Cloud Removal Based on SAR-Optical Remote Sensing Data Fusion via a Two-Flow Network

Ruihan Mao, Hua Li, Gaofeng Ren, and Zhangcai Yin

Abstract—Optical remote sensing imagery plays an important role in observing the Earth’s surface today. However, it is not easy to obtain complete multitemporal optical remote sensing images because of the cloud cover, how reconstructing cloud-free optical images has become a big challenge task in recent years. Inspired by the remote sensing fusion methods based on the convolutional neural network model, we propose a two-flow network to remove clouds from optical images. In the proposed method, synthetic aperture radar images are used as auxiliary data to guide optical image reconstruction, which is not influenced by cloud cover. In addition, a novel loss function called content loss is introduced to improve image quality. The ablation experiment of the loss function also proves that content loss is indeed effective. To be more in line with a real situation, the network is trained, tested, and validated on the SEN12MS-CR dataset, which is a global real cloud-removal dataset. The experimental results show that the proposed method is better than other state-of-the-art methods in many indicators (RMSE, SSIM, SAM, and PSNR).

Index Terms—Cloud removal, data fusion, deep learning, optical, remote sensing, synthetic aperture radar (SAR).

I. INTRODUCTION

WITH the development of remote sensing technology, the number of remote sensing images dramatically increased in recent years. However, a common problem hinders the application of remote sensing data: cloud cover. According to statistics, about one-third of optical remote sensing images contain clouds, the rate is even higher in the area of the sea [1]. Cloud cover makes it harder to obtain a series of high-quality remote sensing data over multiple time periods in the same area. However, in recent studies, multitemporal data are very important for remote sensing applications, especially for land use detection, disaster detection, and other detection of Earth surface [2], [3], [4], [5], [6], [7]. For this, exploring a robust cloud removal approach is valuable for the development of remote sensing.

Generally, the cloud removal task can be regarded as a special case of image repair tasks, repairing the part of images covered by clouds. The existing remote sensing image repair approaches can be roughly divided into three types: 1) spatial-based methods, 2) spectral-based methods, and 3) temporal-based methods [8].

Spatial-based methods assume that the different parts of the image have similar geometric structures [9], and the missing part of the image can be completed through the remaining part. This approach is the most traditional approach of the three kinds of methods, and typical spatial-based methods include interpolation [10], [11] and propagated diffusion methods [12]. Due to the lack of other information reference, spatial-based methods are difficult to generate clear images. In some special cases, the premise does not even hold, for this reason, spatial-based methods are almost impossible to reconstruct the thick cloud image because it is completely blocked by clouds. Spectral-based methods use information complementarity between different spectral to reconstruct the image, which requires the existence of a complete band of multispectral data but thick clouds can block spectral information at all wavelengths, so this approach does not apply to cloud removal. Temporal-based methods use remote sensing images at different time points in the same area to provide supplementary information and this method is also known as mosaicking [13], [14]. When the observation conditions of selected remote sensing images of different time points are similar, this approach can usually get an excellent result. For the temporal-based methods, selecting an appropriate time interval is the key to constructing images. If the time interval is too short, the scene may still be covered by clouds. However, if the time interval is too long, the correlation between multitemporal images may disappear; therefore, temporal-based methods are not easy to be applied because of their strict requirements on data.

In addition to the abovementioned three methods, multisource data fusion is also used to remove clouds in recent years, especially based on synthetic aperture radar (SAR)-optical fusion. Compared with the other methods, multisource-based approaches have the following significant advantages.

The method based on multisource data is easier to get the reference image with a closer time interval to the cloud-contaminated image. If the reference image is the SAR image, which is not affected by climate, the reference image can even be provided in real time. A shorter time gap means less chance of changes, making reference images more reliable. So, cloud removal approaches based on SAR-optical fusion have nature superiority over other methods.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
Although multisource-based methods have some advantages over other methods, there are still some challenges in realizing the conversion of remote sensing data from different sources. Due to the different imaging methods of remote sensing images from different sources, they cannot be fused directly, so it is important to find a suitable method for converting or fusing SAR and optical images.

Latest remote sensing data fusion methods have been gradually transformed from traditional statistical methods to deep learning. In some recent studies, a two-flow network has been proved to have better performance than a single-flow network for the fusion of heterogeneous data [15], [16]. In this article, we propose a robust two-flow network for cloud removal, and the main contributions of the proposed cloud removal approach are as follows.

1) A novel two-flow network is proposed for cloud removal by fusing optical and SAR remote sensing data.
2) Training a network directly with L1 or L2 loss functions is easy to produce fuzzy images; therefore, a novel loss function called content loss is introduced to mitigate this problem.
3) The proposed network is trained on a big real dataset, the experimental results are more in line with the real situation.

