CDnetV2: CNN-Based Cloud Detection for Remote Sensing Imagery With Cloud-Snow Coexistence

Jianhua Guo, *Student Member, IEEE*, Jingyu Yang, *Senior Member, IEEE*, Huanjing Yue, *Member, IEEE*, Hai Tan, Chunping Hou, and Kun Li, *Member, IEEE*

***Abstract*— Cloud detection is a crucial preprocessing step for optical satellite remote sensing (RS) images. This article focuses on the cloud detection for RS imagery with cloud-snow coexistence and the utilization of the satellite thumbnails that lose considerable amount of high resolution and spectrum information of original RS images to extract cloud mask efficiently. To tackle this problem, we propose a novel cloud detection neural network with an encoder–decoder structure, named CDnetV2, as a series work on cloud detection. Compared with our previous CDnetV1, CDnetV2 contains two novel modules, that is, adaptive feature fusing model (AFFM) and high-level semantic information guidance flows (HSIGFs). AFFM is used to fuse multilevel feature maps by three submodules: channel attention fusion model (CAFM), spatial attention fusion model (SAFM), and channel attention refinement model (CARM). HSIGFs are designed to make feature layers at decoder of CDnetV2 be aware of the locations of the cloud objects. The high-level semantic information of HSIGFs is extracted by a proposed high-level feature fusing model (HFFM). By being equipped with these two proposed key modules, AFFM and HSIGFs, CDnetV2 is able to fully utilize features extracted from encoder layers and yield accurate cloud detection results. Experimental results on the ZY-3 satellite thumbnail data set demonstrate that the proposed CDnetV2 achieves accurate detection accuracy and outperforms several state-of-the-art methods.**

***Index Terms*— Adaptive feature fusing, cloud detection, deep convolutional neural network (DCNN), high-level semantic information, ZY-3 satellite thumbnails.**

I. Introduction

With the universal application of remote sensing (RS) technology, optical RS images play an important role in geosciences, military, agriculture, forestry, hydrology, and environmental protection [2]. Nevertheless, most of the Earth surface is usually covered with clouds. Under cloudy conditions, imaging paths are severely obstructed, and the satellite sensor cannot directly sense the real Earth's surface. Therefore, detecting cloud in RS images is important to assess the quality of RS images [1]. Moreover, cloudage information is also useful in meteorological observer and weather forecast [3].

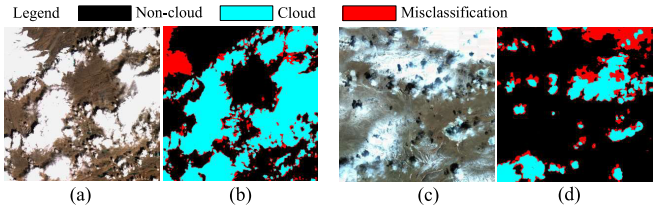

Fig. 1. Cloud detection results of CDnetV1 [1] for tough cases with cloud-snow coexisting areas. (a) Thumbnail. (b) Detection result. (c) Thumbnail. (d) Detection result.

In practice, satellite thumbnails of smaller sizes than original RS images are usually used for fast cloud detection [1] to further estimate cloudage information. However, thumbnail images usually lose considerable amount of spectrum information, making cloud detection from thumbnails more difficult and challenging than that from hyper/multispectral RS images with high resolution. Representative cloud detection methods, such as ISCCP [4], CLAVR [5], and APOLLO [6] based on spectral information are inapplicable to thumbnails. Over the past few years, increasing attention is attracted to do related work with only limited spectral information. Cloud detection from RGB color images [7]–[9] has achieved promising cloud detection results. However, by directly applying these methods to RS thumbnails, the cloud-snow coexistence problem cannot be handled as they have similar pattern, such as color and local texture [10]. Therefore, it is challenging to accurately detect cloud from cloud-snow coexistence image through color/texture-based approaches. In order to increase the accuracy of cloud detection from RS thumbnails covered with snow, it is necessary to design discriminative high-level features and exploit advanced classification methods.

In our previous work [1], a CNN-based method (named CDnetV1) was developed. Compared with several other state-of-the-art methods, this method shows promising cloud detection performance on several types of land covers. However, this method does not explicitly address the coexistence of cloud and snow in cloud detection. As shown in Fig. 1, a significant amount of snow is falsely classified as cloud.

Manuscript received November 27, 2019; revised March 15, 2020; accepted April 27, 2020. Date of publication May 18, 2020; date of current version December 24, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61771339, Grant 61672378, and Grant 61520106002, and in part by the Tianjin Research Program of Application Foundation and Advanced Technology under Grant 18JCYBJC19200. *(Corresponding author: Jingyu Yang.)*

Jianhua Guo, Jingyu Yang, Huanjing Yue, and Chunping Hou 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; hcp@tju.edu.cn).

Hai Tan is with Land Satellite Remote Sensing Application Center, MNR, Beijing 100048, China (e-mail: tanhai001@139.com).

