An All-scale Feature Fusion Network with Boundary Point Prediction for Cloud Detection

Wenjing Wang, and Zhenwei Shi*, *Member, IEEE*

***Abstract*—Cloud detection is a significant pre-processing for remote sensing images. In recent years, many methods based on deep learning are proposed to detect clouds and multi-scale feature fusion is often used in these methods. However, most existing methods fuse features through concatenation and element-wise summation, which are simple and can be improved in spatial information recovery. Therefore, we explore the way of fusing features to recover the missing spatial information more sufficiently. Besides, we also observe that some cloud detection results are not accurate enough near the boundary of clouds. In view of the above observations, in this paper, we propose a cloud detection network, ABNet, which includes All-scale feature Fusion modules and a Boundary point Prediction module. The All-scale feature Fusion module can optimize the features and recover spatial information by integrating features of all scales. And the Boundary point Prediction module further remedies cloud boundary information by classifying the cloud boundary points separately. Experimental results demonstrate that our method improves the accuracy of cloud detection compared with other methods.**

***Index Terms*—cloud detection, feature fusion, boundary points**

I. Introduction

Cloud detection is an essential pre-processing for remote sensing images. Clouds prevent optical satellite sensors from obtaining clear ground information and decrease the visibility of images, affecting the subsequent processing and application of remote sensing images. Therefore, it is worthwhile to investigate cloud detection to address cloud coverage problems [1], [2].

In the past few decades, many researchers have studied and developed cloud detection methods based on spectral threshold [3], [4]. These methods calculate threshold by various characteristics such as cloud reflectivity and brightness. Besides, machine learning is also being used for cloud detection [5], which extracts features such as cloud texture, color, and geometric, and then designs and trains classifiers. However, these methods require setting appropriate thresholds or designing various features manually for different instances, which demand a lot of specialized knowledge and are not robust enough.

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 62125102 and the Beijing Natural Science Foundation under the Grant 4192034 (Corresponding author: Zhenwei Shi).

Wenjing Wang, 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:wenjingwang@buaa.edu.cn; shizhenwei@buaa.edu.cn).

CNN(Convolutional Neural Network) possesses a powerful ability to learn proper features, then does not require manual feature selection. Accordingly, many methods using CNN are proposed to detect the clouds, mainly based on fully convolutional networks(FCN) [6]. Most cloud detection methods using FCN carry out down-sampling and up-sampling operations many times, sometimes known as the encoding stage and decoding stage. The down-sampling operations result in the loss of spatial information. Consequently, researchers often consider how to recover spatial information in the up-sampling stage.

As a classical encoder-decoder structure, the U-Net [7] fuses the corresponding features of the encoding and decoding processes by skip-connection to retrieve spatial information. CS-CNN [8], RS-Net [9], CloudFCN [10] are cloud detection models based on the U-net architecture.

In addition to skip-connection, multi-scale feature fusion is also widely used to optimize features and remedy the missing spatial information. MF-CNN [11] and CDnet [12] use feature pyramid modules. FECN [13] and MSCFF [14] resample different scale features to output sizes and then fuse them.

In multi-scale feature fusion methods, however, most existing methods fuse features through concatenation [11], [13], [14] or element-wise summation [12], [14]. These feature fusion methods merely perform fixed linear aggregations of feature maps [15], which are simple and could be less effective [16]. Therefore, they could not be the best choice and still have room for improvement in the degree of spatial information recovery. Moreover, we also observe that some cloud mask results are not accurate enough near the boundary of the clouds. Compared with ground truth, some boundary detection results have missed or redundant detection errors. As a result, it is necessary to investigate how to fuse features to recover the missing spatial information more sufficiently and further remedy cloud boundary information.

Inspired by the previous research and our observations, in this paper, we propose a novel cloud detection network, ABNet, which includes All-scale feature Fusion modules and a Boundary point Prediction module. The AF modules can fill in the missing information of the down-sampling process. They are based on the error feedback mechanism from the deep back projection technology. By effectively integrating the features of all levels, the modules make better use of information of each resolution. The BP module also serves to remedy the information of cloud boundary, which extracts boundary point features and classifies them. The module improves the cloud boundary accuracy by predicting the boundary points separately.

