevant to clouds. The final output is obtained by merging the 1 × 1 convolution of the outputs from SAM and MCAM. The specific structure of the FFM is shown in Fig. 7 and Eq. (9)

$$\mathrm{FFM} = \mathrm{conv}_{1\times 1}(\mathrm{concat}((x_1 \cdot \mathrm{SAM}(x_1)), \mathrm{MCAM}(x_2))), \tag{9}$$

where x_1 represents the output of the CNN branch, x_2 represents the output of the Swin branch, and $\mathrm{conv}_{1\times 1}$ is used to change the channel dimension.

The SAM module applies max pooling and average pooling operations to the input features to extract various spatial features. Subsequently, the concatenated results undergo a 1 × 1 convolution and sigmoid activation function to obtain the spatial attention weight matrix. The specific equation of SAM is shown in Eq. (10)

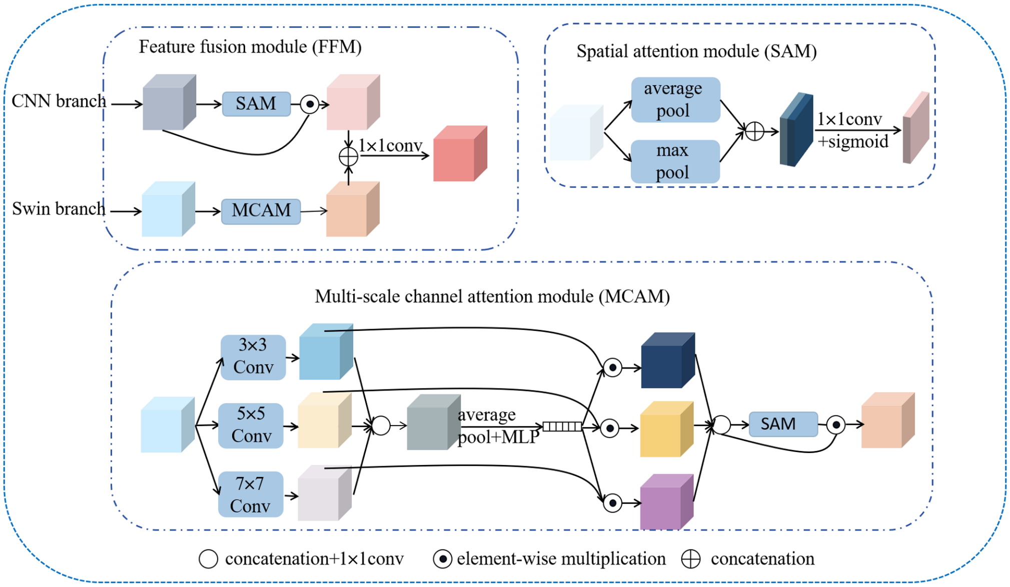

Fig. 7 Architecture of FFM.

$$\text{SAM} = \text{sigmoid}(\text{conv}_{1\times1}(\text{concat}(\text{maxpool}(x), \text{averagepool}(x)))), \tag{10}$$

where $x$represents the input features of SAM.

The MCAM module passes through three parallel convolutional modules for feature extraction at different scales. The convolutional kernel sizes are respectively set to 3×3, 5×5, and 7×7, with a stride of 1. Then, the features at multiple scales are added together to obtain the fused feature. The fused feature is then subjected to average pooling and passed through an MLP to obtain a weight vector. Next, the weight vector is multiplied element-wise with the features extracted by the three convolutions. Finally, the fused feature undergoes the SAM to achieve the final result. The specific equation of MCAM is shown in Eq. (11)

$$\begin{cases} f_1 = \text{conv}_{3\times3}(x) \\ f_2 = \text{conv}_{5\times5}(x) \\ f_3 = \text{conv}_{7\times7}(x) \\ f_4 = \text{MLP}(\text{averagepool}(\text{conv}_{1\times1}(\text{concat}(f_1, f_{2,}f_3)))) \\ f_5 = \text{conv}_{1\times1}(\text{concat}(f_1 * f_4, f_2 * f_4, f_3 * f_4)) \\ \text{MCAM} = \text{SAM}(f_5) * f_5 \end{cases}, \tag{11}$$

where x represents the input features, $f_{1\tilde{3}}$ correspond to the output features of convolutional kernels of sizes 3×3, 5×5, and 7×7, f_4 represents the weight vector, and f_5 represents the fused feature.

