Edge-SAR-Assisted Multimodal Fusion for Enhanced Cloud Removal

Zhenyu Wen, *Senior Member, IEEE*, Jiahui Suo, Jie Su, Bingning Li, and Yejian Zhou

***Abstract*— In Earth observation activities, cloud severely affects the interpretation of high-resolution imagery, generated by optical satellites. Therefore, removing clouds from optical imagery becomes a topic of interest in the remote sensing field. Currently, most methods use auxiliary synthetic aperture radar (SAR) images to reconstruct optical images by merging SAR and optical images into a deep learning network. However, the speckle noise of the SAR image is not taken into consideration during feature fusion processing, leading to blurry edges in the reconstructed optical images. To get fine-grained optical images, we propose a novel cloud removal framework based on the edge fusion of SAR and optical images. First, the edge feature of SAR images is extracted by the GRHED. As the prior knowledge, it can provide fine-grained edge information for subsequent reconstruction work. Then channels from three modal data are stacked to guide the reconstruction of optical images by exploiting their correlations and interactions. Furthermore, a structural similarity (SSIM) loss function is introduced to optimize the training network and improve the coherence of the image structure. Experimental results confirm its advantages on the SEN12MS-CR dataset.**

***Index Terms*— Cloud removal, deep learning, edge extraction, feature fusion.**

I. Introduction

WITH the development of remote sensing technology, high-resolution images generated by optical sensors are available to support some comprehensive Earth observation applications, such as micro-object detection and fine-grained semantic segmentation [1], [2]. However, coupled with spaceborne optical imaging, the cloud cover becomes a big challenge to interpret this kind of image. From the existing works, there is more than 55% land covered by the cloud in spaceborne optical images [3]. As a result, the contrast of optical imagery reduces, and the crucial area in the image is obscured. For many remote sensing applications that rely on continuously monitored data streams, such as agricultural measurement and disaster monitoring [4], [5], cloud cover is a serious obstacle. Therefore, to ensure data quality and availability, cloud removal is obviously essential for various applications.

The electromagnetic wave used for synthetic aperture radar (SAR) imaging is not affected by clouds due to its excellent penetration capability [6], [7], [8], thus largely mitigating the challenge of cloud removal. SAR images can provide texture and structural information of cloud-covered areas, aiding in the reconstruction of contaminated optical images. However, since SAR lacks spectrally resolved measurements, certain domain-specific potentials and peculiarities remain uncompensated. On the other hand, optical images contribute spectral and spatial information to the contaminated regions, making the generated images more consistent with ground truth. Therefore, the complementary nature of SAR and optical images in remote sensing enables more comprehensive, accurate, and clear image restoration results in cloud removal tasks. Bermudez et al. [9] trained cGANs to learn the mapping between SAR and optical images, attempting to generate optical images directly from SAR images. However, due to the different imaging principles of the two modalities, the quality of the generated optical images could not be guaranteed. Grohnfeldt et al. [10] introduced a novel cGAN-based approach for declouding by fusing SAR and optical imagery. They expanded the network's ability to read and fuse multimodal remote sensing data. Gao et al. [11] used a two-step approach to remove cloud, where they first transformed SAR images into simulated optical images using CNN and then fused SAR images, simulated optical images, and corrupted optical images to reconstruct cloud-covered areas using GAN. But the reconstructed image has spectrum deviation and loss of texture. Meraner et al. [12] used a deep residual neural network to eliminate clouds by fusing SAR and optical images. However, this approach is inadequate for areas with thick cloud where local information is entirely lost. It will lead to low-quality image reconstruction. While these methods can reconstruct contaminated images, they ignore the fact that speckle noise in SAR images may distort edge information in optical data. This can cause blurred edges in cloud-free images and hinder the recovery of fine details, especially for complex textures.

