CDnet: CNN-Based Cloud Detection for Remote Sensing Imagery

Jingyu Yang, *Senior Member, IEEE*, Jianhua Guo, *Student Member, IEEE*, Huanjing Yue, *Member, IEEE*, Zhiheng Liu, Haofeng Hu, and Kun Li, *Member, IEEE*

***Abstract*— Cloud detection is one of the important tasks for remote sensing image (RSI) preprocessing. In this paper, we utilize the thumbnail (i.e., preview image) of RSI, which contains the information of original multispectral or panchromatic imagery, to extract cloud mask efficiently. Compared with detection cloud mask from original RSI, it is more challenging to detect cloud mask using thumbnails due to the loss of resolution and spectrum information. To tackle this problem, we propose a cloud detection neural network (CDnet) with an encoder–decoder structure, a feature pyramid module (FPM), and a boundary refinement (BR) block. The FPM extracts the multiscale contextual information without the loss of resolution and coverage; the BR block refines object boundaries; and the encoder–decoder structure gradually recovers segmentation results with the same size as input image. Experimental results on the ZY-3 satellite thumbnails cloud cover validation data set and two other validation data sets (GF-1 WFV Cloud and Cloud Shadow Cover Validation Data and Landsat-8 Cloud Cover Assessment Validation Data) demonstrate that the proposed method achieves accurate detection accuracy and outperforms several state-of-the-art methods.**

***Index Terms*— Cloud detection, cloud detection neural network (CDnet), deep convolutional neural network (DCNN), satellite imagery, thumbnails.**

I. Introduction

With the rapid development of remote sensing technology, high-resolution satellite imagery is readily available and has been widely used in agriculture engineering [1], environmental protection [2], land or mineral resource exploration [3], geographical survey [4], and military reconnaissance [5]. Since nearly 66% earth surface is covered by clouds [6], most remote sensing imageries would inevitably be contaminated by clouds. Cloud coverage degrades the quality of satellite imagery by disabling satellite sensor to obtain clear views of the earth's surface, thus affecting imagery postprocessing, such as remote sensing image (RSI) classification and segmentation [7], image matching [8], and 3-D surface generation [9]. Hence, it is important to quickly and accurately detect cloud mask to assess the quality of remote sensing imagery.

Most previous works utilized spectral information (far infrared and thermal infrared information) in hyperspectral/multispectral remote sensing imagery to identify and segment clouds. Typical methods include International Satellite Cloud Climatology Project (ISCCP) [10], Clouds from the Advanced Very High Resolution Radiometer (CLAVR) [11], and AVHRR Processing scheme Over clouds, Land and Ocean (APOLLO) [12]. However, some high-resolution remote sensing imageries, e.g., China's ZY-3 multispectral imagery [7], have only four bands (blue, green, red, and near infrared), which are challenging to reliably detect cloud. To improve cloud detection performance, it is necessary to incorporate discriminate features, such as texture, geometry, and ground objects' size [13]. However, it is difficult to design discriminative features because the high complexity of targets and the large coverage of high-resolution remote sensing imagery usually lead to tremendous computational complexity.

The most straightforward way to reduce computational complexity is working on subsampled images or even thumbnail images (i.e., preview image) [14]. Thumbnail images contain necessary ground objects' information of original multispectral/hyperspectral or panchromatic images for preview and have smaller sizes. But a thumbnail image generally contains only an RGB image with three bands or even a gray image with only one band. Hence, cloud detection from thumbnail images is more difficult than that from high-resolution hyperspectral or multispectral RSI [15], [16], especially for images with cloud-snow coexistence as shown in Fig. 1. Research work on this line [15], [17]–[20] has achieved high accuracy for cloud detection, but most of them fail to distinguish between cloud and snow. Therefore, it is desirable to develop more powerful feature description and classification techniques in order to obtain accurate cloud detection from thumbnails.

In recent years, the deep neural network has achieved tremendous success in image analysis and recognition, significantly outperforming traditional machine learning across

Manuscript received September 18, 2018; revised January 23, 2019 and March 1, 2019; accepted March 9, 2019. Date of publication April 3, 2019; date of current version July 22, 2019. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61771339, Grant 61571322, Grant 61672378, and in part by the Tianjin Science and Technology Program under Grant 17ZXRGGX00160 and Grant 18JCYBJC19200. *(Corresponding author: Kun Li.)*

J. Yang, J. Guo, and H. Yue are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: yjy@tju.edu.cn; g_j_h@tju.edu.cn; huanjing.yue@tju.edu.cn).

Z. Liu is with the School of Geology Engineering and Geomatics, Chang'an University, Xi'an 710054, China (e-mail: liuzhiheng@chd.edu.cn).

H. Hu is with the School of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin 300072, China (e-mail: haofeng_hu@tju.edu.cn).

K. Li is with the Tianjin Key Laboratory of Cognitive Computing and Application, School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: lik@tju.edu.cn).

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

Digital Object Identifier 10.1109/TGRS.2019.2904868

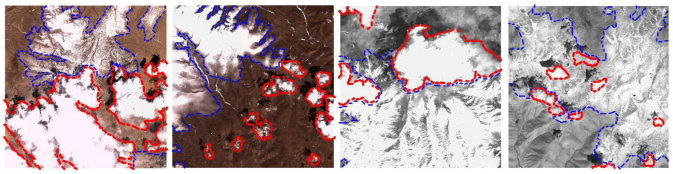


Fig. 1. Illustration of cloud–snow coexistence in remote sensing imagery, where cloud and snow regions are sketched by red and blue dotted lines, respectively.

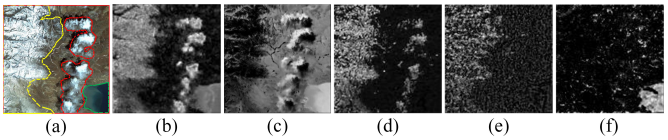


Fig. 2. Discriminative features extracted by CDnet from a ZY-3 satellite thumbnail. (a) Input thumbnail. (b)–(f) Five exemplar feature maps. The cloud, snow, and water regions are sketched by red, yellow, and green dotted lines, respectively.

many vision tasks [21]. Based on the powerful deep learning, we propose a new cloud detection method, cloud detection neural network (CDnet), to segment cloud regions from thumbnails of RSI. Different from previous deep learning-based semantic segmentation models, such as global convolutional network (GCN) [22], Deeplab V2 [23], Deeplab V3 [24], and pyramid scene parsing network (PSPnet) [25], the CDnet simultaneously considers the multiscale and global contextual information, object boundaries, and score map resolution. Specifically, we first proposed a *feature pyramid module* (FPM) to extract multiscale and global contextual information for category recognition of image regions. Then, we introduce a *boundary refinement* (BR) module [22] to capture sharp and detailed object boundaries. Most importantly, the CDnet has an encoder–decoder network structure, which exploits features at multilevel layers to generate sharp boundaries and gradually recovers score map resolution. As shown in Fig. 2, the extracted features from the three channel thumbnails are discriminative for semantic regions, e.g., snow, cloud, and water, which would yield excellent cloud detection performance without spectral information. Specifically, in the feature map Fig. 2(b), snow, cloud, and water regions are activated. In Fig. 2(c), cloud, water, and land regions are activated. In Fig. 2(d), snow and cloud regions are activated. In Fig. 2(e), snow regions are activated. In Fig. 2(f), the water region is activated. By combining these discriminative feature maps in the powerful network, cloudy regions are able to be reliably identified without spectral information even for challenging cloud–snow coexistence areas. Experimental results show that the CDnet is able to achieve excellent cloud detection performance for RSI.

The main contribution of this paper is summarized as follows. First, compared with the work [22], we propose an FPM module to extract the multiscale and global context information. The proposed FPM consists of three kinds of features, i.e., four parallel dilated convolution layers to extract multiscale features, one global average pooling (GAP) block to extract global features, and the original input features. In contrast, the GCN module in [22] only extracts single-scale features. Compared with ASPP+GAP proposed in DeepLabV3 [24], the proposed FPM introduces a short connection between the input and the output layer. This makes the FPM block has the ability in alleviating the vanishing-gradient problem, strengthening feature propagation, and encouraging feature reusing. Experimental results show that the CDnet with the FPM module achieves better cloud detection performance than that with the ASPP+GAP module. Second, the proposed CDnet for cloud segmentation from remote sensing imagery scales well to various data with different spatial resolutions and spectral components. On the one hand, the CDnet achieves accurate detection performance using only partial information, i.e., thumbnails or panchromatic imagery captured by the ZY-3 satellite, which significantly saves memory and computational resources. On the other hand, the CDnet also provides accurate results for the full data captured by the GF-1satellite and the Landsat-8 satellite.

The remainder of this paper is organized as follows. In Section II, we briefly review related work on detection. In Section III, we introduce the CDnet framework and present key modules in details. In Section IV, we evaluate the modules and variants of the CDnet and present experimental results on ZY-3 satellite thumbnails cloud cover validation data set and two other cloud cover validation datasets. This paper is concluded in Section V.

II. Related Work

A. Hand-Crafted Feature-Based Methods

In the past few decades, many cloud detection and segmentation methods have been proposed. In this paper, we roughly divided these methods into three categories: 1) simple threshold methods; 2) multiple image-based methods; and 3) learning-based methods.

1) Simple Thresholding Methods: Thresholding methods are widely used in cloud detection. Typical methods include ISCCP [10] and APOLLO [12]. Recently, thresholding strategies are often used as basic ingredients in more advanced methods. In [26], daytime cloud detection algorithm based on multispectral thresholds discriminated clouds from clear skies. Wei *et al.* [27] proposed a dynamic thresholding algorithm for cloud detection on the MODIS land surface reflectance database. Zhong *et al.* [28] developed a modified automatic cloud cover assessment (ACCA) method, including strict and loose threshold, to produce a cloud map with relatively high accuracy. Using only thresholding schemes may not be able to achieve satisfactory performance across various types of land surfaces. Therefore, Li *et al.* [16] used thresholding segmentation and guided filtering to generate a preliminary cloud mask and then fused geometric features and texture features to improve cloud detection results. Fisher [29] combines the thresholding-based method with morphological features to detect cloud and shadow from SPOT5 High-Resolution Geometric (HRG) imagery. Although thresholding-based methods are simple and efficient, for some complex ground objects, such as cloud–snow coexistence areas, classification performance is not satisfactory for practical applications.

2) Multiple Image-Based Methods: To tackle the limitation of simple thresholding techniques, using multiple images with temporal information is much more informative. Such methods show that temporal information is able to improve cloud detection results. Zhu and Woodcock [30] estimated a time series model for each pixel based on the robust iteratively reweighted least squares (RIRLS) method. Qian *et al.* [31] proposed to detect cloud on optical RSI time series using the mean shift algorithm. Gmezchova *et al.* [32] proposed the regularized least squares and kernel regression methods for cloud detection. Hagolle *et al.* [33] and Goodwin *et al.* [34] detected cloud using multiple temporal images, assuming that underlying landscape has little change within a short time period, and the pixel values of cloud areas are obviously different. These methods show significant improvement in cloud estimation and achieve high accuracy for the cloud and cloud shadow detection. But image data are not always available for multiperiod observations [35]. Moreover, multitemporal analysis is very sensitive to data quality [28].

