Scene Aggregation Network for Cloud Detection on Remote Sensing Imagery

Xi Wu, and Zhenwei Shi*, *Member, IEEE,*

***Abstract*—There has been a breakthrough in cloud detection by using convolutional neural networks during these years. However, there are still weaknesses among current cloud detection algorithms because only cloud mask information is used. As clouds represent differently in different scenes, the scene information may give hints to improve cloud detection performance. Therefore, different from the previous cloud detection literature, in this letter, we propose an end-to-end new deep learning network named Scene Aggregation Network (SAN), which aggregates the scene information in the framework. Specifically, basic features are firstly extracted by utilizing all levels of network features. Then, the aggregated features used to produce the final cloud masks are created by fusing the basic features and the specially introduced scene information. Experimental results have demonstrated that with scene information aggregated, our proposed method can be robust on images with different scenes. Additionally, as SAN outperforms other state-of-the-art methods, our proposed method suits for cloud detection and can achieve improvement on this task.**

***Index Terms*—convolutional neural networks, scene information, cloud detection**

I. Introduction

Cloud detection is a very significant application in remote sensing image processing. On the one hand, clouds are common in remote sensing images since they impede the earth's surface in a large area [1]. On the other hand, clouds can be challenges in many remote sensing applications [2]. Therefore, to utilize the remote sensing images, it is necessary to add cloud detection preprocessing before any task-specific applications in remote sensing.

In recent years, this topic has been hotly discussed among researchers, and many cloud detection methods have been proposed. In tradition, physical methods and statistical methods are mainly adopted in cloud detection. For physical methods, spectral reflectance of the image bands is mainly considered [3]–[5]. A series of thresholds of band reflectance and the relationships between bands are manually designed. Another group of traditional cloud detection approaches is based on statistics [1], [6]. These methods use the image pixel's embedding features to solve a pixel-wise binary classification problem for cloud detection. Both ways can produce cloud masks with high accuracy. However, it is not easy to set proper thresholds or designing proper features for all kinds of situations. As a result, these methods may face challenges in difficult scenes.

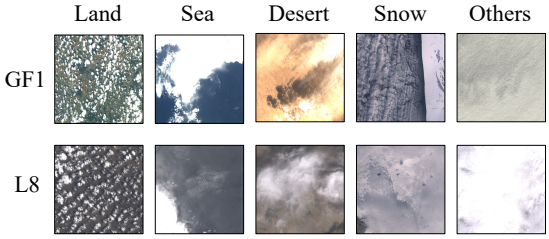

Fig. 1. Clouds represent in different forms on different scenes such as land, sea, desert, snow, and others. Scene information from the remote sensing images may give hints in the cloud detection tasks. The image sources of the first and second row are Gaofen-1(GF1) [7] and Landsat-8(L8) [8], respectively.

Recently, deep learning based methods have been widely used in remote sensing applications [9], such as road detection [10] and raft detection [11]. In the research field of cloud detection, many researchers have also been enlightened to create cloud detection methods [7], [12]–[19] based on fully convolutional networks (FCN) [20]. Generally, FCN is a framework that helps convolutional neural networks to become pixel-wise classifiers. It replaces its fully connected layers to a 1×1 convolutional layer and maintains the other parts of the networks. CloudFCN [12], MFCNN [13] and CloudSegNet [14] try to use several convolutional layers to build convolutional networks for cloud detection. Among these methods, as the network goes deep, the extracted feature maps' sizes are often shrunk. Therefore, features are upsampled and fused after deep features are obtained. Similarly, in [16], [17], the style of U-Net [21] is chosen. High-level features and low-level features are combined through concatenation layer by layer in the U-Net design. In [7], [15], the basic classification networks [22] are applied as the network backbone. Different from the previous designs, feature maps from all levels are upsampled finally before they are all fed into the final decision layer. There are also methods DeeplabV3+ [18] and DAN [19] which concentrate on the post-processing of the feature maps. In DeeplabV3+ [18], atrous spatial pyramid pooling is used to fulfill the multi-scale information of the features. In DAN [19], global features are obtained in both the spatial dimension and the channel dimension through attention mechanism. As the features extracted by CNN can be robust, therefore, the produced cloud masks are often in high accuracy.