The rest of this article is organized as follows. In Section II, we briefly review the development of CNN and the methods of image reconstruction based on CNN. Section III describes the two-flow network and loss function in detail. In Section IV, the characteristics of the SEN12MS-CR dataset are introduced. The experimental results and discussion are given in Section V. Finally, Section VI concludes this article.

II. RELATED WORKS

A. Convolutional Neural Networks for Cloud Removal

Recent years have witnessed the transition from traditional approaches to data-driven deep learning approaches in the field of computer vision (CV), in which convolutional neural network (CNN) is one of the most widely used architectures, which facilitates many CV tasks. And influenced by this trend, cloud removal methods based on CNN began to emerge in large numbers in recent years and achieved amazing results.

Before CNN was widely used in image restoration, most studies of cloud removal were about prior-based methods. Although prior-based methods can get a good result in some situations, it needs professionals to design different cloud removal methods for different imaging conditions or different sensors. Compared with prior-based methods, CNN-based methods can solve the problem well by the end-to-end learning. Recent studies have divided cloud types into thick clouds and thin clouds, and the CNN-based removal methods are also different for different cloud types. This is because thick clouds can completely block ground information of the images, which leads to CNN-based methods usually needing to use auxiliary data as the additional input to reconstruct high-quality cloud-free images. Recent studies most commonly used multitemporal data as the auxiliary data for thick cloud removal [17], [18], [19], which used CNN to fuse remote sensing images with different time stamps in the same area to get cloud-free images. Although these multitemporal-based cloud methods can produce very clear cloud-free images, they usually have strict requirements for data. To mitigate this problem, Zhang et al. [20] proposed a novel deep learning framework, which can use arbitrary numbers of temporal images to reconstruct images. Except for multitemporal data, multisource data, especially SAR data, can also be used as auxiliary data for thick cloud removal. The studies about SAR-optical fusion-based methods will be discussed in detail in Section II-B.

Compared with thick cloud removal, thin cloud removal methods usually focus on suppressing the cloud influence instead of replacing the cloudy pixels. Therefore, the thin cloud removal methods based on CNN can directly obtain the cloud-free images from the cloud-cover images without additional auxiliary data [21], [22]. In addition, because different wavelengths of the spectrum are affected differently by clouds, the spectral-based method is also promising for thin clouds removal. Li et al. [23] proposed a spectral-based method to remove thin clouds by using information complementarity between different spectral. Although these methods have shown the great potential of the CNN for cloud removal, the lack of data has been a major challenge in the field. To solve the problem, Tao et al. [24] proposed a self-paced learning method to reduce the network’s dependence on data, which does not need paired images to train the network.

B. Cloud Removal Based on SAR-Optical Fusion

With the development of space technology, researchers can get remote sensing images from different sources more easily than before. Although these images come from different sensors, and the imaging mechanisms of these satellites are different, the corresponding gray values of different images from different sources are linearly dependent. For this reason, it is easy to fuse heterogeneous optical images. However, the linear dispense between SAR images and optical images become very weak due to the special imaging mode and speckle noise in SAR images, which makes the fusion of optical images and SAR images difficult. Eckardt et al. [25] for the first time combined with multifrequency SAR image, in the method, geographical weighting was used to remove cloud image pixel by pixel. Huang et al. [26] used a sparse representation method to remove clouds and SAR images used as auxiliary data. Both of these representative methods require cloud detection and the reconstruction of the part of the image covered by the cloud. This requires professional researchers to design different algorithms for different cloud cover types, which is undoubtedly a big challenge. However, with the development of deep learning in the field of computer vision, this problem can be solved implicitly by end-to-end learning of deep neural networks.

In general, a cloud removal algorithm based on SAR-optical fusion can be regarded as a remote sensing images fusion or restoration task. In recent years, inspired by the success of CNN in the field of natural images, many remote sensing images restoration and fusion methods based on CNN are proposed. For these remote sensing studies based on CNN, fully convolutional
network (FCN) [27] and conditional generative adversarial network (cGAN) [28] are the two most widely used CNN architectures. In [21] and [52], a deep residual FCN network is used for cloud removal, which proves that residual learning is beneficial to remote sensing image restoration. For pan-sharpening tasks, a special FCN named a two-flow network is introduced in [15] to fuse remote sensing image information from different sources. When FCN is used for image reconstruction, $L_1$ loss or $L_2$ loss is usually used as the loss function, although these two losses can reconstruct the pixelwise information of images, they lack the overall perception of images and, therefore, the lose part of the texture information of images. To solve this problem, cGAN adds a discriminant network to evaluate the truthfulness of generated images, so cGAN-based methods can generate more realistic images than FCN-based methods. The cGAN-based cloud removal method needs to adjust the model so that its inputs are SAR images and cloud optical images, and its target images are cloud-free optical images [29]. Another more direct way to realize a cGAN-based cloud removal method is to convert SAR images into optical images to get cloud-free optical images. In [30] and [31], a special cGAN called cycle-GAN [32] is introduced to translate SAR images into optical images, which can convert two different styles of images to each other, and it can be trained without using paired images. Although the cloud removal methods based on cGAN have great potential, its drawbacks are also obvious: cGAN is difficult to train and lacks robustness when fed with bad data (especially input image covered by a large cloud) [33].