3) Learning-Based Methods: To further improve the performance, more and more machine learning methods, including support vector machine (SVM) [36], neural network [37], random forest (RF) [38], maximum likelihood (MaxLike) [39], Markov random field [40], K-nearest neighbor (K-NN) [20], and decision tree theory [41], are used in RSI processing. SVM is the most popular one due to its prominent classification performance. It is applied to many cloud detection cases [36], [42]. As the input of a classifier, hand-crafted features, such as texture/color information [18] and morphological features [29], are difficult to accurately capture the cloud characteristics under complex environment [35]. This leads us to seek more effective and accurate feature representation methods.

B. Deep Feature-Based Methods

The rapid development of a deep convolutional neural network (DCNN) provides us new available approaches for image processing [21]. DCNNs are able to extract high-level abstract features from input images and significantly improve the accuracy of image classification or recognition. Image semantic segmentation techniques based on deep learning have also achieved amazing performance. Segmentation results in PASCAL VOC2012 challenge [43], and the highest mean intersection over union (MIOU) is achieved by DeeplabV3+ [44] up to 89%. Similarly, DCNNs have also been introduced to RSI processing, such as ground objects' classification [45], feature extraction [46], scene classification [47], object detection [17], and super-resolution reconstruction [48]. For cloud detection, Xie *et al.* [17], Goff *et al.* [49], and Chen *et al.* [50] used the fully connected DCNN to detect cloud regions from superpixels obtained by the simple linear iterative cluster (SLIC) method. However, the performance is limited by the presegmental superpixels. Recent efforts tried to transfer classification and objects' recognition networks, such as AlexNet [51], GoogLeNet [52], VGG [53], and ResNet [54], into fully convolutional ones by replacing the fully connected layers with convolutional ones. These algorithms achieve significant improvement in terms of segmentation accuracy over traditional methods. The strategies used to further improve cloud detection accuracy can be roughly divided into the following three aspects.

1) Exploit Context Information: Contextual correlation is important for complex scene understanding [25]. To enlarge the receptive field of neural networks, Yu and Koltun [55] used dilated convolution to systematically aggregate multiscale contextual information without losing resolution. Zhan *et al.* [56] enhanced the network VGG-16 [53] with dilated convolution for cloud detection from satellite images. For its promising performance, dilated convolution has been used in many networks for image semantic segmentation, such as PSPnet [25], RefineNet [57], Deeplab, and its variants [23], [24], [44].

2) Preserve Score Map Resolution: Preserving score map resolution is another main research direction in semantic segmentation. To obtain a score map with the same size as the input image, Kalia *et al.* [58] proposed the Cloud-CNN network for cloud/shadow detection based on the encoder–decoder architecture [59] evaluated on Himawari-8 AHI and GOES-16 ABI multispectral data. Ozkan *et al.* [60] proposed a deep pyramid network (DPN) with encoder and generator filter blocks (decoder architecture) for cloud detection from RGB color RSIs. Recently, Zhang *et al.* [61] proposed a lightweight neural network based on the U-Net model [62] for on-board pixelwise cloud detection on small satellites.

3) Refine Object Boundaries: High-quality segmentation results should be coherent with object boundaries. To refine cloud boundaries, Yue *et al.* [63] utilized the discrete conditional random field (CRF) [64] to refine segmentation boundaries by exploiting contextual information in cloud segmentation. Zhan *et al.* [56] exploited low-level visual features to generate sharp and detailed cloud boundaries. To further improve the localization capability near cloud boundaries, Yuan *et al.* [65] proposed an edge-aware segmentation network with an encoder–decoder structure for cloud detection.

Most existing deep neural networks for cloud detection addressed only one of these aspects. In this paper, we propose the CDnet for cloud detection by simultaneously exploiting multiscale and global contextual information, preserving score map resolution, and refining object boundaries. To this end, we introduce the FPM and BR modules. The FPM combines multiple parallel dilated convolution layers and the GAP block. Dilated convolution with different sampling rates enlarges the field of view of filters and effectively incorporates multiscale context [23]. The GAP block is able to extract image-level features, which helps capture long-range information beyond the capability of dilated convolutional layers. BR is a boundary refinement residual block, which helps refine object boundaries. These strategies make the segmentation results more accurate and reliable.

III. Proposed Method

In this section, we describe the overall framework of the proposed CDnet, as shown in Fig. 3. The key components, i.e., modified ResNet-50, FPMs, BR blocks, classification

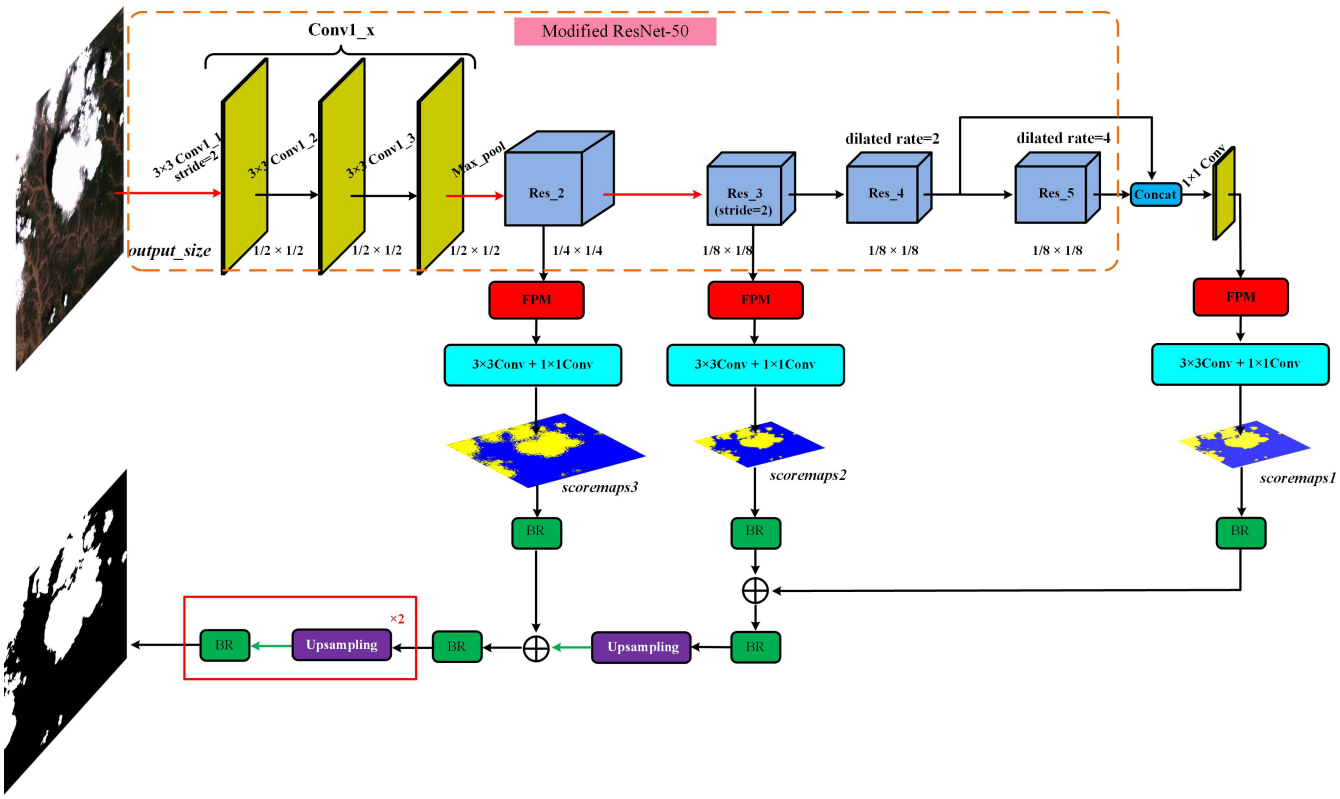


Fig. 3. Framework of the proposed CDnet. Red and green arrows represent the 2× downsampling and 2× upsampling operators, respectively. The red rectangular box with "×2" represents the upsampling and BR operations are implemented twice. The operator ⊕ represents the elementwise summation operator. The first three convolution layers in the modified ResNet-50 use convolutions with stride 2, 1, and 1, respectively, and the filter size is 3 × 3.

layers, and loss function involved in the CDnet, are presented in detail.

A. Overall Framework of CDnet

The framework of the proposed CDnet is shown in Fig. 3. We first use the pretrained and modified ResNet-50 network (detailed network structure is shown in Fig. 4) to extract features. Then, the proposed FPMs extract multiscale information at different stages of the modified network. For each FPM, a 3 × 3 convolution (including 512 filters, batch normalization, and Relu) is followed to reduce the number of channels. Finally, a 1 × 1 convolutional layer is used to generate semantic score maps for each class. Three low-resolution score maps are generated in the middle stages of the CDnet: *scoremaps1* and *scoremaps2* are 1/8 × 1/8 size of the input image, while *scoremaps3* is 1/4 × 1/4 size of the input image. We fuse *scoremaps1* with *scoremaps2* by an elementwise summation operation and refine the fusion results by a BR operation. The refined fusion results are upsampled to the same resolution as scoremaps3.[1] Subsequently, we further fuse the upsampled results with *scoremaps3* by an elementwise summation operation followed by a BR operation to obtain a new scoremap, which is 1/4 × 1/4 size of the input image. Finally, the new scoremap is upsampled to the same size as the input image by two upsampling and BR operations. These key modules of the CDnet are described in Sections III-B–III-D.

[1]The upsampling layer in this paper uses bilinear interpolation.

B. Modified ResNet-50 Feature Network

Residual networks [54] with skip connection in each block are easier to train and optimize particularly for very deep networks, which achieves an impressive performance in various vision tasks. For its promising performance in extracting discriminative features, we stand on ResNet-50 as the main structure of the proposed network. As shown in Fig. 3, instead of using a 7 × 7 receptive field with a stride of 2 in the first convolutional layer of original ResNet-50, we use three small filters with 3 × 3 receptive fields (Conv1_x block), i.e., conv1_1 with stride 2, conv1_2, and conv1_3 with stride 1. Such a modification greatly reduces the number of parameters: the three 3 × 3 convolutional layers have 27 parameters and the 7 × 7 convolutional layer has 49 parameters. In addition, the incorporated three nonlinear 3 × 3 convolutional layers also make the modified ResNet-50 deeper and more discriminative [53] than ResNet-50 (see results in Table I). Besides, all the strides in the convolutional layers of modified ResNet-50 are set to 1 except that the stride of the first convolution layer of the whole network and that of the first convolution layer in Res_3 block are set to 2, as shown in Fig. 3. In Fig. 3, the sizes of output feature maps at the intermediate layers of modified ResNet-50, marked by *output_size*, are given as normalized ratios against the original input size.

