Simultaneous Cloud Detection and Removal From Bitemporal Remote Sensing Images Using Cascade Convolutional Neural Networks

Shunping Ji, *Member, IEEE*, Peiyu Dai, *Student Member, IEEE*, Meng Lu, and Yongjun Zhang

***Abstract*— Clouds and cloud shadows heavily affect the quality of the remote sensing images and their application potential. Algorithms have been developed for detecting, removing, and reconstructing the shaded regions with the information from the neighboring pixels or multisource data. In this article, we propose an integrated cloud detection and removal framework using cascade convolutional neural networks, which provides accurate cloud and shadow masks and repaired images. First, a novel fully convolutional network (FCN), embedded with multiscale aggregation and the channel-attention mechanism, is developed for detecting clouds and shadows from a cloudy image. Second, another FCN, with the masks of the detected cloud and shadow, the cloudy image, and a temporal image as the input, is used for the cloud removal and missing-information reconstruction. The reconstruction is realized through a self-training strategy that is designed to learn the mapping between the clean-pixel pairs of the bitemporal images, which bypasses the high demand of manual labels. Experiments showed that our proposed framework can simultaneously detect and remove the clouds and shadows from the images and the detection accuracy surpassed several recent cloud-detection methods; the effects of image restoring outperform the mainstream methods in every indicator by a large margin. The data set used for cloud detection and removal is made open.**

***Index Terms*— Convolutional neural networks (CNNs), cloud detection, cloud removal, multitemporal remote sensing images.**

I. Introduction

As THE most important technology to obtain the geometric and physical information of the Earth surface [1], remote sensing has raised growing attention and achieved wide applications in a variety of disciplines and industries. However, due to the thick atmosphere of the Earth, the quality of the remote sensing data is severely influenced by the clouds and cloud shadows. The clouds in the remote sensing images cause a series of problems in localization, image interpretation, data fusion [2], and object detection [3]. Consequently, it is important to detect the clouds and shadows and restore the shaded regions from the remote sensing images.

A. Review of Cloud Detection

The fractal structures in the geometry and the diversity in the spectrum of the clouds cause challenges in detecting clouds from the remote sensing data. Some studies focus on the spectral bands that are sensitive to the clouds and extract their morphological, biological, or physical information to discriminate the cloud from the background [4]–[9]. For example, Braaten *et al.* [4] used the brightness and the normalized difference between the red and green bands to identify the clouds, and the near-infrared (NIR) band for distinguishing the cloud shadows. Fisher [5] used the NIR and shortwave-infrared bands, and watershed-from-markers transform to detect the cloud and shadow regions from the SPOT5 HRG imagery. Li *et al.* [8] introduced a spatial–spectral domain cloud-detection model for the GF-1 imagery, in which a coarse cloud mask was refined using the spectral, geometric, and texture features. Zhu and Woodcock [9] attempted to detect clouds by several spectral tests including the top of atmosphere (TOA) reflectance and brightness temperature from thermal bands and find cloud shadows according to the darkening effect in the NIR bands. There are also studies designed for detecting common shadows from a single image [10]–[12].

Other studies attempted to detect clouds from the multitemporal images. Some early methods detected clouds through generating a change map between different temporal images with an empirical threshold that discriminates the cloud and noncloud areas. Wang *et al.* [13] developed the image-fusion and wavelet-transform technologies for the detection and removal of clouds and cloud shadows from the multitemporal Landsat TM images, but designed the detection and removal as two independent procedures. Bian *et al.* [14] calculated the difference in the blue bands between the bitemporal HJ-1 images to discover the clouds. Hagolle *et al.* [15] detected clouds based on a sudden increase in the reflectance in the blue wavelength on a pixel-based change map from the bitemporal images. As false changes are inevitable due to different imaging conditions, sensors, seasons, and so on, the generality ability of these change-detection-based methods is largely limited.

A group of widely applied methods are based on conventional supervised machine learning, which can detect clouds

Manuscript received February 20, 2020; revised April 21, 2020; accepted May 10, 2020. Date of publication May 22, 2020; date of current version December 24, 2020. This work was supported by the National Key Research and Development Program of China under Grant 2018YFB0505003. *(Corresponding author: Shunping Ji.)*

Shunping Ji, Peiyu Dai, and Yongjun Zhang are with the School of Remote sensing and Information Engineering, Wuhan University, Wuhan HB 430079, China (e-mail: jishunping@whu.edu.cn; pydai@whu.edu.cn; zhangyj@whu.edu.cn).

Meng Lu is with the Department of Physical Geography, Utrecht University, 3584 CE Utrecht, The Netherlands (e-mail: m.lu@uu.nl).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.2994349

from a single image or multitemporal images. For example, a support vector machine (SVM) and its variations were applied to discriminate the cloud from the background using the luminance and texture features [16]–[18]. Azimi-Sadjadi and Zekavat [17] introduced a hierarchical SVM structure to classify six types of clouds and four land cover classes with the mean and deviation of the temporal images as input. Lee *et al.* [19] designed a multicategory SVM framework for cloud classification.

The mainstream cloud-detection methods are deep learning-based [20], [21]. As a powerful representation learning method, deep learning, especially convolutional neural network (CNN), has been widely applied to object detection, segmentation, and denoising, as well as cloud detection. Zhan *et al.* [22] introduced a multiscale prediction strategy, integrating the low-level and high-level features from a CNN to classify cloud and snow. Xie *et al.* [23] used the improved simple linear iterative clustering (SLIC) to segment the remote sensing images and then applied a CNN for cloud detection. Very recently, several new studies appeared [24]–[31], and all of them are the variants of a fully convolutional network (FCN) and are featured with different optimization strategies such as adopting a multiscale aggregation [27] or using atrous convolutions [28].

B. Review of Cloud Removal

The study of cloud removal belongs to missing-information reconstruction. The methodologies can be roughly classified into four categories: spatial-based, spectral-based, temporal-based and multisource-based methods.

The spatial-based methods attempt to recover the shaded regions only from the clean pixels of an individual cloudy image, that is, the pixels below the cloud and the shadow are considered to share the similar textures as the cloud-free neighborhood and can be restored from them. However, the shortcoming of this type of methods [32]–[36] is obvious. The assumption of similarity between the shaded area and its neighborhood on the large-size and complex remote sensing images is rarely met. For example, a house may be located at the center of a large farmland; the heterogeneity of the land covers is more complex in urban areas.

The spectral-based methods are based on the complementary information between different spectral bands of a multispectral image. They rely on the prerequisite that parts of the multispectral bands can penetrate thin clouds [37]–[41]; therefore, they are incompetent to restore images with thick clouds.

Temporal-based methods introduce additional observations from different time series at the same area to reconstruct the corrupted region instead of using only the cloudy image, which are more generic and commonly show a better performance. According to the use of multitemporal images, it can be classified into three categories [42]: temporal replacement, temporal interpolation, and temporal learning.

The temporal replacement-based methods replace the missing regions in the cloudy images with the corresponding clean pixels in the temporal images [43]–[45]. The temporal interpolation methods are similar to the temporal replacement methods but introduce some interpretation strategies, such as ordinary cokriging interpolation [46] and neighborhood similar-pixel interpolation [47].

Temporal learning methods aim to establish the relationship between the temporal images by using the learning-based methods. For example, a dictionary learning algorithm was developed for the recovery of the cloud and shadow regions from the multitemporal images [48]; random forest and CNN were introduced to different tasks of missing-information reconstruction [49], [50].

The temporal-based methods are sensitive to the geometrical registration errors and the abrupt spectral changes between the images, both of which will impact the effect of image-to-image mapping.

The multisource-based methods focus on fusing the remote sensing data from different types of sensors. In addition to the fusion of the optical remote sensing images obtained from different sensors, synthetic aperture radar (SAR) data are recently considered as a new type of auxiliary data source for the recovery of optical cloudy images, as microwave signals are more capable of penetrating clouds [51], [52].

C. Objective and Contribution

Although cloud detection and cloud removal have been extensively studied, they are treated independently with different methods: specifically, the former is handled with the classification and segmentation technologies and the latter is based on fusion and missing-information reconstruction. As the cloud detection and removal are highly related and complementary, an integrated framework processing the two tasks simultaneously is favorable but currently missing. The other problem of previous studies is that the advanced deep learning method has not been fully introduced to the cloud-removal task. For example, Zhang *et al.* [50] used a time- and memory-consuming CNN structure instead of a mainstream and efficient FCN structure.

In this article, we put forward a novel integrated framework for simultaneous cloud detection and removal using cascaded CNNs. The main idea and contribution of this article can be summarized as follows.

1) The framework we proposed is the first deep learning-based framework for integrated cloud detection and removal, to the best of our knowledge. The framework is based on multitemporal remote sensing images and is generic for any cloud-detection algorithms and cloud-repairing methods with multitemporal data. It is also demonstrated that the integrated framework outperformed recent CNN-based methods in either cloud detection or cloud removal.
2) An FCN structure is designed for pixelwise cloud and cloud shadow detection. The FCN focuses on the multiscale effects of the remote sensing data through introducing a multiscale aggregation that combines features from different scales of a densely connected feature pyramid and a channel-attention mechanism that seeks

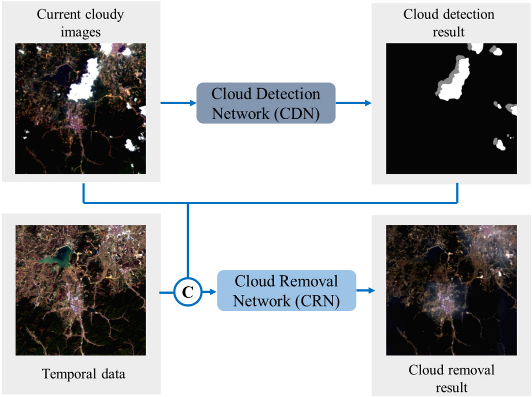


Fig. 1. Integrated framework of cloud detection and removal. The circled C indicates a concatenation operator. The cloud-detection result is one of the inputs for the CRN only in the training stage.

global consistency between these features, which boosts the segmentation effects on remote sensing images from those classic FCNs for general semantic segmentation.

3) In cloud removal, we introduced a self-training strategy, i.e., without using any manually labeled samples, to train another FCN for mapping a temporal image to a current cloudy image. The strategy uses the relations of clean-pixel pairs between the bitemporal images, where invalid pixels (cloud and shadow mask) have been excluded from the cloud-detection network (CDN). The self-training strategy is a great benefit for a supervised deep learning-based method that commonly demands sufficient labeled samples.
4) The data set (http://gpcv.whu.edu.cn/data/) we used for experiment has been made open, as we have found there lacks open data sets for considering both cloud detection and cloud removal.

The rest of this article is organized as follows. In Section II, the integrated cloud-detection and removal framework and the specific CNN networks for cloud detection and removal are elaborated. The experimental results and the discussion are presented in Sections III and IV, respectively. A conclusion is given in Section V.

II. Method

A. Integrated Cloud-Detection and Removal Framework

The proposed framework for simultaneous cloud detection and cloud removal is shown in Fig. 1. The input is a cloudy image and a corresponding temporal image, which is processed by two cascaded FCNs. The first FCN, called the CDN, is developed for cloud and shadow detection. The second FCN, the cloud-removal network (CRN), is developed for restoring the cloud-shaded region with the input of the bitemporal images. The detected clouds and shadows from the CDN are the input of the CRN only in the training stage. Basically, the framework can be seen as an integrated spatial–temporal–spectral model, as it takes different spectral bands and temporal images as inputs and uses CNNs to process the spatial and spectral information.