However, since only cloud mask information is used in this task, there is still space for improving these cloud detection algorithms. These algorithms are in the typical image segmen-

The work was supported by the National Key R&D Program of China under the Grant 2019YFC1510905, the National Natural Science Foundation of China under the Grant 61671037 and the Beijing Natural Science Foundation under the Grant 4192034 (Corresponding author: Zhenwei Shi).

Xi Wu and Zhenwei Shi (Corresponding author) are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: xiwu1000@buaa.edu.cn; shizhenwei@buaa.edu.cn).

tation solutions where the information to be utilized is fixed to the original remote sensing images and their target labels. This paradigm limits the neural network only to concentrate on the relationships between the remote sensing images and the target label maps. As a result, other important information, such as scene information, is neglected. Fig. 1 illustrates that on the scenes of land and sea, clouds are often thick and without cloud shadows, while in the desert, cloud shadows may accompany with the clouds. When snow exists in the remote sensing images, clouds seem to be slightly transparent. Besides, sometimes clouds covers almost the whole image. In this case, the ground cover is unknown. Therefore, if we cannot point out the type of scene information, it indicate that the cloud cover ratio may be rather high. There also exists the previous work [23] which shows that clouds represent differently on different scenes. Above all, scene information can be used in cloud detection tasks.

Motivated by the previous works and the observations, in this paper, we introduce scene information to our proposed Scene Aggregation Network (SAN) to learn. Different from the previous literature which only process the cloud detection task, our proposed network can simultaneously process two tasks. One is providing cloud masks, and the other is classifying the corresponding scene category. With the full use of scene information, SAN can obtain improvement in cloud detection accuracy.

The contributions of our work are summarized as follows,

1)We propose a novel cloud detection network SAN that can aggregate the scene information in the cloud detection process;

2)Ablation studies are implemented to evaluate the network design of SAN and its cloud detection improvement on remote sensing images with different scene types;

3)Our proposed SAN can achieve better performances compared with other state-of-the-art cloud detection methods on two datasets.

The remainder of this paper is organized as follows. In Section.II, we then present our framework SAN. In section.III, experimental results are displayed. Finally, we conclude in Section.IV.