3.4 Loss Function

The binary cross-entropy (BCE) provides pixel-level loss supervision, ensuring accurate classification of individual pixels. The intersection over union (IOU) loss provides image-level loss supervision, reinforcing cloud pixels that were missed at the pixel level and weakening falsely detected cloud pixels. Therefore, we utilize the BCE loss in combination with the IOU loss to form our loss function for cloud detection, distinguishing between cloud and non-cloud. The equation of BCE loss is defined as Eq. (12)

$$\text{Loss}_{\text{BCE}}(y, y') = \sum_{w=1}^{W} \sum_{h=1}^{H} (y_{\text{hw}} \log(y'_{\text{hw}}) + (1 - y_{\text{hw}}) \log(1 - y'_{\text{hw}})), \tag{12}$$

where $y \in R^{W\times H\times 1}$ denotes the ground truth cloud mask and $y' \in R^{W\times H\times 1}$denotes the produced cloud mask. The equation of the IOU Loss is defined as Eq. (13)

$$\text{Loss}_{\text{IoU}}(y, y') = 1 - \frac{\sum_{w=1}^{W} \sum_{h=1}^{H} y_{\text{hw}} y'_{\text{hw}}}{\sum_{w=1}^{W} \sum_{h=1}^{H} (y_{\text{hw}} + y'_{\text{hw}} - y_{\text{hw}} y'_{\text{hw}})}, \tag{13}$$

The final Eq. (13) combines these two loss functions so that the loss function can perform loss supervision at the pixel and image level, and the specific equation is defined as Eq. (14)

$$\text{Loss}_{\text{final}} = \lambda_1 \text{Loss}_{\text{BCE}}(y, y') + \lambda_2 \text{Loss}_{\text{IoU}}(y, y'), \tag{14}$$

where λ_1 and λ_2 represent the corresponding loss function weights and we set both of them to 1.

3.5 Evaluation Metrics

To evaluate the effectiveness of our model, we utilize metrics such as accuracy, precision, recall, and $F1$. The following are the equations for the evaluation metrics

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{15}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{16}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{17}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{18}$$

where TP represents the prediction is the cloud pixel and the actual mask is also the cloud pixel, TN represents that the prediction is the non-cloud pixels and the real mask is also the non-cloud pixels, FP represents that the prediction is the cloud pixel and the real mask is non-cloud pixels, and FN represents that the prediction is non-cloud pixels and the true mask is the cloud pixels.

4 Implementation Details

All the experiments were implemented on a workstation with a single graphics processing unit (GPU) device (NVIDIA GV102: GeForce RTX 3090) running the Windows operating system. The code was written in Python 3.7 and used Torch version 1.9.1. We employed the Adam optimizer to minimize the difference between the network's predicted outputs and the ground truth values and set the hyperparameters as $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate was set to 10^{-6}. For each experiment, we trained the model for 100 epochs to obtain the final results.

5 Experimental Results and Analysis

5.1 Comparison Methods

To validate the effectiveness of the model, this study compares STCNet with BoundaryNet,[17] PspNet,[32] CDNetv2,[33] RsNet,[34] and ManfanNet.[35] BoundaryNet is a model specifically designed for accurate cloud detection, aiming to address issues of unclear cloud boundaries and low accuracy. CDNetv2 is a cloud detection model that employs attention mechanisms and multi-scale feature fusion to achieve high-precision cloud detection on satellite image thumbnails. PspNet is a neural network model used for semantic segmentation tasks, which utilizes multi-scale fusion to combine features from different levels for more comprehensive and accurate semantic information. RsNet is a cloud detection model based on UNet,[36] which effectively integrates semantic information from deep and shallow layers through skip connections, improving the effectiveness of cloud detection. MafanNet enhances the ability to extract features and deeply mine spatial information through its designed attention modules, boundary refinement enhancement model, and bilateral FFM, thereby improving the detailed repair of cloud and cloud shadow boundaries.