Kun 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 article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.2991398

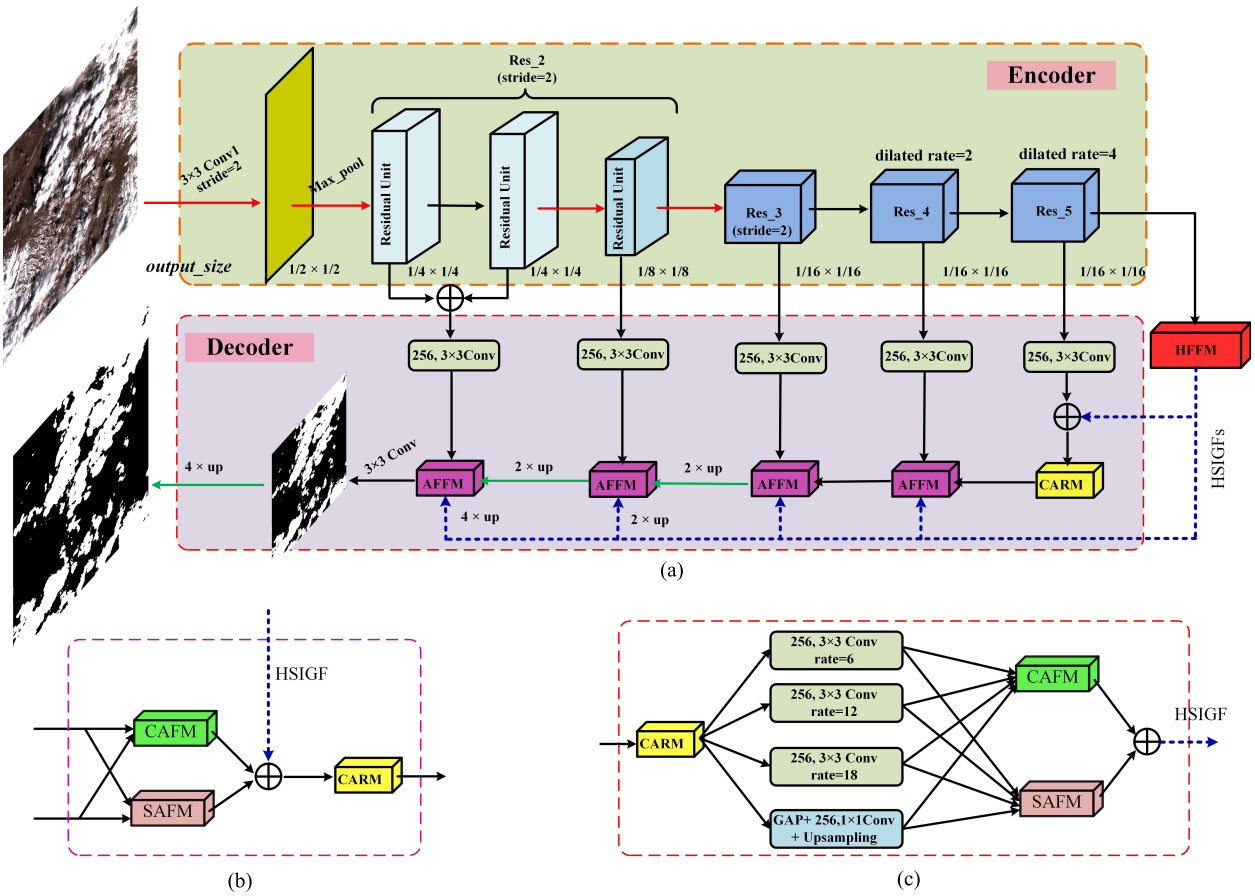

Fig. 2. Framework of the proposed CDnetV2. Red and green arrows: downsampling and upsampling operators, respectively. Blue dotted arrow: HSIGFs. ⊕: elementwise summation operation. CAFM: channel attention fusion model, SAFM: spatial attention fusion model, and CARM: channel attention refinement model. (a) Proposed CDnetV2 structure. (b) AFFM. (c) HFFM.

The erroneous cloudage information is unfavorable for subsequent high-level tasks.

To improve cloud detection accuracy on thumbnails with snow/ice-cloud coexistence, this article proposes a new cloud detection neural network, named CDnetV2, with an encoder-decoder U-shape structure. Fig. 2 shows the detailed network structure. CDnetV2 focuses on adaptively fusing multilevel feature maps and remedying diluted high-level semantic information features to improve cloud detection accuracy. To be specific, we propose an adaptive feature fusing model (AFFM) consisting of a channel attention fusion model (CAFM) and a spatial attention fusion model (SAFM) to adaptively fuse multilevel features through channel-attention and spatial-attention fusion. Except for the two feature fusion models in AFFM, a channel attention refinement model (CARM) is designed to further refine the fused feature maps. In addition, as analyzed in [11], high-level semantic information features would be gradually diluted when transmitting through the layers of the decoder side. We introduce a series of high-level semantic information guidance flows (HSIGFs) to remedy the semantic dilution problem. The high-level semantic information in HSIGFs is extracted by a proposed high-level feature fusion block embed at the last layer of the backbone network. Experimental results show that CDnetV2 equipped with HSIGFs and AFFM is able to explicitly explore and leverage discriminative features to reliably identify cloudy regions from thumbnails covered with snow.

In general, the main contributions of this article can be summarized as follows.