Inspired by [13], this letter proposes a cloud removal method based on the fusion of multimodal data (SAR-edge-optical). We use the GRHED [14] edge extractor to extract SAR edges and incorporate multimodal data as the new input for a deep residual neural network. This algorithm can

Manuscript received 4 March 2023; revised 3 August 2023; accepted 17 August 2023. Date of publication 21 August 2023; date of current version 4 September 2023. This work was supported in part by the National Natural Sciences Foundation of China under Grant 62101494 and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY23F010012. *(Corresponding author: Yejian Zhou.)*

Zhenyu Wen is with the Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China, and also with the School of Information Science and Technology, University of Science and Technology of China, Anhui 230026, China.

Jiahui Suo is with the Zhejiang University of Technology, Hangzhou 310023, China.

Jie Su is with the School of Computing, Newcastle University, NE45AX Newcastle Upon Tyne, U.K.

Bingning Li is with the Xi'an Satellite Control Center, Xi'an 710071, China.

Yejian Zhou is with the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China, and also with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: yjzhou25@zjut.edu.cn).

Digital Object Identifier 10.1109/LGRS.2023.3307240

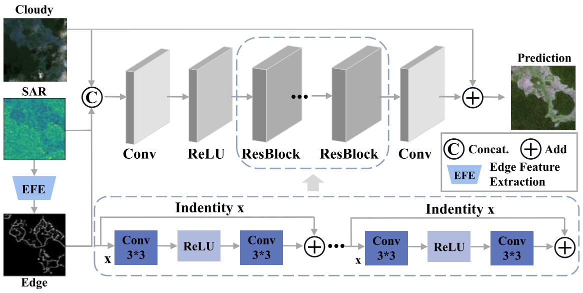

Fig. 1. Overview of the proposed framework.

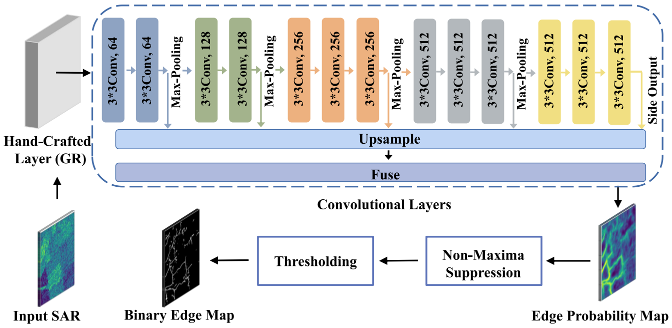

Fig. 2. Architecture of EFE. The ReLU layer following the convolutional layer is not displayed in an effort to simplify the network.

effectively transfer complementary information from SAR and edge maps to optical images, overcoming the impact of SAR speckle noise and generating reliable texture details.

The main contributions of this letter are as follows.

1) We propose a novel framework for cloud removal. The fine-grained SAR edge is first extracted and then multimodal data are fused to reduce the negative impact of speckle noise of SAR images on image reconstruction.
2) To mitigate the issue of producing blurred images when training a network with the L_1 loss function, we introduce a loss function called structural similarity (SSIM) loss, which preserves fine-grained edge information by minimizing the difference in structure between the original and reconstructed images.

II. Methodology

The proposed cloud removal method consists of two parts: edge feature extraction (EFE) of SAR images and a ResNet-based deep feature extraction network. In EFE, we adopt a CNN-based edge detector GRHED and obtain binary edge maps with suitable thresholds. For the feature extraction network, the multimodal data are fused and fed into the convolutional network to extract hyperfeature maps, which are then input into the ResNet-based network to reconstruct the cloud-covered areas. Fig. 1 illustrates the overall framework.

A. SAR Edge Extraction

The pipeline of EFE is shown in Fig. 2. The SAR images are input into a hand-crafted layer to generate gradient feature maps. The hand-crafted layer is defined by GR [15], i.e., a ratio-based gradient computation method. Computing GR can be seen as a method of data augmentation, which enables learning fewer distribution types while keeping the total amount of data unchanged.