5.2 Comparison Experiments

All experiments were conducted on the GF1-WHU dataset. The results of different methods are shown in Fig. 8 and Table 1. In regions covered by thick clouds, as shown in Figs. 8(a) and 8(b), different models exhibit similar detection performance, except for differences in texture details. RsNet and PSPNet suffer from detail loss and edge smearing. MafanNet suffers from the adhesion of segmentation areas. CDNetv2 and BoundaryNet experience varying degrees of thin cloud and detail information loss. STCNet performs the best, reducing cloud omissions and producing clearer edges compared with other models.

In regions covered by thin clouds, as depicted in Figs. 8(c) and 8(d), DB-Net outperforms other models in detecting thin clouds and exhibits fewer instances of mistaking the background for clouds. BoundaryNet shows numerous cloud omissions, whereas CDNet and PSPNet not only have cloud omissions but also suffer from false cloud detections.

In addition, in regions covered by small clouds, as shown in Figs. 8(e) and 8(f), DB-Net exhibits better segmentation results and closer alignment with the ground truth in terms of details. BoundaryNet suffers from the loss of small clouds, whereas other models experience more severe omissions.

To further validate the effectiveness of DB-Net, we conducted quantitative analysis in different scenes. Table 1 presents the quantitative comparison results for thick, thin, and small cloud scenes. Figures 9 and 10 display the accuracy and $F1$_score for different scenes, respectively. In the three scenes, our model achieves the highest accuracy. In the thick cloud scene, the accuracy is 97.48%. In the thin cloud scene, the accuracy is 94.68%. In the small cloud scene, the accuracy is 96.55%. Compared with the highest accuracy, the accuracy improved by ~0.53%, 2.7%, and 0.46% in the above scenes.

Due to the distinct features of thick clouds and the reduced likelihood of foreground-background confusion, all metrics in the thick cloud scene are the highest among the three

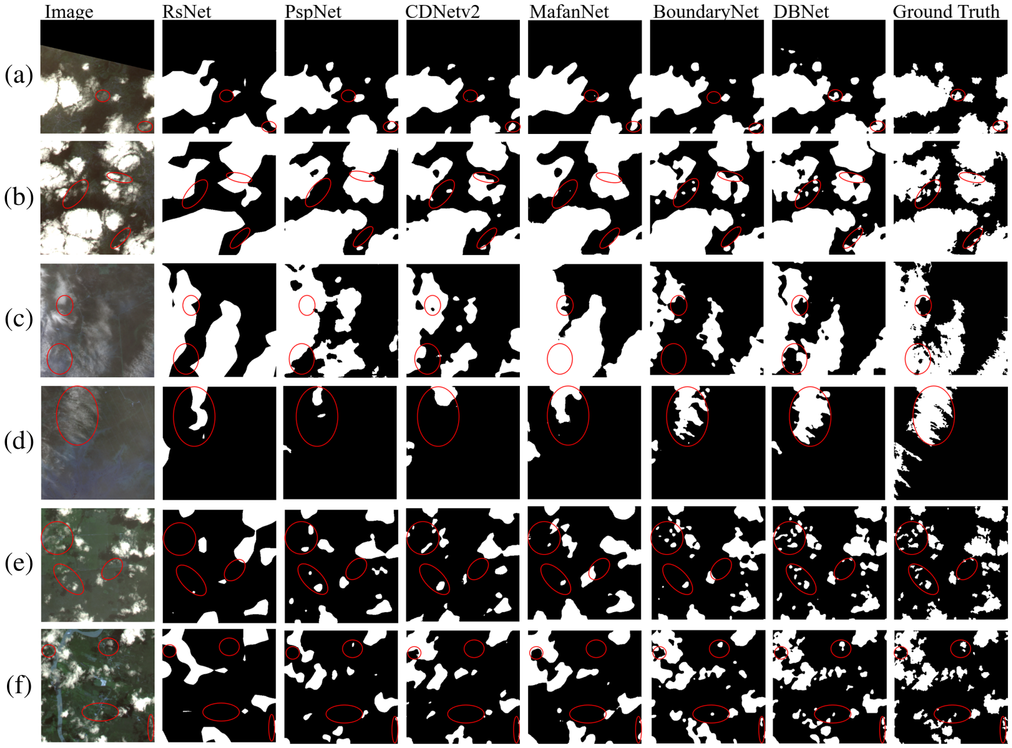

Fig. 8 Comparison of three scene images using different methods. The types of three scenes are as follows: panels (a) and (b) represent scenes with thick clouds, panels (c) and (d) represent scenes with thin clouds, and panels (e) and (f) represent scenes with small clouds.