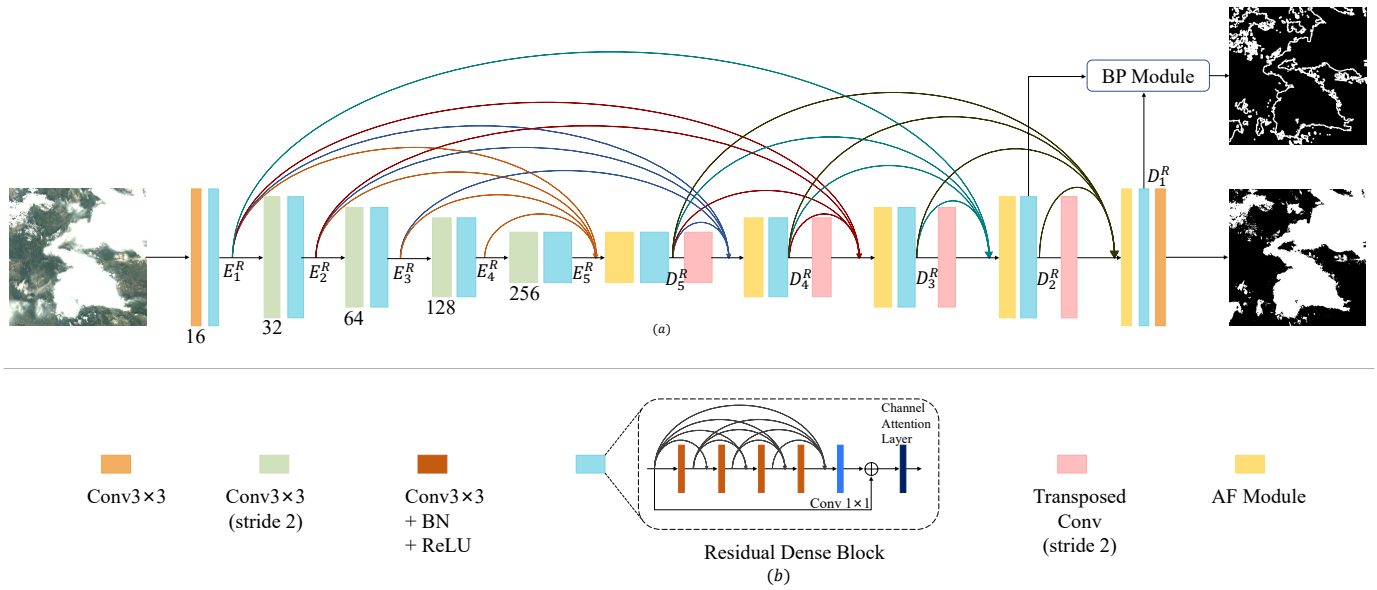

Fig. 1. The workflow of the proposed network. Fig. 1(a) indicates the network with the encoding stage and decoding stage. Fig. 1(b) shows a basic convolution block, Residual Dense Block(RDB). The AF module exists in each decoder level, and its inputs come from the features of the encoder levels lower than it and the features of the decoding levels higher than it. 16-256 represents the number of channels in each level.

The contributions of this work are summarized as follows,

1)We propose a novel cloud detection network called ABNet, design a new feature fusion method named All-scale Feature Fusion module and enhance the utilization of cloud boundary information by a Boundary point Prediction module;

2)We experiment and analyze that the AF modules and the BP module. The experimental results have shown the AF modules can improve the accuracy of cloud detection by recovering spatial information more sufficiently. The BP module can remedy erroneous boundary points and generate more accurate cloud masks;

3)Our proposed network performs better than other state-of-the-art methods on two cloud detection datasets.

II. PROPOSED METHOD

A. Overview

Fig. 1 shows the workflow of the proposed method. The ABNet contains an encoder and a decoder containing the AF modules and the BP module. The encoder is mainly composed of Residual Dense Blocks and convolutional layers. In the encoding stage, the convolutional layers are gradually deepened and the feature maps learn high-level semantic information. The decoder is mainly comprised of the AF modules, Residual Dense Blocks, transposed convolutional layers, and the BP module. In the decoding stage, the feature maps gradually recover and integrate the information of full scales to help the recovery of details. The AF module exists in each decoder level, and its inputs come from the features of the encoder levels lower than it and the features of the decoding layers higher than it. Take for example the AF module of D_3 level, which fuses the inputs from $E_1^R, E_2^R, D_4^R, D_5^R$. The curved arrows in Fig. 1(a) represent the inputs of the AF modules, i.e., the features to be fused by the AF modules. As a basic block of convolution, the Residual Dense Block(RDB) is composed of four dense connected convolutional layers and a channel attention layer [17]. Besides, the outputs of RDB in D_1^R and D_2^R layers are entered into the BP module, then concatenate and obtain the prediction of cloud boundary points.