For a given pixel located at position (x, y) in the image I, the horizontal and vertical gradient components (GR) can be computed as follows:

$$\begin{aligned} G^h(x, y) &= \log\big(R^h(x, y)\big) \\ G^v(x, y) &= \log(R^v(x, y)). \end{aligned} \tag{1}$$

$R^h(x, y)$ is calculated as follows, and the computation of $R^v(x, y)$ can be performed in a similar manner:

$$R^h(x, y) = \frac{m_1^h(x, y)}{m_2^h(x, y)} \tag{2}$$

where

$$\begin{aligned} m_1^h(x, y) &= \sum_{x'=-W}^{W} \sum_{y'=1}^{W} u\big(x + x', y + y'\big) \times e^{-\frac{|x'|+|y'|}{\alpha}} \\ m_2^h(x, y) &= \sum_{x'=-W}^{W} \sum_{y'=-W}^{-1} u\big(x + x', y + y'\big) \times e^{-\frac{|x'|+|y'|}{\alpha}} \end{aligned} \tag{3}$$

and where W is the upper integer part of $\log(10) \times \alpha$.

The magnitude $G_{\mathrm{gr}}(x, y)$ and orientation $\mathrm{ang}_{\mathrm{gr}}(x, y)$ of GR can be defined as follows:

$$\begin{aligned} G_{\mathrm{gr}}(x, y) &= \sqrt{G^h(x, y)^2 + G^v(x, y)^2} \\ \mathrm{ang}_{\mathrm{gr}}(x, y) &= \arctan \frac{G^v(x, y)}{G^h(x, y)}. \end{aligned} \tag{4}$$

Subsequent testing is performed on gradient feature maps computed with GR. The network backbone is derived from the convolutional layers of HED [16], which are transformed from the VGG16 network [17]. Specifically, the original fully connected layers and the final max-pooling layer are discarded, and the side outputs are added after five convolutional layers. The convolutional layers could automatically learn abundant hierarchical representations guided by deep supervision of the side outputs, which is crucial for solving the challenging ambiguity in the detection of SAR edges. As the network goes more profound, the scales of side outputs become smaller. To fuse features from multiple scales, bilinear interpolation [18] is used for upsampling the side outputs to the desired size. Simultaneously, a weighted-fusion layer is introduced to learn automatically how to fuse these five side outputs optimally and obtain a fused output.

The final output is the average of all the outputs which include five side outputs and one fused output. Afterward, we use the nonmaximum suppression method [19] and then obtain a binary edge map with an appropriate threshold.

As the dataset used does not have an edge ground truth (refer to Section III-A for dataset details), training could not be carried out. Therefore, we use pretrained weights to extract edge maps from the SAR images.

B. Image Reconstruction Network

The edge map obtained in EFE will be merged with the SAR and cloudy optical image as the new input to the network.

As shown in Fig. 1, we use ResNet [20] as the network backbone, which consists of multiple ResBlocks. Each ResBlock is composed of two cascaded convolutional layers, a ReLU function, and a shortcut connection. The shortcut connection achieves an additive identity mapping. In addition, we can stack a variety of ResBlocks to achieve a deeper network, extracting more deep-layer features.

Let x be the input, with $H(x)$ being an expected output, the residual mapping learned by multiple convolutional layers can be defined as follows:

$$F(x) = H(x) - x. \tag{5}$$

The residual mapping learned by the network is a correction to the pixels of the input cloud-covered image. For scenarios with thick clouds, i.e., clouds with high visual opacity [21], the correction increases as the cloud thickness grows. In cloud-free conditions, the information can be directly propagated from the input to the output through the residual skip connections without modification, maximizing the preservation of the original image information.

In general, at the beginning of the network, the image's SAR channels and SAR edge channels are simply concatenated to the other channels of the input optical image. Shallow features are first extracted using convolutional layers. Then, the network is transformed into nonlinearity with the use of the ReLU activation function. Rich global features are extracted through densely connected ResBlocks. SAR images with their edge maps help compensate for the missing information in the blurred areas. Finally, a 3 × 3 convolution restores the dimension of features to match the optical image dimension.