Table 1 Accuracy, precision, recall, and $F1$ scores for each testing method in the thick, thin, and small cloud regions.

Method	Accuracy	Precision	Recall	$F1$
The thick cloud region				
DB-Net	**0.9748**	0.9676	**0.9737**	**0.9706**
BoundaryNet	0.9695	0.9545	0.9724	0.9633
CDNetv2	0.9609	**0.9781**	0.9277	0.9522
MafanNet	0.9539	0.9317	0.9607	0.9460
PspNet	0.9523	0.9203	0.9705	0.9447
RsNet	0.9498	0.9408	0.9395	0.9402
The thin cloud region				
DB-Net	**0.9468**	0.6790	0.8018	**0.7353**
MafanNet	0.9447	**0.7194**	0.6521	0.6841
BoundaryNet	0.9198	0.5425	0.7966	0.6454
CDNetv2	0.9138	0.5180	0.8654	0.6481
PspNet	0.8882	0.4452	0.8847	0.5923
RsNet	0.8868	0.4427	**0.9017**	0.5938
The small cloud region				
DB-Net	**0.9655**	0.8405	**0.8942**	**0.8665**
BoundaryNet	0.9609	0.8061	0.8847	0.8436
CDNetv2	0.9568	**0.9164**	0.7098	0.7999
PspNet	0.9546	0.8129	0.8147	0.8138
MafanNet	0.9466	0.7750	0.7915	0.7831
RsNet	0.9425	0.7773	0.7395	0.7579

Note: bold values indicate that the model performs the best on that metric.

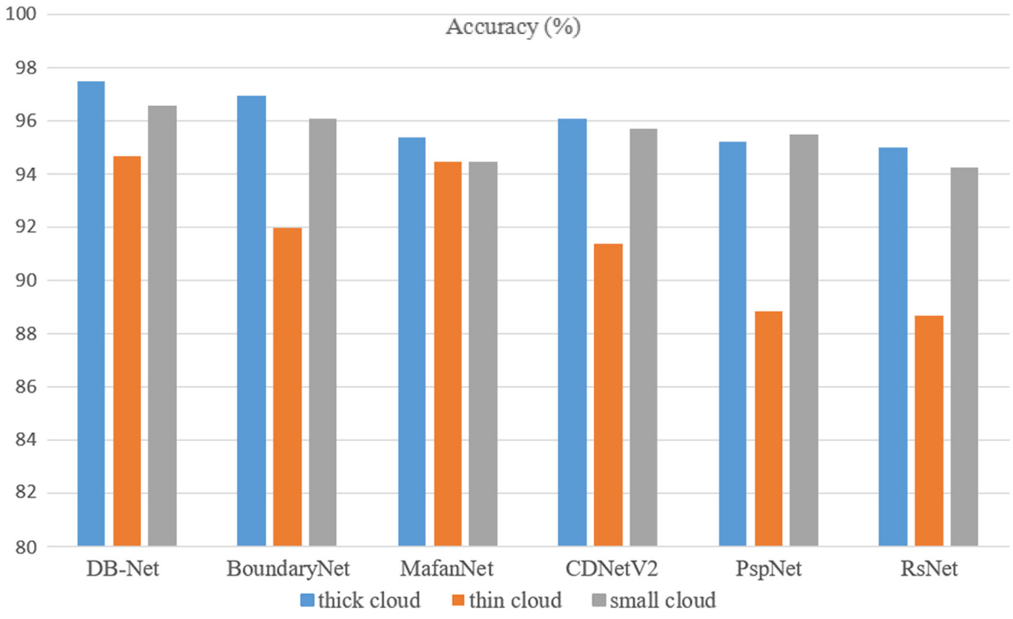

Fig. 9 Accuracy comparison for thick, thin, and small clouds under different models.