To sum up, cGAN-based methods are difficult to train and do not have sufficient robustness for cloud removal, and FCN-based methods using $L_1$ loss or $L_2$ loss are difficult to generate high-quality images. So, a more robust and high-performance method needs to be explored for cloud removal.

III. METHODOLOGY

A. Two-Flow Network

We use a two-flow network instead of a single-flow network for cloud removal. The difference between them is shown in Fig. 1. The two-flow network adds two prefusion networks to extract image features from optical and SAR images before fusion them, and it is worth noting that the prefusion networks do not share parameters. The two-flow network can extract image features from optical and SAR images by different network parameters, and due to the discrepancy in optical and SAR data distribution, the two-flow network is more effective in image feature extraction. In addition, this special structure enables the network to fuse images with different spatial resolutions during training. The design details of the network will be shown in Section II-B.

B. TF-CRNet Network

The proposed network, called TF-CRNet, is based on a pan-sharpening model presented by Liu et al. [15]. Similar to the SAR-optical fusion algorithm, pan-sharpening requires the fusion of gray values and texture features from two different sources to reconstruct high-resolution images. To reconstruct the cloud-free optical imagery, the TF-CRNet incorporates SAR imagery to guide the optical images covered by thick clouds while preserving the part of the optical image not to be covered by clouds. The specific design of the network is shown in Fig. 2. The details on TF-CRNet are as follows.

1) Convolution: There are three kinds of convolution layers in the network. All convolution kernels have a size of $3 \times 3$. The stride of the first type convolution is 1, which is used to ensure the feature map size remains unchanged. To reduce the loss of image details caused by the pooling layer, the convolution layer with a stride of 2 is used for down-sampling instead of the pooling layer. Different from the former two, transpose convolution is used for upsampling. Transpose convolution is a special convolution operation, which can be regarded as the reverse process of ordinary convolution. In recent studies, transpose convolution is used to replace traditional linear interpolation in FCN, the implementation details of it are described in [35].

2) Residual learning: Residual learning was first proposed in the field of image classification [36] to solve the problem that the accuracy of the deep network decreases with the increase of depth. It has been pointed out in recent studies [37] that residual learning can effectively solve the problem of the shattering gradient problem in DNN training, residual learning is applied to many CNN architectures to improve performance [21], [38]. In general, the operating principle of a residual block can be described as follows:

$$x_{l+1} = h(x_l) + F(x_l, W_l).$$ (1)
A common form of the $h(x_I)$ is $h(x_I) = x_I$, a residual block example is shown in Fig. 3. It also is applied to the structure of the TF-CRNet. The convolution kernel size of the residual block is set to $3 \times 3$ and the convolution stride size is 1. In addition, the input feature map size and output size of the residual block are the same.

3) **Proposed network uses parametric satisfaction linear unit (PReLU):** PReLU [39] to replace RELU, which is most frequently used in DNN. PReLU uses nonzero activation to activate negative values, which can avoid information loss during CNN inference.

4) **Concatenation:** Inspired by the “crop and copy” operation of U-Net [40], in TF-CRNet, long skip connection architecture is introduced so that the network architecture can reuse DNN low-level feature map semantic information.

The TF-CRNet input SAR image with 2 bands and cloud optical image with 13 bands, respectively. The size of these images is $256 \times 256$. Downsampling convolution is used to increase the number and reduce the size of feature maps while upsampling convolution is used to reduce the channel number and expand the size of feature maps. Batch normalization structure is added to the model before every convolution operation to accelerate network convergence. Experiments on network configuration on real datasets also verify the effectiveness of these structures.

### C. Loss Function

$L_1$ loss or $L_2$ loss is usually used for image reconstruction. Although $L_2$ loss is more robust than $L_1$ loss, it tends to generate a smooth image and appear blur effect. For this reason, in recent studies, $L_1$ loss becomes a better choice than $L_2$ loss [41], [42]. The $L_1$ loss for cloud-free images construction is defined as follows:

$$L_1 = \frac{1}{N} \| TF_{-} CRNet(SAR, optical) - R \|_1. \quad (2)$$

$TF_{-} CRNet(SAR, optical)$ is the cloud-free image of the network output. $R$ is the real target image and $N$ is the total number of pixels.