The repeated pooling and subsampling operators of the ResNet-50 usually led to the reduction of spatial resolution. The loss of spatial information may be harmful to produce

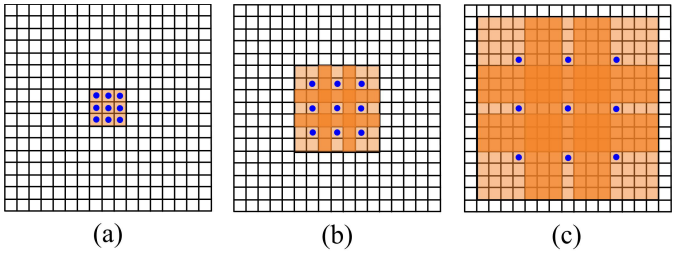


Fig. 4. Illustration of the receptive field of 3 × 3 dilated convolutional kernels at dilated rates (a) 1, (b) 2, and (c) 4, respectively. (a) 1-dilated convolution receptive field F_1 has a size of 3 × 3. (b) 2-dilated convolution receptive field F_2 has a size of 7 × 7. (c) 4-dilated convolution receptive field F_3 has a size of 15 × 15. This chart illustrates the exponential expansion of receptive field without loss of resolution in dilated convolution.

denser features and score maps. Meanwhile, traditional networks usually used different convolutional kernels to extract multiscale features, but they are associated with exponential growth of learnable parameters. Hence, in order to enlarge the receptive field of filters without significantly increasing the amount of computation and avoiding the loss of spatial information, we replace the traditional convolution with the dilated convolution [55] in our modified Resnet-50. In the modified ResNet-50, the dilated rates for the convolutional layers in Res_4 and Res_5 blocks are set to 2 and 4, respectively. For a standard 3×3 convolution, let F_1, F_2, and F_3, represent the receptive fields of the 1-dilated convolution, 2-dilated convolution, and 4-dilated convolution, respectively. The receptive field is a square of exponentially increasing size, which is shown in Fig. 4. Let the receptive field of an element p denoted by F_{i+1} (i=0, 1, 2,..., n-2) and the size of the receptive field of p in F_{i+1} be the number of these elements. It is clear that the size of the receptive field of each element in F_{i+1} is $(2^{i+2}-1) \times (2^{i+2}-1)$.

C. Feature Pyramid Module

As discussed earlier, dilated convolution allows us to significantly enlarge the receptive field of filters at deep convolutional neural network (DCNN) layers [55]. We incorporate multiple parallel dilated convolutional layers with different sampling rates similar to the atrous spatial pyramid pooling (ASPP) model [24] to capture multiscale context information. To further capture global context, we combine the multiple parallel dilated convolution layers with the GAP block [66]. To alleviate the vanishing-gradient problem, strengthen feature propagation, and encourage feature reusing, we introduce a short connection by concatenating the multiscale features, global context features, and input features of the FPM block into the final pyramid features.

As shown in Fig. 5, the FPM consists of four parallel dilated convolutional layers and one GAP block followed by an upsampling layer. To be specific, the four parallel dilated convolutional layers have the dilated rates of 1, 6, 12, and 18, respectively. Each dilated convolution layer has 512 filters, followed by batch normalization and Relu, and the dropout layer with the dropout ratio of 0.3 is used to alleviate the overfitting. GAP combined with an upsampling layer,

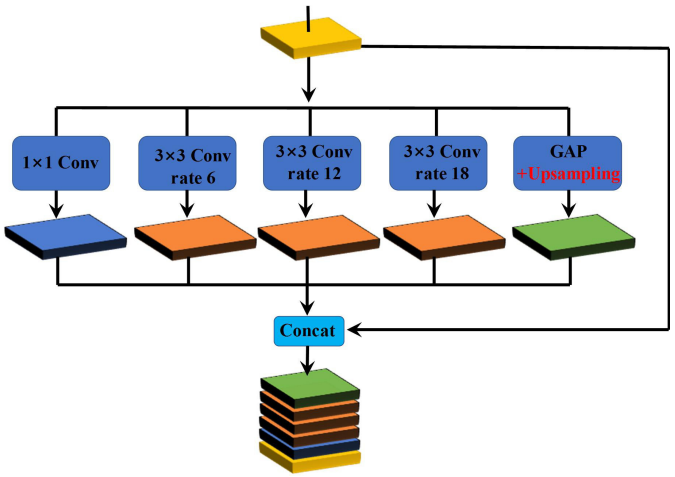


Fig. 5. FPM structure, including multiple parallel dilated convolution layers and a GAP block followed by a upsampling layer.

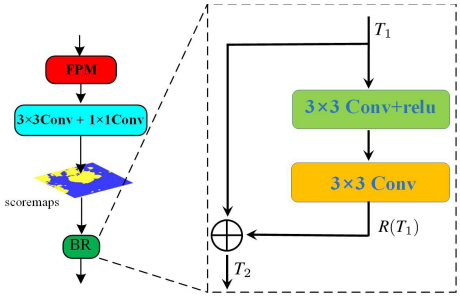


Fig. 6. BR block has a residual structure.

i.e., bilinear interpolation operations, makes the output has the same size as that of the input. Therefore, a deep convolutional network with FPM blocks can be more accurate and efficient trained. In Section IV-B, we compare the FPM with the ASPP+GAP module in DeeplabV3 [24] to demonstrate the effectiveness of the proposed FPM.

D. Boundary Refinement Module

To preserve sharp and detailed object boundaries, CRF is usually utilized as a postprocessing step to improve segmentation results, e.g., in DeeplabV2+CRF [23]. However, such a postprocessing scheme could be suboptimal for the overall inference process. In our CDnet, we adopt a BR module [22] to refine object boundaries. The BR module is concatenated into fully convolutional networks (FCNs) in a unified framework and trained in an end-to-end manner. As shown in Fig. 6, the BR module has a residual structure. Let T_1 represent the coarse scoremap, and the refined scoremap T_2 after the BR module can be represented as $T_2 = T_1 + R(T_1)$, where $R(T_1)$ is the residual branch. Fig. 7 shows an example of the coarse scoremap T_1 and the refined scoremap T_2. It can be observed that the BR module refines object boundaries.

E. Classification Layer and Loss

Semantic image segmentation is a multilabel classification problem, and softmax regression is the most effective choice

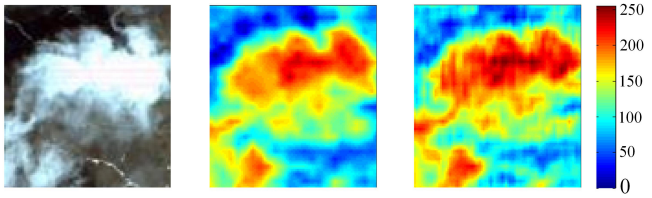


Fig. 7. Example of (a) cloud image, (b) coarse score map T_1, and (c) refined score map T_2.

for multiclass prediction [23], [25], [67]. Therefore, softmax is used to classify features extracted in the CDnet. In the output of the fully convolutional layer, let x_i denote a C-dimensional vector representing unnormalized scores for the location i, and softmax function is applied to each x_i to generate a probability label vector as follows:

$$p(y_i = j|x_i) = \frac{\exp(x_{ij})}{\sum_{c=1}^{C} \exp(x_{ic})}. \quad (1)$$

To avoid overfitting, the loss function $J(\Theta)$ of softmax with a regularization term is formulated as

$$J(\Theta) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{C}\left[\mathbf{1}\{y_i = j\}\log\frac{\exp(x_{ij})}{\sum_{c=1}^{C}\exp(x_{ic})}\right] + \frac{\lambda}{2}\|\Theta\|^2 \quad (2)$$

where M is the number of samples, C is the number of categories, Θ contains network parameters, $\mathbf{1}\{\cdot\}$ represents the indicative function, λ is a coefficient to balance these two terms, and x_{ic} indicates the ith sample grouped into the cth class. y_i is the label of x_i. The parameters Θ are updated through iterations aimed at minimizing $J(\Theta)$.

IV. Experimental Results

In this section, we comprehensively evaluate the proposed CDnet on ZY-3 satellite thumbnails. Specifically, we first present data preparation and experimental settings. Then, we discuss the performance of modules and variants of the CDnet. Third, we further investigate the performance of the CDnet qualitatively and quantitatively. Finally, we also evaluate the proposed CDnet on other two cloud cover assessment validation data sets.

A. Data Set and Experimental Settings

1) ZY-3 Satellite Cloud Cover Assessment Validation Data: The data set consists of 475 scenes thumbnails, including 280 RGB thumbnails and 195 gray ones, whose sizes are 1k × 1k and 3k × 3k, respectively. Thumbnails in the data set contain typical terrain information, including grassland, farmland, cities, mountain areas, snowy regions, and so on. For robust performance, images of different seasons were sampled into the data set. The data set is divided into three parts: 200 (train), 195 (val), and 80 (test) pixel-level labeled images for training, validation, and testing, respectively. Ground-truth segmentation masks (reference images) are obtained by manually marking the cloud regions in satellite thumbnails. As shown in Fig. 8, the selected scenes are also evenly distributed across the territory of China. All the thumbnails used in this paper can be previewed online: http://clouds.sasmac.cn/query.

Fig. 8. Distribution of ZY-3 satellite imagery.

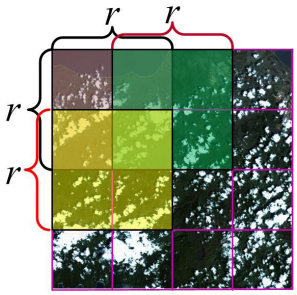


Fig. 9. Image cropping at the stepsize of half image size ($r = 321$).

2) Data Augmentation and Crop: Data augmentation, including image rotation and flipping, is performed to compensate a limited number of images in the data set. Since the input size of the CDnet is 321 × 321, we divide thumbnails into subimages of the same size. As shown in Fig. 9, we divide images into subimages of size 321 × 321 at the step size of 160. The strategy of overlapping division is adopted to achieve translational invariance and enrich the appearance patterns of prominent features in the sampled images. A large step size would reduce the number of extracted patches. Meanwhile, a small step size would lead to too much overlap of the extracted patches. The diversity and amount of the extracted subimages can well fit the training requirements by setting the step size to half of the subimage size. After data augmentation and cropping, there are about 46k subimages in the training data set.

3) Comparison Methods and Evaluation Metrics: We compare our proposed CDnet with three CNN-based cloud detection methods, i.e., DPN [60], MVGG-16 [56], and L-unet [61]. DPN is a deep pyramid network with encoder and generator filter blocks (decoder architecture) for cloud detection from RGB RSIs. MVGG-16 is a fully convolutional version of VGG-16 adapted for cloud and snow detection on remote sensing imagery. L-unet is a lightweight neural

network based on U-Net model [62] tailored for on-board cloud detection in small satellites.

For comprehensive evaluation, we also compare with five representative segmentation networks for generic images, i.e., FCN-8 [67], DeeplabV2 [23], DeeplabV3 [24], DeeplabV3+ [44], and PSPnet [25]. FCN-8 is a well-known FCN, where the output scoremap is 1/8 × 1/8 size of the input image. DeeplabV2 is a semantic segmentation network with atrous convolution and fully connected CRF. DeepLabV3 improves DeepLabV2 using filters at multiple sampling rates and effective field of views. DeepLabV3+ further improves DeepLabV3 using an encoder–decoder structure with atrous separable convolution for semantic image segmentation. PSPnet is a pyramid scene parsing network and ranked the first on the ImageNet scene parsing challenge 2016, PASCAL VOC 2012 benchmark, and Cityscapes benchmark.

Besides, two traditional machine learning methods, i.e., MaxLike classifier [7] and SVM [36], are also compared. To comprehensively measure the segmentation results, we utilize five widely used quantitative metrics, i.e., overall accuracy (OA), MIoU, kappa coefficient (Kappa), producer accuracy (PA), and user accuracy (UA) [7], [67].