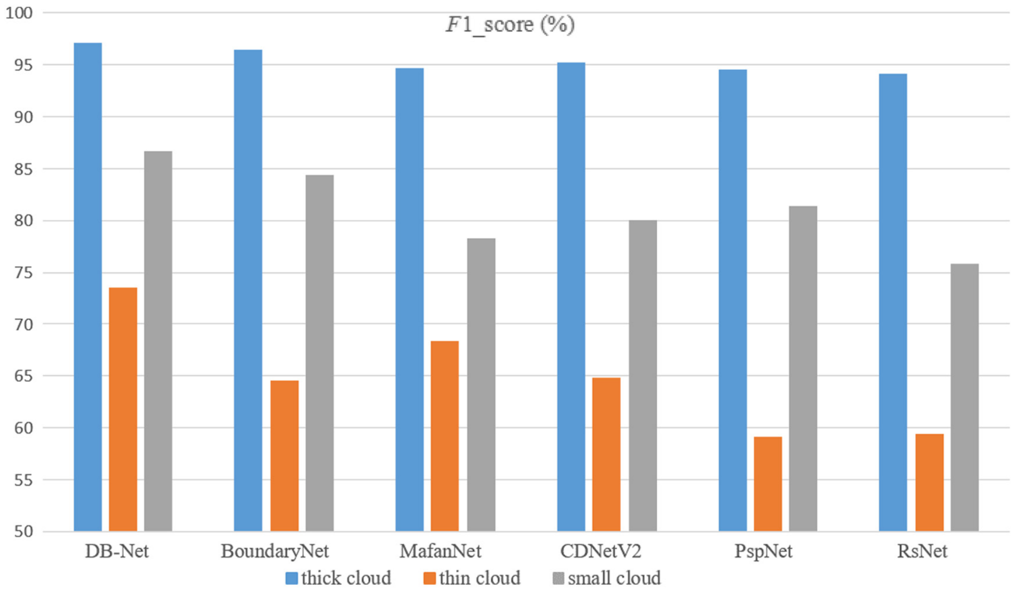

Fig. 10 $F1$_score comparison for thick, thin, and small clouds under different models.

scenes. In the thick cloud scene, CDNetv2 achieves the highest precision, albeit at the cost of sacrificing recall. This trade-off reduces false positive detections of non-cloud pixels but also decreases the accurate detection of cloud pixels.

Compared with thick clouds, thin clouds have lower reflectance and less obvious features. Thin clouds also exhibit translucency and are easily confused with background features. These factors make it difficult to differentiate between foreground and background information. As a result, all metrics in the thin cloud scene are the lowest among the three scenes. In the thin cloud scene, RsNet achieves the highest recall, but this comes at the cost of increased detection of background information as clouds. ManfanNet achieves the highest precision rate, reducing the background information while also diminishing the effectiveness of cloud detection.

Small clouds occupy very few pixels and are easily overlooked during the feature extraction process. The metrics in the small cloud scene lie between those of the thin and thick cloud scenes. In the small cloud scene, although DB-Net has lower precision compared with CDNetv2, it

Table 2 Comparative experimental results of different methods.

Method	Accuracy (%)	Precision	Recall	*F*1
DB-Net	0.9627	0.8980	0.9348	0.9160
BoundaryNet	0.9507	0.8499	0.9314	0.8903
MafanNet	0.9485	0.8751	0.8857	0.8803
CDNetv2	0.9445	0.8654	0.8969	0.8711
PspNet	0.9327	0.7925	0.9286	0.8851
RsNet	0.9268	0.8079	0.8629	0.8345

achieves a balance between precision and recall, resulting in an overall improvement in the detection of small clouds.

Table 2 presents the quantitative evaluation results for all scenes, and it can be observed that the proposed DB-Net outperforms other methods in all metrics. DB-Net achieves an accuracy of 96.27%, a precision of 89.80%, a recall of 93.48%, and an $F1$ score of 91.60%. BoundaryNet exhibits a similar recall to DB-Net, but its performance in other metrics is significantly lower compared with DB-Net. DB-Net effectively suppresses non-cloud background information while preserving the information of thin and small clouds, achieving a balance between precision and recall. The metrics of other models are considerably lower compared with DB-Net.