As shown in (2), $L_1$ loss focuses more on the pixelwise change while ignoring the whole image’s texture. To solve the problem, a loss function called content loss was introduced, which was proposed by Johnson et al. [43]. The content loss is calculated by a pretrained VGG network, and in previous studies, the VGG network was usually pretrained on the ImageNet dataset. However, the three-channel input of the general color images is inconsistent with the 13-channel sentinel-2 optical images, so it is essential to retrain another VGG network to calculate the content loss. The new VGG network is trained on the Eurosat dataset, which is a novel dataset based on Sentinel-2 satellite images covering 13 spectral bands and consisting of 10 classes within a total of 27 000 labeled and geo-referenced images [44]. The resulting VGG network has a classification accuracy of 96% on the test dataset. Fig. 4. shows
the calculation process of content loss. Content loss can also be defined as follows:

$$L_{\text{content}}^{0,j} = \frac{1}{N} \| \phi_j(R) - \phi_j(F) \|_1. \quad (3)$$

As shown in (3), the content loss can be regarded as the $L_1$ loss of feature maps generated by VGG. $R$ and $F$ denote target and predicted cloud-free images, respectively. $N$ are the total number of pixels and $\phi_j$ are the activations of the $j$th layer of the network $\phi$. The content loss is beneficial to image reconstruction because it is calculated by feature maps, which contain some more abstract image information than just pixelwise level information.

For cloud removal tasks, $L_1$ loss can reduce the loss of spectral information while content loss is used to improve the visual perception of images. Spectral information is very important for quantitative remote sensing, and a better visual perception is important for object recognition. Therefore, the final loss function should take into account these two different application fields, which can be expressed as follows:

$$L_{\text{total}} = L_1 + \lambda_1 L_{\text{content}}. \quad (4)$$

After a lot of experimental finetuning, $\lambda_1 = 0.5$ and $j = 7$ is the optimal configuration.

D. Evaluation Indicators

Four evaluation indicators are used to evaluate the similarity between the output image of the model and ground truth, which are root-mean-squared error (RMSE), peak signal-to-noise ratio (PSNR), spectral angle mapping (SAM) [45], and structure similarity index measurement (SSIM) [46]. The dynamic range of SSIM and PSNR was set as the maximum pixel value of Sentinel-2 images (10 000). The unit of RMSE is top-of-atmosphere reflectance ($\rho_{\text{TOA}}$) and SAM is measured in degrees.

E. Data Preprocessing and Training Settings

To remove outliers, the range of sentinel-2 images grey value is clamped to $[0, 10 000]$, and for the Sentinel-1 VV and VH polarizations the clamped range is $[-25, 0]$ and $[-32.5, 0]$, respectively. Then, these images were normalized, respectively, before feeding to the network.

The network’s parameters are initialized by Kaiming initialization [47] and optimized by Adam optimization [48]. The initial learning rate is 0.0001 and gradually decays with epochs, and the batch size was set to 16 to take full advantage of batch normalization. Models are implemented in PyTorch and trained on NVIDIA GeForce RTX 2080Ti GPUs.

IV. DATASET

The proposed model was trained, tested, and validated on the SEN12MS-CR dataset [49], which is an evolution of the SEN12MS dataset [50], it is a publicly available, large dataset for cloud removal. SEN12MS-CR dataset consists of coregistered triplet image, which are Sentinel-1 SAR images, 13-band cloud-free, and cloud-covered Sentinel-2 optical images. The dataset sampled 169 nonoverlapping ROIs from four seasons, and the coverage area of these ROIs is about $52 \times 40$ km$^2$. On average, each ROI consists of more than 700 patches, and the size of each patch is $256 \times 256$. To get more patches, adjacent patches have a 50% overlap and the total number of patches sampled is 157 521. The part of each optical patch covered by cloud is about 47.93%±36.08%, which means about half of the optical image is affected by cloud cover, and there is a huge discrepancy in land cover types and geometric structures among these patches, which ensures the diversity of the dataset without causing overfitting of CNN. The pixel values of the optical images are in the range of $[0, 10 000]$, and the two bands of SAR data of sentinel-1 come from two polarization channels (VV and VH) acquired in IW mode.