B. All-scale Feature Fusion Module

As shown in Fig. 1, we add the AF module in each level of the decoder. The AF module enables each level in the decoder to integrate the features of all resolutions. And it is designed based on the error feedback mechanism from the deep back projection technology [18]. The deep back projection, derived from super-resolution reconstruction. It guides image reconstruction by learning the reconstruction error between the low-resolution input l_{t-1} and the generated high-resolution result h_0^t. The reconstructed image is h^t, and the reconstruction process is shown in Eq. 2.

$$h_0^t = U(l^{t-1}) \tag{1}$$

$$h^t = h_0^t + U(D(h_0^t) - l^{t-1}) \tag{2}$$

where U means up-sampling and D means down-sampling.

Inspired by this, we use the error feedback mechanism to design the fusion way of different scale features. Specifically, the module first obtains the differences between the current level and other levels through convolution or deconvolution and then integrates these differences back to the original features. The process can enhance the features of the current level by fusing low-level features and high-level features. As a result, the AF module achieves better feature fusion and makes better use of the feature information. And the ablation experiment also proves the effectiveness of the module.

The AF module of the i-th layer in the decoder is defined by Eq. 3.

$$D_i^4 = AF(D_i^0, \{D_5^R, \ldots, D_{i+1}^R, E_{i-1}^R, \ldots, E_1^R\}) \tag{3}$$

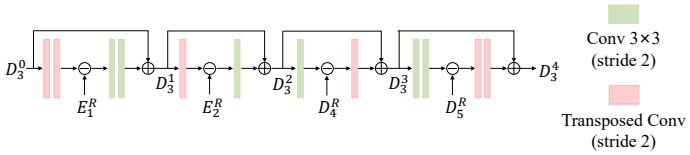

Fig. 2. The feature fusion process of the AF module from the third layer decoder D_3. In D_i^t, t represents the t-th feature fusion ($t = 0, \ldots, 4$), and i represents the feature level ($i = 1, \ldots, 5$). $E_1^R, \ldots, E_{i-1}^R, D_{i+1}^R, \ldots, D_5^R$ are the RDB outputs of the encoder and decoder.

where D_i^0 is the feature before feature fusions. D_i^4 is the feature through four feature fusions by the AP module ($i = 1, \ldots, 5$, i is the feature level). The levels of feature fusion include the high-level features ($D_5^R, \ldots, D_{i+1}^R$) from the RDB output of decoders and the low-level features ($E_{i-1}^R, \ldots, E_1^R$) from the RDB output of encoders.

$$D_i^t = \begin{cases} D_i^{t-1} + [(D_i^{t-1}) \uparrow_{i-j} - E_j^R] \downarrow_{i-j} & , i > j \\ D_i^{t-1} + [(D_i^{t-1}) \downarrow_{j-i} - D_j^R] \uparrow_{j-i} & , i < j \end{cases} \tag{4}$$

Eq. 4 shows how D_i^t fuses the features in the AP module, where $\uparrow_{i-j}$ represents the up-sampling $i - j$ times through deconvolution, $\downarrow_{j-i}$ represents the down-sampling $j - i$ times through convolution whose stride is 2 ($j = 1, \ldots, 5, j \neq i$). And t means the t-th fusion in the AF module ($t = 1, 2, 3, 4$). There is an example of D_3 layer, which is illustrated in Fig. 2.

In the AF module, the features of each resolution are combined with low-level features of the encoder and high-level features of the decoder. The features of each scale in the decoding stage get information from the features of other scales to recover the missing spatial information. The exchange of information can utilize information more adequately, and both high-resolution and low-resolution feature representations in the decoder are strengthened.

C. Boundary Point Prediction Module

The segmentation results of cloud boundary regions are affected by the quality of labels and the network segmentation performance. In addition to trying to keep the labels accurate, we also seek to improve network performance to achieve more accurate cloud boundary detection.

To increase boundary accuracy, we can pay more attention to the cloud boundaries than the interior areas. Accordingly, the Boundary point Prediction module is introduced to predict the boundaries individually.