4) Experimental Setting: All CNNs were trained under the Caffe framework and optimized by the stochastic gradient descent (SGD) algorithm [68]. The operating system is Ubuntu 14.04 equipped with NVIDIA GTX 1080 Ti GPU. The proposed CDnet is trained in an end-to-end manner. Learning rate started with 2×10^{-6}, whose decay policy is "poly" [23]. The number of minibatch size, momentum, and total iteration is 8, 0.9, and 2×10^{6}, respectively. Competing CNN-based methods are trained with the same parameter settings as the CDnet and fine-tuned with their corresponding pretrained CNN weights.[2]

B. Evaluation of the Proposed FPM

In this section, we evaluate the proposed FPM block by comparing with the ASPP+GAP module in DeeplabV3 [24]. The structure of the ASPP+GAP module is shown in Fig. 10(a). It consists of one 1 × 1 convolution, three 3 × 3 convolutions, and a GAP block. The four parallel dilated convolutional layers have dilated rates of 1, 6, 12, and 18, respectively. Each dilated convolution layer has 256 filters, followed by batch normalization and Relu. Different from ASPP+GAP in DeeplabV3 [24], each convolution layer in the proposed FPM block has 512 filters. Therefore, the FPM block is able to learn features containing more contextual information. In addition, the proposed FPM has three kinds of features, i.e., four parallel dilated convolution layers to extract multiscale features, one GAP block to extract global features, and the original input features introduced by a short connection. As a result, FPM is able to alleviate the vanishing-gradient problem, strengthen feature propagation, and encourage feature reusing.

[2]CDnet: https://github.com/tornadomeet/ResNet. FCN-8: https://github.com/shelhamer/fcn.berkeleyvision.org. DeeplabV2: https://bitbucket.org/aquariusjay/deeplab-public-ver2.PSPnet: https://hszhao.github.io/projects/pspnet/. DPN and MVGG-16: http://www.robots.ox.ac.uk~vgg/research/very_deep/. DeeplabV3: https://github.com/rishizek/tensorflow-deeplab-v3. DeeplabV3+: https://github.com/rishizek/tensorflow-deeplab-v3-plus.

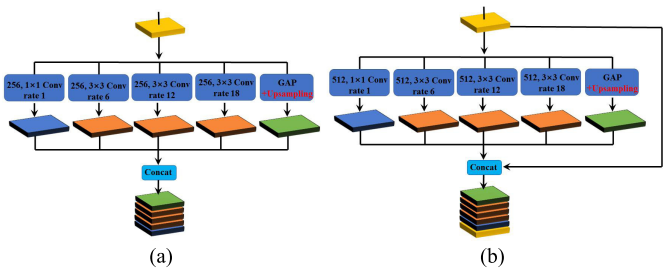


Fig. 10. Key module structure, where (a) ASPP+GAP is first proposed in DeeplabV3 and (b) FPM is proposed in this paper.

TABLE I
CLOUD EXTRACTION ACCURACY (%)

Method	OA	MIoU	Kappa	PA	UA
CDnet(ASPP+GAP)	95.41	89.38	82.05	87.82	89.85
CDnet(FPM)	96.47	91.70	85.06	89.75	90.41

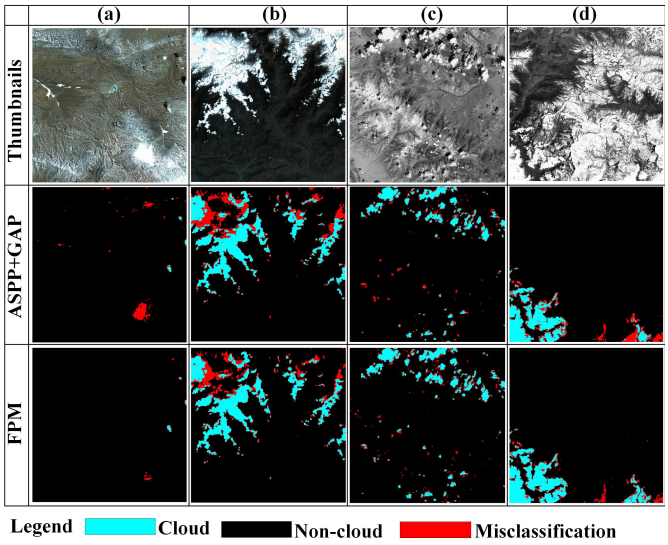


Fig. 11. Cloud detection results for tough cases with cloud–snow coexisting areas. (a) and (b) Two RGB thumbnails. (c) and (d) Two gray thumbnails.

In Table I, we present the quantitative results of our CDnet with different modules, i.e., FPM and ASPP+GAP. It can be seen that the proposed CDnet with the FPM module achieves better cloud detection performance than CDnet with the ASPP+GAP module. Fig. 11 presents the visual comparison results for four typical tough cases with cloud–snow coexistence areas. The results detected by FPM present less misclassified pixels than those detected by ASPP+GAP. This attributes to the powerful capability of FPM in extracting more rich, representative, and discriminative features than ASPP+GAP.

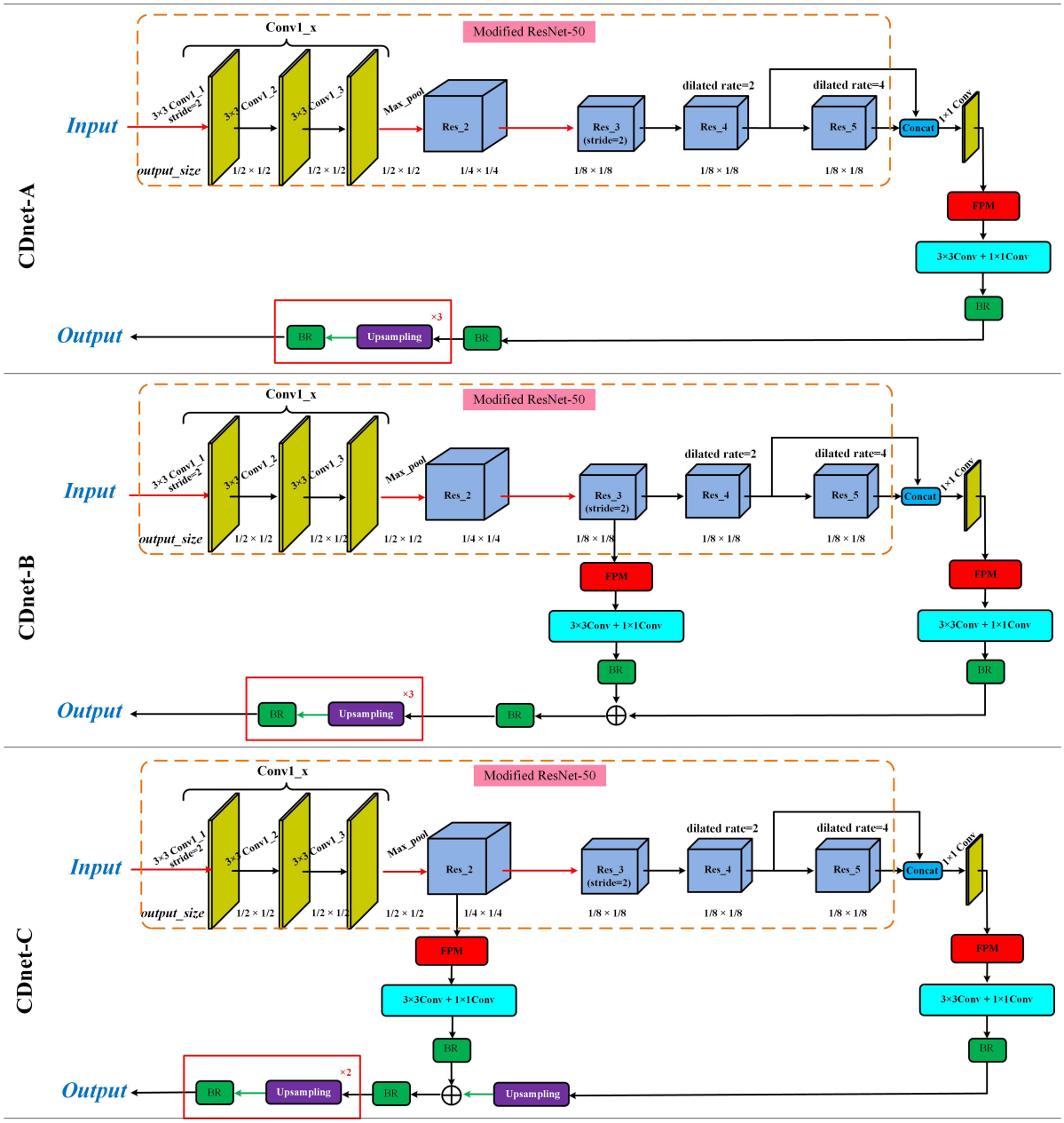


Fig. 12. Detailed structures of the three network structure variants. The red and green arrows represent the downsample and upsample operators, respectively. The red rectangular box represents upsample, and BR implemented twice in CDnet-C and three times in CDnet-A and CDnet-B. The operator $\bigoplus$ represents the elementwise sum operators. The first three convolution layers in the modified ResNet-50 use convolutions with stride 2, 1, and 1, respectively, and the filter size is 3 × 3.

C. Ablation Experiments

1) Network Structures: To demonstrate the effectiveness of the proposed CDnet, we first evaluate MRN+FPM and MRN+FPM+BR to investigate the performance of two key components, i.e., FPM and BR block. Specifically, in MRN+FPM, an FPM module is appended to the modified ResNet-50 (MRN), where the scoremap is directly upsampled to the same size as the input image without BR. While in MRN+FPM+BR, the scoremap is enhanced by a BR operation before upsampling. Second, we design CDnet-FPM, CDnet-BR, and CDnet-FPM-BR to investigate ablation experiments for FPM and BR. In CDnet-FPM, the FPM module is removed from the CDnet. In CDnet-BR, the BR module is removed from the CDnet. In CDnet-FPM-BR, both FPM and BR modules are removed from the CDnet. Third, in order to investigate the effectiveness of the proposed CDnet, we design three variant network structures, i.e., CDnet-A, CDnet-B, and CDnet-C, as shown in Fig. 12.