C. Loss Function

Assume that the predicted image is X, the ground truth is Y and the input image is I. L_1 (average absolute error) is used as the basic error function, as defined below

$$L_1 = \frac{\|X - Y\|_1}{N} \tag{6}$$

where N is the total number of pixels.

To preserve the original information of the noncovered areas to the greatest extent, the cloud mask (M) of the cloud-covered optical image is extracted and it is incorporated into the calculation of the loss function. This loss function is defined as cloud-adaptive regularized loss [12], i.e., L_{CARL}

$$L_{\text{CARL}} = \frac{\|M \odot (X - Y) + (1 - M) \odot (X - I)\|_1}{N} + L_1. \tag{7}$$

SSIM measures the similarity between two images and evaluates quality based on the degradation of structural information

$$\text{SSIM} = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{\left(\mu_X^2 + \mu_Y^2 + C_1\right)\left(\sigma_X^2 + \sigma_Y^2 + C_2\right)} \tag{8}$$

where μ, σ, and σ_{XY} denote the mean, variance, and covariance of X and Y, respectively. In addition, use $C_1 = 0.01^2$ and $C_2 = 0.03^2$ to avoid zero numerator or denominator.

The SSIM loss is highly sensitive to local structural changes and assigns a higher weight to the boundaries, thus resulting in higher losses in the vicinity of the boundaries. The SSIM loss function can be expressed as follows:

$$L_{\text{SSIM}} = 1 - \frac{1}{N}\sum_{p=1}^{\text{N}} \text{SSIM}(p) \tag{9}$$

where p is the center pixel of an image patch. The size of the patch and Gaussian filter is 11 × 11.

To obtain cloud-free images with clear boundaries, a custom loss function L_D is defined as follows:

$$L_D = L_{\text{CARL}} + L_{\text{SSIM}}. \tag{10}$$

III. Experiments

A. Dataset and Evaluation Metrics

The SEN12MS-CR [22] dataset is adopted in this work, which is the first publicly available dataset for Earth observation cloud removal. It provides large-scale global and seasonal coverage. The dataset comprises four seasonal subdatasets and a total of 169 regions of interest. It includes the corresponding Sentinel-1 dual-pol SAR data, Sentinel-2 13-band cloud-free optical data, and cloud-afflicted optical data, each with a size of 256 × 256. The SAR data include two polarizations, namely, VV and VH. To validate the effectiveness of our framework, we select 40 regions of interest from the spring subdataset and divide them into train, validation, and test sets at the ratio of 36:2:2. We evaluate the cloud removal performance by four widely used metrics: peak signal-to-noise ratio (PSNR), SSIM, mean absolute error (MAE), and spectral angle mapper (SAM). The pixel-level reconstruction performance of the image is evaluated with two metrics, PSNR and MAE. SSIM reflects the spatial structure recovery based on visual perception principles, while SAM indicates the preservation of spectral information in the reconstruction results. Higher PSNR and SSIM values, as well as lower MAE and SAM values, indicate higher image quality in the reconstruction.

B. Implementation Details

The proposed framework is implemented using publicly available PyTorch. The batch size is set to 28, and the maximum epoch of training iterations is set to 100 to reach convergence. The learning rate is 7×10^{-5}, and the number of ResBlock is 16. The experiments are all performed on an NVIDIA A100 80 GB PCIe.

C. Comparative Analysis and Visualization

To assess the performance of the proposed method with varying cloud presence, optical images with cloud coverage and shadow areas ranging from 0% to 20%, 20% to 40%, 40% to 60%, and 60% to 80% are selected.