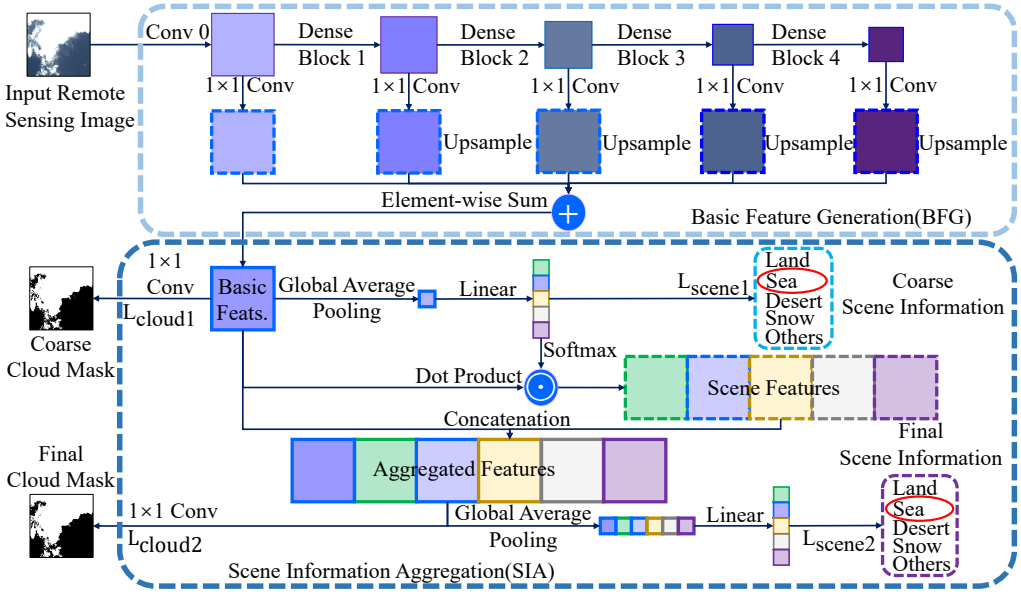

Fig. 2. (Best view in zoom and color.) An illustration of SAN. SAN combines two parts, basic feature generation(BFG) and scene information aggregation(SIA). In BFG, the basic features F_b are generated. They are the element-wise sum of the features extracted by the first convolutional layer and different dense blocks [24]. Then, in SIA, considering the scene information of the image can be hints for the cloud detection task (see Fig.1), we introduce this special information into our proposed SAN. SAN produces two types of tasks, the cloud mask generating tasks and the scene information classification tasks. During the SIA process, the basic features F_b is utilized to form scene features F_s and aggregated features F_a. The final cloud mask is generated according to F_a.

II. Scene Aggregation Network

A. Overview

The proposed cloud detection method SAN can be divided into two parts, Basic Feature Generation(BFG) and Scene Information Aggregation(SIA). In BFG, basic features F_b are generated. They are the sum of the features extracted by the first convolutional layer and different dense blocks [24]. Then, in SIA, scene information, which may give hints in cloud detection (see Fig.1), will be integrated to form the aggregated features F_a. To make our proposed SAN in an end-to-end manner, dual-classification on both the cloud and the scene information is adopted. Based on the aggregated features, the final cloud masks are generated. Fig.2 shows the framework of our proposed SAN.

B. Basic Feature Generation

The module BFG encodes the input remote sensing image into basic features. As Fig.2 shows, features are generated through five levels, Convolution 0(Conv 0) and Dense Block 1-4. In each dense block, feature maps which are generated by all the inside convolutional layers are kept through concatenation. In format, the feature maps F_l of the l^{th} layer can be denoted

as,

$$F_l = \begin{cases} concate(M_l, M_{l-1}, ..., M_1) & l \geq 2, \\ M_1 & l = 1, \end{cases} \quad (1)$$

$$M_l = g_{l-1}(F_{l-1}), \quad (2)$$

where M_l is obtained by a non-linear function $g(.)$ as Eq.2 shows. In detail, $g(.)$ is designed in the form of Bottleneck [25], which is a group of operations: BN [26]-ReLU [27]-Conv(1×1)-BN-ReLU-Conv(3×3). Fig.3. is a dense block example. In each dense block, for every single layer except the first layer, all the previous features are reused by concatenation in the channel dimension. Therefore, the problem of vanishing gradient is alleviated and the model is easy to train [24], even if the network becomes very deep.

Downsampling modules are set in BFG. After Conv0, A 3×3 max pooling layer with stride 2 follows Conv0, while transition layers are set behind Dense Block $1 - 3$. The configuration of transition layers is BN-ReLU-1×1 Conv-2×2 average Pool (stride 2). After different levels of features are generated, they are upsampled to the size of the original input remote sensing image. Finally, basic features F_b are obtained by the operation of the element-wise sum of all levels of features.

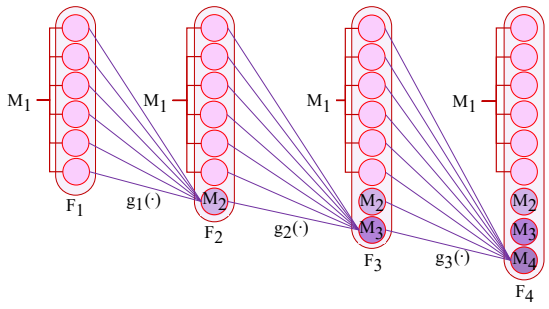

Fig. 3. An example of a 3-layer dense block. M_i and F_i $(i = 1, 2, 3, 4)$ are feature maps or parts of the feature maps. In the forward pass, they are reused in the calculation by concatenation in the channel dimension. In the training processes, the parameters of a composite function $g_i(.)$ is updated.