1) A novel end-to-end deep convolutional neural network (DCNN), named CDnetV2, with encoder-decoder U-shape structure is proposed for efficient cloud detection from satellite thumbnails with cloud-snow coexistence.
2) An adaptive feature fusing model (AFFM), which consists of a CAFM, an SAFM, and a CARM, is designed to fuse multilevel feature maps seamlessly and improve the representation power of the proposed network.
3) A series of HSIGFs is introduced to remedy high-level semantic information diluted at decoder layers, which explicitly makes feature maps at each level aware of the locations of the cloud objects.

We organize this article as follows. In Section II, we introduce the related work on CNN-based cloud detection from RS images. The proposed CDnetV2 framework, feature fusion module AFFM, and high-level feature guidance flow HSIGFs are detailed in Section III. Data set and experimental settings are illustrated in Section IV. The experimental results are presented in Section V. Section VI discusses and analyzes the channel-attention and spatial-attention weights in

CAFM and SAFM respectively, followed by conclusions in Section VII.

II. Related Work

In the past few decades, cloud detection from RS imagery has attracted increasing attention. Cloud detection methods can broadly be categorized into traditional rule-based methods on the spectral and/or spatial domain [12]–[15] and machine learning-based methods [16]–[18]. Rule-based methods focus on exploiting reflectance variations in different bands, such as visible, shortwave-infrared and thermal bands, to distinguish clouds from clear sky pixels. However, rule-based methods strongly depend on particular sensor models. For example, Fmask algorithm [12], [13] is developed for Landsat 4-8 and Sentinel-2 satellite imagery, and the multifeature combined (MFC) method [14] is mainly developed for GF-1 MFV satellite imagery. Learning-based methods have significantly advanced cloud detection from RS imagery due to powerful data adaptability. However, early machine learning-based methods, for example, using support vector machine (SVM) [16], neural network [17], or maximum likelihood [18], heavily rely on hand-crafted features, such as color, texture, and morphological features. These hand-crafted features are difficult to have sufficient discriminability to distinguish cloud for complicated cases such as the cloud-snow coexistence [1].

As a subset of machine learning, deep learning has attracted tremendous attention in image processing. In the field of RS, cloud detection based on CNN methods has also attracted increasing attention. In this article, we roughly divide the CNN-based methods into two categories: 1) cloud detection based on image-patch classification and 2) cloud detection based on semantic segmentation.

A. Cloud Detection Based on Image-Patch Classification

Fully connected deep convolutional neural networks (DCNN) achieving great success in image classification and recognition, have also been brought to cloud detection from RS images. Early methods [19]–[21] predict a category for each pixel by testing the surrounding image patch, which are time consuming and present annoying noise in their cloud detection results. To improve spatial coherence, some methods [22], [23] grouped image pixels to superpixels before cloud detection. However, the performance of superpixel-based cloud detection methods is limited by the accuracy of superpixel clustering [24]. Zi *et al.* [25] further used the fully connected conditional random field (CRF) to refine the cloud detection results. However, such a postprocessing optimization is very time-consuming particularly for RS image of large sizes.

B. Cloud Detection Based on Semantic Segmentation

Inspired by the high performance semantic segmentation of fully convolutional neural networks (FCNNs), cloud detection based on FCNNs has achieved remarkable results. Long *et al.* [26] apply FCN-8 directly for Landsat-8 satellite imagery cloud detection [27]. But, FCN-8 [26] is not able to provide accurate edges in segmentation results. For these reasons, recent works mainly adopt the encoder–decoder structure, such as U-net [28] and SegNet [29], as the architecture for cloud detection from multispectral RS imagery. Typical works, such as [30]–[34] modified the U-net model by fusing multilevel feature for RS cloud detection, and there also some work that use the lightweight U-net variants for onboard RS cloud detection [35], [36]. Chai *et al.* [37] and Lopez *et al.* [38] utilized modified SegNet [29] for multispectral RS image cloud detection, and their encoder structures are based on the VGG model [39]. Besides, a large number of other network structures, such as multiscale/multilevel feature fusion networks [10], [40]–[42], were also proposed. Meanwhile, the powerful semantic segmentation network for nature images, Deeplab V3+ [43], was also utilized for multispectral RS cloud detection [44]. Recently, Liu *et al.* [45] proposed a new deep learning architecture, named CloudNet, in which no downsampling was applied for cloud detection with accurate boundary details.

Previous CNN-based cloud detection methods, such as deep pyramid network (DPN) [40] and modified U-net [28], focus on elementwise summation and multilayer concatenation to fuse multilevel feature maps to improve cloud detection accuracy. Our method investigates how to adaptively fuse multilevel feature maps seamlessly through channel and spatial attention, respectively. In addition, we introduce the HSIGFs to remedy the diluted high-level semantic information, which makes the layers at decoder side aware of the locations of the cloud objects. In this article, CDnetV2 also uses ResNet-50 [46] as a backbone to extract multilevel/abstract features, but differs significantly in the following three aspects. First, CDnetV1 extracts multiscale/level features using the feature pyramid module (FPM); on the contrast, multiscale/level features in CDnetV2 directly comes from the backbone network and HSIGF, which decreases computational complexity and retains multiscale/level semantic features at same time. Second, CDnetV2 not only exacts extraction multiscale/level features but also cares for feature fusion ways that through channel-attention and spatial attention way to fuse multilevel/multiscale feature maps. Third, CDnetV2 introduces a series of HSIGF to remedy the semantic dilution problem at the layers of the decoder side, which further improves cloud detection results.