B. Cloud and Shadow Detection

The CDN is implemented by a pixel-to-pixel FCN with a DenseNet-style building block [53], embedded with a multiscale feature-fusion strategy and a channel-attention mechanism to handle the clouds and shadows of various sizes and shapes, and outputs a cloud mask map with the same size of the input image. As is shown in Fig. 2, the structure consists of an encoder, a decoder, and lateral connections between them. In each scale of the encoder, the densely connected architecture is used, i.e., the features of all previous scales are concatenated to the current features. In each DenseNet block, the input feature (denoted as dotted parallelogram) is concatenated to the feature maps after each of the first and second convolutions, respectively, both of which have been activated by the ReLU. The downsampling layer in the encoder and the upsampling layer in the decoder are realized by using a 2× max pooling and a 2× deconvolution, respectively. To make full use of the multiscale features, we upsample the last features of each scale to the original resolution, apply a 1×1 convolution for compression and an ReLU for activation, and then concatenate them channelwisely to form a multiscale feature map (the concatenated four gray layers). Since the multiscale feature map is inconsistent between the channels as they come from different scales, we further apply a channel-attention module (CAM) on the map to achieve global consistency between the channels.

Our CAM is inspired by the work of Vaswani *et al.* [54] and Fu *et al.* [55], and combines the former's multiheaded self-attention strategy and the latter's between-channel interdependence calculation. As shown in Fig. 3, the CAM takes a feature map (denoted as F with dimension $H \times W \times C$) from a CNN as input and processes it with three independent 3×3 convolutions to produce three parallel feature maps K, Q, and V. The reshaped K and Q $(N = H \times W)$ are multiplied to produce a $C \times C$ feature map, which is then multiplied by the reshaped V. Finally, the result of the dot product is reshaped, convoluted with a 1×1 kernel, and added to the input feature map to be the channel-attention boosted feature map.

The specific parameters and superparameters in the CDN are set as follows: the stride of the convolution layer is 1 and the kernel size is 3. The rate of max pooling and the stride in the deconvolution layers are 2. The growth rate of a DenseNet block in encoder is set to 256. The feature dimension of each scale in the decoder is 1024, 512, 256, 128, and 64, respectively.

A multicategory cross-entropy loss function is used to produce a probability map indicating how likely a pixel belongs to the cloud and cloud shadow. We set a relatively low threshold to translate the probability map to a mask map, which means most of the cloud and cloud shadow regions will be covered with a high recall. This guarantees that the rest of the images are clean for training the following CRN, which will learn an optimal radiometric transformation between the noncloudy regions of the bitemporal images.

C. Cloud Removal

The cloud-removal process is a dual task: a relative radiometric transformation from the temporal image to the current

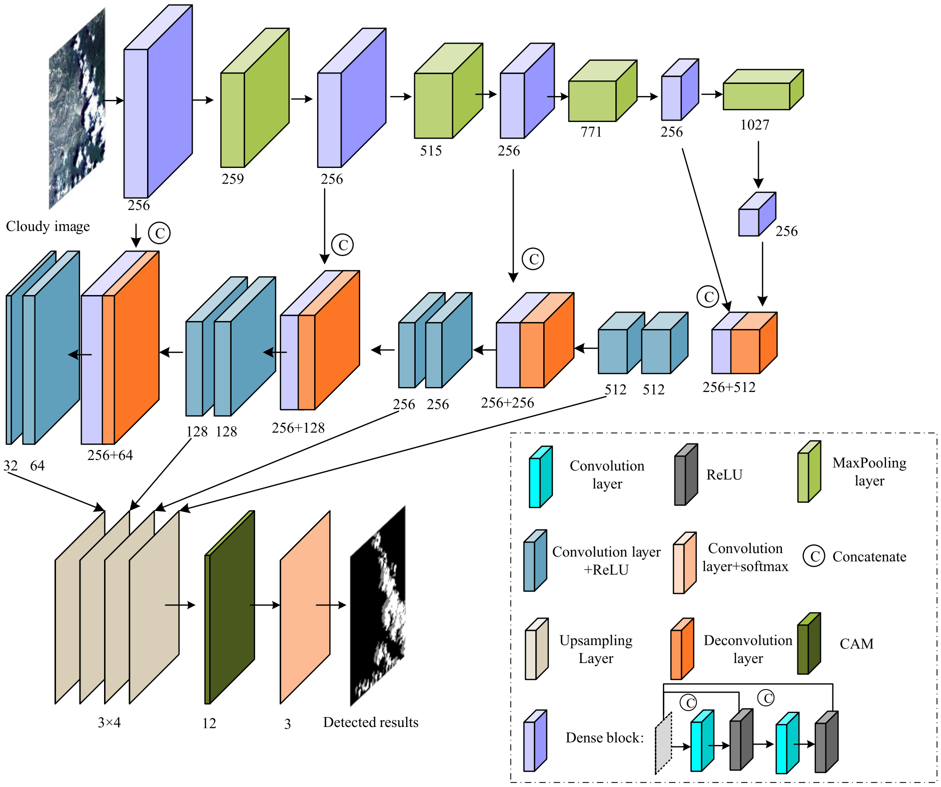


Fig. 2. Structure of the CDN. CAM is the channel-attention module.

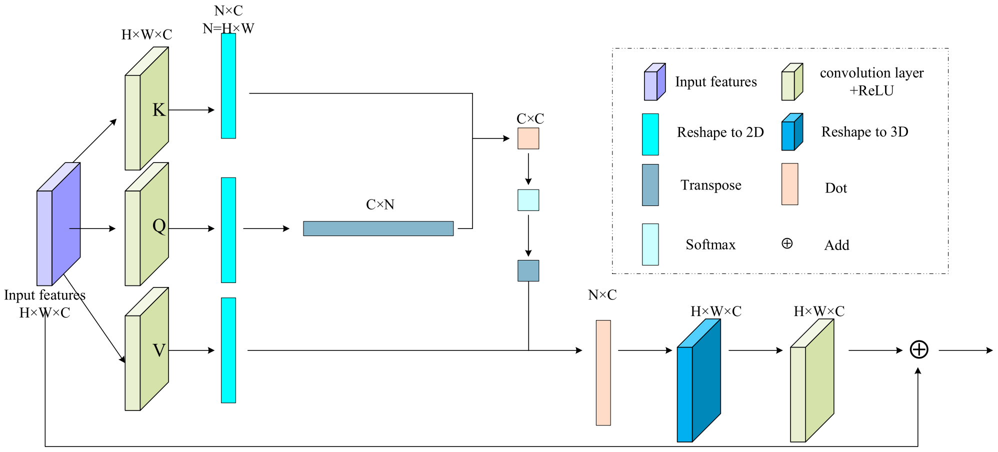


Fig. 3. Structure of the proposed CAM for cloud and cloud shadow segmentation. The F, K, and Q are the feature maps. H and W indicate the height and width of the features. C is the dimension of the features. $N = H \times W$.

image, which ensures the radiometric consistency of the two images, and a restoration of the shaded area using the learned transformation parameters. The dual task is simultaneously handled by a self-trained CRN based on an FCN structure.

In the training stage, the CRN learns the relative radiometric transformation through the corresponding pixels of the two-time images after having automatically excluded the shaded pixels, which were detected by the CDN, from the bitemporal images. The core idea of the CRN is to simulate randomly the cloud regions as the training samples, which avoids the requirement of true samples. A highly accurate recovering model is trained with these samples to restore an image that approaches the original cloudy image with real clouds excluded.

Fig. 4 shows the cloud-removal process. In the training stage, the input images of the CRN are the simulated damaged image (A2) and the temporal clean image (B1). For generating the two images, the cloud (marked in green) and cloud shadow (marked in red) masks detected from the CDN are required. First, the masked pixels are excluded from the bitemporal images. Then, new clouds (black pixels) are randomly simulated in large amount in the current image (A2). The CRN is trained with A2 and B1 to adapt to any possible missing information that is simulated and to produce

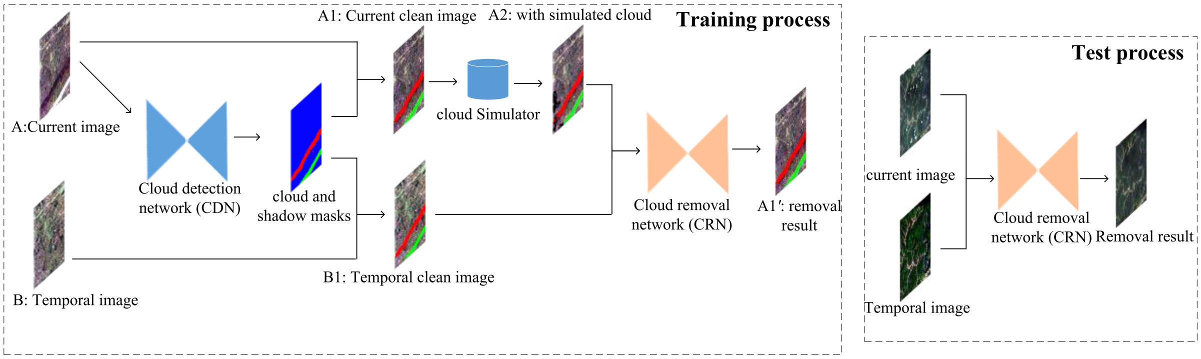


Fig. 4. Cloud-removal process. In the training process, real cloud and shadows (green and red masks) are detected from the CDN with high recall; simulated clouds (black mask) are randomly generated to simulate the arbitrary clouds. In the test, the input is the original bitemporal images.

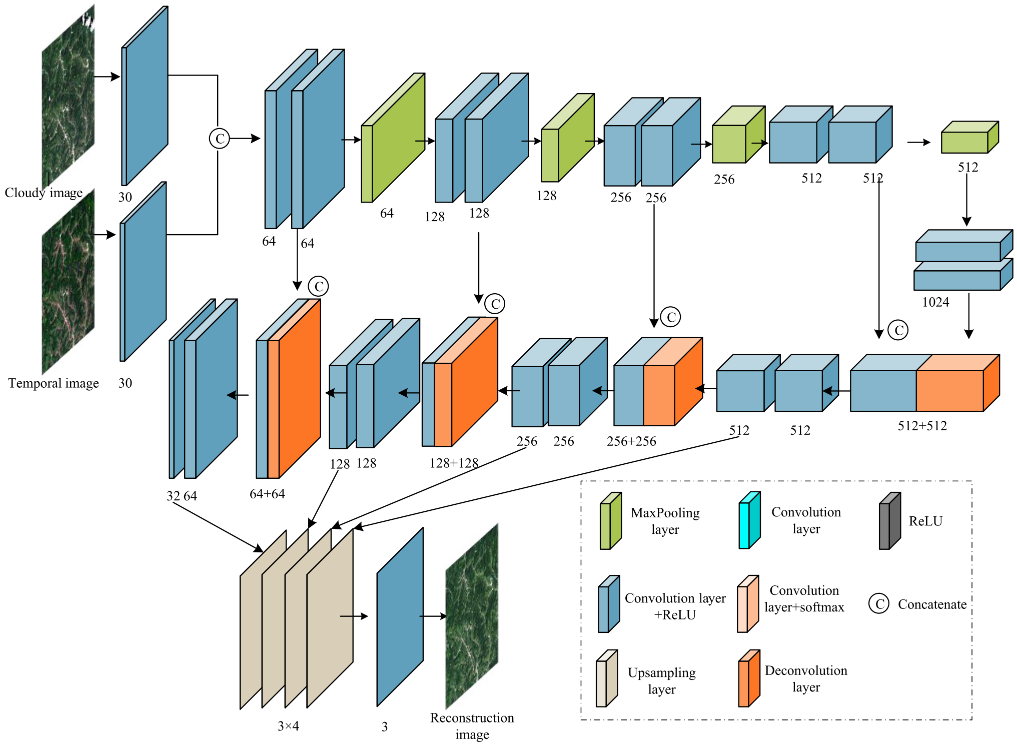


Fig. 5. Structure of a CRN.

an image approaching to the current image with the real clouds excluded, i.e., the current clean image (A1).