C. Scene Information Aggregation

The module SIA introduces scene information in the cloud detection task. Basic features obtained by the previous BFG module can deal with two tasks. Firstly, by setting a cross-entropy loss L_{cloud1} of each pixel, basic features can be used to get coarse cloud masks after a 1×1 convolution. Secondly, after a global average pooling layer and a linear layer, the input image's scene category can be classified by the basic features. Similarly to the cloud detection task, scene information classification loss L_{scene1} is also set as a cross-entropy loss. Therefore, coarse scene information can be obtained.

However, scene information participates in the calculation only in the training phase. In the testing phase, the cloud detection task and the scene category classification task are separated. Therefore, we aggregate the scene information in further processes. In the calculation of the coarse scene information, we can obtain the scores s_c of different types of scenes, where $c \in 1, 2, ..., C$ and C is the number of types of scenes. After s_c is normalized by a softmax operation to form the probability of each type of scene p_c, shown as Eq.3. The scene probability p_c can be viewed as the knowledge of the scene categories. Therefore, the scene features F_s, which can be viewed as the integration of scene information and the basic features, can be defined as Eq.4 shows.

$$p_c = \frac{exp(s_c)}{\sum_{t=1}^{C} exp(s_t)}. \quad (3)$$

$$F_s = concate(p_1 \cdot F_b, p_2 \cdot F_b, ..., p_C \cdot F_b). \quad (4)$$

Then, aggregated features F_a, which is a combination of the basic features F_b and the scene features F_s, can be calculated in Eq.5.

$$F_a = concate(F_b, F_s). \quad (5)$$

Aggregated features F_a can be used to produce the final cloud masks and scene information like basic features F_b. However, the former will be more representative since scene information is deeply merged in these features. To generate the final cloud masks and scene information, in the training phase, the loss function of cloud detection L_{cloud2} and scene classification L_{scene2} are cross-entropy losses as the previous loss settings. Although L_{scene1} and L_{scene2} are similar in the formula, L_{scene2} goes a step further to help F_a to learn a feature representation of generating cloud masks of different scenes. Experiments of ablation studies (Section III.C) have shown that adding L_{scene2} in the training phase can improve the cloud detection results.

The total loss of our SAN is a linear combination of the above four losses, L_{cloud1}, L_{scene1}, L_{cloud2}, L_{scene2}:

$$\begin{aligned} L_{total} =& \lambda_{cloud1} \cdot L_{cloud1} + \lambda_{scene1} \cdot L_{scene1} \\ &+ \lambda_{cloud2} \cdot L_{cloud2} + \lambda_{scene2} \cdot L_{scene2}, \end{aligned} \quad (6)$$

where λ_{cloud1}, λ_{scene1}, λ_{cloud2}, λ_{scene2} are weight balance parameters. Our SAN is an end-to-end cloud detection framework, and both two tasks, the cloud detection task and the scene classification task, can be trained simultaneously.

III. Experiments

A. Dataset and Evaluation Metrics

TABLE I
STATISTICS OF THE TWO DATASETS FOR CLOUD DETECTION: 1)GF-1 WFV [7], 2)LANDSAT-8 [8]

Item	GF-1 WFV [7]	Landsat-8 [8]
# images in training set	6432	3431
# images in testing set	1600	919
# images of catagory: land, sea, desert, snow, others	2408,948,3017, 750,909	1180,500,547, 233,1890

In this study, two datasets are used to conduct the experiments and evaluate our proposed method SAN. The first dataset is collected from [7], and its source is the Gaofen-1 wide field of view images(GF-1 WFV). The source of the other dataset is Landsat-8 [8]. All the images on both datasets

are cropped to 300×300 and are manually classified into five types of scenes: *land, sea, desert, snow, and others*. Statistics of these two datasets are shown in Table I. 'Intersection over Union' (IoU), F1-score, OA and AA are widely used metrics in pixel-wise classification and can be used to evaluate the cloud detection accuracy [7], [15], therefore, they are chosen to be our evaluation metric.