III. Proposed Method

Convolutional neural networks (CNNs) have been proved to be highly effective in RS cloud detection. In this article, we propose a novel CNN framework, named CDnetV2, for cloud detection from satellite thumbnails covered with snow. In the following sections, we first introduce the overall framework, then give details of key modules involved in CDnetV2.

A. Overall Workflow

As shown in Fig. 2, we present the framework of the proposed CDnetV2. To be specific, we first use ResNet-50 [46] as an encoder to extract multilevel/abstract features. Then we add a decoder after the encoder to construct

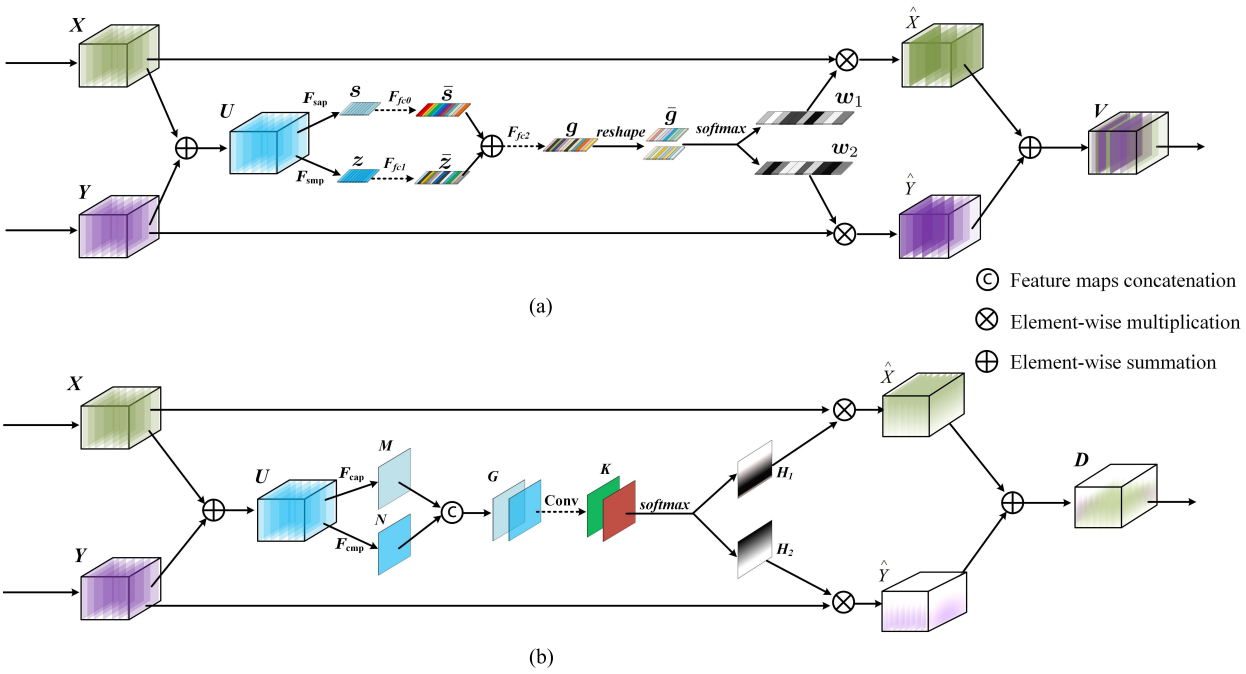

Fig. 3. Detail structure of (a) CAFM and (b) SAFM. It is a two branches feature maps fusion case. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the input feature maps. $\boldsymbol{V}$ and $\boldsymbol{D}$ are the output fused feature maps in CARM and SAFM, respectively.

an encoder–decoder network structure to utilize different level features, especially low-level feature, to generate sharp cloud detection boundaries. As shown in Fig. 2, we propose an adaptive feature fusion model (AFFM) in CDnetV2 to ensure that feature maps at different levels are seamlessly merged. As we know, low-level feature maps contain rich spatial information while high-level feature maps contain rich semantic information [10], [40]. Integration of them provides more detailed cloud information so as to achieve pixel-level cloud detection accuracy. In addition, since high-level semantic information features would be gradually diluted when transmitting through the layers of the decoder side [11], we introduce a series of HSIGFs to remedy the semantic dilution problem. The high-level semantic information in HSIGFs are extracted by using a proposed high-level feature fusing model (HFFM). Recently, typical segmentation works, such as FCN [26], and PSPnet [47], directly upsampled eight times from the final feature map to generate the output masks. In this article, we use four times upsampling to further improve cloud detection performance while maintaining reasonable computing and memory complexity. These key modules are described in detail in the following.