In this paper, the proposed CDnet consists of MRN and multiple FPM+BR modules, which utilizes the features extracted from different scales and levels. In contrast, the network of MRN+FPM+BR and CDnet-A only contains one FPM+BR module, which can only take advantage of the features at one scale and level. MRN+FPM+BR directly upsamples the scoremap refined by BR to the resolution of the original input image, while CDnet-A gradually upsamples the scoremap

TABLE II
CLOUD EXTRACTION ACCURACY (%) FOR MODULES AND VARIANTS OF THE CDNET

Method	OA	MIoU	Kappa	PA	UA
ResNet50	91.13	82.83	73.38	81.99	80.34
MRN*	93.03	85.24	77.51	82.59	82.82
MRN+FPM	93.89	88.50	81.82	87.10	85.51
MRN+FPM+BR	94.31	88.97	82.59	87.12	87.04
CDnet-FPM	93.14	88.14	80.44	87.64	84.46
CDnet-BR	95.04	89.63	83.78	87.36	88.67
CDnet-FPM-BR	93.10	87.91	80.01	87.01	83.84
CDnet-A	94.84	89.41	82.91	87.32	88.07
CDnet-B	95.27	90.51	84.01	88.97	89.71
CDnet-C	96.09	90.73	84.27	88.74	90.28
CDnet	**96.47**	**91.70**	**85.06**	**89.75**	**90.41**

∗ MRN is the abbreviation of modified ResNet-50.

to the resolution of the original input image by performing upsampling and BR operations for three times. In addition, the main difference among three variant networks and CDnet lies in the utilization of scoremap derived from different feature maps of modified ResNet-50. Specifically, the proposed CDnet fuses the end feature maps of Res_2, Res_3, and modified ResNet-50. In contrary, CDnet-A only utilizes the end feature maps of modified ResNet-50; CDnet-B utilizes the end feature maps of Res_3 and modified ResNet-50; CDnet-C utilizes the end feature maps of Res_2 and modified ResNet-50.

2) Quantitative Results: As shown in Table II, MRN has a better performance than original ResNet50 [54]. A stack of three 3×3 convolutional layers instead of 7×7 convolutional layer makes network deeper and decision function more discriminative. Replacing the traditional convolution with the dilated convolution also enlarges the receptive field without significantly increasing the amount of computation. These strategies make the detection results of MRN more accurate than those of original ResNet50. Using upsampling as a naive decoder, MRN+FPM outperforms MRN by 3.26% and 4.31% in terms of MIoU and Kappa, respectively. The results verify the capability of the FPM block that captures more discriminative features at different scales than single-scale filters. By adding the BR block, MRN+FPM+BR outperforms MRN+FPM by almost 0.47% and 0.77% in terms of MIoU and Kappa, respectively.

In ablation experiments, CDnet-BR provides the best results, which suggests that the FPM module contributes the most to the detection performance. CDnet outperforms CDnet-BR, CDnet-FPM, and CDnet-FPM-BR, which demonstrates that FPM, BR, and encoder–decoder network structure all contribute to the segmentation results. CDnet variants show a better performance than ResNet50, MRN, MRN+FPM, MRN+FPM+BR, CDnet-FPM, CDnet-BR, and CDnet-FPM-BR for most cases, excepting that CDnet-A is slightly inferior to CDnet-BR. The proposed CDnet outperforms its three variants, which demonstrates that features at midlevel and low-level layers also contribute to segmentation accuracy, and fusion of these different layers' features provides better cloud segmentation results than a particular layer alone.

D. Cloud Detection Results on ZY-3 Data Set

1) Qualitative Results: CDnet is the best in performance as verified by the results in Section IV-C and is further compared with other methods on ZY-3 satellite thumbnails cloud cover validation data. Visual results for typical thumbnails are shown in Fig. 13 for CNN-based methods and in Fig. 14 for two traditional learning-based methods. In Fig. 13, four RGB thumbnails and four gray thumbnails are selected, including three snow-free (only clouds cover) thumbnails [see Fig. 13(a), (e), and (g)], three cloud–snow coexisting thumbnails [see Fig. 13(b), (c), and (f)], and two only snow cover region thumbnails [see Fig. 13(d) and (h)]. Fig. 14 also uses the same set of thumbnails. For visual inspection, correctly detected cloud pixels are marked in bright cyan, while noncloud pixels are marked in black. Misclassified pixels are marked in red.

Results in Figs. 13 and 14 show that CNN-based methods are far more accurate than two traditional machine learning-based methods, as the results of MaxLike [7] and SVM [36] contained more misclassified pixels marked in red. Cloudless images with heavy snow are detected as cloudy regions by MaxLike and SVM, while CNN-based methods are able to provide better results although they also have different performance. For example, OA values of MaxLike and SVM averaged over the exemplar images are 0.5897 and 0.6310, respectively, which are significantly lower than that of L-unet, i.e., 0.8022, the poorest performer among CNN-based methods. The proposed CDnet has the higher average OA value, i.e., 0.9691, thanks to its capability in distinguishing could regions even from challenging snow–cloud coexistence cases.

Among CNN-based methods, L-unet produces more misclassified pixels in snow cover regions. FCN-8 has a deeper neural network structure and it produces better segmentation results than L-unet. Modified VGG-16 (MVGG-16) simultaneously exploits low-level and high-level features and significantly outperforms FCN-8. Both using VGG-16 as the network backbone, DPN and DeeplabV2 achieve a better performance than MVGG-16. The PSPnet exploits global context information via a pyramid scene parsing network and achieves s better performance than FCN-8, MVGG-16, DPN, and DeeplabV2. DeepLabV3 and DeepLabV3+ further improve semantic segmentation accuracy by incorporating multiscale filters or encoder–decoder network structure.[3] Being a departure from the above eight CNN-based methods, the proposed CDnet integrates the advantages of FPM and BR block, achieving the best segmentation results. As shown in Fig. 13, CDnet successfully distinguishes cloud and snow pixels from RGB and gray images. The results demonstrate that the CDnet is able to capture discriminative features of clouds and greatly improve accuracy of cloud mask extraction in snow-covered areas.

[3]In this paper, we set ResNet-50 as the network backbone of PSPnet, DeepLabv3, and Deeplab V3+ for fair comparison with CDnet.

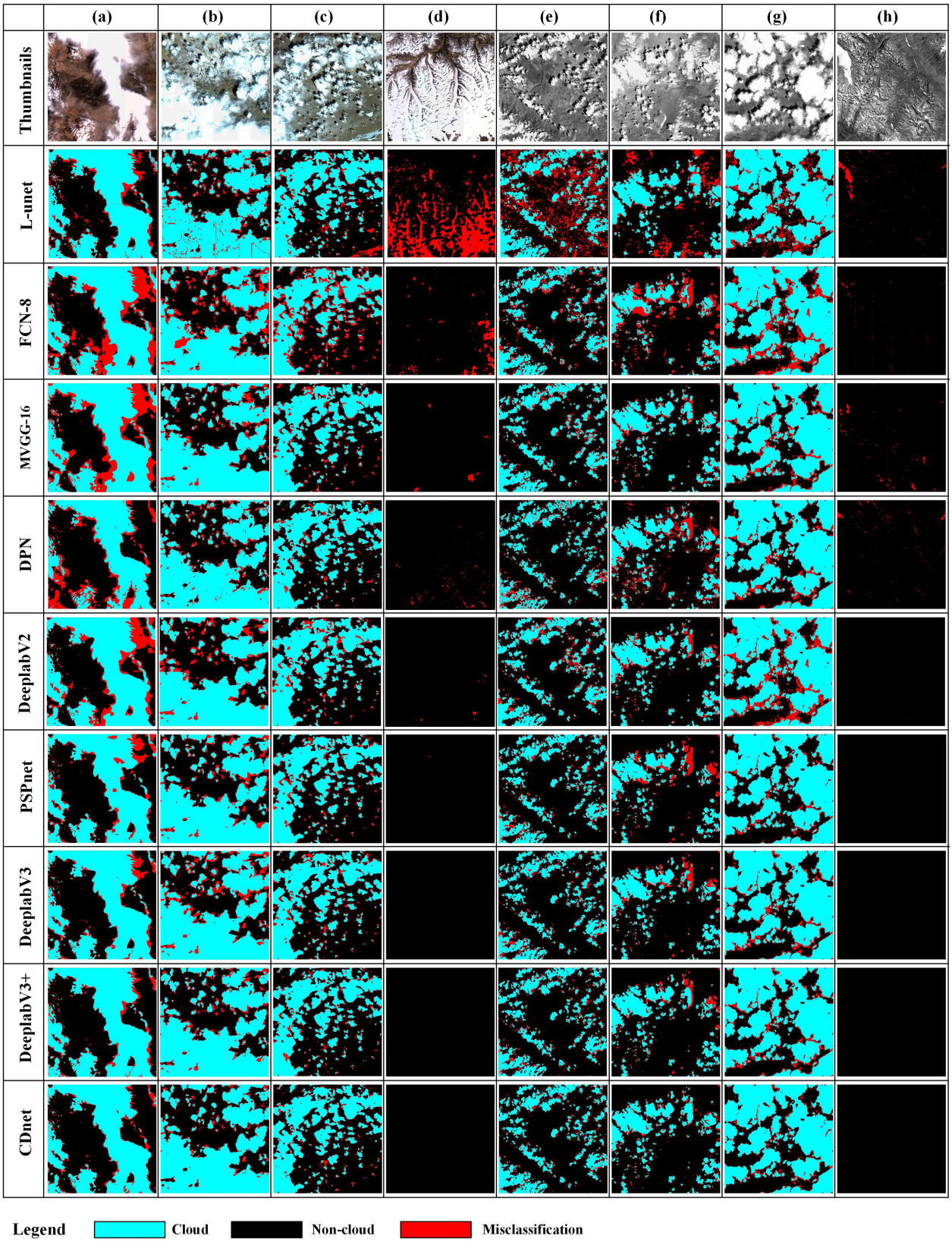


Fig. 13. Comparison of cloud extraction results of CNN-based methods in thumbnails of ZY-3 satellite imagery. (a)–(d) Four RGB thumbnails. (e)–(h) Four gray thumbnails. Among of them, (a), (e), and (g) are for cloud-only cases, (b), (c), and (f) are for cloud-snow co-existing cases, and (d) and (h) are for snow-only case. The sizes of RGB thumbnails and gray thumbnails are 1k × 1k and 3k × 3k, respectively.

2) Quantitative Results: Table III presents quantitative results in terms of OA, MIOU, Kappa, PA, and UA. We note that a low cloud coverage of percentage (less than 5%) may cause an apparent reduction in the cloud PA and UA [69]. Therefore, images with cloud coverage less than 5% are not included in the evaluation. Results in Table III indicate

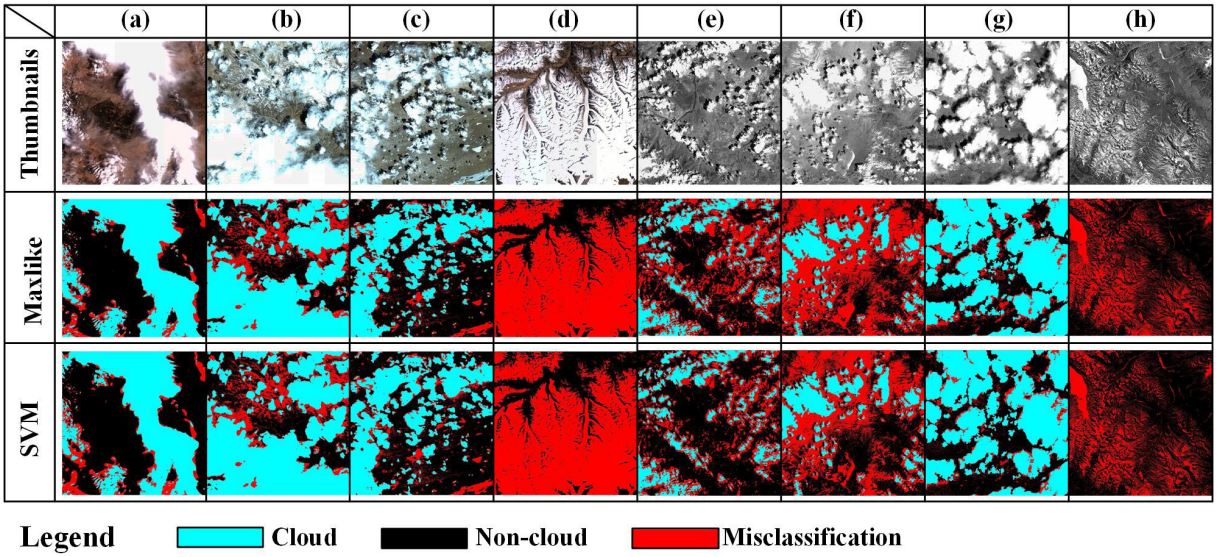


Fig. 14. Comparison of cloud extraction results of classic learning-based methods in thumbnails of ZY-3 satellite imagery. Thumbnails are the same as in those in Fig. 13.