B. Experiment Setup

Our proposed method SAN is implemented with PyTorch 1.4 on Ubuntu 16.04 and an NVIDIA Geforce GTX 1080Ti GPU card. The BFG module of our proposed SAN is mainly built on dense blocks. We refer to the configuration of Densenet-169 [24] to setup Conv0 and Dense Block 1-4 in BFG. In the training phase, we fine-tuned layers in the module of BFG by Densenet-169 [24] pre-trained models on Imagenet [28] and parameters of the other layers of SAN are randomly initialized. As for the weights balancing parameters λ_{cloud1}, λ_{scene1}, λ_{cloud2}, λ_{scene2}, we set them all to 1. The other training settings, which are the same as other comparison methods, are listed as follows. The training method is the stochastic gradient descent method (SGD). The initial learning rate is 0.001 and its policy is set as 'poly'. The value of momentum is set as 0.9. The training batch size is set as 4 and the number of epochs is 100. In the testing phase, only the final cloud masks are used for evaluation.

C. Ablation Studies and Comparison with Other Methods

The ablation study aims to analyze the significance of designed modules of our proposed SAN, which is a combination of BFG and SIA. It should be noticed that after the module of BFG, the extracted basic features F_b can be used for cloud detection only through minimizing L_{cloud1}, with the other parts of SIA neglected. Therefore, we evaluate four types of design on both datasets: 1) using only BFG through minimizing L_{cloud1}; 2) using only BFG through minimizing L_{cloud1} and L_{scene1}; 3) using both BFG and SIA, but not minimizing L_{scene2}; 4) using both BFG and the whole SIA. Cloud detection results of the above four types of network design are shown in Table.II. We can observe the effectiveness by using BFG and the whole SIA with minimizing all the losses.

TABLE II
ABLATION STUDY ON THE NETWORK DESIGN.(IOU%)

Design	GF1 [7]	L8 [8]
BFG + L_{cloud1}	86.89	87.38
BFG + L_{cloud1} + L_{scene1}	86.97	86.86
BFG + L_{cloud1} + L_{scene1} + L_{cloud2}	87.44	87.69
BFG + SIA	**87.75**	**88.18**

Furthermore, we also extend the evaluation of the proposed SAN's cloud detection robustness improvement on different kinds of scenes. Table.III shows the results, which implies that with scene information aggregated, the aggregated features F_a can be more robust for the cloud detection tasks.

TABLE III
CLOUD DETECTION IMPROVEMENT ON DIFFERENT KINDS OF SCENES.(IOU%)

	Land	Sea	Desert	Snow	Others	All
Dataset: GF1-WFV [7]						
BFG+L_{cloud1}	**83.26**	77.37	**72.51**	69.70	95.03	86.89
BFG+SIA	81.78	**80.20**	70.56	**76.23**	**96.74**	**87.75**
Δ	-1.52	2.83	-2.05	6.53	1.71	0.86
Dataset: Landsat-8 [8]						
BFG+L_{cloud1}	**85.20**	**71.06**	81.67	45.15	**91.18**	87.38
BFG+SIA	85.16	70.67	**86.83**	**69.74**	91.06	**88.18**
Δ	-0.04	-0.39	5.16	24.59	-0.12	0.80

We compare our SAN with some other state-of-the-art cloud detection methods finally. CloudSegNet [14], UNet [16], CloudFCN [12] are three cloud detection methods based on deep learning. DeeplabV3+ [18] and DAN [19] are popular image segmentation methods which can also be used to detect clouds. Besides, we change the form of BFG as a layer-by-layer decoding structure similar to U-Net [21] and denote this structure as "SAN_L". The same DenseNet-169 [24] backbone is used in DAN [19], DeeplabV3+ [18], SAN_L and SAN. Pretrained models on ImageNet [28] are also used in these methods. Quantitative results are recorded in Table.IV and visual comparisons are illustrated in Fig.4. All these outcomes suggest SAN has a good performance on cloud detection.