The cloud segmentation mask is a binary image. After considering the characteristic, we use the Sobel operation [19] and dilation to get a bunch of boundary point coordinates. Then the point features are extracted by the dot product of the boundary point image and the feature map D_1^R and D_2^R. Then they concatenate and pass through two 1×1 convolution layers to predict the corresponding outputs point by point. By the above steps, the BP module corrects the classification of boundary details. The flow chart is shown in Fig. 3. As a result, the loss function of the network contains two parts, the loss function L_{cloud} for cloud detection of the image and the loss

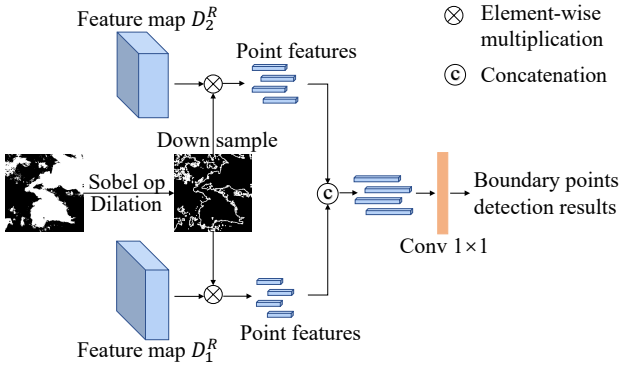

Fig. 3. The flow chart of the BP module which predicts the boundaries individually. The point features extracted from boundary point coordinates concatenate and predict boundary point detection results. Subsequently, we calculate the loss with the boundary points extracted by ground truth.

function $L_{boundary}$ for boundary points features. They both are cross-entropy losses, and the formula is as follows:

$$L_{total} = \lambda_1 L_{cloud} + \lambda_2 L_{boundary} \tag{5}$$

where λ_1 and λ_2 are weight parameters.

III. Experiments

A. Datasets

We evaluate the proposed algorithm with two datasets. The first dataset is from [13], which is collected from GF-1 images, and the source of the second dataset [20] is Landsat8. Both datasets contain forests, grasslands, deserts, coastal, snow, and so on. All of the images are cropped to 256×256. Finally, we obtain 5796 patches in the first data set and 6054 patches in the other data set. 80 percent of these images were used for training and 20 percent for testing.

B. Experiment Setup

Our method is implemented with PyTorch 1.4 on CentOS 7.6 and a Tesla V100 GPU card and optimized by the adaptive moment estimation (Adam [21]). We use 'poly' as the learning rate policy with the initial learning rate of 0.0001. Besides, the number of epochs is 100 in training, where the batch size is set to 4. For the weight parameters λ_1 and λ_2, we set them both to 1. In the experiments, mean Intersection Over Union (mIOU), Overall Accuracy (OA), F1 Score, and Kappa are selected as evaluation matrices to quantitatively evaluate the performance of cloud detection networks.

C. Ablation Studies

The purpose of the ablation study is to evaluate the effectiveness and contribution of the two modules. We set up the following comparison experiments:1) Neither module is used, but simple skip connections between encoder and decoder replace all-scale feature fusion at each level. 2) Only the BP module is used. 3) Only the AF modules are used. 4) Both the AF and BP modules are used. The experimental results of the above design schemes on GF1-WFV dataset [13] are shown in Table. I, which proves the effectiveness of the two modules.

The AF modules improve the cloud detection results more by effective feature fusion.

TABLE I
ABLATION STUDY FOR TWO DESIGN MODULES (%)

Number	Method	mIoU	OA	F1	Kappa
1	Our CNN	85.03	95.15	91.67	83.28
2	Our CNN+BP	85.86	95.43	92.19	84.31
3	Our CNN+AF	88.70	96.47	93.87	87.72
4	Our CNN+BP+AF	**89.44**	**96.77**	**94.29**	**88.58**

Figure. 4 shows three examples from Experiment 3 and 4. Judging by the visual performance, the detection of cloud boundary has been improved with the BP module. By learning and predicting the point features near the cloud boundary, the BP module reduces missed or redundant detection errors. Consequently, the network with added the BP module achieves better cloud boundary detection and predicts results closer to the ground truth.

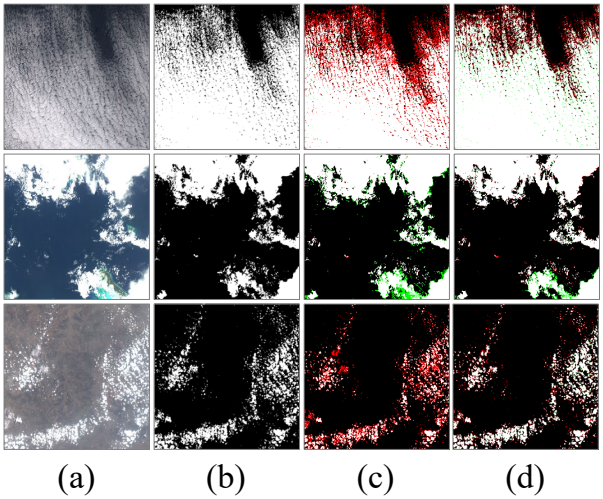

Fig. 4. Visual comparisons of the effects of the BP module. (a) RGB image. (b) Ground truth labels. (c) The cloud masks of Experiment 3, Our CNN+AF. (d) The cloud masks of Experiment 4, Our CNN+BP+AF. Black means the background, white represents the cloud, red means the cloud misclassified as the background and green means the background misclassified as the cloud.