The self-training strategy has two obvious advantages. First, the pixels in the simulated cloud region can be used as a part of ground truth to assess quantitatively the performance of an algorithm, which cannot be accomplished with the real cloud regions. Second, the simulation of a large amount of random clouds helps build a robust model that can recover arbitrary cloud-shaded area in the real remote sensing images, including those cases in the test images; otherwise, the usage of the small set of real cloudy images can hardly train an applicable model.

In the prediction stage, the network takes the original bitemporal images as input and produces a repaired image. Note that the cloud location and region in the cloudy image is not required, as the model has learned how to fix arbitrary clouds in an image through the training process with simulated clouds.

It should also be noted that, to achieve the optimal performance of the CRN model, the model should better be trained on the current bitemporal images to be repaired. Hence, the simultaneous cloud-detection and removal framework plays a key role in providing the mask of cloud and shadows.

The structure of the CRN is shown in Fig. 5. It can be seen as a simplified version of the CDN, where the DenseNet-style building block is replaced with the simpler VGG building block and the CAM in the decoder is removed. Each VGG block contains two 3 × 3 convolution layers followed by an ReLU. The features after multiscale fusion are used to reconstruct the final cloud-free images. The feature dimensions from the original to the lowest scale in the encoder and decoder are symmetrically set to 64, 128, 256, 512, and 1024.

III. Experimental Results and Analysis

To demonstrate the effectiveness of our integrated framework for cloud and shadow detection and removal, several

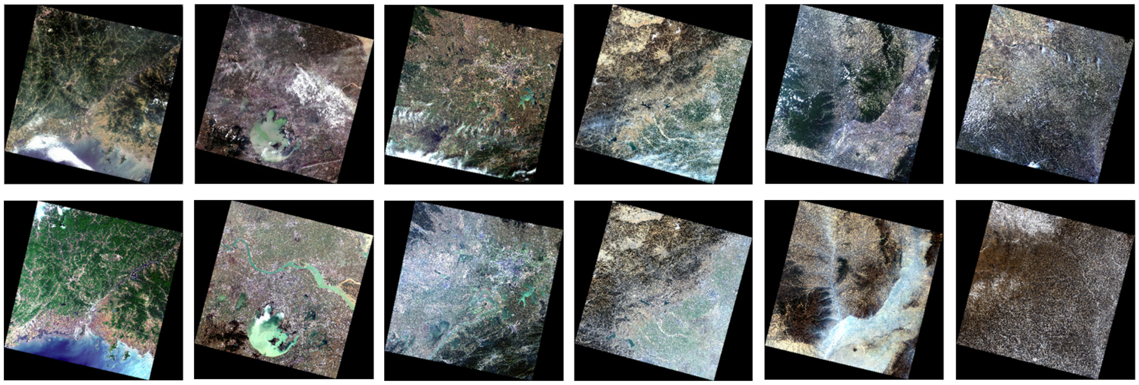


Fig. 6. Six different areas of the WHU Cloud data set. The first row is the cloudy images to be processed and the second row is the temporal auxiliary images.

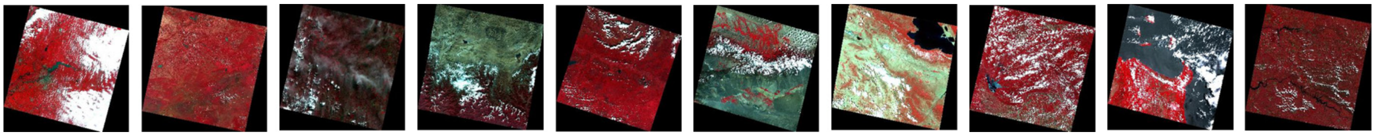


Fig. 7. GF data set, which is very different from the WHU data set, contributing to a reliable comparison between the cloud-detection methods.

TABLE I
SUMMARY OF THE STUDY SITES OF WHU CLOUD DATA SET

Data	Path/row	Location	Covers	Acquired Time	Temporal data Time
I	118/032	Liaoning Province & North Korea	forests, residential areas and sea	2018/06/24	2018/05/23
II	119/038	Jiangsu Province	flat coastal area and large residential areas	2018/05/14	2018/02/23
III	123/039	Hubei Province	Mountains, plains and lakes	2018/04/08	2017/10/30
IV	124/033	Hebei Province	mountains and cities	2018/04/15	2017/12/23
V	126/035	Shanxi province	Mountainous	2018/05/13	2018/03/12
VI	127/034	Shaanxi province	Mountainous	2018/06/23	2018/01/14

experiments are performed and evaluated quantitatively and visually. In Section III-A, the data set is introduced; cloud-detection and removal experiments are described in Sections III-B and III-C, respectively; in Section III-D, the integrated cloud-detection and removal test is analyzed.

A. Experimental Data Set

A new Landsat-8 data set is proposed for simultaneous cloud detection and removal (called WHU Cloud data set). The data set consists of six cloudy and cloud-free image pairs in different areas (Fig. 6). The original data are downloaded from USGS (https://earthexplorer.usgs.gov/). To avoid real land cover changes, the bitemporal images were acquired at a similar time. We manually delineated the areas of clouds and shadows as ground truth. To make up the scarcity of such a data set designed for both cloud detection and removal, we open the data set to facilitate relevant supervised deep learning-based methods (http://gpcv.whu.edu.cn/data/).

Table I lists the location, ground cover, and acquisition time of the six images. The tile of path/row 118/032 is located at the junction of Liaoning Province, China, and North Korea, and covers forests, residential areas, and sea. The image obtained on May 23, 2018 is used as the temporal reference data, and the cloud and shadows in the image on June 24, 2018 is to be detected and repaired. The second tile, path/row 119/38, is located at the eastern flat coastal area of China with a large residential area. The tile of path/row 123/39 consists of mountains and plains with lakes. The fourth area is in Hebei Province, northern China (path/row 124/33). Half of this image covers a plain area with cities and the other half mountains. The tiles path/row 126/35 and path/row 127/34 are mountainous and located in Shanxi and Shaanxi provinces, China.

To evaluate comprehensively our method, an additional GF cloud-detection data set [24], [31] is employed in the cloud- and cloud-shadow-detection task. The data set consists of ten GF cloudy images and covers a wide range of land covers including seaside, lakes, mountains, urbans, cropland, and deserts (Fig. 7).

To make the algorithm compatible with most of the optical satellite images, only red, green, and blue bands, which correspond to bands 4, 3, and 2 of the Landsat 8 images, and 4, 3, and 2 of the GF images, are chosen. In practice, the images were seamlessly cropped into 512 × 512 patches for training and prediction, considering the GPU memory capacity.

For the cloud-detection task, on the WHU Cloud data set, 680 patches are prepared for training, 50 patches for validation, and 129 for test. From the GF data set, 1604 patches are selected for training, 1200 for validation, and 1046 for test.

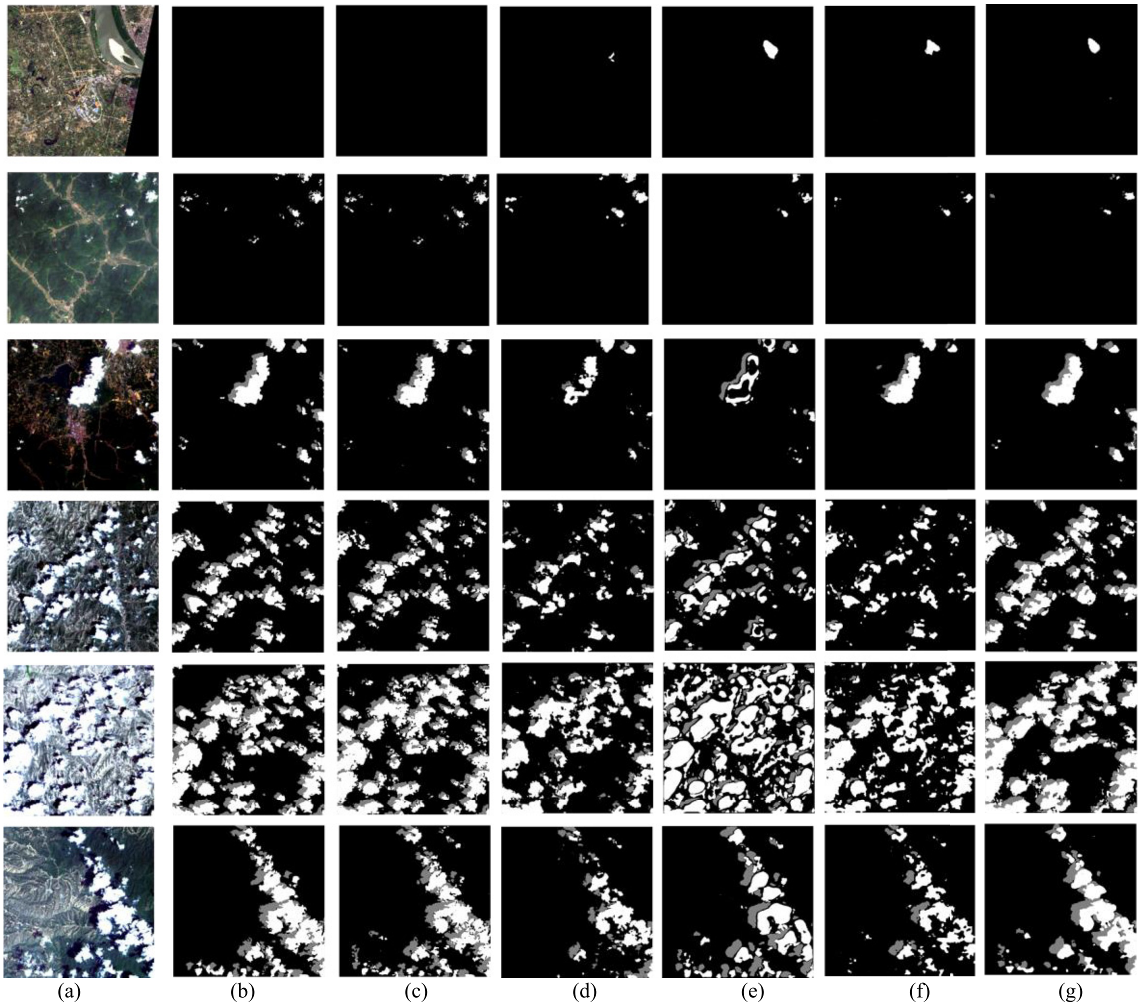


Fig. 8. Comparison of the cloud- and shadow-detection results predicted from different methods on the WHU-Cloud data set. (a) Image. (b) Label. (c) CDN. (d) MSCN. (e) MF-CNN. (f) DeeplabV3+. (g) U-Net. White mask is cloud and the gray one is the cloud shadow.