The results of McGAN, SpA GAN, DSen2-CR, and Ours on the SEN12MS-CR dataset are shown in Fig. 3. From left to right, each column represents cloudy optical images, SAR images, the images generated by McGAN, SpA GAN, DSen2-CR, Ours, and the cloud-free optical images. McGAN and SpA GAN tend to remove smaller and scattered thin

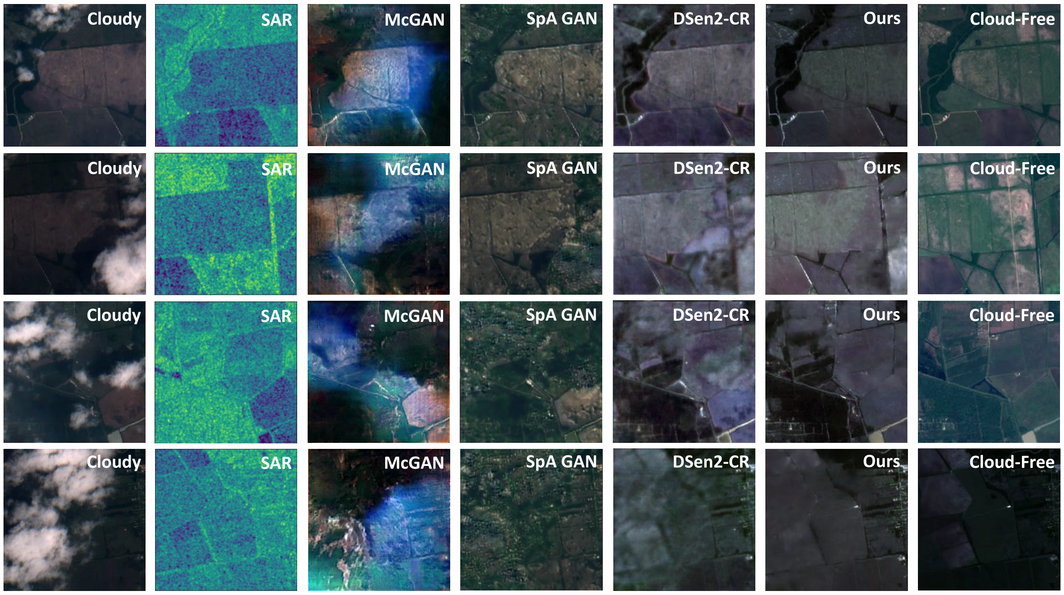

Fig. 3. Qualitative results of cloud removal for from different scenes. The first column on the far left, from top to bottom represents cloudy images with cloud and cloud shadow coverage ranging from 0% to 20%, 20% to 40%, 40% to 60%, and 60% to 80%, respectively.

TABLE I
QUANTITATIVE COMPARISONS OF PROPOSED METHOD TO OTHER METHODS

Method	PSNR ↑	SSIM ↑	MAE ↓	SAM ↓
McGAN[23]	11.9742	0.4074	0.2174	27.6590
SpA GAN[24]	14.7166	0.4830	0.1951	15.9155
DSen2-CR[12]	15.6773	0.5069	0.1516	18.1004
Ours	**18.1306**	**0.5108**	**0.1175**	**13.9684**

clouds, as shown in the first row of Fig. 3, enabling the restoration of relatively blurry textures beneath the clouds. However, they cannot recover useful geospatial information under thick clouds. Furthermore, the color fidelity of McGAN is significantly poor, and the spectral information deviates notably from the actual image. DSen2-CR outperforms previous methods significantly in terms of texture information reconstruction and color fidelity. It effectively restores the overall outline of the entire land in all the four scenarios, but the brick structures within the land are not distinct enough. Moreover, there are noticeable artifacts in the recovered images, and the cloud removal effect is not satisfactory, especially in scenes with higher cloud coverage. However, our model can effectively remove clouds in various scenarios and demonstrates greater advantages in recovering fine-scale features of the land.

The quantitative results are presented in Table I. The results of metric values are consistent with our previous visualization analysis. McGAN exhibits a larger SAM value, indicating significant spectral shifts and color distortions. In scenarios with thick cloud cover, both McGAN and SpA GAN exhibit substantial deviations from the ground-truth images, resulting in notably lower SSIM values and higher MAE values. Due to the presence of artifacts in the recovered images, the performance of DSen2-CR in terms of PSNR and SSIM is affected. Overall, our model exhibits clear superiority in spectral recovery and texture reconstruction, with better quantitative metrics compared with other methods. It is worth noting that different weather conditions and long time intervals between the acquisition of cloudy optical data and cloud-free optical data in SEN12MS-CR may lead to variations in color and details between the corresponding images. This can potentially affect the measurement of quantitative metrics.