Overall, it can be observed that DB-Net performs the best in terms of results, effectively preserving small and thin cloud information while suppressing background information, thus achieving the best detection performance. BoundaryNet achieves detection performance that is close to DB-Net, but it has some shortcomings in terms of missing thin and small clouds. CDNetv2 and PspNet exhibit good detection performance in thick cloud scenes; however, they have significant instances of missed and false detections of clouds. MafanNet's detection performance in thin cloud areas is second only to DBNet, but there is regional adhesion in thick cloud areas, and a large amount of detail is lost in small cloud areas. In contrast, RsNet performs the worst, exhibiting significant cloud omissions, false detections, and unclear edges.

5.3 Ablation Experiments

To investigate the contributions of the three key modules in our proposed network, we conducted ablation experiments on the GF1-WHU dataset using Swin Transformer as the baseline. The results are shown in Table 3.

From Table 3, it is evident that different schemes result in improvements in model performance. Scheme 1 is based on baseline and incorporates the CFEM module, which aims to capture multi-scale information of clouds and focus on important channel information. Taking into account the local feature extraction capability of CNNs, we further add the CNN branch to scheme 1, forming scheme 2, which enhances local details and texture features. The FFM is added to scheme 2, forming scheme 3, known as DB-Net. DB-Net adaptively focuses on the semantic information of small and thin clouds while ignoring the background information of non-cloud objects.

Table 3 Quantitative analysis results of the ablation experiments.

Method	FEM	CFEM	FFM	Accuracy	Precision	Recall	*F*1
Baseline	—	—	—	0.9485	0.8764	0.9347	0.9046
Scheme 1	—	√	—	0.9532	0.8575	0.9366	0.8953
Scheme 2	√	√	—	0.9549	0.8579	0.9458	0.8997
Scheme 3	√	√	√	0.9627	0.8980	0.9348	0.9160

6 Conclusion

The accurate detection of clouds is essential in the field of weather forecasting and satellite images. In addition, cloud cover significantly affects the Earth's energy balance, and in the field of climate change monitoring, by more precisely monitoring cloud layers, scientists can better understand the impact of climate change on the global climate system. In this paper, we propose a deep learning model called DB-Net. It effectively extracts the global contextual information and local texture details of clouds, improving the accuracy of cloud detection and enhancing its ability to detect small and thin clouds. Furthermore, DB-Net mitigates the interference of background information and reduces the occurrence of false detections of certain regions as clouds. We conducted experiments on the GF1-WHU dataset and compared our results with existing cloud detection methods. The following conclusions were drawn.

1. Compared with, BoundaryNet and MafanNet, the proposed method has improved the accuracy by 1.2% and 1.42%, respectively, and the $F1$ score by 2.57% and 3.57%, respectively, demonstrating the effectiveness of our approach.
2. To address the issue of global or local information loss when using CNN or transformer individually, the FEM module is designed to extract global information while avoiding the loss of local information. To mitigate the problem of information loss in conventional multi-scale feature extraction, the CFEM module is designed to reduce the loss of multi-scale cloud information. To tackle the issue of missing thin and small clouds during the cloud detection process, the FFM module is designed to fuse dual-branch information and employs attention mechanisms to distinguish background information, focusing on small and thin cloud information.

The limitation of this research is that the proposed DB-Net adopts the dual-branch structure, which leads to a higher number of parameters and increased model complexity. This issue will be addressed in our future work.

Disclosures

The authors declare that they have no conflicts of interest regarding this research.

Code and Data Availability

The GF-1 WFV images used in this paper were provided by the China Centre for Resources Satellite Data and Application (CRESDA) and the China Land Surveying and Planning Institute (CLSPI).

Author Contributions

Conceptualization, R.Z., C.Z., H.P., and H.H.; data curation, R.Z., C.Z., and Z.H.; methodology, R.Z., C.Z., H.P., H.H., Y.H., Y.Z., and Z.H.; validation, R.Z., C.Z., and H.P.; formal analysis, R.Z., C.Z., H.P., and Z.H.; investigation, R.Z., H.P., C.Z., H.H., Y.H., J.W, S.Y., and H.Z.; writing—original draft preparation, R.Z., C.Z., H.P., and Z.H.; writing—review and editing, R.Z., C.Z., H.P., H.H., and Z.H.; supervision, R.Z., H.P., H.H., Y.Z., Y.H., J.W., S.Y., and Z.H. All authors have read and agreed to the published version of the paper.