In the cloud-detection process, the adaptive moment estimation (Adam) optimization algorithm is used and the learning rate is set to $10e^{-5}$. In the cloud-removal process, only randomly simulated samples on the current image pairs are used for training, 86 patches for validation, and 100 for testing on average. The stochastic gradient descent (SGD) is used, and the learning rate is set to $10e^{-4}$. The training processes for both tasks were iterated 1000 epochs for our model and other deep learning-based methods that are applied for comparison. The algorithm is implemented using the Keras framework of Windows 10 environment, with an NVIDIA 11G 1080-Ti GPU.

B. Cloud and Shadow Detection

The proposed CDN for cloud and shadow detection is compared with several general segmentation methods, DeeplabV3+ [56] and U-Net [57], and two very recent CNN-based cloud-detection methods, MF-CNN and MSCN [24], [31], both quantitatively and visually. IoU, recall, precision, and overall accuracy (OA) are employed to evaluate the quantitative results. The OA assesses the pixel-level accuracy including foreground (cloud and cloud shadow treated as two types) and background, and other indicators are for the assessment of foreground detection. In Table II, where different methods are tested on the WHU data set, S&C indicates that the segmentation results of cloud and cloud shadow are counted together. The best result of each index is stressed in bold, and the second-optimal is underlined. It is observed that all the indices of the CDN we proposed surpassed all the other methods in the cloud detection. The IoU of our method is 10% higher than that of the second-best method. In the integrated cloud and cloud shadow detection, the IoU of our method is 9% higher than the second-best one. The DeeplabV3+ performed worse than the U-Net, and the MF-CNN and MSCN performed the worst.

Fig. 8 shows the examples of the cloud and shadow detection using different methods. The MSCN and DeepLabV3+ can only distinguish a part of fragmented clouds. The deeplabV3+ has weak detection ability for shadows, which may be caused by the scarce training samples and coarse upsampling process in its decoder. The MSCN showed poorer performance than the U-Net, likely due to the introduction of batch normalization. According to [58], the performance of

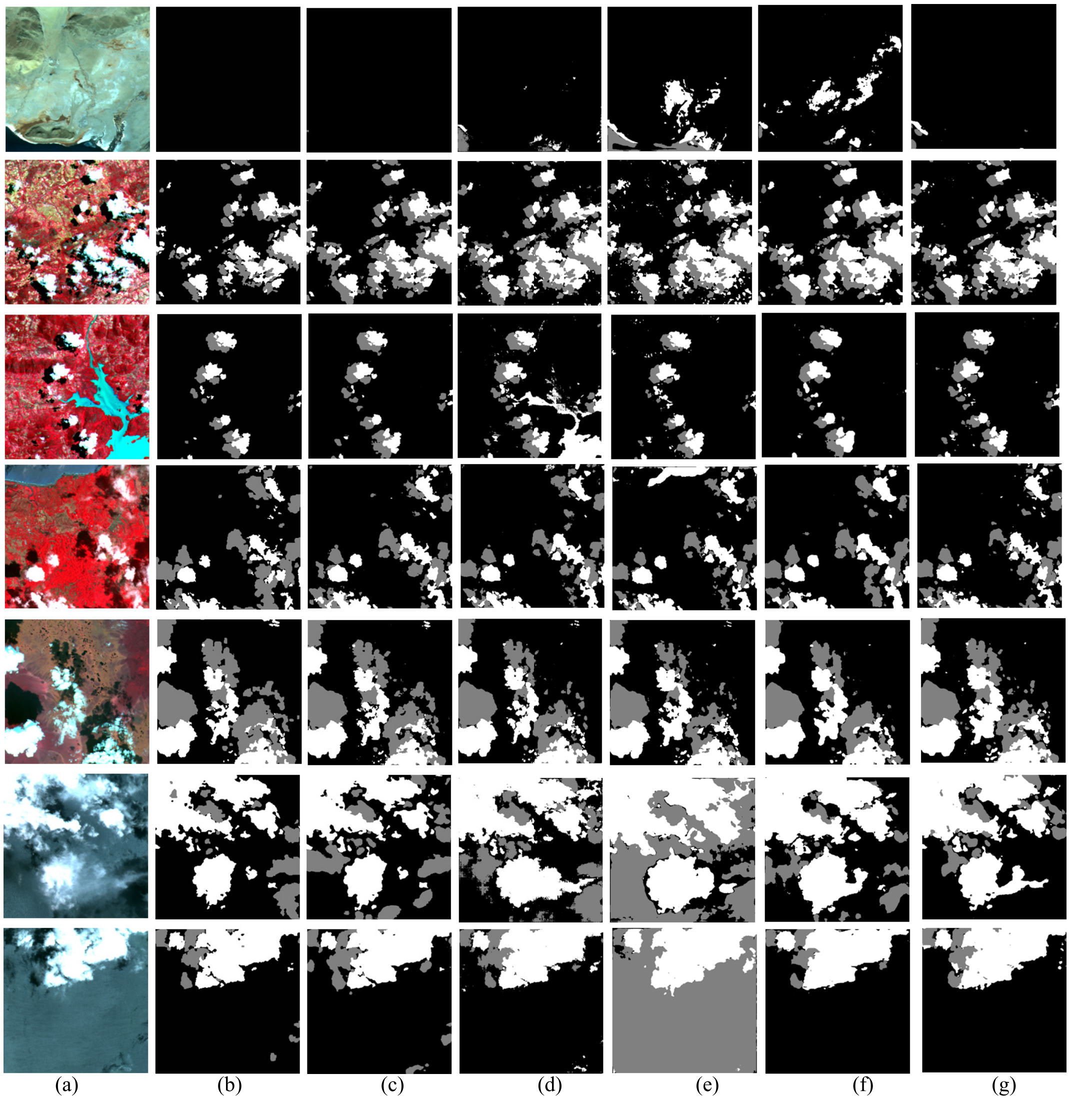


Fig. 9. Comparison of the cloud- and shadow-detection results predicted from different methods on the GF data set. (a) Image. (b) Label. (c) CDN. (d) MSCN. (e) MF-CNN. (f) DeeplabV3+. (g) U-Net.

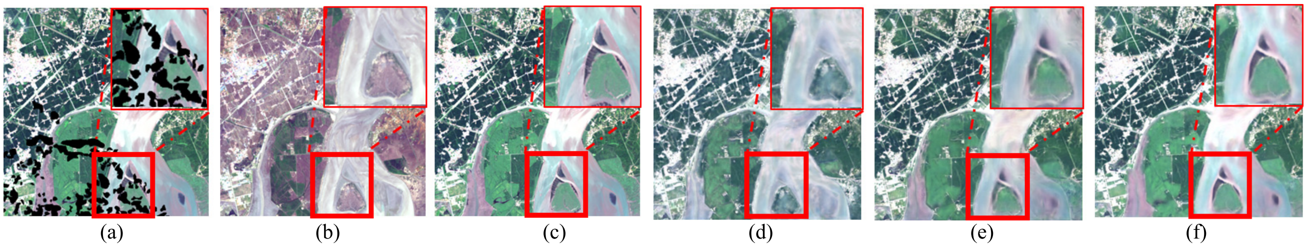


Fig. 10. Comparison between different cloud-removal methods based on the simulated clouds on the WHU data set. The region in the red thick box is enlarged at the top-right corner. (a) Simulated cloudy image (black mask). (b) Temporal image. (c) Ground truth. (d) Result of STSCNN. (e) U-Net. (f) CRN.

batch normalization relies on large batch size. While in our experiment, the batch size is set to 2 due to the restriction of the GPU memory. The results of U-Net are oversmooth due to the lack of the multiscale aggregation strategy, as used in our method. The MF-FCN produced oversharp boundaries. Our CDN performed obviously better than the other methods and predicted a mask map close to the ground truth. With our method, it is guaranteed that clean pixels can be segmented from clouds and shadows to train the CRN that follows.

Table III shows the quantitative results of different methods on the GF data set. Our method performed the best again. It outperformed the second-best DeeplabV3+ over 9% IoU at the cloud and shadow detection. The MSCN and MF-CNN performed almost the same as DeeplabV3+, and U-Net was the worst. Fig. 9 shows some examples of various land cover scenes covering the clouds. Our method exhibited an overall advantage against the other methods at various terrain and atmospheric conditions.

TABLE II

QUANTITATIVE EVALUATION OF THE CLOUD AND SHADOW DETECTION ON WHU CLOUD DATA SET. S&C IS THE JOINT SEGMENTATION RESULT OF CLOUD AND SHADOW

Method	Type	IoU↑	Recall↑	Precision↑	Accuracy↑
MSCN [24, 31]	Shadow	0.2881	0.3608	0.5886	0.9771
	Cloud	0.5828	0.6840	0.7976	0.9735
	S&C	0.4836	0.5798	0.7445	0.9753
MF-CNN [27]	Shadow	0.3963	0.6670	0.4941	0.9738
	Cloud	0.4531	0.8535	0.4913	0.9443
	S&C	0.4361	0.7933	0.4920	0.9501
deeplabV3+ [56]	Shadow	0.3445	0.4212	0.6543	0.9794
	Cloud	0.6267	0.7521	*0.7898*	0.9758
	S&C	0.5345	0.6454	*0.7568*	0.9776
U-Net [57]	Shadow	*0.5372*	*0.6990*	*0.6990*	*0.9845*
	Cloud	*0.6678*	*0.8819*	0.7334	*0.9763*
	S&C	*0.6261*	*0.8229*	0.7236	*0.9804*
CDN	Shadow	**0.6104**	**0.7637**	**0.7525**	**0.9875**
	Cloud	**0.7763**	**0.8863**	**0.8622**	**0.9862**
	S&C	**0.7194**	**0.8468**	**0.8272**	**0.9868**

TABLE III

QUANTITATIVE EVALUATION OF THE CLOUD AND SHADOW DETECTION ON GF DATA SET

Method	Type	IoU↑	Recall↑	Precision↑	Accuracy↑
MSCN [24, 31]	Shadow	*0.4556*	*0.7087*	0.4395	0.9491
	Cloud	0.7844	0.9102	0.8503	0.9565
	S&C	0.6803	*0.8584*	0.7663	0.9528
MF-CNN [27]	Shadow	0.4285	0.5616	*0.6440*	*0.9550*
	Cloud	0.7877	0.8912	0.8715	0.9582
	S&C	0.6850	0.8065	*0.8197*	0.9566
deeplabV3+ [56]	Shadow	0.4347	0.6890	0.5408	0.9536
	Cloud	*0.8230*	*0.9138*	*0.8922*	*0.9697*
	S&C	*0.6970*	0.8572	0.7885	*0.9617*
U-Net [57]	Shadow	0.2416	0.5916	0.2900	0.9038
	Cloud	0.7233	0.8688	0.8120	0.9488
	S&C	0.5273	0.7990	0.6080	0.9263
CDN	Shadow	**0.5878**	**0.7676**	**0.7151**	**0.9676**
	Cloud	**0.8779**	**0.9360**	**0.9341**	**0.9774**
	S&C	**0.7917**	**0.8927**	**0.8749**	**0.9725**

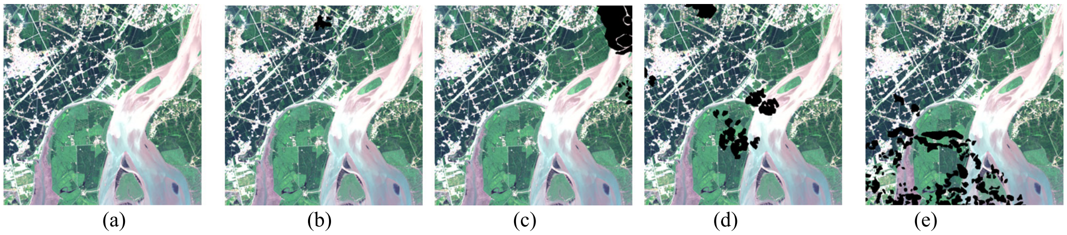


Fig. 11. Simulated masks with different sizes and types. (a) Image with no shaded pixels. (b)–(e) Images with increasing missing pixels. The percentages of the missing pixels in (b)–(e) is 0.7%, 8%, 18%, and 22%, respectively.