TABLE III
CLOUD EXTRACTION ACCURACY (%)

Method	OA	MIoU	Kappa	PA	UA
Maxlike	77.73	66.16	53.55	91.30	54.98
SVM	78.21	66.79	54.87	**91.77**	56.37
L-unet	86.51	73.67	63.79	83.15	64.79
FCN-8	90.53	81.08	68.08	82.91	78.87
MVGG-16	92.73	86.65	78.94	88.12	81.84
DPN	93.11	86.73	79.05	87.68	83.96
DeeplabV2	93.36	87.56	79.12	87.50	84.65
PSPnet	94.24	88.37	81.41	86.67	89.17
DeeplabV3	95.03	88.74	81.53	87.63	89.72
DeeplabV3+	96.01	90.45	83.92	88.47	90.03
CDnet	**96.47**	**91.70**	**85.06**	89.75	**90.41**

that the proposed CDnet consistently outperforms eight other CNN-based comparison methods in terms of OA, MIoU, Kappa, PA, and UA. Moreover, these CNN-based methods are significantly better than the two traditional methods, since hand-crafted features are not as discriminative as those extracted by CNN-based methods. Nevertheless, the PA of the two traditional methods is higher than that of CNN-based methods, since they tend to classify all white pixels as cloud regions, including snow coverage areas. As a result, their UA values are significantly lower than those of the CNN-based methods.

E. Experiments on GF-1 and Landsat-8 Cloud Cover Assessment Validation Data

1) GF-1 Satellite Image: GF-1 WFV Cloud and Cloud Shadow Cover Validation Data released by the SENDIMAGE Lab includes 108 GF-1 wide field-of-view (WFV) level-2A scenes and their corresponding cloud and cloud shadow reference masks.[4] In this experiment, we use 40 (train), 40 (val), and 28 (test) scenes for training, validation, and testing, respectively. Data are composed of channels 4, 3, and 2. In the training stage, we divide the large GF-1 satellite imagery into subimages with the size of 321×321. In the testing stage, we divide the original image into subimages of size 513×513, since our GPU memory is not enough to process large size images. The final result for the whole image is generated by stitching the results of subimages together. The settings for the training and testing are the same as those for ZY-3 satellite thumbnails. We compare our proposed CDnet with the eight CNN-based methods. In addition, the automatic multifeature combined (MFC) cloud detection method [16], which is the baseline method for GF1 data set, is also compared. In this paper, we do not use the low-accuracy traditional methods, i.e., Maxlike [7] and SVM [36], as comparison methods in the additional tests on GF-1 Cloud Cover Assessment Validation Data. The cloud extraction results of four typical GF-1 Satellite imageries are shown in Fig. 15. Table IV shows quantitative results on the testing data set. Both results in Fig. 15 and Table IV suggest that the proposed CDnet achieves the best segmentation accuracy.

2) Landsat-8 Cloud Cover Assessment Validation Data: Original Landsat-8 Cloud Cover Assessment Validation Data contains 96 operational land imager (OLI) thermal infrared sensor (TIRS) terrain-corrected (Level-1T) scenes.[5] In this experiment, we select 22 (22)-scene Landsat-8 satellite images, whose cloud percentages lie in 35%∼100%, for training (validation). The data for training and validation include barren, grass/crops, forest, shrubland, urban, snow/ice, wetlands, water area, and so on. In addition, 11-scene images with the cloud percentage of 35%∼100% and another 20-scene images with the cloud percentage of 5%∼35% are selected for testing.[6]

[4]http://sendimage.whu.edu.cn/en/mfc-validation-data/

[5]https://landsat.usgs.gov/landsat-8-cloud-cover-assessment-validation-data

[6]In order to construct a balanced distributed data set, 21-scene images with the cloud percentage of 0%∼5% in data set are excluded for training, validation, and testing [69].

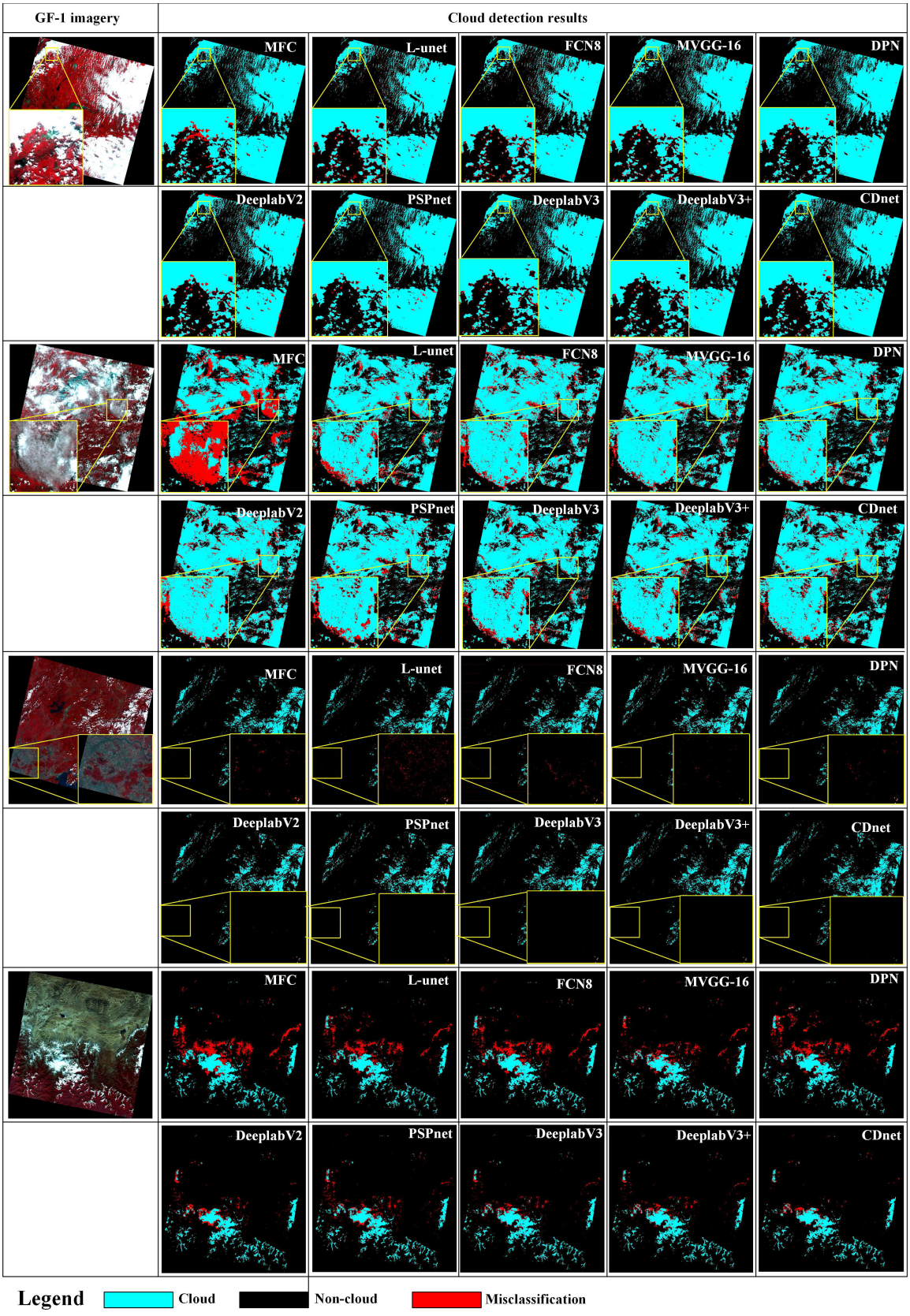


Fig. 15. Comparison of cloud extraction results of different methods in three GF-1 Satellite imageries. From top to bottom, they are thick cloud, thin cloud, cloud, and snow–cloud coexistence images, respectively.

To verify the performance of CNN-based methods with partial spectral information, channels 3–5 of Landsat-8 satellite image are used in the training and testing. Data preprocessing and experimental setting for training and testing are the same as GF-1 satellite image validation data. In addition, Fmask algorithm (Fmask) [70] is the baseline method for Landsat-8