As far as the author knows, SEN12MS-CR is the first big real dataset constructed for optical-SAR-fusion-based cloud removal networks. The dataset collects a large number of scenes in different regions and seasons around the world and has very complete spectral information for experimenters to use. Some example images of the dataset are shown in Fig. 5. These examples show three different types of cloud cover, namely thin cloud cover, partially covered by thick cloud and completely covered by thick cloud. To avoid the network overfitting the dataset and
TABLE I
TEST RESULTS ON DIFFERENT METRICS (THE BIGGER VALUES OF SSIM AND PSNR ARE, THE BETTER THE RESULTS ARE. THE SMALLER VALUES OF RMSE AND SAM ARE, THE BETTER THE RESULTS ARE)

| Method     | RMSE_{PTDA} | SSIM    | SAM (°) | PSNR |
|------------|-------------|---------|---------|------|
| Proposed model | 0.03505    | 0.8799  | 7.285   | 30.17|
| pix2pix    | 0.03513    | 0.8641  | 7.865   | 29.65|
| dsen2-cr   | 0.03491    | 0.8782  | 7.552   | 30.12|

The boldface values means that this quantitative indicator in bold of the method outperforms.

affecting the evaluation results of the model, it is necessary to strictly implement the complete separation between the training dataset, verification dataset, and test dataset. Therefore, these three datasets consist of different images from different ROIs and different seasons, in which 139 ROIs are for training, 15 are for verification, and 15 are for testing. The data of the four seasons are evenly distributed in the datasets.

V. EXPERIMENTAL RESULTS AND ANALYSIS

In recent studies, the pix2pix model is one of the most commonly used models, so it is used as a baseline for comparison with TF-CRNet. In addition, the dsen2-cr model [52] as the SOTA method on the SEN12MS-CR dataset is also used as the baseline. The experiment results are shown in Fig. 6. It can be seen that there is an obvious spectral difference between the output images of pix2pix and the ground truth. In contrast, TF-CRNet can better reconstruct images influenced by clouds and cloud shadows than pix2pix. The conclusion can also be verified by comparing their image metrics in Table I, the proposed model is better than the baseline for these four quantitative indexes. By observing the loss function of the pix2pix model, it is not difficult to analyze the reasons for the results. The loss function of pix2pix can be defined as

\[
G^* = \arg \min_G \max_D L_{cGAN}(G, D) + \lambda L_{L1}(G). \tag{5}
\]

The loss function of pix2pix contains two items, namely \(L_1\) loss and cGAN loss, in which the cGAN loss is calculated by the discriminator network. For a cloud-removal network, the generated image should be as close to the real image as possible, which is important for remote sensing applications. However, the cGAN loss is usually used to evaluate the change of image style, and it is not sensitive to the change of image color, which is not conducive to the true restoration of the ground information. In addition, in the processing of training, mode collapse and failure of convergence rarely happen for the proposed network, which is unlike pix2pix and other cGANs.

Similar to the proposed method, the dsen2-cr model also uses \(L_1\) loss function to get better results than the pix2pix model. Although the dsen2-cr model shows comparable performance to the proposed method, its most indicators are slightly lower than the proposed model. Moreover, the number of parameters of the dsen2-cr model is much larger than the proposed model.

VI. DISCUSSION

A. Influence of the Content Loss

To improve images’ quality, the content loss is introduced in the proposed method. In this section, an ablation experiment is designed to verify its effectiveness of the content loss, and
Fig. 7. Examples of SEN12MS-CR dataset triplet, the size of these patches is all $256 \times 256$. (a) Optical images covered by clouds. (b) SAR images. (c) Target cloud-free optical images. Optical images are synthesized in true RGB color using B4-B3-B2 of Sentinel-2. SAR images are shown by two channels ($R=VH$, $G=VH$, and $B=VV$). (d) Output images obtained only using L1 loss. (e) Output images obtained using L1 loss and content loss. (f) Output images obtained only using content loss.

![Figure 7](image-url)

The explement results are tabulated in Table II. We compare the effect of using content loss versus not using it on the quality of the generated image by four indicators, and it goes without saying that using the content loss function is better than not using it. The only exception is the RMSE metric, the reason is that the optimization goal of the $L_1$ loss function is to reduce the difference between a single pixel value and true value, and the RMSE is also a pixelwise metric.

As is shown in Fig. 7, the differences between content loss and without content loss are shown in a real case. The road in Fig. 7(d) generated by the model without content loss is very blurry. However, the model with content loss can utilize the attentional mechanisms of CNN, which allow it to restore significant areas of the image in Fig. 7(e).

It is undoubtful that the content loss is helpful for restoring the texture of optical images. However, as shown in Fig. 7(f), if train the network only with content loss, the color information is completely lost. So, using both $L_1$ loss and content loss and finding the appropriate weights for them are essential to train the cloud removal network.