Acknowledgments

We would like to express our gratitude for the support provided by the National Natural Science Foundation of China (Grant No. 42241164).

References

1. K. Anderson et al., "Earth observation in service of the 2030 agenda for sustainable development," *Geo-sp. Inf. Sci.* **20**(2), 77–96 (2017).
2. A. Boluwade, "Remote sensed-based rainfall estimations over the East and West Africa regions for disaster risk management," *ISPRS J. Photogramm. Remote Sens.* **167**, 305–320 (2020).
3. Y. Hu and Y. Dong, "An automatic approach for land-change detection and land updates based on integrated NDVI timing analysis and the CVAPS method with GEE support," *ISPRS J. Photogramm. Remote Sens.* **146**, 347–359 (2018)

4. J. Gao et al., "A general deep learning based framework for 3D reconstruction from multi-view stereo satellite images," *ISPRS J. Photogramm. Remote Sens.* **195**, 446–461 (2023).
5. Y. Zhang et al., "Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: refinements of the radiative transfer model and the input data," *J. Geophys. Res.: Atmos.* **109**(D19), 105–132 (2004).
6. Z. Hong et al., "Efficient global color, luminance, and contrast consistency optimization for multiple remote sensing images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **16**, 622–637 (2022).
7. F. Xie et al., "Multilevel cloud detection in remote sensing images based on deep learning," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **10**(8), 3631–3640 (2017).
8. Z. Shao et al., "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Trans. Geosci. Remote Sens.* **57**(6), 4062–4076 (2019).
9. D. Chai et al., "Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks," *Remote Sens. Environ.* **225**, 307–316 (2019)
10. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 7132–7141 (2018).
11. Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.* (2021).
12. Q. Wang et al., "ECA-Net: efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.* (2020).
13. S. Woo et al., "CBAM: convolutional block attention module," *Lect. Notes Comput. Sci.* **11211**, 3–19 (2018).
14. J. Zhang et al., "Deep network based on up and down blocks using wavelet transform and successive multi-scale spatial attention for cloud detection," *Remote Sens. Environ.* **261**, 112483 (2021).
15. J. Yu et al., "An effective cloud detection method for Gaofen-5 images via deep learning," *Remote Sens.* **12**(13), 2106 (2020).
16. Q. He et al., "DABNet: deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.* **60**, 5601216 (2021).
17. K. Wu et al., "Cloud detection with boundary nets," *ISPRS J. Photogramm. Remote Sens.* **186**, 218–231 (2022).
18. G. Guowei et al., "Multi-path multi-scale attention network for cloud and cloud shadow segmentation," *IEEE Trans. Geosci. Remote Sens.* **62**, 5404215 (2024).
19. E. Zhang et al., "Multilevel feature context semantic fusion network for cloud and cloud shadow segmentation," *J. Appl. Remote Sens.* **16**(4), 046503 (2022).
20. X. Zhang et al., "CIFNet: context information fusion network for cloud and cloud shadow detection in optical remote sensing imagery," *J. Appl. Remote Sens.* **17**(1), 016506 (2023).
21. A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," arXiv:2010.11929 (2010).
22. Z. Liu et al., "Swin Transformer: hierarchical vision Transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 9992–10002 (2021).
23. J. Chen et al., "TransUNet: transformers make strong encoders for medical image segmentation," arXiv:2102.04306 (2021).
24. C. Zhao et al., "Boundary-aware bilateral fusion network for cloud detection," *IEEE Trans. Geosci. Remote Sens.* **61**, 5403014 (2023).
25. Z. Zhang et al., "Cloudformer V2: set prior prediction and binary mask weighted network for cloud detection," *Mathematics* **10**(15), 2710 (2022).
26. Y Tan et al., "Cloud and cloud shadow detection of GF-1 images based on the Swin-UNet method," *Atmosphere* **14**(11), 1669 (2023).
27. N. Ma et al., "CNN-TransNet: a hybrid CNN-transformer network with differential feature enhancement for cloud detection," *IEEE Geosci. Remote Sens. Lett.* **20**, 1001705 (2023).
28. Y. Wu et al., "Ultrasound image segmentation method for thyroid nodules using ASPP fusion features," *IEEE Access* **8**, 172457–172466 (2020).
29. Z. Li et al., "Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery," *Remote Sens. Environ.* **191**, 342–358 (2017).
30. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. and Pattern Recognit.* (2016).
31. J. Dai et al., "Deformable convolutional networks," in *IEEE Conf. Comput. Vis. and Pattern Recognit.* (2017).
32. Z. Hengshuang et al., "Pyramid scene parsing network," in *IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 6230–6239 (2017).
33. G. Jianhua et al., "CDnetV2: CNN-based cloud detection for remote sensing imagery with cloud-snow coexistence," *IEEE Trans. Geosci. Remote Sens.* **59**(1), 700–713 (2020).