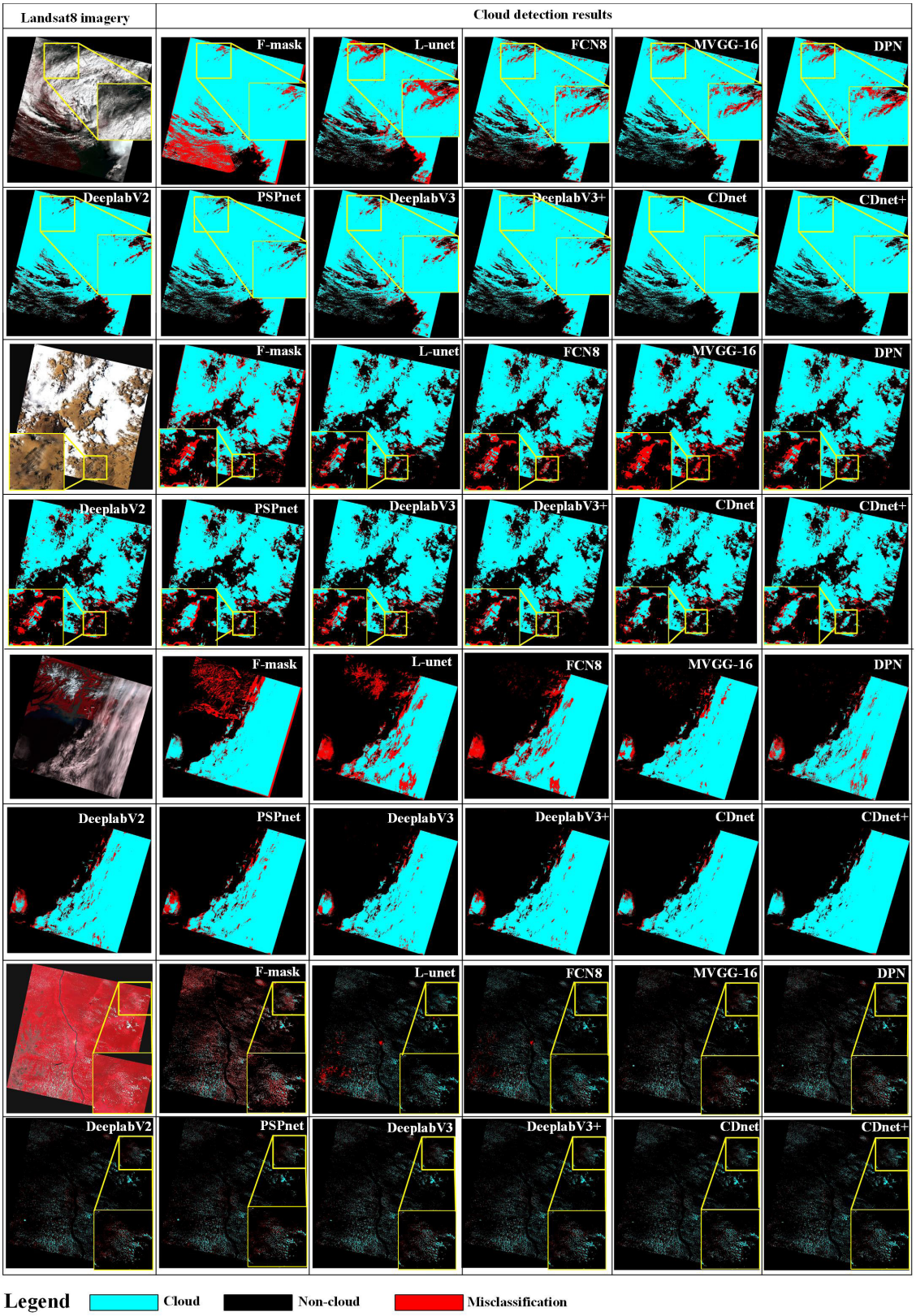


Fig. 16. Comparison of cloud extraction results of different methods in three Landsat8 Satellite imageries. From top to bottom, they are typhoon eye cloud, inland desert cloud, snow/ice and wetlands area cloud, and rainforest cloud images, respectively.

Satellite image data set and is also included in the comparison. The cloud extraction results of four typical Landsat-8 satellite images are shown in Fig. 16. Table V shows the quantitative results of the testing data set. Both results in Fig. 16 and Table V suggest that the proposed CDnet network achieves promising cloud detection performance for Landsat-8 satellite image, and it outperforms Fmask and other eight CNN-based methods.

TABLE IV
CLOUD EXTRACTION ACCURACY (%) OF GF-1 SATELLITE IMAGERY

Method	OA	MIoU	Kappa	PA	UA
MFC	92.36	80.32	74.64	83.58	75.32
L-unet	92.44	82.39	76.26	87.61	74.98
FCN-8	92.61	82.71	76.45	87.45	75.61
MVGG-16	93.07	86.17	77.13	87.68	79.50
DPN	93.19	86.32	77.25	86.85	80.93
DeeplabV2	95.07	87.00	80.07	86.60	82.18
PSPnet	95.30	87.45	80.74	85.87	83.27
DeeplabV3	95.95	88.13	81.05	86.36	88.72
DeeplabV3+	96.18	89.11	82.31	87.37	89.05
CDnet	**96.73**	**89.83**	**83.23**	**87.94**	**89.60**

TABLE V
CLOUD EXTRACTION ACCURACY (%) OF LANDSAT-8 SATELLITE IMAGERY

Method	OA	MIoU	Kappa	PA	UA
Fmask	85.21	71.52	63.01	86.24	70.38
L-unet	90.56	77.95	68.79	79.32	78.94
FCN-8	90.88	78.84	71.32	76.28	82.31
MVGG-16	93.28	81.83	76.90	77.29	83.00
DPN	93.31	81.59	77.08	78.80	88.57
DeeplabV2	93.40	86.13	81.20	83.46	90.92
PSPnet	94.11	86.34	81.52	84.61	89.93
DeeplabV3	94.67	86.90	81.63	84.93	89.87
DeeplabV3+	95.43	88.29	83.12	86.98	90.59
CDnet	96.38	90.32	84.31	89.52	91.92
CDnet+	**97.16**	**90.84**	**84.91**	**90.15**	**92.08**

Spectral information in channels 6, 7, and 9 of Landsat-8 satellite image is informative differentiating between clouds and snow. Data with channel compositions 3–7 and 9 are also used for training and testing. Meanwhile, we also present the results of CDnet trained on Landsat-8 satellite image with channels 3–7 and 9 in Fig.16, denoted by CDnet+. However, CDnet+ only slightly improves cloud detection performance at the price of more computation and memory usage. The results verify that the proposed CDnet is able to successfully extract discriminative features from spatial information (texture information) with limited spectral information. As a result, the CDnet is capable of differentiating clouds and snow in cloud–snow coexistence images.

3) Summary: By analyzing the cloud extraction results on GF-1 and Landsat-8 data set, the proposed CDnet can reliably extract cloud masks on various remote sensing imageries. Meanwhile, the experimental results show that the CDnet is able to obtain kappa coefficient larger than 83%, which suggests almost perfect consistency with the ground truth according to the interpretation kappa criterion, i.e., 0∼20% extremely low consistency, 21%∼40% general consistency, 41%∼60% moderate consistency, 61%∼80% high consistency, and 81%∼100% almost perfect [71]. Most notably, the CDnet has achieved the best performance, outperforming other CNN-based methods, including the recent Deeplab series [23], [24], [44] and the PSPnet [25]. These results show that the proposed CDnet has a powerful semantic segmentation capability, which can be well used for remote sensing imagery clouds' detection.

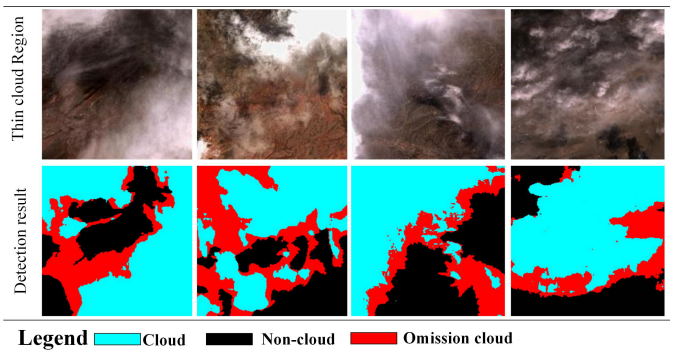


Fig. 17. Boundary localization capability of thin cloud on ZY-3 satellite thumbnails. (Top) Thin-cloud region. (Bottom) Cloud detection results of the CDnet.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a neural network (CDnet) for cloud mask extraction from ZY-3 satellite thumbnails. The CDnet has three advantages: 1) it extracts multiscale contextual information without loss of resolution and coverage; 2) it exploits high-level semantic features and mid-to-low-level visual features for category recognition of image regions and generates relatively detailed boundaries; 3) it captures sharper object boundaries by refining boundary operation and gradually recovers score maps resolution with an encoder–decoder network structure. Experimental results on ZY-3 satellite thumbnails show that the CDnet is able to achieve promising performance for generic RSIs, even for tough cases such as those containing cloud–snow coexistence areas. In addition, the CDnet has strong robustness and adapts well to other satellite imageries. Experimental results on GF-1 WFV Cloud Cover Validation Data and Landsat-8 Cloud Cover Assessment Validation Data show that the CDnet is able to achieve equally excellent cloud detection performance using only three-band information of the multispectral imageries.

Although the CDnet obtains satisfactory cloud detection results, the boundary localization for thin cloud needs to be further improved. As shown in Fig. 17, the thin-cloud areas have many omission pixels in the segmentation results. This may be due to the insufficient data of this type. We will gather more thin-cloud samples in our future work.

ACKNOWLEDGMENT

The authors would like to thank the Satellite Surveying and Mapping Application Center (SASMAC) of China for providing thumbnails of ZY-3 satellite imagery. They would also like to thank the editors and reviewers for their valuable suggestions.

REFERENCES

[1] T. S. Magney, L. A. Vierling, J. U. H. Eitel, D. R. Huggins, and S. R. Garrity, "Response of high frequency photochemical reflectance index (PRI) measurements to environmental conditions in wheat," *Remote Sens. Environ.*, vol. 173, pp. 84–97, Feb. 2016.

[2] Z. Li *et al.*, "Remote sensing of atmospheric particulate mass of dry $PM_{2.5}$ near the ground: Method validation using ground-based measurements," *Remote Sens. Environ.*, vol. 173, pp. 59–68, Feb. 2016.

[3] A. A. Fenta *et al.*, "The dynamics of urban expansion and land use/land cover changes using remote sensing and spatial metrics: The case of mekelle city of northern ethiopia," *Int. J. Remote Sens.*, vol. 38, no. 14, pp. 4107–4129, Jul. 2017.

[4] M. G. Benson and J. L. Faundeen, "The U.S. geological survey, remote sensing, and geoscience data: Using standards to serve us all," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGRSS)*, vol. 3, Jul. 2000, pp. 1202–1204.

[5] Z. G. Wang, Q. Kang, Y. J. Xun, Z. Q. Shen, and C. B. Cui, "Military reconnaissance application of high-resolution optical satellite remote sensing," in *Proc. Int. Symp. Optoelectronic Technol. Appl.*, vol. 9299, Nov. 2014, Art. no. 9299195.

[6] Y. Zhang, W. B. Rossow, A. A. Lacis, V. Oinas, and M. I. Mishchenko, "Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: Refinements of the radiative transfer model and the input data," *J. Geophys. Res. Atmos.*, vol. 109, no. D19, pp. 1–27, Oct. 2004.

[7] F. Yang, J. Guo, H. Tan, and J. Wang, "Automated extraction of urban water bodies from ZY-3 multi-spectral imagery," *Water*, vol. 9, no. 2, p. 144, Feb. 2017.

[8] J. Guo, F. Yang, H. Tan, J. Wang, and Z. Liu, "Image matching using structural similarity and geometric constraint approaches on remote sensing images," *J. Appl. Remote Sens.*, vol. 10, no. 4, Oct. 2016, Art. no. 045007.

[9] Y. Zhang, F. Du, and C. Zhu, "DEM extraction and accuracy assessment based on ZY-3 stereo images," in *Proc. Int. Conf. Comput. Sci. Netw. Technol.*, Dec. 2012, pp. 1439–1442.

[10] W. B. Rossow and L. C. Garder, "Cloud detection using satellite measurements of infrared and visible radiances for ISCCP," *J. Climate*, vol. 6, no. 12, pp. 2341–2369, Dec. 1993.

[11] L. L. Stowe *et al.*, "Global distribution of cloud cover derived from NOAA/AVHRR operational satellite data," *Adv. Space Res.*, vol. 11, no. 3, pp. 51–54, 1991.

[12] G. Gesell, "An algorithm for snow and ice detection using AVHRR data an extension to the APOLLO software package," *Int. J. Remote Sens.*, vol. 10, nos. 4–5, pp. 897–905, Apr. 2007.

[13] X. Huang, H. Chen, and J. Gong, "Angular difference feature extraction for urban scene classification using ZY-3 multi-angle high-resolution satellite imagery," *ISPRS J. Photogram. Remote Sens.*, vol. 135, pp. 127–141, Jan. 2018.

[14] T. Suga, K. Muto, K. Yagi, T. Onodera, Y. Nakada, and K. Takano, "Thumbnail image size for image transmission system using satellite communication in disaster," *Int. Inf. Inst.*, vol. 18, no. 3, pp. 1019–1027, Mar. 2015.