In order to explore the influence of the weight of the two loss functions on the results, we designed a control experiment and showed the results in Fig. 8. As $\lambda_1$ varies from $10^{-2}$ to $10^3$, which was shown in Fig. 8, the model achieves the best result when $\lambda_1$ is around $10^{-1}$. After more precise finetuning, the value of $\lambda_1$ is set to 0.5 finally. According to Table II and Fig. 8, it can be concluded that content loss can enrich the texture information of the predicted images, which is important for generating high-quality images.

In addition, as shown in (3), the depth of feature maps ($j$) used by content loss can also affect the training results. The value of $j$ usually has two options, such as $j = 4$ or $j = 7$, and the results of taking different values of $j$ are tabulated in Table III. The results show that when $j = 7$, all quantitative indexes except RMSE are better than those when $j = 4$. And it can be seen from (3) that the content loss is the $L_1$ loss of feature maps, and when $j = 0$, the $L_1$ loss is exactly equivalent to content loss. In addition, it is not difficult to see from Fig. 4 that content loss and $L_1$ loss are closer when $j$ is smaller. Therefore, the results can verify that $L_1$ loss is more beneficial to pixelwise metric RMSE, which is consistent with the results in Table II.
Fig. 8. Weight of content loss is set as $10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}$, and $10^{3}$, respectively, to research the sensitivity of parameters. The model achieves the best result when $\lambda_1$ is around $10^{-1}$, after more precise fine-tuning, the value of $\lambda_1=0.5$ is the optimal configuration. (a) RMSE. (b) SSIM.

### TABLE III

| $j$ | RMSE(µµµµµµµµµµααaαaαaα) | SSIM | SAM (°) | PSNR |
|-----|----------------|-------|---------|------|
| $j = 4$ | 0.03474 | 0.8741 | 7.358 | 29.96 |
| $j = 7$ | 0.03505 | 0.8799 | 7.285 | 30.17 |

The boldface values means that this quantitative indicator in bold of the method outperforms.

### B. Influence of the Cloud Mask Loss

In [29], a cloud mask loss was used to make the cloud removal approach more focused on recovering the part of the image covered by clouds. In [51], an intensity normalization is used to make the cloud-free pixels that can be used as a reference to further improve the estimated results. These two methods can make the model focus more on restoring the part of the image covered by clouds and get better experiment results on the synthetic dataset. In this article, we also explore whether the trick works for a real dataset. For this, a cloud-shadow mask loss is designed for the proposed method so that it can pay more attention to the part of the image affected by the cloud, it can be defined as

$$L_{\text{mask}} = \frac{1}{N} \| M \cdot R - M \cdot F \|_1. \quad (6)$$

In (6), $M$ is the mask matrix of the same spatial dimension as the images, the cloud pixels and shadow pixels are set for value 1 and uncorrupted pixels are set for value 0. $R$ is the target image and $F$ is the predicted image. To detect clouds and cloud shadows of the optical images, a detection method proposed in [52] is used.

Taking cloud-shadow mask loss into consideration, the total loss can be defined as

$$L_{\text{total}} = L_1 + \lambda_1 L_{\text{content}} + \lambda_2 L_{\text{Mask}}. \quad (7)$$

In (7), the value of $\lambda_1$ is still 0.5 and $\lambda_2 = 0.001$. As the results tabulated in Table IV, although $\lambda_2$ is set to a very small value, cloud-shadow mask loss also has a negative effect on the predicted images in the real dataset.

It can be easily concluded from Table IV that mask loss does not play an active role in the real dataset, the reason for this is that the detection of cloud and shadow in a real dataset is much more difficult than in a synthetic dataset. In the real situation, the cloud removal network needs to remove both clouds and cloud shadows, which are difficult to simulate in a synthetic dataset. So, although the mask loss does work well on the synthetic dataset, it does not work on the real dataset as easily as it. The result also shows that the network trained on the real dataset is more practical.

### C. Two-Flow Network Versus Single-Flow Network

The difference between a two-flow network and a single-flow network is shown in Fig. 1. In contrast to the proposed network that feeds images into two different branches separately, the input of a single-flow network is a 15-band image, which is the stack of SAR images and optical images on the channel dimension. To demonstrate the superiority of the proposed method, we designed a controlled experiment to compare the two-flow network with the one-flow network. It is worth noting that in order to more fairly compare their performance, these two networks have the same kernel size, layer number, skip connection, residual block, and active function. In addition, the hyperparameters and loss function are also the same for training. The qualitative results of these two architectures are compared in Table V. The
results show that the evaluation indexes of the two-flow network are all higher than the single-flow network except SAM. As can be seen, the results indicating the overperformance of two-flow architecture than a single-flow one.