Figure. 5 shows the visualization of feature maps with or without AF modules. The contrast between Figure. 5 (b) and Figure. 5 (c) demonstrates that the feature representations are optimized and the spatial information is recovered more sufficiently after all-scale feature fusion.

D. Comparisons with Other Methods

We evaluate our ABNet with some other representative methods. UNet [9], DeeplabV3+ [22] and HRNet [23] are popular semantic segmentation methods that can achieve cloud detection. MF-CNN [24], RSNet [25] and CDnetV2 [26] are cloud detection algorithms based on deep learning in recent years. We train and test the above methods on the same datasets with the same parameter settings.

Table. II shows the quantitative assessment results on the GF1-WFV dataset [13]. And Table. III shows the results on the Landsat-8 dataset [20]. Fig. 6 shows the visual performance of

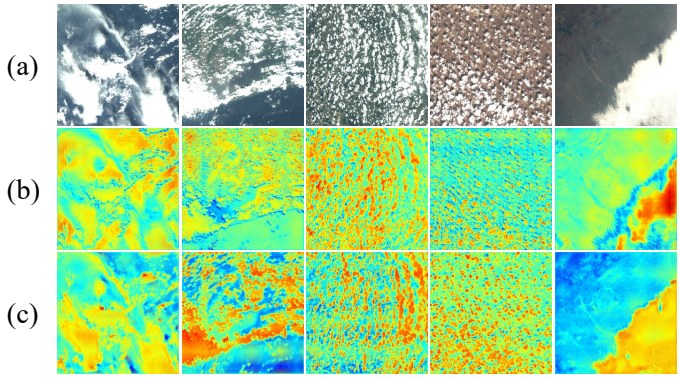

Fig. 5. Visual comparisons of feature maps with or without AF modules. (a) RGB image. (b) The feature map without all-scale feature fusion from D_1 of Experiment 2. (c) The feature map after all-scale feature fusion from D_1 of Experiment 4. The brighter the color, the greater the value. The feature representations are optimized after all-scale feature fusion.

different methods. All these methods can detect most clouds, but our method performs better for the texture and details of clouds and is more accurate near cloud boundary. Besides, our method has fewer missed or false detection results for remote sensing images with complex conditions. Therefore, it can be seen that our method has a better performance in cloud detection.

TABLE II
QUANTITATIVE COMPARISONS WITH OTHER STATE-OF-THE-ART CLOUD DETECTION METHODS ON GF1-WFV DATASET [13].(%)

Method	mIoU	OA	F1	Kappa
UNet [9]	84.88	95.18	91.54	83.08
MF-CNN [24]	82.50	94.18	90.09	80.09
RSNet [25]	88.07	96.15	93.59	87.00
DeeplabV3+ [22]	88.51	96.49	93.75	87.49
CDnetV2 [26]	88.88	96.67	94.00	87.92
HRNet [23]	89.06	96.54	94.13	88.16
Ours	**89.44**	**96.77**	**94.29**	**88.58**

TABLE III
QUANTITATIVE COMPARISONS WITH OTHER STATE-OF-THE-ART CLOUD DETECTION METHODS ON LANDSAT-8 DATASETS [20].(%)

Method	mIoU	OA	F1	Kappa
UNet [9]	77.16	87.16	87.36	74.28
MF-CNN [24]	79.48	88.88	88.79	77.13
RSNet [25]	87.16	93.22	93.14	86.27
DeeplabV3+ [22]	84.05	91.37	91.52	82.69
CDnetV2 [26]	83.04	90.87	90.72	81.44
HRNet [23]	88.40	93.88	93.92	87.69
Ours	**88.94**	**94.20**	**94.19**	**88.29**

IV. CONCLUSION AND PERSPECTIVES

In this paper, we propose a cloud detection network that contains All-scale feature Fusion modules and a Boundary point Prediction module. The AF modules based on the error feedback mechanism integrate the features of full scales. And they recover spatial information more adequately than concatenation or element-wise summation. The BP module achieves more accurate cloud boundary detection by calculating the extra loss with the cloud boundary points. The experiment results have proved the effectiveness of our network.