[15] Q. Zhang and C. Xiao, "Cloud detection of RGB color aerial photographs by progressive refinement scheme," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7264–7275, Nov. 2014.

[16] 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.

[17] 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.

[18] E. Başeski and A. Cenaras, "Texture and color based cloud detection," in *Proc. 7th Int. Conf. Recent Adv. Space Technol. (RAST)*, Jun. 2015, pp. 311–315.

[19] Z. Chen, T. Deng, H. Zhou, and S. Luo, "Cloud detection based on HSI color space and SWT from high resolution color remote sensing imagery," *Proc. SPIE*, vol. 8919, Oct. 2013, Art. no. 891907.

[20] M. Xia, W. Lu, J. Yang, Y. Ma, W. Yao, and Z. Zheng, "A hybrid method based on extreme learning machine and k-nearest neighbor for cloud classification of ground-based visible cloud image," *Neurocomputing*, vol. 160, pp. 238–249, Jul. 2015.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[22] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters-improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2017, pp. 1743–1751.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[24] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam. (2017). "Rethinking atrous convolution for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/1706.05587

[25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Mar. 2017, pp. 6230–6239.

[26] Q. J. He, "A daytime cloud detection algorithm for FY-3A/VIRR data," *Int. J. Remote Sens.*, vol. 32, no. 21, pp. 6811–6822, Jul. 2011.

[27] J. Wei *et al.*, "Dynamic threshold cloud detection algorithms for MODIS and Landsat 8 data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 566–569.

[28] B. Zhong, W. Chen, S. Wu, L. Hu, X. Luo, and Q. Liu, "A cloud detection method based on relationship between objects of cloud and cloud-shadow for Chinese moderate to high resolution satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 11, pp. 4898–4908, Nov. 2017.

[29] 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.

[30] Z. Zhu and C. E. Woodcock, "Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change," *Remote Sens. Environ.*, vol. 152, pp. 217–234, Sep. 2014.

[31] J. Qian, Y. Luo, Y. Wang, and D. Li, "Cloud detection of optical remote sensing image time series using mean shift algorithm," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 560–562.

[32] L. Gmezchova, J. Amorslpez, and G. Campsvalls, "Cloud masking and removal in remote sensing image time series," *J. Appl. Remote Sens.*, vol. 11, no. 1, Jan. 2017, Art. no. 015005.

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

[34] N. R. Goodwin, L. J. Collett, R. J. Denham, N. Flood, and D. Tindall, "Cloud and cloud shadow screening across Queensland, Australia: An automated method for Landsat TM/ETM + time series," *Remote Sens. Environ.*, vol. 134, pp. 50–65, Jul. 2013.

[35] Z. Shao, J. Deng, L. Wang, Y. Fan, N. Sumari, and Q. Cheng, "Fuzzy autoencode based cloud detection for remote sensing imagery," *Remote Sens.*, vol. 9, no. 4, p. 311, Mar. 2017.

[36] C. Latry, C. Panem, and P. Dejean, "Cloud detection with SVM technique," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGRSS)*, Jan. 2008, pp. 448–451.

[37] M. J. Hughes and D. J. Hayes, "Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing," *Remote Sens.*, vol. 6, no. 6, pp. 4907–4926, May 2014.

[38] J. Deng, H. Wang, and J. Ma, "An automatic cloud detection algorithm for landsat remote sensing image," in *Proc. Int. Workshop Earth Observ. Remote Sens. Appl. (EORSA)*, Jul. 2016, pp. 395–399.

[39] L. Xu, A. Wong, and D. A. Clausi, "A novel Bayesian spatial-temporal random field model applied to cloud detection from remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 4913–4924, Sep. 2017.

[40] Q. Li, W. Lu, J. Yang, and J. Z. Wang, "Thin cloud detection of all-sky images using Markov random fields," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 417–421, May 2012.

[41] B. B. Barnes and C. Hu, "A hybrid cloud detection algorithm to improve MODIS sea surface temperature data quality and coverage over the eastern gulf of Mexico," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 6, pp. 3273–3285, Jun. 2013.

[42] H. Ishida, Y. Oishi, K. Morite, K. Moriwaki, and T. Y. Nakajima, "Development of a support vector machine based cloud detection method for MODIS with the adjustability to various conditions," *Remote Sens. Environ.*, vol. 205, pp. 309–407, Feb. 2018.

[43] M. Everingham, L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2012). *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[44] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. (2018). "Encoder-decoder with atrous separable convolution for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/1802.02611

[45] Y. Tao, M. Xu, F. Zhang, B. Du, and L. Zhang, "Unsupervised-restricted deconvolutional neural network for very high resolution remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 6805–6823, Dec. 2017.

[46] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016.

[47] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[48] Y. Luo, L. Zhou, S. Wang, and Z. Wang, "Video satellite imagery super resolution via convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2398–2402, Dec. 2017.

[49] M. L. Goff, J. Y. Tourneret, H. Wendt, M. Ortner, and M. Spigai, "Deep learning for cloud detection," in *Proc. Int. Conf. Pattern Recognit. Syst. (ICPRS)*, vol. 10, Oct. 2017, pp. 1–6.

[50] Y. Chen, R. Fan, M. Bilal, X. Yang, J. Wang, and W. Li, "Multilevel cloud detection for high-resolution remote sensing imagery using multiple convolutional neural networks," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 5, p. 181, May 2018.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst. (ICONIP)*, Aug. 2012, pp. 1097–1105.

[52] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, May 2015, pp. 1–9.

[53] K. Simonyan and A. Zisserman. (2016). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

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

[55] F. Yu and V. Koltun. (2015). "Multi-scale context aggregation by dilated convolutions." [Online]. Available: https://arxiv.org/abs/1511.07122

[56] 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.

[57] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5168–5177.

[58] S. G. Subodh Kalia, S. Li, and R. R. Nemani, "DeepSAT's cloudCNN: A deep neural network for rapid cloud detection from geostationary satellites," in *Proc. Fall Meeting*, Aug. 2017, pp. 1589–1596.

[59] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[60] S. Ozkan, M. Efendioglu, and C. Demirpolat. (2018). "Cloud detection from RGB color remote sensing images with deep pyramid networks." [Online]. Available: https://arxiv.org/abs/1801.08706

[61] G. X. Zhaoxiang Zhang, A. Iwasaki, and J. Song. (2018). *Small Satellite Cloud Detection Based on Deep Learning and Image Compression*. [Online]. Available: https://www.preprints.org/manuscript/201802.0103/v1

[62] 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 (MICCAI)*, 2015, pp. 234–241.

[63] Z. Yue, X. Fengying, and Z. Jiang, "A cloud detection method for landsat 8 images based on PCANet," *Remote SensingPattern Recognit.*, vol. 10, no. 6, p. 877, Jun. 2018.

[64] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2014). "Semantic image segmentation with deep convolutional nets and fully connected CRFs." [Online]. Available: https://arxiv.org/abs/1412.7062

[65] 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.

[66] M. Lin, Q. Chen, and S. Yan. (2013). "Network in network." [Online]. Available: https://arxiv.org/abs/1312.4400

[67] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[68] Y. Bengio, *Practical Recommendations for Gradient-Based Training Deep Architectures*, vol. 7700. Berlin, Germany: Springer, 2012.

[69] S. Qiu, B. He, Z. Zhu, Z. Liao, and X. Quan, "Improving fmask cloud and cloud shadow detection in mountainous area for Landsats 4–8 images," *Remote Sens. Environ.*, vol. 199, pp. 107–119, Jun. 2017.

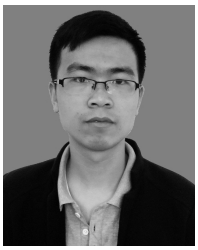
[70] Z. Zhu, S. Wang, and C. E. Woodcock, "Improvement and expansion of the fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and sentinel 2 images," *Remote Sens. Environ.*, vol. 159, pp. 269–277, May 2015.

[71] S. Simon. *StATS: What is A Kappa coefficient? (Cohen's Kappa)*. [Online]. Available: http://www.pmean.com/definitions/kappa.htm

Jingyu Yang (M'10–SM'17) received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2003, and the Ph.D. degree (Hons.) from Tsinghua University, Beijing, in 2009.

He has been a Faculty Member with Tianjin University, Tianjin, China, since 2009, where he is currently a Professor with the School of Electrical and Information Engineering. He was with Microsoft Research Asia (MSRA), Beijing, in 2011, within the MSRAs Young Scholar Supporting Program, and with the Signal Processing Laboratory, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2012 and from 2014 to 2015. His research interests include image video processing, 3-D imaging, and computer vision.

Dr. Yang served as the Special Session Chair in the International Conference on Visual Communications and Image Processing 2016 and the Area Chair in the International Conference on Image Processing 2017. He was selected in the program for New Century Excellent Talents in University (NCET) from the Ministry of Education, China, in 2011, the Reserved Peiyang Scholar Program of Tianjin University in 2014, and the Tianjin Municipal Innovation Talent Promotion Program in 2015.



Jianhua Guo (S'18) received the B.E. degree in surveying and mapping engineering from Anhui Jianzhu University, Hefei, China, in 2014, and the M.A. degree in geodesy and surveying engineering from Liaoning Technical University, Fuxin, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. He was a jointly educates student with the Satellite Surveying and Mapping Application Center, Beijing, China, from 2015 to 2017.

His research interests include remote sensing image matching, classification, and segmentation. More details about his work can be found at https://www.researchgate.net/profile/Jianhua_Guo11.

Huanjing Yue (M'17) received the B.S. and Ph.D. degrees from Tianjin University, Tianjin, China, in 2010 and 2015, respectively.

She was an intern with Microsoft Research Asia, Beijing, China, from 2011 to 2015. She visited the Video Processing Laboratory, University of California at San Diego, La Jolla, CA, USA, from 2016 to 2017. She is currently an Assistant Professor with the School of Electrical and Information Engineering, Tianjin University. Her research interests include image processing and computer vision.

Dr. Yue received the Microsoft Research Asia Fellowship Honor in 2013.

Zhiheng Liu received the B.E. degree in surveying and mapping engineering from the Henan University of Engineering, Zhengzhou, China, in 2014, and the master's degree from the School of Geology Engineering and Geomatics, Chang'an University, Xi'an, China, where he is currently pursuing the Ph.D. degree. He was a jointly educates student with the Satellite Surveying and Mapping Application Center, Beijing, China, from 2015 to 2016.

His research interests include remote sensing geology, image processing, and environmental protection.

Haofeng Hu received the B.S. and Ph.D. degrees from Nankai University, Tianjin, China, in 2002 and 2011, respectively.

He visited the Institute of Optics, French National Center for Scientific Research, Paris, France, from 2011 to 2013. He is currently an Assistant Professor with the School of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin. His research interests include optical imaging and polarization imaging technologies.

Kun Li (M'12) received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2006, and the master's and Ph.D. degrees from Tsinghua University, Beijing, in 2011.

She visited the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2012 and from 2014 to 2015. She is currently an Associate Professor with the School of Computer Science and Technology, Tianjin University, Tianjin, China. Her research interests include dynamic scene 3-D reconstruction and image/video processing.

Dr. Li was selected in the Peiyang Scholar Program of Tianjin University in 2016. She received the Platinum Best Paper Award in IEEE ICME 2017.