VII. Conclusion

This article proposes a cloud removal network based on the fusion of optical-SAR images and achieves an ideal result. The network is trained on a real dataset to ensure that the results are more in line with the real situation. To get more realistic images, a novel loss function is designed to optimize the network. By comparing four quantitative indicators, the results show that the proposed method is superior to the baseline. In addition, ablation experiments were designed to verify the influence of different settings on experimental results, which also confirmed the feasibility and validity of the proposed method. Finally, we also explore the influence of the cloud-shadow loss, and the results show that for the real dataset, a cloud-shadow loss does not always play a positive role. Due to the sentinel-1 and sentinel-2 data being readily available, the proposed method is helpful for the more temporally observation of the Earth observation.

Although the proposed method shows superior performance on the dataset, it is still difficult to generate a very clear image. In the future, we will take advantage of the design of a super-resolution network to make the network restore more details of images.

REFERENCES

[1] C. H. Lin, P. H. Tsai, K. H. Lai, and J.-Y. Chen, “Cloud removal from multitemporal satellite images using information cloning,” IEEE Trans. Geosci. Remote Sens., vol. 51, no. 1, pp. 232–241, Jan. 2013.
[2] A. Lin et al., “A Big Data-driven dynamic estimation model of relief supplies demand in urban flood disaster,” Int. J. Disaster Risk Reduction, vol. 49, 2020, Art. no. 101682.
[3] H. Wu, A. Lin, X. Xing, D. Song, and Y. Li, “Identifying core driving factors of urban land use change from global land cover products and POI data using the random forest method,” Int. J. Appl. Earth Observ. Geoinf., vol. 103, 2021, Art. no. 102475.
[4] P. Du et al., “Information fusion techniques for change detection from multi-temporal remote sensing images,” Inf. Fusion, vol. 14, no. 1, pp. 19–27, 2013.
[5] A. Y. M. Abdullah et al., “Spatio-temporal patterns of land use/land cover change in the heterogeneous coastal region of Bangladesh between 1990 and 2017,” Remote Sens., vol. 11, no. 7, 2019, Art. no. 790.
[6] C. Zhang et al., “A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images,” ISPRS J. Photogrammetry Remote Sens., vol. 166, pp. 183–200, 2020.
[7] S. Ghaffarian et al., “Post-disaster building database updating using automated deep learning: An integration of pre-disaster OpenStreetMap and multi-temporal satellite data,” Remote Sens., vol. 11, no. 20, 2019, Art. no. 2427.
[8] H. Shen et al., “Missing information reconstruction of remote sensing data: A technical review,” IEEE Geosci. Remote Sens. Mag., vol. 3, no. 3, pp. 61–85, Sep. 2015.
[9] C. Guillenmot and O. L. Meur, “Image inpainting: Overview and recent advances,” IEEE Signal Process. Mag., vol. 31, no. 1, pp. 127–144, Jan. 2014.
[10] C. Zhang, W. Li, and D. Travis, “Gaps-fill of SLC-off Landsat ETM+ satellite image using a geostatistical approach,” Int. J. Remote Sens., vol. 28, no. 22, pp. 5103–5122, 2007.
[11] C. Yu et al., “Kriging interpolation method and its application in retrieval of MODIS aerosol optical depth,” in Proc. 19th Int. Conf. Geoinformat., 2011, pp. 1–6.
[12] A. Maalouf, P. Carre, B. Augereau, and C. Fernandez-Maloigne, “A bandelet-based inpainting technique for clouds removal from remotely sensed images,” IEEE Trans. Geosci. Remote Sens., vol. 47, no. 7, pp. 2363–2371, Jul. 2009.
[13] M. Li, S. C. Liew, and L. K. Kwoh, “Producing cloud free and cloud-shadow free mosaic from cloudy IKONOS images,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2003, pp. 3946–3948.
[14] E. H. Helmer and B. Ruefelnacht, “Cloud-free satellite image mosaics with regression trees and histogram matching,” Photogrammetric Eng. Remote Sens., vol. 71, no. 9, pp. 1079–1089, 2005.
[15] X. Liu, Q. Liu, and Y. Wang, “Remote sensing image fusion based on two-stream fusion network,” Inf. Fusion, vol. 55, pp. 1–15, 2020.
[16] L. H. Hughes, M. Schnitt, L. Mou, Y. Wang, and X. X. Zhu, “Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 5, pp. 784–788, May 2018.
[17] Y. Chen, Q. Weng, L. Tang, X. Zhang, M. Bilal, and Q. Li, “Thick clouds removing from multitemporal Landsat images using spatiotemporal neural networks,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2020.3043980.
[18] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei, “Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 8, pp. 4274–4288, Aug. 2018.
[19] Y. Chen, L. Tang, X. Yang, R. Fan, M. Bilal, and Q. Li, “Thick clouds removing from multitemporal ZY-3 satellite images using deep learning,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 143–153, 2020, doi: 10.1109/JSTARS.2019.2954130.
[20] Q. Zhang et al., “Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning,” ISPRS J. Photogrammetry Remote Sens., vol. 162, pp. 148–160, 2020.
[21] W. Li et al., “Thin cloud removal with residual symmetrical concatenation network,” ISPRS J. Photogrammetry Remote Sens., vol. 153, pp. 137–150, 2019.
[22] Y. Zi, F. Xie, N. Zhang, Z. Jiang, W. Zhu, and H. Zhang, “Thin cloud removal for multispectral remote sensing images using convolutional neural networks combined with an imaging model,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 3811–3823, 2021, doi: 10.1109/JSTARS.2021.3068166.
[23] J. Li et al., “Deep learning based thin cloud removal fusing vegetation red edge and short wave infrared spectral information for Sentinel-2 A imagery,” Remote Sens., vol. 13, no. 1, 2021, Art. no. 157.
[24] C. Tao, S. Fu, J. Qi, and H. Li, “Thick cloud removal in optical remote sensing images using a texture complexity guided self-paced learning method,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–12, 2022, doi: 10.1109/TGRS.2022.3157917.
[25] R. Eckardt et al., “Removal of optically thick clouds from multi-spectral satellite images using multi-frequency SAR data,” Remote Sens., vol. 5, no. 6, pp. 2973–3006, 2013.
[26] B. Huang, Y. Li, X. Han, Y. Cui, W. Li, and R. Li, “Cloud removal from optical satellite imagery with SAR imagery using sparse representation,” IEEE Geosci. Remote Sens. Lett., vol. 12, no. 5, pp. 1046–1050, May 2015.
[27] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3431–3440.
[28] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” 2014, arXiv:1411.1784.
[29] J. Gao et al., “Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks,” Remote Sens., vol. 12, no. 1, 2020, Art. no. 191.
[30] M. F. Reyes et al., “SAR-to-optical image translation based on conditional generative adversarial net-works–optimization, opportunities and limits,” Remote Sens., vol. 11, no. 17, 2019, Art. no. 2067.
[31] L. Wang et al., “SAR-to-optical image translation using supervised cycle-consistent adversarial networks,” IEEE Access, vol. 7, pp. 129136–129149, 2019.
[32] Y. J. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2223–2232.
[33] L. Mescheder, A. Geiger, and S. Nowozin, “Which training methods for GANs do actually converge?,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 3481–3490.
[34] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in Proc. 4th Int. Conf. Learn. Representations, San Juan, Puerto Rico, May 2-4, 2016. [Online]. Available: http://arxiv.org/abs/1511.06434