C. Cloud Removal

In this section, the experiments of different cloud-removal methods, our CRN, the recent spatial–temporal–spectral CNN (STSCNN) [50], and the U-Net [57] are executed for quantitative and visual comparisons.

Several representative indicators are employed to evaluate the reconstruction result quantitatively, including the structural similarity index measurement (SSIM), peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM), and correlation coefficient (CC). The SSIM measures the similarity in the structures and pixels between the prediction result and the ground truth. The PSNR is the ratio between the maximum power of a signal and the power of noise that affects the fidelity of its representation. The SAM reveals the spectral distortion of the reconstruction result. The CC assesses the correlation of the prediction and ground truth in pixel level.

For quantitative assessment, we randomly simulated the missing pixels in a current image without clouds, as the pixel values beneath the true clouds are inaccessible to evaluate any indicator. The goal is to train a network with the input of the bitemporal images to produce a repaired image approaching to the current image. The ground truth is the whole current image instead of only the shaded area so that the reconstruction effect of an algorithm could be evaluated comprehensively.

After trained with the simulated masks, the prediction results of the three methods are displayed in Table IV. In all the indicators, our CRN considerably exceeded the other two methods. As the U-Net and the CRN share a similar FCN structure, it is concluded that the multiscale strategy through

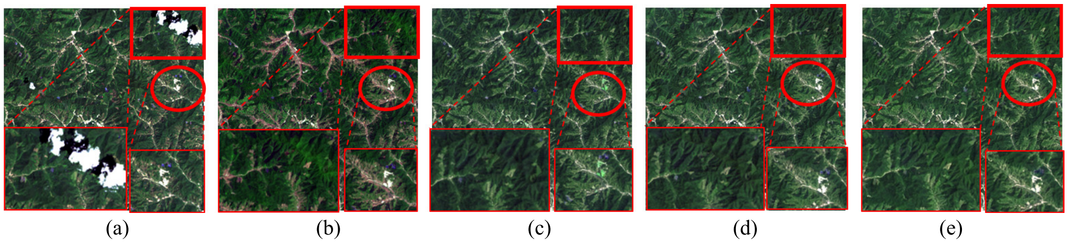


Fig. 12. Comparison of different methods on reconstructing real cloudy area from the WHU data set. (a) Cloudy image. (b) Temporal image. (c) STSCNN. (d) U-Net. (e) CRN.

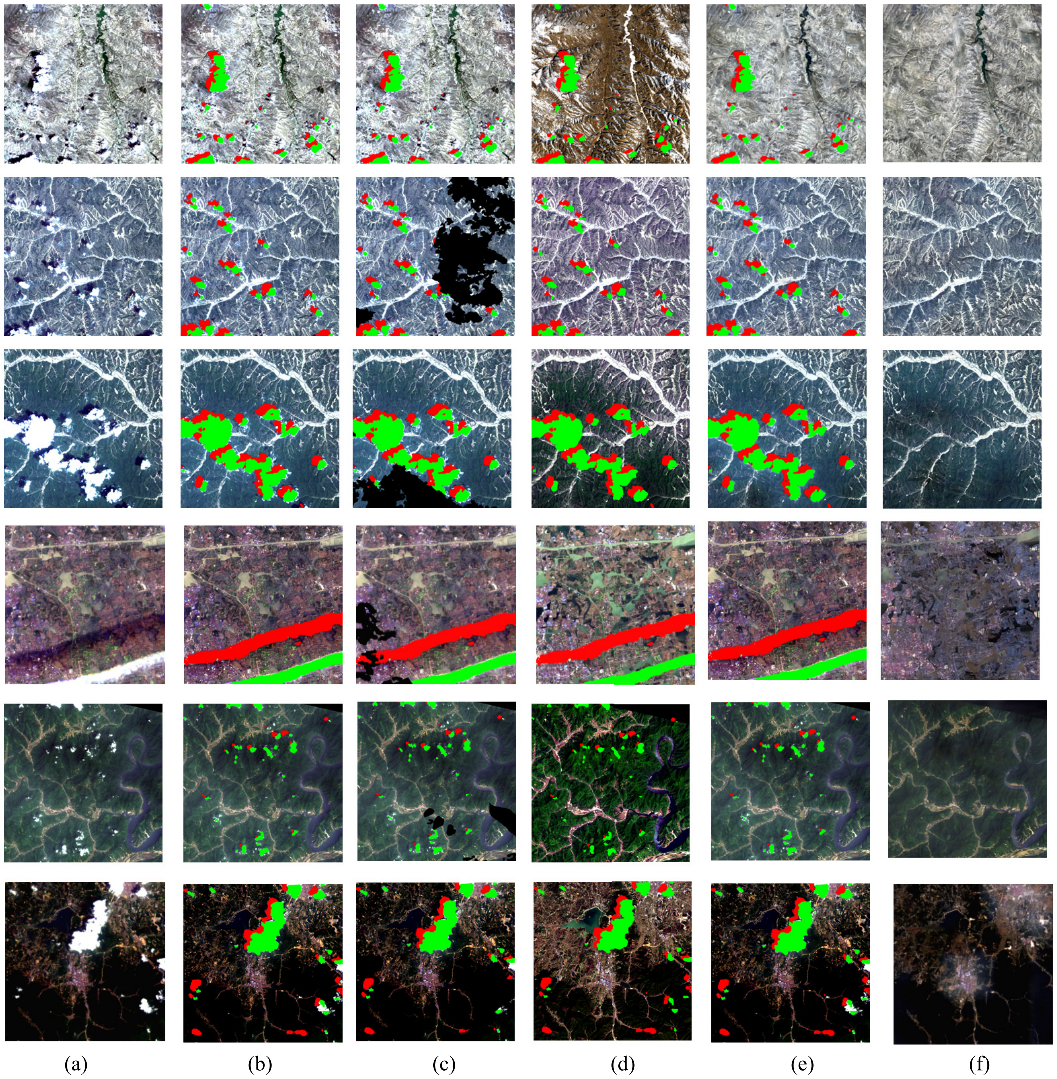


Fig. 13. Prediction results of integrated cloud and shadow detection and removal. (a) Original cloudy data. (b) Cloud (green) and shadow (red) masks detected from the CDN. (c) Simulated random masks (black) for training the CRN. (d) Temporal data excluding the shaded regions. (e) Reconstruction results of simulated masks. (f) Final cloud-removal results. Note that the first row is a special case where the simulated mask is set empty.

a concatenation of the features from different scales improves the performance of an FCN, considering the size differences of the objects on the remote sensing data. Although the STSCNN introduced dilated convolutions to extract features with different-size receptive fields, its inferior performance is resulted from two factors. First, the size of all feature maps is fixed and the feature channels are shallow. It is a time- and memory-consuming CNN structure compared with the mainstream FCN structure. Second, different feature maps extracted from the dilated convolutions cannot completely reflect the scale robustness, because an aggregation of multiscale information is lacking.

Fig. 10 shows an area covering a large river. The missing information, i.e., the simulated black mask, is poorly predicted by the STSCNN [Fig. 10(d)]. The enlarged window contains obvious raw textures from the temporal image, indicating that

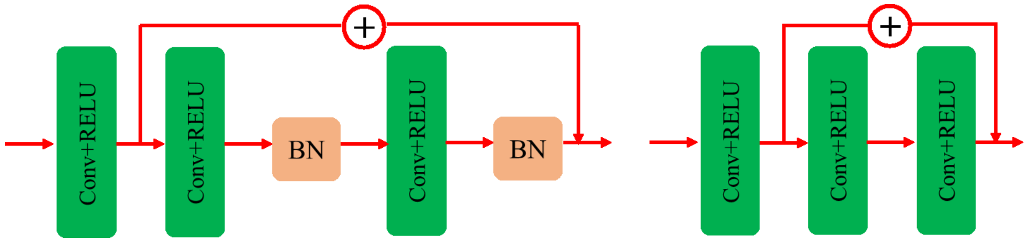


Fig. 14. ResNet-style building block with (Left) BN and (Right) without BN layer.

TABLE IV

Quantitative Evaluation of Different Cloud-Removal Algorithms on Simulated Clouds on Data I of the WHU Data Set

Method	mPSNR ↑	mSSIM↑	SAM↓	CC↑
STSCNN [50]	19.2090	0.7430	4.0797	0.8321
U-Net[57]	*21.6150*	*0.7923*	*3.4299*	*0.9077*
CRN	**22.7103**	**0.8201**	**3.3872**	**0.9277**

TABLE V

Quantitative Evaluation of Our CRN on Different Masks

mask percentage	mPSNR ↑	mSSIM↑	SAM↓	CC↑
0%	21.9531	0.7991	3.3464	0.9517
0.7%	21.9527	0.7991	3.3463	0.9517
8%	21.9425	0.7988	3.3516	0.9516
18%	21.9523	0.7991	3.3479	0.9517
22%	21.9449	0.7987	3.3496	0.9517

the radiometric transformation between the bitemporal images was not properly learned. The U-Net [Fig. 10(e)] and CRN [Fig. 10(f)] show much better results both in textural and spectral information preservations. However, it can be clearly observed that the CRN reconstructed more details, preserved finer and clearer textures, and was the closest to the ground truth. The U-Net blurred the image, as it is predicted only through one upsampling path from the lowest spatial resolution layer.

The algorithm's stability is further evaluated through varying sizes and types of the simulated masks (Fig. 11). Table V shows the metrics stay almost the same with respect to the increased amount of missing pixels, demonstrating the robustness of our algorithm against different sizes and types of masks.

The trained models on the simulated cloud masks can then be applied to reconstruct the real cloudy area. In Fig. 12, we visually investigate the cloud-removal effects of different methods, as the values of the shaded pixels are unavailable for quantitative evaluation. It is observed that the cloud-shaded area in the top-right rectangle is considerably blurred by the STSCNN, which is also confused by the snow in the eclipse. The U-Net and our CRN can preserve the color and texture of the snow area. In the top-right cloudy area, our method preserved more details than the U-Net, demonstrating again the effectiveness of the multiscale strategy.

D. Integrated Cloud Detection and Removal

This experiment evaluates the cloud-removal performance of our integrated cloud-detection and cloud-removal framework. A high recall rate up to 95% of cloud and shadow detection is realized through adaptively adjusting the threshold that translates the probability map obtained from the CDN to a mask map, which ensures most of the cloud areas are located and the rest of the pixels are clean. Then, the mask map is fed into the CRN along with the bitemporal images. Although the real pixels underneath the clouds and shadows are unavailable, the quantitative cloud-removal performance of the integrated framework can be assessed by the simulated masks in the prediction stage.