B. Adaptive Feature Fusion Module (AFFM)

In most previous work, the elementwise summation is one of the most widely used methods for fusion multilevel/multiscale feature maps [10], [26]. However, this simple summation operation does not take into account the level/scale differences between the different feature maps. Inspired by Li *et al.* [48] and Woo *et al.* [49], in order to effectively fuse multilevel/multiscale feature maps, we propose an adaptive feature fusion module (AFFM) to ensure that feature maps at different level/scales are seamlessly merged. The detailed structure of the AFFM is shown in Fig. 2(b). It consists of three basic build modules, that is, CAFM, SAFM, and CARM. As shown in Fig. 2(b), we first use CAFM and SAFM to adaptively fuse multilevel/multiscale feature maps through channel and spatial attention, respectively. Then, these fused features are further merged together with the high-level semantic feature from HSIGFs through elementwise summation operation. Finally, we use CARM to further refine the merged feature map. In other words, CAFM and SAFM focus on adaptive fusing features while CARM focuses on improving the representation of features. In what follows, we describe the structures of the three modules in detail.

1) CAFM: This model integrates the feature maps of multiple branches in a channel-attention way. As shown in Fig. 3(a), we produce channel attention weights by exploiting the cross-branch relationship between the two input branch feature maps, that is, $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ and $\boldsymbol{Y} \in \mathbb{R}^{H\times W\times C}$, to adaptively fuse cross-branch feature maps with different feature levels. To achieve this goal, we first integrate the two branches with an elementwise summation operation described as follows:

$$\boldsymbol{U} = \boldsymbol{X} + \boldsymbol{Y} \tag{1}$$

where $\boldsymbol{U} \in \mathbb{R}^{H\times W\times C}$. Inspired by [49], we use max-pooled and average-pooled operations simultaneously to generate channelwise statistics as $\boldsymbol{s} \in \mathbb{R}^{C}$ and $\boldsymbol{z} \in \mathbb{R}^{C}$, respectively. Here, the cth element of $\boldsymbol{s}$ and $\boldsymbol{z}$ are denoted as

$$s_c = F_{\text{sap}}(\boldsymbol{U}_c) \tag{2}$$

$$z_c = F_{\text{smp}}(\boldsymbol{U}_c) \tag{3}$$

where $\boldsymbol{U}_c \in \mathbb{R}^{H\times W}$ is the cth channel feature map of $\boldsymbol{U}$. $F_{\text{sap}}(\cdot)$ and $F_{\text{smp}}(\cdot)$ are global average pooling and global

max pooling operations for aggregating spatial information, respectively. Subsequently, both channelwise statistics are then forwarded to a multilayer perception network before aggregation, that is

$$\bar{\boldsymbol{s}} = \boldsymbol{F}_{fc0}(\boldsymbol{s}, \boldsymbol{W}_{01}, \boldsymbol{W}_{02}) = \delta(\boldsymbol{W}_{02}\delta(\boldsymbol{W}_{01}\boldsymbol{s})) \tag{4}$$

$$\bar{\boldsymbol{z}} = \boldsymbol{F}_{fc1}(\boldsymbol{z}, \boldsymbol{W}_{11}, \boldsymbol{W}_{12}) = \delta(\boldsymbol{W}_{12}\delta(\boldsymbol{W}_{11}\boldsymbol{z})) \tag{5}$$

where $\boldsymbol{W}_{01}$ and $\boldsymbol{W}_{11} \in \mathbb{R}^{(C/r)\times C}$, $\boldsymbol{W}_{12}$ and $\boldsymbol{W}_{02} \in \mathbb{R}^{nC\times(C/r)}$, $\bar{\boldsymbol{s}} \in \mathbb{R}^{nC}$, and $\bar{\boldsymbol{z}} \in \mathbb{R}^{nC}$. $\delta(\cdot)$ represents the nonlinear ReLU operation [50]. In this article, we set $r = 8$ to achieve promising performance [51]. Here, n represents the number of branches. $n = 2$ in Fig. 3(a). After the FC layers, we perform elementwise summation operation to further fuse these two branches information, that is

$$\boldsymbol{e} = \bar{\boldsymbol{s}} + \bar{\boldsymbol{z}} \tag{6}$$

where $\boldsymbol{e} \in \mathbb{R}^{nC}$. The following three FC layers are introduced to enable the guidance for the adaptive channel weights selections, that is:

$$\boldsymbol{g} = \boldsymbol{F}_{fc2}(\boldsymbol{e}, \boldsymbol{W}_{21}, \boldsymbol{W}_{22}, \boldsymbol{W}_{23}) = \delta(\boldsymbol{W}_{23}\delta(\boldsymbol{W}_{22}\delta(\boldsymbol{W}_{21}\boldsymbol{e}))) \tag{7}$$

where $\boldsymbol{W}_{21} \in \mathbb{R}^{2nC\times nC}$, $\boldsymbol{W}_{22} \in \mathbb{R}^{2nC\times 2nC}$, $\boldsymbol{W}_{23} \in \mathbb{R}^{nC\times 2nC}$, and $\boldsymbol{g} \in \mathbb{R}^{nC}$. We reshape $\boldsymbol{g} \in \mathbb{R}^{2C} \rightarrow \bar{\boldsymbol{g}} \in \mathbb{R}^{2\times C}$. Then we use a softmax operation to obtain the final channel attention weights, that is