34. J. H. Jeppesen et al., “A cloud detection algorithm for satellite imagery based on deep learning,” *Remote Sens. Environ.* **229**, 247–259 (2019).
35. K. Chen et al., “Multiscale attention feature aggregation network for cloud and cloud shadow segmentation,” *IEEE Trans. Geosci. Remote Sens.* **61**, 5612216 (2023).
36. O. Ronneberger et al., “U-net: convolutional networks for biomedical image segmentation,” *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).

Ruyan Zhou received her PhD in agricultural bio-environment and energy engineering from Henan Agricultural University in 2007. Currently, she is an associate professor in the College of Information Technology, Shanghai Ocean University, Shanghai, China. Her research interests include photogrammetry and deep learning.

Chenhao Zhang is a master’s student at Shanghai Ocean University. His main research focus is deep learning–based cloud detection.

Haiyan Pan received her PhD in surveying and mapping from Tongji University, Shanghai, China, in 2020. Currently, she is a lecturer at the College of Information Technology, Shanghai Ocean University. Her research interests include multi-source data processing, multitemporal remote sensing data analysis, and deep learning.

Haiyang He received his master of science in advanced control and system engineering from the University of Manchester (U.K.). During 2019–2022, he was a vision-developing engineer in ASEA Brown Boveri (ABB). Since 2023, he has been an AI algorithm engineer at Shanghai Sensetime Intelligent Technology Co., Ltd. His research includes AI quality inspection, computer vision, multi-axis robot path planning, navigation system, and control system optimization.

Yun Zhang received his PhD in applied marine environmental studies from Tokyo University of Maritime Science and Technology, Tokyo, Japan, in 2008. From 2011 to the present, he has been a professor at the College of Information and Technology, Shanghai Ocean University, Shanghai, China. His research interests include the study of navigation system reflection signal technique and its maritime application.

Yanling Han received his BE degree in mechanical design and manufacturing, his ME degree in mechanical automation from Sichuan University, Sichuan, China, and his PhD in engineering and control theory from Shanghai University, Shanghai, China. She is a professor and is currently employed with the Shanghai Ocean University, Shanghai, China. Her research interests include the study of ocean remote sensing, flexible system modeling, and deep learning.

Jing Wang received her PhD in biomedical engineering in the Department of Biomedical Engineering of Shanghai Jiaotong University in 2014. From 2015 to the present, she has been a lecturer at the College of Information Technology, Shanghai Ocean University. Her research interests include computer vision and medical image processing.

Shuhu Yang received his PhD in physics of physics from the School of Physics, Nanjing University. Currently, he has been a lecturer at the College of Information Technology, Shanghai Ocean University since 2012. His research interests include hyperspectral remote sensing, the evolution of the Antarctic ice sheet, and the use of navigational satellite reflections.

Zhonghua Hong received his PhD in GIS from Tongji University, Shanghai, China, in 2014. Currently, he has been a professor at the College of Information Technology, Shanghai Ocean University since 2022. His research interests include satellite/aerial photogrammetry, high-speed videogrammetric, planetary mapping, 3D emergency mapping, GNSS-R, deep learning, and processing of geospatial big data.