TABLE IV
QUANTITATIVE COMPARISONS WITH OTHER STATE-OF-THE-ART CLOUD DETECTION METHODS ON BOTH DATASETS.(%)

Method	IoU	F1	OA	AA
Dataset: GF1-WFV [7]				
CloudSegNet [14]	73.34	84.62	94.71	92.55
UNet [16]	82.54	90.44	96.59	94.73
CloudFCN [12]	83.40	90.95	96.71	94.29
DAN [19]	78.12	87.72	95.60	92.84
DeeplabV3+ [18]	85.15	91.98	97.16	95.90
SAN_L(ours)	86.93	93.01	97.45	95.53
SAN(ours)	**87.75**	**93.48**	**97.65**	**96.16**
Dataset: Landsat-8 [8]				
CloudSegNet [14]	71.08	83.10	85.07	87.76
UNet [16]	76.95	86.98	89.11	89.35
CloudFCN [12]	77.89	87.57	88.69	88.46
DAN [19]	84.23	91.44	92.35	92.10
DeeplabV3+ [18]	87.94	93.58	94.37	94.21
SAN_L(ours)	87.04	93.07	93.79	93.54
SAN(ours)	**88.18**	**93.72**	**94.54**	**94.45**

IV. CONCLUSION AND PERSPECTIVES

In this paper, we propose Scene Aggregation Network(SAN) that detects clouds on remote sensing images. Our SAN is an end-to-end cloud detection framework and can process both tasks, the cloud detection task and the scene classification task. With scene information of the remote sensing image aggregated, the cloud detection accuracy can be improved. Experimental results have verified the effectiveness of SAN. In future work, we will continue improving the efficiency of cloud detection.

REFERENCES

[1] Z. An and Z. Shi, "Scene learning for cloud detection on remote-sensing images." *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 4206–4222, 2015.