[35] V. Dumoulin and F. Visin, “A guide to convolution arithmetic for deep learning,” stat, vol. 1050, p. 11, 2018.

[36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[37] D. Baldazzi et al., “The shattered gradients problem: If ResNets are the answer, then what is the question?,” in Proc. Int. Conf. Mach. Learn., 2017, pp. 342–350.

[38] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, “PanNet: A deep network architecture for pan-sharpening,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 5449–5457.

[39] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1026–1034.

[40] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervention, 2015, pp. 234–241.

[41] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, “Loss functions for image restoration with neural networks,” IEEE Trans. Image. Imag., vol. 3, no. 1, pp. 47–57, Mar. 2017.

[42] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2017, pp. 136–144.

[43] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 694–711.

[44] P. Helber, B. Bischke, A. Dengel, and D. Borghi, “Introducing Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2018, pp. 204–207.

[45] R. H. Yuhas, A. F. H. Goetz, and J. W. Boardman, “Discrimination among semiarid landscape endmembers using the spectral angle mapper (SAM) algorithm,” in Proc. Sammaries 3rd Annu. JPL Airborne Geosci. Workshop, 1992, pp. 147–149.

[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[47] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1026–1034.

[48] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. 3rd Int. Conf. Learn. Representations, San Diego, CA, USA, May 7–9, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[49] 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.

[50] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu, “SEN12MS—A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion,” ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., vol. 4, pp. 153–160, 2019.

[51] W. Li, Y. Li, and J. C. W. Chan, “Thick cloud removal with optical and SAR imagery via convolutional-mapping-deconvolutional network,” IEEE Trans. Geosci. Remote Sens., vol. 58 no. 4, pp. 2865–2879, Apr. 2020.

[52] A. Meraner et al., “Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion,” ISPRS J. Photogrammery Remote Sens., vol. 166, pp. 333–346, 2020.