$$(\boldsymbol{w}_1; \boldsymbol{w}_2) = \sigma(\bar{\boldsymbol{g}}) \tag{8}$$

where $\sigma(\cdot)$ is the softmax operator. $\boldsymbol{w}_1 \in \mathbb{R}^C$ and $\boldsymbol{w}_2 \in \mathbb{R}^C$ represent the soft channel weight for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Finally, the fused feature $\boldsymbol{V} \in \mathbb{R}^{H\times W\times C}$ is calculated by the following formula:

$$\boldsymbol{V}_c = w_{1c} \cdot \boldsymbol{X}_c + w_{2c} \cdot \boldsymbol{Y}_c, \quad w_{1c} + w_{2c} = 1 \tag{9}$$

where $\boldsymbol{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_C]$, $\boldsymbol{V}_c \in \mathbb{R}^{H\times W}$. Scalar w_{1c} and w_{2c} are the cth attention weight of $\boldsymbol{w}_1 = [w_{11}, w_{12}, \ldots, w_{1C}]$ and $\boldsymbol{w}_2 = [w_{21}, w_{22}, \ldots, w_{2C}]$, respectively. $\boldsymbol{X}_c \in \mathbb{R}^{H\times W}$ and $\boldsymbol{Y}_c \in \mathbb{R}^{H\times W}$ are the cth channel of input feature maps $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. There are two branches in Fig. 3(a), and it is easy to extend to multiple branches case, such as the proposed HFFM module [in Fig. 2(c)] with four branches.

2) SAFM: Different from the CAFM, SAFM focuses on spatial attention to fuse multilevel/multiscale feature maps, which is complementary to CAFM. To be specific, SAFM uses the learned spatial attention maps to weight the across-branch feature maps to fuse multilevel/multiscale feature maps adaptively. As shown in Fig. 3(b), to achieve features fusion through spatial-attention, we apply max and average pooling operations along the channel axis and concatenate the pooled results to produce a feature descriptor $\boldsymbol{G} \in \mathbb{R}^{H\times W\times 2}$. After that we further apply five convolution layers, as shown in Fig. 4, to generate encoded feature maps $\boldsymbol{K} \in \mathbb{R}^{H\times W\times 2}$, followed by a softmax layer to generate two spatial attention maps $\boldsymbol{H}_1 \in \mathbb{R}^{H\times W\times 1}$ and $\boldsymbol{H}_2 \in \mathbb{R}^{H\times W\times 1}$ for the two input branch feature maps $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. In this two branches example case as shown in Fig. 3(b), the spatial-attention fused feature map $\boldsymbol{D} \in \mathbb{R}^{H\times W\times C}$ is obtained by

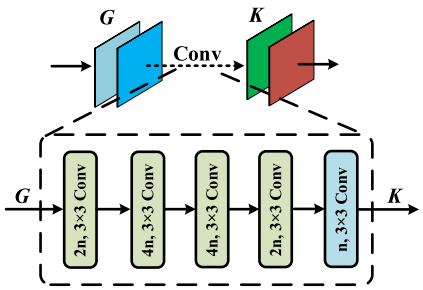

Fig. 4. Five convolution layers structure. The five layers except the last layer are the standard convolution layers have nonlinearity ReLU [50] and BN) [52] layers. In this two branches example case, $n = 2$.

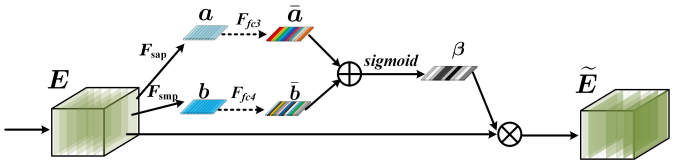

Fig. 5. Detail structure of CARM.

$$\boldsymbol{D} = \boldsymbol{H}_1 \otimes \boldsymbol{X} + \boldsymbol{H}_2 \otimes \boldsymbol{Y} \tag{10}$$

where $\otimes$ represents elementwise multiplication operation. During multiplication, the spatial attention values are broadcasted (copied) along the channel dimension.

3) CARM: The CARM is widely used in visual tasks [49], [51]. CARM is able to increase network sensitivity to promote informative channels and suppress uninformative ones, thus improving the representation power of network. In this article, we propose a modified CARM to refine feature maps. Its structure is shown in Fig. 5. To be specific, we embed two branches similar to the proposed CAFM to generate two channelwise statistics defined by $\bar{\boldsymbol{a}} \in \mathbb{R}^C$ and $\bar{\boldsymbol{b}} \in \mathbb{R}^C$, respectively. Then we performed elementwise summation operation to further fuse these two branches. The channel attention weight factor $\boldsymbol{\beta} \in \mathbb{R}^C$ is obtained with a sigmoid activation operation. Finally, the channel attention refinement operation is performed by a channelwise multiplication layer. As shown in Fig. 5, let $\boldsymbol{E} = [\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_C]$ denote the input feature maps. The channel refined result of $\boldsymbol{E}$ is denoted as $\widetilde{\boldsymbol{E}} = [\widetilde{\boldsymbol{E}}_1, \widetilde{\boldsymbol{E}}_2, \ldots, \widetilde{\boldsymbol{E}}_C]$. The cth feature map of $\widetilde{\boldsymbol{E}}_C$ can be represented as

$$\widetilde{\boldsymbol{E}}_c = \beta_c \cdot \boldsymbol{E}_c \tag{11}$$

where scalar β_c is the cth attention weight of $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_C]$.