TABLE II
QUANTITATIVE COMPARISONS OF DIFFERENT LOSSES

Method	PSNR ↑	SSIM ↑	MAE ↓	SAM ↓
L_{CARL}	16.2951	0.4891	0.1898	15.8016
L_{CARL}+L_{SSIM}	18.1306	0.5108	0.1175	13.9684

D. Ablation Study

To examine the role of the SSIM loss function on image restoration performance, we conduct experiments with the L_{CARL} loss used alone and the L_{CARL} loss combined with the L_{SSIM} loss. Table II shows the results. The results indicate that adding the L_{SSIM} loss function can increase quantitative metrics and improve cloud removal performance. It also demonstrates that the multidata fusion method can reduce the noise in the generated images and facilitate the subsequent reconstruction.

IV. CONCLUSION

In this letter, we propose a novel cloud removal framework that can preserve fine-grained edge structures in images. To mitigate the adverse effects of speckle noise in SAR images on image reconstruction, we fuse the SAR edge maps with SAR and optical data as the new input for the network. In addition, a custom loss function is introduced to optimize the network for reconstructing images with clear edge structures. The experimental results demonstrate that our method produces clearer and more detailed images with better quantitative indices.

References

[1] M. Zha, W. Qian, W. Yang, and Y. Xu, "Multifeature transformation and fusion-based ship detection with small targets and complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[2] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.

[3] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3826–3852, Jul. 2013.

[4] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[5] T. Lei, J. Wang, X. Li, W. Wang, C. Shao, and B. Liu, "Flood disaster monitoring and emergency assessment based on multi-source remote sensing observations," *Water*, vol. 14, no. 14, p. 2207, Jul. 2022.

[6] M. A. Iqbal, A. Anghel, and M. Datcu, "Coastline extraction from SAR data using Doppler centroid images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[7] Y. Zhou, L. Zhang, Y. Cao, and Y. Huang, "Optical-and-radar image fusion for dynamic estimation of spin satellites," *IEEE Trans. Image Process.*, vol. 29, pp. 2963–2976, 2020.

[8] Y. Huang et al., "HRWS SAR narrowband interference mitigation using low-rank recovery and image-domain sparse regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5217914.

[9] J. D. Bermudez, P. N. Happ, D. A. B. Oliveira, and R. Q. Feitosa, "SAR to optical image synthesis for cloud removal with generative adversarial networks," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 4, no. 1, pp. 5–11, Sep. 2018.

[10] C. Grohnfeldt, M. Schmitt, and X. Zhu, "A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Valencia, Spain, Jul. 2018, pp. 1726–1729.

[11] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, "Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks," *Remote Sens.*, vol. 12, no. 1, p. 191, Jan. 2020.

[12] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 333–346, Aug. 2020.

[13] J. Guo, C. He, M. Zhang, Y. Li, X. Gao, and B. Song, "Edge-preserving convolutional generative adversarial networks for SAR-to-optical image translation," *Remote Sens.*, vol. 13, no. 18, p. 3575, Sep. 2021.

[14] C. Liu, F. Tupin, and Y. Gousseau, "Training CNNs on speckled optical dataset for edge detection in SAR images," *ISPRS J. Photogramm. Remote Sens.*, vol. 170, pp. 88–102, Dec. 2020.

[15] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2015.

[16] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1395–1403.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, Banff, AB, Canada, 2014, pp. 1–15.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 8, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[19] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, Aug. 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[21] J. Li et al., "Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 373–389, Aug. 2020.

[22] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5866–5878, Jul. 2021.

[23] K. Enomoto et al., "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1533–1541.

[24] H. Pan, "Cloud removal for remote sensing imagery via spatial attention generative adversarial network," 2020, *arXiv:2009.13015*.