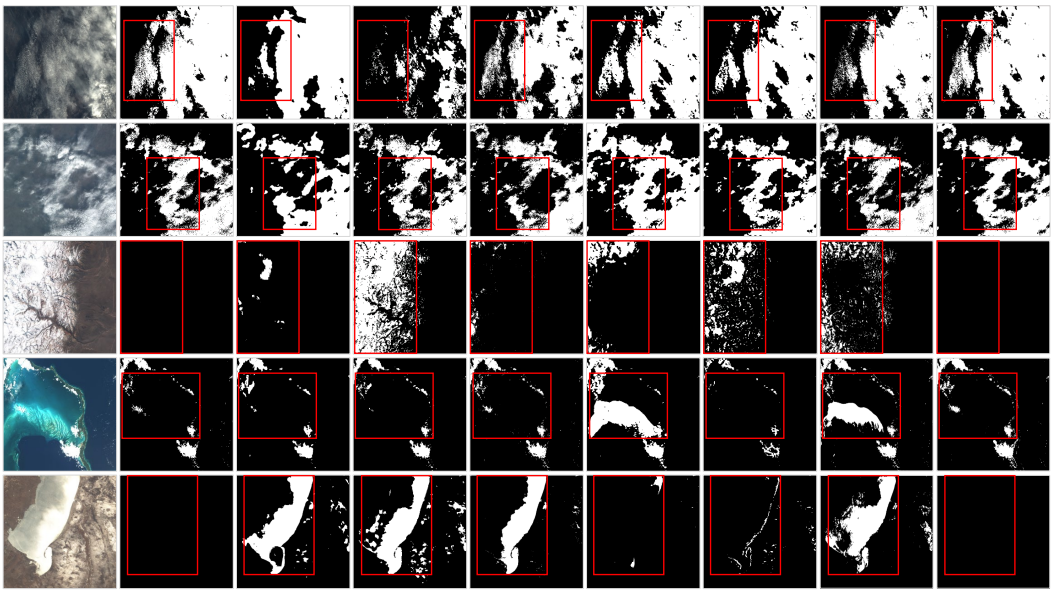

Fig. 6. Visual comparisons of different cloud detection methods. (a) RGB image. (b) Ground truth labels. (c) UNet [9]. (d) MF-CNN [24]. (e) RSNet [25]. (f) DeeplabV3+ [22].(g)CDnetV2 [26]. (h) HRNet [23]. (i) Ours.

REFERENCES

[1] X. Wu, Z. Shi, and Z. Zou, "A geographic information-driven method and a new large scale dataset for remote sensing cloud/snow detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 174, pp. 87–104, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271621000290

[2] W. Li, Z. Zou, and Z. Shi, "Deep matting for cloud detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–13, 2020.

[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] 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.

[6] 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.

[7] 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.

[8] J. Drönner, N. Korfhage, S. Egli, M. Mühling, B. Thies, J. Bendix, B. Freisleben, and B. Seeger, "Fast Cloud Segmentation Using Convolutional Neural Networks," *Remote Sensing*, vol. 10, no. 11, 2018.

[9] 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.

[10] 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.

[11] 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.

[12] J. Yang, J. Guo, H. Yue, Z. Liu, H. Hu, and K. Li, "CDnet: CNN-Based Cloud Detection for Remote Sensing Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6195–6211, 2019.

[13] X. Wu and Z. Shi, "Utilizing Multilevel Features for Cloud Detection on Satellite Imagery." *Remote Sens.*, vol. 10, no. 11, p. 1853, 2018.

[14] Z. Li, H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He, "Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 197–212, 2019.

[15] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3560–3569.

[16] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[17] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," pp. 11 531–11 539, 2020.

[18] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep Back-Projection Networks for Single Image Super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[19] O. R. Vincent, O. Folorunso *et al.*, "A descriptive algorithm for sobel image edge detection," in *Proceedings of informing science & IT education conference (InSITE)*, vol. 40. Informing Science Institute California, 2009, pp. 97–107.

[20] 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.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] 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.

[23] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5686–5696.

[24] 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.

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

[26] J. Guo, J. Yang, H. Yue, H. Tan, and K. Li, "CDnetV2: CNN-Based Cloud Detection for Remote Sensing Imagery with Cloud-Snow Coexistence," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–14, 2020.