C. HSIGFs

To remedy the diluted high-level semantic information and improve the accuracy of cloud detection, we introduce a series of HSIGFs to explicitly make each layer at decoder side aware of the locations of the cloud objects. The high-level semantic information in HSIGFs is extracted by the proposed HFFM.

As shown in Fig. 2(a), we place the proposed HFFM after the last layer of the backbone network, Resnet-50 [46], to capture global semantic guidance information. The detailed structure of HFFM is shown in Fig. 2(c). To be specific, we first use the proposed CARM to refine the input convolutional features. Then we capture multiscale and global context information by multiple dilated convolutional layers [53] and one global average pooling operation. Finally, the multiscale and global context information are adaptively fused by using the proposed CAFM and SAFM modules. The fused feature map contains the most important semantic information for cloud accurate detection. As shown in Fig. 2(b), we introduce the high-level semantic features through guidance flows to remedy semantic information dilution at decoder layers.

D. Loss Function

Similar to [1], the softmax loss function is utilized as the principle loss, denoted by $\mathcal{L}_p$ for multiclass prediction [54]. We also add an auxiliary loss function $\mathcal{L}_a$ to supervise the output of CDnetV2. The auxiliary loss function is applied to the last residual block of backbone network. As shown in [47] and [55], the joint loss helps optimize the learning process and promotes the network to achieve a promising segmentation state. In addition, we introduce a regularization term in loss function to avoid over fitting. Therefore, the final loss function is defined as

$$\mathcal{L}(\Theta) = \lambda_1 \mathcal{L}_p(\Theta) + \lambda_2 \mathcal{L}_a(\Theta) + \frac{\lambda_3}{2}\|\Theta\|^2 \tag{12}$$

where λ_1, λ_2, and λ_3 are equally set to 1. Network parameter Θ are updated through iterations to minimize the final loss $\mathcal{L}(\Theta)$.

IV. Data Set and Experimental Settings

A. Data Set

In this article, experiments are conducted on a set of ZY-3 satellite thumbnails.[1] ZY-3 satellite is China's first civilian high-resolution optical transmission stereomapping satellite launched on January 9, 2012 [56]. The optical sensor on ZY-3 satellite can obtain multispectral and panchromatic images at a resolution of 5.8 and 2.1 m, respectively. The satellite data are mainly used in land surveys, natural resource survey, and monitoring, and other related applications. More information about the ZY-3 satellite is referred to official website of the Landsat Remote Sensing Application Center of China.[2]

The data set used in this article is different from the one in CDnetV1 [1] which contains only a few thumbnails with cloud-snow coexistence. This new data set consists of 432 scenes thumbnails,[3] including 232 snow-covered thumbnails and 200 snow-free ones. In this article, we develop an efficient image labeling approach. For each image, we first draw closed boundaries for cloud areas using an image editing software such as Adobe Photoshop. Then, the closed boundaries are filled by morphological operations to generate label. We also conducted a labeling error analysis. We randomly sample 20 thumbnails, whose cloudage is in the range between 10% and 90% from the data set. Specifically, five experienced labelers are invited to relabel these 20 thumbnails independently, obtaining five label masks for each thumbnail. Then, the winner-take-all strategy is used to generate a more reliable label mask as the ground truth for error analysis. The number of mislabeled pixels is used as the evaluation criterion. Finally, the average mislabeling rate of sampled thumbnails is about 0.18%. In addition, a largest mislabeled pixel rate of 0.61% was found in a thumbnail that is partly covered by thin cloud. Overall, most of the labeled masks show perfect consistency with the groundtruth.

In this article, the data set is divided into two parts: 332 pixel-level labeled thumbnails for training and 100 pixel-level labeled thumbnails for testing. The 100 testing images consists of 80 snow covered images and 20 snow-free covered ones. In addition, we conduct data augmentation operations including image rotation and flipping [57] to improve the robust of the network. Finally, there are about 40k subimages of size 321 $\times$ 321 in the training data set. In Figs. 6 and 7, we show some of thumbnail examples used in this article. It can be seen that cloud has color distributions and local texture patterns similar to these of snow/ice, which makes them difficult to be distinguished between each other with our eyes.

B. Training Details

In this article, the proposed CDnetV2 are trained under TensorFlow deep learning framework [58] and optimized by SGD (stochastic gradient descent) and momentum optimization method [59]. Similar to [1], we adopt "poly" learning rate strategy to update the learning rate. We set each iteration with batch size 12, momentum 0.9, batch norm decay 0.9997, initial learning rate 3×10^{-5}, and weight decay 1×10^{-5} in training. The total epoch is 60 in all experiments. All of the experiments are carried out on Ubuntu 14.04 operating system and ran within NVIDIA GeForce RTX 2080 Ti.