To demonstrate the advantage of our integrated cloud-detection and removal framework (Fig. 4) and the self-training strategy, we compare it with a CRN that is trained with the available cloud-detection data set. In Table VI, "pretrain" means the model is pretrained on the WHU training set, i.e., all the pixels excluding clouds and shadows are used to train the CRN. "Self-train" means the model is self-trained on the randomly simulated masks of the current image, excluding the detection results from the CDN. From Table VI, our "Self-train" method is comprehensively better than the pretrained CRN. The reason is explicit: due to the diversity of images, e.g., the time interval between the current and temporal data, imaging condition, and regional terrain, which also result in significant varying of the values of PSNR, SSIM, SAM, and CC, the pretrained model has not enough generalization ability against the various situations to recover a new cloudy image that is not trained on. In contrast, our integrated detection and recover model is highly advantageous and can significantly improve the performance from a general pretrained model, as the self-training is specified for only the current image pairs and could be implemented without any labels.

In Fig. 13, the detected clouds and shadows are marked in green and red, respectively [Fig. 13(b)], on the original cloudy images [Fig. 13(a)]. Fig. 13(c) simulates random clouds

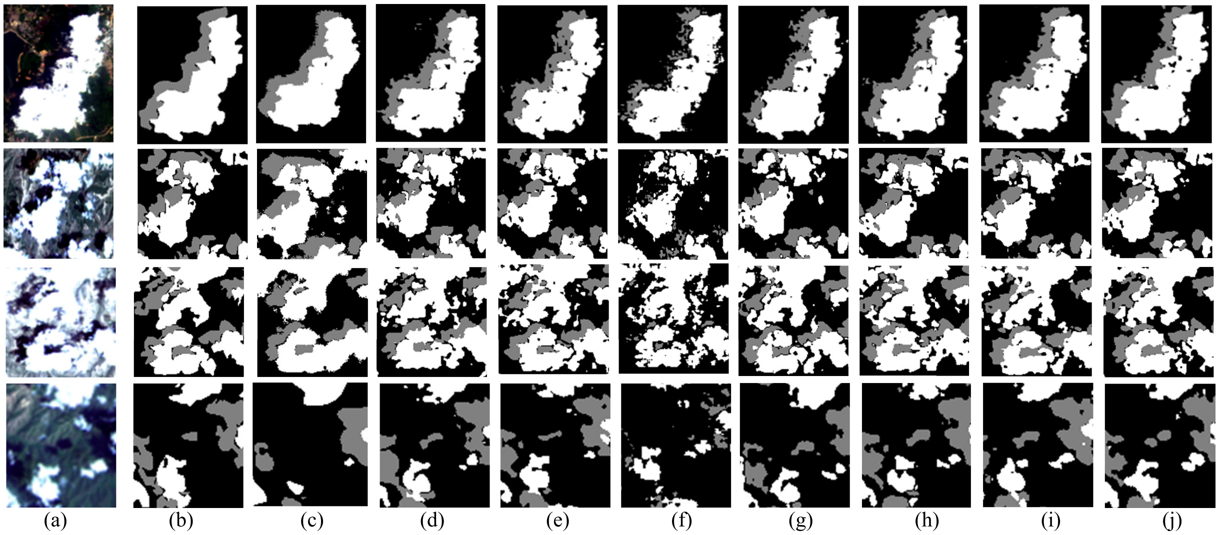


Fig. 15. Comparison of different FCN structures in cloud and shadow detection. (a) Cloudy image. (b) Label. (c) U-Net. (d) CRN(VGG). (e) CRN(ResNet). (f) CRN(ResNet-BN). (g) CRN(DenseNet). (h) CRN(VGG) with CAM. (i) CRN(ResNet) with CAM. (j) CRN(DenseNet) with CAM (i.e., our CDN).

TABLE VI
EVALUATION OF CLOUD REMOVAL ON SIMULATED MASKS USING THE INTEGRATED CDN AND CRN

Indicator / Data	mPSNR ↑		mSSIM ↑		SAM ↓		CC ↑	
	Pre-train	Self-train	Pre-train	Self-train	Pre-train	Self-train	Pre-train	Self-train
I	17.9167	**23.9390**	0.6876	**0.8647**	4.8303	**2.5090**	0.8565	**0.9434**
II	10.9072	**17.9530**	0.4960	**0.7188**	6.7481	**5.1483**	0.6722	**0.6956**
III	10.5489	**19.3818**	0.4902	**0.6814**	10.7707	**4.9251**	0.5560	**0.8217**
IV	14.4284	**19.7801**	0.6023	**0.7513**	7.0706	**2.9932**	0.6669	**0.87212**
V	15.4095	**21.6067**	0.7148	**0.8144**	5.7859	**2.6808**	0.7606	**0.8950**
VI	11.2677	**19.1590**	0.3455	**0.7081**	Nan	Nan	0.4565	**0.8170**

(black) to evaluate quantitatively the CRN along with the temporal image [Fig. 13(d)]. Fig. 13(e) and (f) shows the reconstruction results of only the simulated cloudy masks and the complete cloudy images, respectively. It is observed that the reconstructed images preserved the color and texture of the original images, and the shaded areas were well repaired.

IV. Discussion

The novel deep learning framework for simultaneous cloud detection and removal integrates the tasks of cloud and cloud shadow detection and cloud removal. It not only improves from the recent methods to enable integrated cloud detection and removal but also is a general framework for accommodating different algorithms or tasks, e.g., a traditional spectral-based cloud detection, and other missing-information-reconstruction problems, e.g., the dead lines in the Aqua MODIS band 6 and the Landsat SLC-off problem. The other contribution is that we advanced the CDN from a conventional FCN by introducing a multiscale strategy, a densely connected encoder, and a CAM. In this section, ablation experiments are executed to demonstrate explicitly their effects. The performance of the alternative CNN building blocks and the effect of batch normalization are also discussed. Finally, the limitations and extensions of our method are analyzed.

A. Cloud Detection With Different Building Blocks

The ResNet-style building block is another popular block for a CNN-based feature extractor in addition to the VGG and DenseNet blocks, both of which have been used in our networks. Although the ResNet block is applied in the MSCN [31], the BN layers and small batch size hindered its performance. For the ablation study, we start from the U-Net and the CRN (Fig. 5), and the latter can be seen as a multiscale-aggregation-boosted U-Net. Based on the relatively simple CRN, we replace the VGG block with the ResNet block with and without BN layers (Fig. 14) and DenseNet block, respectively, and introduce the CAM block to assess the performance of the building block variations as well as the effects of the CAM. Note that our CDN for cloud detection is exactly equal to the combination of CRN(DensNet256) and CAM.

Table VII lists the detection results of different methods, which lead to several implications. First, the side effect of the BN layer when the batch size is small is verified, which is consistent with the observation from [58]. The results of the CDN without BN (ResNet) outperformed the CDN with BN (ResNet-BN) more than 15% IoU. Second, the importance of the multiscale strategy is clearly demonstrated. Without multiscale aggregation, the U-Net performed the worst (62.6% IoU in S&C) In contrary, our multiscale-boosted CRN structures with the three different building blocks all obtained much better results, i.e., 68.8% (VGG), 70.1% (ResNet), and 71.4% (DenseNet) IoU on the combined cloud and shadow detection. Third, the CRN (DenseNet), with the least parameters, shows to be marginally advantageous over the other models. The shallow and middle-level features may be important to identify

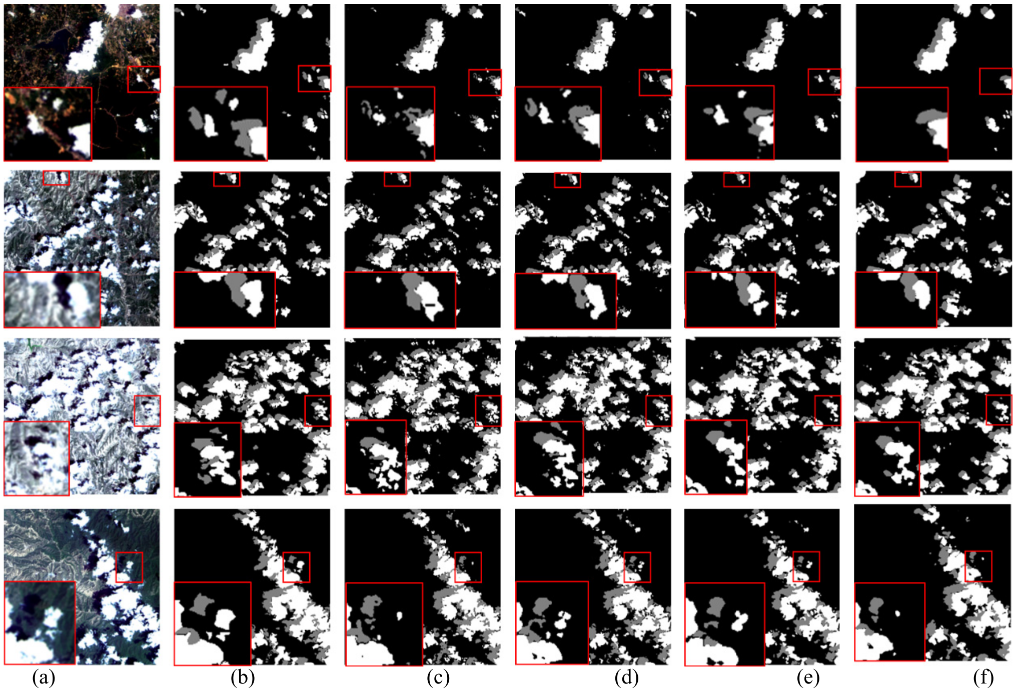


Fig. 16. Examples of cloud and shadow detection with different growth rates based on CDN on the WHU Cloud data set. (a) Image. (b) Label. (c) CRN(VGG). (d) CRN(DenseNet 256). (e) CRN(DenseNet 128). (f) CRN(DenseNet 64).

TABLE VII
QUANTITATIVE EVALUATION OF DIFFERENT FCN STRUCTURES ON THE WHU DATA SET

Method	Type	IoU↑	Recall↑	Precision↑	Accuracy↑	parameters
U-Net [37]	Shadow	0.5372	0.6990	0.6990	0.9845	5,491,427
	Cloud	0.6678	0.8819	0.7334	0.9763	
	S&C	0.6261	0.8229	0.7236	0.9804	
CRN (VGG)	Shadow	0.5736	0.7779	0.6860	0.9851	27,715,279
	Cloud	0.7512	0.8865	0.8312	0.9841	
	S&C	0.6884	0.8515	0.7824	0.9846	
CRN (ResNet-BN)	Shadow	0.2604	0.2896	0.7213	0.9788	33,993,807
	Cloud	0.6618	0.7654	0.8302	0.9789	
	S&C	0.5357	0.6120	0.8115	0.9789	
CRN (ResNet)	Shadow	0.5941	0.7403	0.7506	0.9870	33,985,871
	Cloud	0.7577	0.8483	**0.8764**	0.9853	
	S&C	0.7011	0.8135	0.8353	0.9862	
CRN (DenseNet256)	Shadow	0.5958	0.6928	**0.8096**	**0.9879**	22,404,815
	Cloud	*0.7716*	**0.8881**	0.8547	*0.9858*	
	S&C	*0.7145*	0.8251	**0.8420**	**0.9869**	
CRN(VGG)+CAM	Shadow	*0.6013*	0.7592	0.7430	0.9871	27,715,359
	Cloud	0.7666	0.8859	0.8505	0.9854	
	S&C	0.7100	0.8451	0.8163	0.9862	
CRN(ResNet)+CAM	Shadow	0.5951	**0.8024**	0.6973	0.9860	33,985,951
	Cloud	0.7648	0.8788	0.8550	0.9854	
	S&C	0.7040	**0.8541**	0.8002	0.9857	
CDN	Shadow	**0.6104**	0.7637	*0.7525*	*0.9875*	22,404,895
	Cloud	**0.7763**	0.8863	*0.8622*	**0.9862**	
	S&C	**0.7194**	0.8468	*0.8272*	*0.9868*	

clouds and shadows from the backgrounds. Accordingly, the densely connected architecture that uses all features at different levels in the encoder performed the best. Finally, the channel-attention mechanism is demonstrated effective. Under different building blocks, the performances of using the CAM were all improved compared with the original ones. The best performance is achieved by the combination of CRN (DenseNet256) and CAM, which is precisely the CDN structure we proposed.