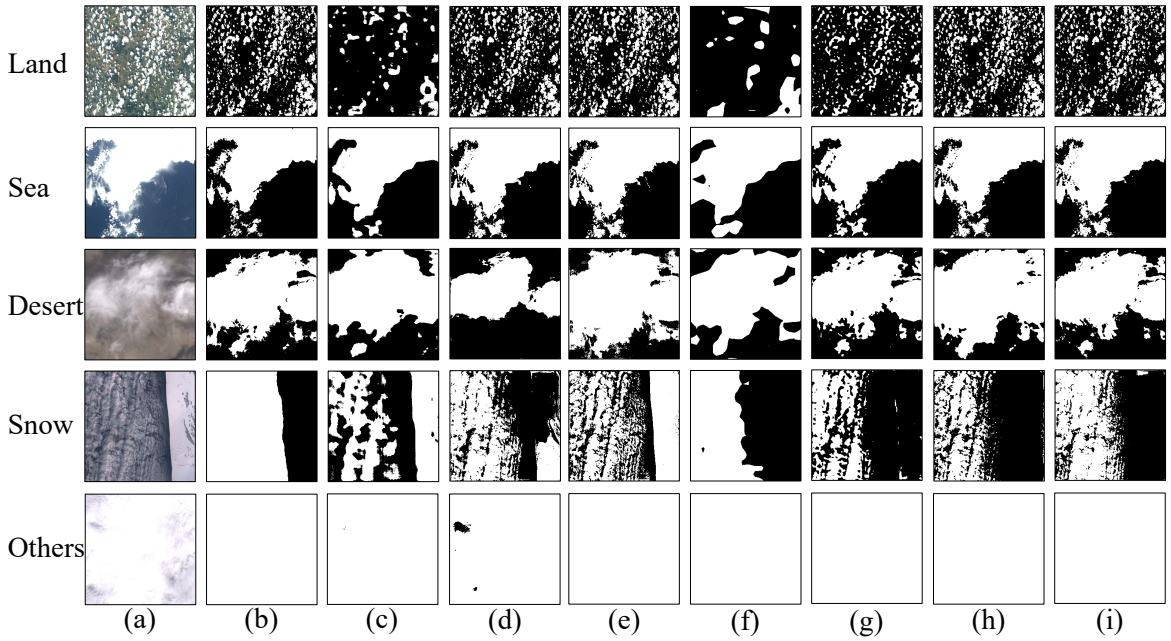

Fig. 4. (Best view in zoom and color). Visual comparisons of using different cloud detection methods. (a) RGB image of the original image. (b) Ground truth labels (black means background while white represents the cloud area.). (c−i) are the cloud masks of CloudSegNet [14], UNet [16], CloudFCN [12], DAN [19], DeeplabV3+ [18], SAN_L and SAN, respectively.

[2] M. Zhou, Z. Zou, Z. Shi, W.-J. Zeng, and J. Gui, "Local attention networks for occluded airplane detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 3, pp. 381–385, 2019.

[3] R. R. Irish, "Landsat 7 automatic cloud cover assessment," in *Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI*, vol. 4049. International Society for Optics and Photonics, 2000, pp. 348–355.

[4] Z. Zhu and C. E. Woodcock, "Object-based cloud and cloud shadow detection in landsat imagery," *Remote Sens. Environ.*, vol. 118, pp. 83–94, 2012.

[5] M. Lu, F. Li, B. Zhan, H. Li, X. Yang, X. Lu, and H. Xiao, "An improved cloud detection method for gf-4 imagery," *Remote Sens.*, vol. 12, no. 9, p. 1525, 2020.

[6] X. Kang, G. Gao, Q. Hao, and S. Li, "A coarse-to-fine method for cloud detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 110–114, 2018.

[7] X. Wu and Z. Shi, "Utilizing multilevel features for cloud detection on satellite imagery." *Remote Sens.*, vol. 10, no. 11, p. 1853, 2018.

[8] S. Foga, P. L. Scaramuzza, S. Guo, Z. Zhu, R. D. Dilley Jr, T. Beckmann, G. L. Schmidt, J. L. Dwyer, M. J. Hughes, and B. Laue, "Cloud detection algorithm comparison and validation for operational landsat data products," *Remote Sens. Environ.*, vol. 194, pp. 379–390, 2017.

[9] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, 2016.

[10] M. Lan, Y. Zhang, L. Zhang, and B. Du, "Global context based automatic road segmentation via dilated convolutional neural network," *Inf. Sci.*, 2020.

[11] T. Shi, Q. Xu, Z. Zou, and Z. Shi, "Automatic raft labeling for remote sensing images via dual-scale homogeneous convolutional neural network," *Remote Sens.*, vol. 10, no. 7, p. 1130, 2018.

[12] A. Francis, P. Sidiropoulos, and J.-P. Muller, "Cloudfcn: Accurate and robust cloud detection for satellite imagery with deep learning," *Remote Sens.*, vol. 11, no. 19, p. 2312, 2019.

[13] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4062–4076, 2019.

[14] S. Dev, A. Nautiyal, Y. H. Lee, and S. Winkler, "Cloudsegnet: A deep network for nychthemeron cloud image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1814–1818, 2019.

[15] Y. Zhan, J. Wang, J. Shi, G. Cheng, L. Yao, and W. Sun, "Distinguishing cloud and snow in satellite images via deep convolutional network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1785–1789, 2017.

[16] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.*, vol. 229, pp. 247–259, 2019.

[17] Y. Guo, X. Cao, B. Liu, and M. Gao, "Cloud detection for satellite imagery using attention-based u-net convolutional neural network," *Symmetry*, vol. 12, no. 6, p. 1056, 2020.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[19] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] L. Sun, X. Mi, J. Wei, J. Wang, X. Tian, H. Yu, and P. Gan, "A cloud detection algorithm-generating method for remote sensing data at visible to short-wave infrared wavelengths," *ISPRS J. Photogramm. Remote Sens.*, vol. 124, pp. 70–88, 2017.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[27] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.