Fig. 15 shows four examples of cloud and shadow detection from different FCN structures. The result of U-Net [Fig. 15(c)] is oversmooth and rough due to the absence of a multiscale strategy. The CRN(VGG) [Fig. 15(d)] is obviously better than the U-Net. The map generated by the CRN(ResNet-BN) [Fig. 15(f)] contains many fragments, and shadows are missing; on the contrary, the map from the CRN(ResNet) [Fig. 15(e)] without BN is much better.

TABLE VIII
QUANTITATIVE EVALUATION OF THE CLOUD DETECTION FOR THE ANALYSIS OF GROWTH RATE ON WHU CLOUD DATA SET

Method	Type	IOU↑	Recall↑	Precision↑	Accuracy↑	Parameter
CRN (VGG)	Shadow	0.5736	0.7779	0.6860	0.9851	27,715,279
	Cloud	0.7512	0.8865	0.8312	0.9841	
	S&C	0.6884	0.8515	0.7824	0.9846	
CDN (DenseNet 256)	Shadow	**0.6104**	*0.7637*	**0.7539**	**0.9877**	22,404,895
	Cloud	**0.7763**	**0.8863**	**0.8679**	**0.9865**	
	S&C	**0.7194**	*0.8468*	**0.8311**	**0.9871**	
CDN (DenseNet 128)	Shadow	*0.6011*	**0.7974**	*0.7525*	*0.9875*	10,827,343
	Cloud	*0.767*	0.8732	0.8622	*0.9862*	
	S&C	*0.7073*	**0.8509**	*0.8272*	*0.9868*	
CDN (DenseNet 64)	Shadow	0.5880	0.7633	0.6913	*0.9875*	6,697,487
	Cloud	0.7031	*0.8814*	*0.8626*	0.9858	
	S&C	0.6990	0.8432	0.8073	0.9867	

TABLE IX
QUANTITATIVE EVALUATION OF DIFFERENT FCN STRUCTURES ON SIMULATED CLOUDS ON DATA I OF WHU DATA SET

Method	mPSNR ↑	mSSIM↑	SAM↓	CC↑
CRN	**22.7103**	**0.8201**	**3.3872**	**0.9277**
CRN(ResNet)	21.5115	0.7887	3.6811	0.9056
CRN(DenseNet)	20.5404	0.7982	3.6040	0.9028
CRN(CAM)	*22.4013*	*0.8132*	*3.3947*	*0.9227*
CRN(ResNet+CAM)	21.3332	0.7920	3.6205	0.9010
CRN(DenseNet+CAM)	22.0024	0.8119	3.3959	0.9171

The maps generated by the CRN(DenseNet) [Fig. 15(g)] are similar to the VGG and ResNet-based structure. Boosted with the CAM that helped reorganize the between-channel consistency of the features from different scales, all the three structures [Fig. 15(h)–(j)] can delineate more details on individual cloud and shadow with more precise boundaries.

B. Growth Rate in Densely Connected Architecture

The dimension of the feature maps in each DenseNet block, which is called growth rate [53], controls the feature dimension and the total amount of parameters. In this section, the effect of growth rate is discussed. Three different cloud detection experiments are conducted, in which the growth rate is set as 256, 128, and 64, respectively.

The quantitative evaluations of cloud and shadow detection on the WHU Cloud data set for the analysis of growth rate are presented in Table VIII. The IoU of the CRN(DenseNet) with growth rate 256 is significantly higher than that of CRN(VGG) by 2% on average. The result with the growth rate equal to 64 also proves that a small growth rate is sufficient to obtain perfect detection performance on the data set. It is worth noting that the consistent conclusion as quantitative evaluation is made from Fig. 16. Local details are amplified in the bottom left of the image, CRN(DenseNet) with the growth rate equal to 256 achieves the most comprehensive and complete detection. While the result of CRN(VGG) is fragmented.

C. Cloud Removal Using Different FCN Structures

We also compare the performance of different building blocks and CAM in cloud removal. Table IX shows the quantitative comparisons of different FCN structures based on the simulated cloudy images. The performances of different FCN structures for cloud removal are generally the opposite of their performances in the cloud detection. The CRN(VGG) outperformed the other models in all the indicators. The SAM score of the CRN(ResNet) is the highest, indicating the result has severe spectral distortion. The CRN(DenseNet) performs poor as well. The indication is that in the cloud-removal task, the simplest VGG block performs well. Moreover, it is observed that the CAM is not necessary, which even slightly downturned the performances. It is uncommon in image classification or semantic segmentation, where an attention mechanism almost always boosts the CNN backbone. These could be explained by the dissimilarities between the two tasks. Cloud removal is a pixel-to-pixel mapping from the temporal image to the current image; in other words, every pixel plays its role and concentrating on parts of pixels seems unnecessary. It is worth noting that although simple network structures seem welcoming in the cloud-removal task, the multiscale strategy is necessary, which is the reason that our CRN outperforms the U-Net on all the indicators in Table IV.

The qualitative comparison of different cloud-removal methods on real clouds is shown in Fig. 17. The CRN, CRN(ResNet), and CRN(DenseNet) all repaired the images well. In the first row of Fig. 17, the spectral distortion of the CRN [Fig. 17(c)] is the smallest, and the zoomed-in region of the other five methods [Fig. 17(d)–(h)] seems darker, which is impacted by the dark temporal image. In the second row, except the CRN(DenseNet) with CAM [Fig. 17(h)] mistook the color in the circled region, and all the other methods [Fig. 17(c)–(g)] obtained satisfactory results.

From the comparison of the cloud-detection and cloud-removal experiments under different network structures and settings, it could be concluded that the performance of a CNN on a specific task is not fully dependent on the complex structures and extensive parameters, even if samples are sufficient,

TABLE X
CLOUD AND SHADOW DETECTION RESULTS (IoU) ON DIFFERENT TYPES OF TERRAINS AND LAND USES ON WHU AND GF DATA SETS

Data	Terrain Type	Bare-land	Mountain	Flat	farmland	Snow/ice	Urban	Desert	Sea/lake
WHU Dataset	Shadow	0.4890	0.5912	0.5844	0.6155	0.5765	0.4670	/	/
	Cloud	0.5712	0.8068	0.7748	0.7978	0.7079	0.9176	/	/
	S&C	0.5503	0.7348	0.7052	0.7369	0.6660	0.7449	/	/
GF Dataset	Shadow	0.5516	0.4867	0.5118	0.4213	0.3558	/	0.8375	0.4691
	Cloud	0.7339	0.8972	0.8404	0.7962	0.7022	/	0.9373	0.9104
	S&C	0.6593	0.7630	0.7239	0.7049	0.4527	/	0.8928	0.7426

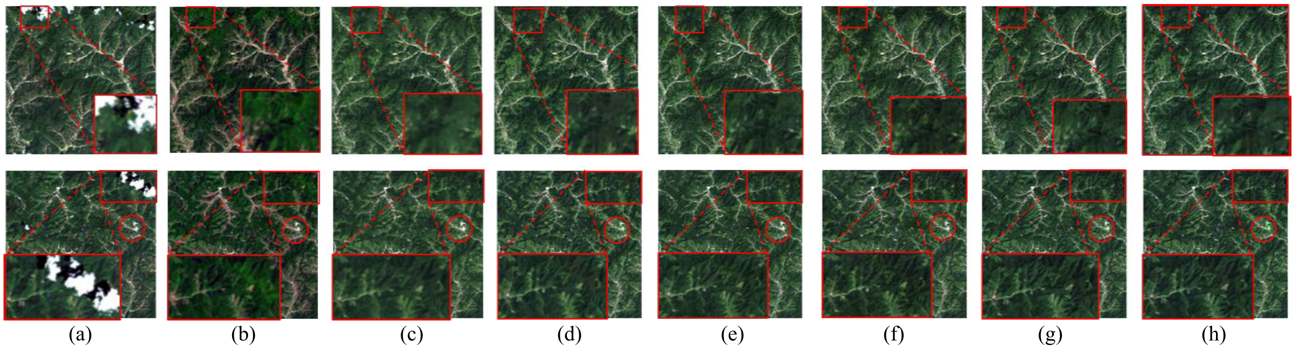


Fig. 17. Comparison of the cloud-removal methods with real clouds. (a) Cloudy image. (b) Temporal image. (c) CRN. (d) CRN(ResNet). (e) CRN(DenseNet). (f) CRN with CAM. (g) CRN(ResNet) with CAM. (h) CRN(DenseNet) with CAM.

for example, in the case of cloud removal, every clean pixel is a sample. This implies that empirical network structure designing is important for a specific remote sensing data processing task, just as handcraft feature designing for a conventional machine learning method. The automated machine learning (AutoML) technology may be introduced in further study, as it may automatically find optimal models for a specific task.

D. Impact of Different Land Types on Cloud Detection

We list the detection results (in IoU score) on different types of terrains or land uses on both WHU and GF data sets using our CDN (Table X). After the CDN model has been pretrained on the training samples covering different types, it is applied to the subsets of the images, which each singles out a terrain or land use type. From Table X, the cloud-detection performance is relatively less optimal in snow or ice regions, which is likely due to the similar patterns shared between snow, ice, and clouds. Bare land also obtained a lower score in both data sets. This is caused by the limited areas of bare land in both data sets, leading to inadequate training samples of this land cover type. In other terrain or land use types, the CDN showed stable performance and the IoU scores all exceeded 70%.

E. Limitation and Extension

Our framework can be directly extended to multi or hyperspectral, medium- or low-resolution remote sensing images, as these structured remote sensing data can be easily regularized, resized, or compacted by a convolution operator to be the input of any CNN structure. The limitation exists in the applications on high-resolution images. The registration accuracy of the high-resolution temporal images, e.g., Worldview images, may hardly reach subpixels. This will impact the learning ability of the CRN from pixel pairs containing inaccurate correspondents. An available high-accuracy digital surface model (DSM) can help the registration to achieve subpixel accuracy, or the images could be cropped smaller to improve the local registration accuracy.

The WHU and GF data sets mainly consist of various types of heavy clouds. However, as the deep learning-based methods are generic (as has been widely demonstrated in various disciplines and applications), we believe when trained with data sets covering light clouds, our model is capable of predicting light clouds.

The percentage of cloudage is not a significant parameter to our CRN. The CRN can be trained with all the noncloud pixel pairs. In Table IV, we have demonstrated that its performance was not affected by cloud coverage that is up to 22%. Taking a Landsat 8 image with 60% cloud coverage for example, there are still millions of samples remained to train a CRN model. However, in practice, an image with more than 30% or 50% cloud coverage may be considered removed instead of repaired.

The result of the CRN depends on the high performance of the CDN, which may not be always guaranteed due to the diversity in the remote sensing data, different imaging situations, and labeled training samples. The strategy in this article is to improve the recall rate as much as possible after the CDN model has been trained with available samples. The precision of cloud detection is correspondingly reduced, but this does not affect the training of the CRN. A possible improvement is to consider the precision–recall curve of the background, from which these clean pixels with a high degree of confidence can be used preferentially.

V. Conclusion

In this article, we proposed a novel deep learning framework for the integrated cloud and cloud shadow detection and removal from the bitemporal images. This integration greatly simplifies the processing of cloudy remote sensing images. Our framework is also generic to different cloud-detection or removal algorithms and other missinformation-reconstruction problems in remote sensing.

We developed two FCNs that formed the main body of the framework, and their performances exceeded those recent methods designed for separate cloud detection or cloud removal on a large margin. In the CDN, multiscale aggregation and densely connected encoder handle the complex and scale effect of the remote sensing data and the CAM leads to the global consistency of features from multiscales and different channels. In the CRN, we applied a self-training strategy to eliminate the need of manual labels. The data set is also opened to facilitate the development of supervised deep learning methods that heavily relay on large and quality-labeled samples.

References

[1] K. Yuan, G. Meng, D. Cheng, J. Bai, S. Xiang, and C. Pan, "Efficient cloud detection in remote sensing images using edge-aware segmentation network and easy-to-hard training strategy," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 61–65.

[2] P. Dai, H. Zhang, L. Zhang, and H. Shen, "A remote sensing spatiotemporal fusion model of landsat and modis data via deep learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 7030–7033.

[3] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[4] J. D. Braaten, W. B. Cohen, and Z. Yang, "Automated cloud and cloud shadow identification in landsat MSS imagery for temperate ecosystems," *Remote Sens. Environ.*, vol. 169, pp. 128–138, Nov. 2015.

[5] A. Fisher, "Cloud and cloud-shadow detection in SPOT5 HRG imagery with automated morphological feature extraction," *Remote Sens.*, vol. 6, no. 1, pp. 776–800, 2014.

[6] S. A. Ackerman, R. E. Holz, R. Frey, E. W. Eloranta, B. C. Maddux, and M. McGill, "Cloud detection with MODIS. Part II: Validation," *J. Atmos. Ocean. Technol.*, vol. 25, no. 7, pp. 1073–1086, Jul. 2008.

[7] G. Mateo-Garcia, L. Gomez-Chova, and G. Camps-Valls, "Convolutional neural networks for multispectral image cloud masking," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 2255–2258.

[8] Z. Li, H. Shen, H. Li, G. Xia, P. Gamba, and L. Zhang, "Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery," *Remote Sens. Environ.*, vol. 191, pp. 342–358, Mar. 2017.

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

[10] E. Maltezos, A. Doulamis, and C. Ioannidis, "Improving the visualisation of 3D textured models via shadow detection and removal," in *Proc. 9th Int. Conf. Virtual Worlds Games for Serious Appl. (VS-Games)*, Sep. 2017, pp. 161–164.

[11] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri, "Automatic shadow detection and removal from a single image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 431–446, Mar. 2016.

[12] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri, "Automatic feature learning for robust shadow detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1939–1946.

[13] B. Wang, A. Ono, K. Muramatsu, and N. Fujiwara, "Automated detection and removal of clouds and their shadows from Landsat TM images," *IEICE Trans. Inf. Syst.*, vol. 82, no. 2, pp. 453–460, 1999.

[14] J. Bian, A. Li, H. Jin, W. Zhao, G. Lei, and C. Huang, "Multi-temporal cloud and snow detection algorithm for the HJ-1A/B CCD imagery of China," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2014, pp. 501–504.

[15] O. Hagolle, M. Huc, D. V. Pascual, and G. Dedieu, "A multi-temporal method for cloud detection, applied to FORMOSAT-2, VEN μS, LANDSAT and SENTINEL-2 images," *Remote Sens. Environ.*, vol. 114, no. 8, pp. 1747–1755, Aug. 2010.

[16] Y. Lee, G. Wahba, and S. A. Ackerman, "Cloud classification of satellite radiance data by multicategory support vector machines," *J. Atmos. Ocean. Technol.*, vol. 21, no. 2, pp. 159–169, Feb. 2004.

[17] M. R. Azimi-Sadjadi and S. A. Zekavat, "Cloud classification using support vector machines," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, vol. 2, Jul. 2000, pp. 669–671.

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

[19] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data," *J. Amer. Stat. Assoc.*, vol. 99, no. 465, pp. 67–81, Mar. 2004.

[20] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, Apr. 2018, Art. no. 7068349.

[21] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016.

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

[23] F. Xie, M. Shi, Z. Shi, J. Yin, and D. Zhao, "Multilevel cloud detection in remote sensing images based on deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3631–3640, Aug. 2017.

[24] 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, Apr. 2019.

[25] 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, Aug. 2019.

[26] M. Wieland, Y. Li, and S. Martinis, "Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network," *Remote Sens. Environ.*, vol. 230, Sep. 2019, Art. no. 111203.

[27] 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, Jun. 2019.

[28] C.-C. Liu *et al.*, "Clouds classification from Sentinel-2 imagery with deep residual learning and semantic image segmentation," *Remote Sens.*, vol. 11, no. 2, p. 119, 2019.

[29] 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, Aug. 2019.

[30] D. Chai, S. Newsam, H. K. Zhang, Y. Qiu, and J. Huang, "Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks," *Remote Sens. Environ.*, vol. 225, pp. 307–316, May 2019.

[31] M. Qin, F. Xie, W. Li, Z. Shi, and H. Zhang, "Dehazing for multispectral remote sensing images based on a convolutional neural network with the residual architecture," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1645–1655, May 2018.

[32] 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.

[33] L. Lorenzi, F. Melgani, and G. Mercier, "Inpainting strategies for reconstruction of missing data in VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 5, pp. 914–918, Sep. 2011.

[34] A. C. Siravenha, D. Sousa, A. Bispo, and E. Pelaes, "Evaluating inpainting methods to the satellite images clouds and shadows removing," in *Signal Processing, Image Processing and Pattern Recognition*. Berlin, Germany: Springer, 2011, pp. 56–65.

[35] C. Yu, L. Chen, L. Su, M. Fan, and S. Li, "Kriging interpolation method and its application in retrieval of MODIS aerosol optical depth," in *Proc. 19th Int. Conf. Geoinf.*, Jun. 2011, pp. 1–6.

[36] Q. Cheng, H. Shen, L. Zhang, and P. Li, "Inpainting for remotely sensed images with a multichannel nonlocal total variation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 175–187, Jan. 2014.

[37] Y. Zhang, B. Guindon, and J. Cihlar, "An image transform to characterize and compensate for spatial variations in thin cloud contamination of landsat images," *Remote Sens. Environ.*, vol. 82, nos. 2–3, pp. 173–187, Oct. 2002.

[38] X. Y. He, J. B. Hu, W. Chen, and X. Y. Li, "Haze removal based on advanced haze-optimized transformation (AHOT) for multispectral imagery," *Int. J. Remote Sens.*, vol. 31, no. 20, pp. 5331–5348, Oct. 2010.

[39] H. Li, L. Zhang, H. Shen, and P. Li, "A variational gradient-based fusion method for visible and SWIR imagery," *Photogramm. Eng. Remote Sens.*, vol. 78, no. 9, pp. 947–958, Sep. 2012.

[40] M. Xu, X. Jia, and M. Pickering, "Automatic cloud removal for landsat 8 OLI images using cirrus band," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2014, pp. 2511–2514.

[41] H. Lv, Y. Wang, and Y. Shen, "An empirical and radiative transfer model based algorithm to remove thin clouds in visible bands," *Remote Sens. Environ.*, vol. 179, pp. 183–195, Jun. 2016.

[42] X. Li, L. Wang, Q. Cheng, P. Wu, W. Gan, and L. Fang, "Cloud removal in remote sensing images using nonnegative matrix factorization and error correction," *ISPRS J. Photogramm. Remote Sens.*, vol. 148, pp. 103–113, Feb. 2019.

[43] D.-C. Tseng, H.-T. Tseng, and C.-L. Chien, "Automatic cloud removal from multi-temporal SPOT images," *Appl. Math. Comput.*, vol. 205, no. 2, pp. 584–600, Nov. 2008.

[44] 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.

[45] C.-H. Lin, K.-H. Lai, Z.-B. Chen, and J.-Y. Chen, "Patch-based information reconstruction of cloud-contaminated multitemporal images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 163–174, Jan. 2014.

[46] C. Zhang, W. Li, and D. J. Travis, "Restoration of clouded pixels in multispectral remotely sensed imagery with cokriging," *Int. J. Remote Sens.*, vol. 30, no. 9, pp. 2173–2195, May 2009.

[47] X. Zhu, F. Gao, D. Liu, and J. Chen, "A modified neighborhood similar pixel interpolator approach for removing thick clouds in landsat images," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 521–525, May 2012.

[48] X. Li, H. Shen, L. Zhang, H. Zhang, Q. Yuan, and G. Yang, "Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7086–7098, Nov. 2014.

[49] S. Tahsin, S. Medeiros, M. Hooshyar, and A. Singh, "Optical cloud pixel recovery via machine learning," *Remote Sens.*, vol. 9, no. 6, p. 527, 2017.

[50] 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.

[51] N. Hoan and R. Tateishi, "Cloud removal of optical image using SAR data for ALOS applications. Experimenting on simulated ALOS data," *J. Remote Sens. Soc. Jpn.*, vol. 29, no. 2, pp. 410–417, Apr. 2009.

[52] R. Eckardt, C. Berger, C. Thiel, and C. Schmullius, "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.

[53] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

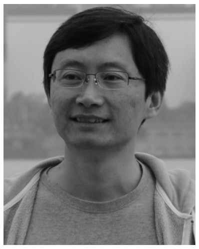
[54] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3146–3154.

[55] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[56] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.

[57] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[58] W. Yuxin and H. Kaiming, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.



Shunping Ji (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2007.

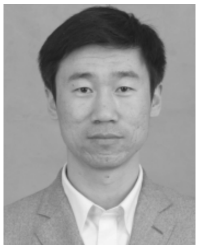
He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include photogrammetry, remote sensing image processing, mobile mapping system, and machine learning.

Peiyu Dai (Student Member, IEEE) received the M.S. degree in surveying and mapping engineering from Wuhan University, Wuhan, China, in 2018, where she is pursuing the Ph.D. degree with the School of Remote Sensing and Information Engineering.

Her research interests include remote sensing and machine learning.

Meng Lu received the M.Sc. degree in Earth science system from the University of Buffalo, SUNY, Buffalo, NY, USA, in 2013, and the Ph.D. degree in geoinformatics with the University of Münster, Münster, Germany, in 2017.

She joined the Department of Physical Geography, Utrecht University, Utrecht, The Netherlands, as a Research Associate, specializing in spatial data analysis, environmental modeling, and geocomputation. Her research interests include geoscientific data analysis, spatiotemporal statistics, machine learning, remote sensing, environmental modeling, and health geography.



Yongjun Zhang received the B.S., M.S., and Ph.D. degrees from Wuhan University (WHU), Wuhan, China, in 1997, 2000, and 2002, respectively.

He is a Professor of photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering, WHU. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource data sets, integration of LiDAR point clouds and images, and 3-D city reconstruction.