C. Comparison Methods and Evaluation Metrics

1) Comparison Methods: In this article, we use three CNN-based cloud detection methods, that is, modified SegNet (MSegNet) [37], modified U-Net (MU-Net) [34], and CDnetV1 [1], as the competing methods. The MSegNet [37] and MU-Net [34] are the typical deep network with encoder–decoder structure. These methods have achieved promising cloud detection results in Landsat 7-8 multispectral images. CDnetV1 [1] achieves a promising cloud detection results in RS images. In addition, two representative segmentation networks, that is, PSPnet [47] and DeeplabV3+ [43],[4] are used as the competing methods.

[1] http://39.105.157.77/EN/query

[2] http://www.lasac.cn/chwx/zrzyzgwx/kyx/index.html

[3] RGB thumbnails and gray ones are of sizes 1 k $\times$ 1 k and 3 k $\times$ 3 k, respectively.

[4] For fair comparison, we set ResNet-50 [46] as the network backbone of PSPnet [47] and Deeplab V3+ [43] to extract context feature maps.

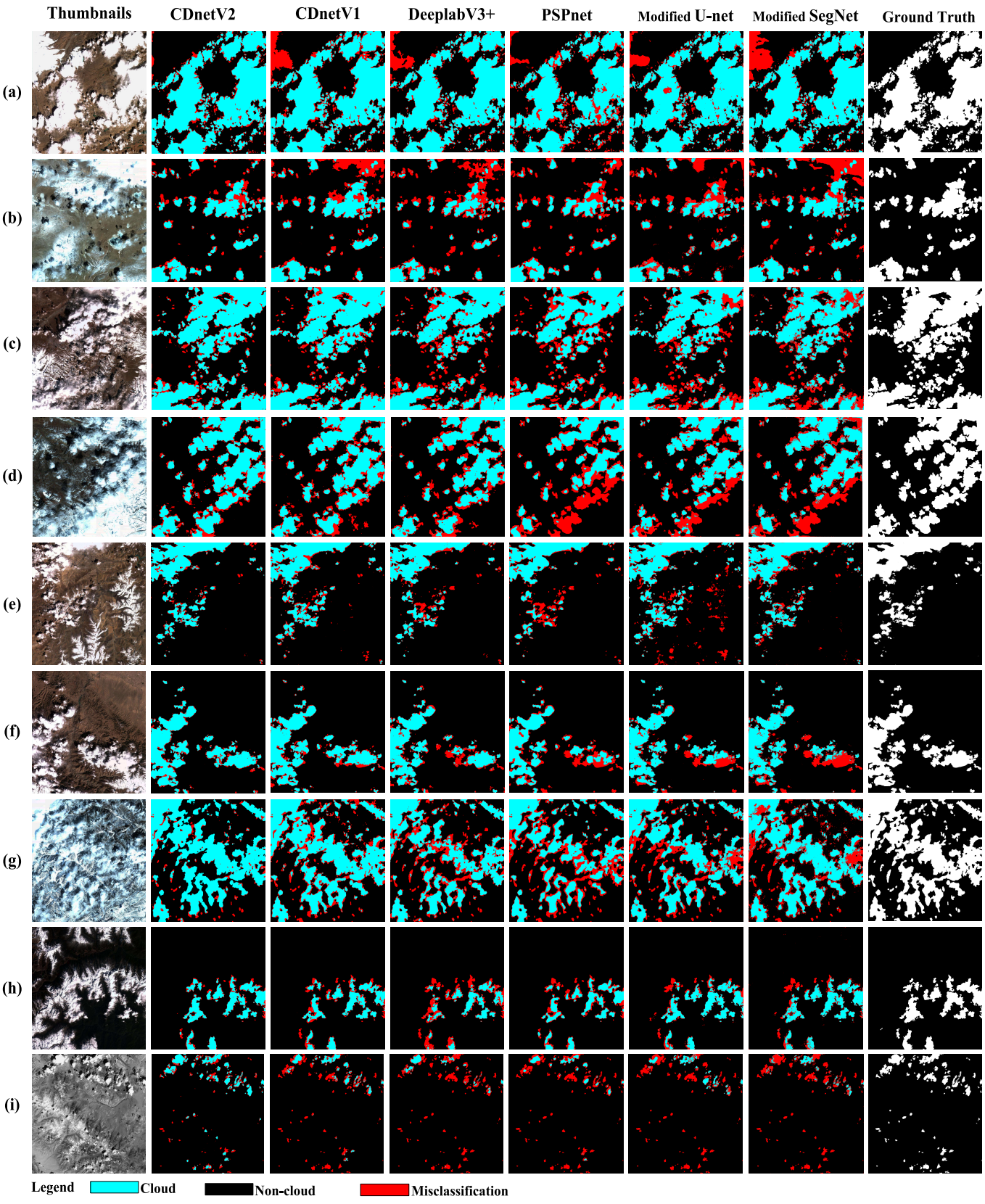

Fig. 6. Visual comparison of cloud extraction results by six CNN-based methods on nine thumbnails (a)∼(i) of size 1k × 1k from ZY-3 satellite imagery. The test thumbnails are selected from areas with heavy cloud-snow coexistence.

2) Evaluation Metrics: In this article, we use the same quantitative metrics as [1], [34], and [37] to comprehensively evaluate the cloud detection results, including overall accuracy (OA), MIo, kappa coefficient (Kappa), producer accuracy (PA), and user accuracy (UA). Here, all the metric scores are calculated by using groundtruth masks as references.

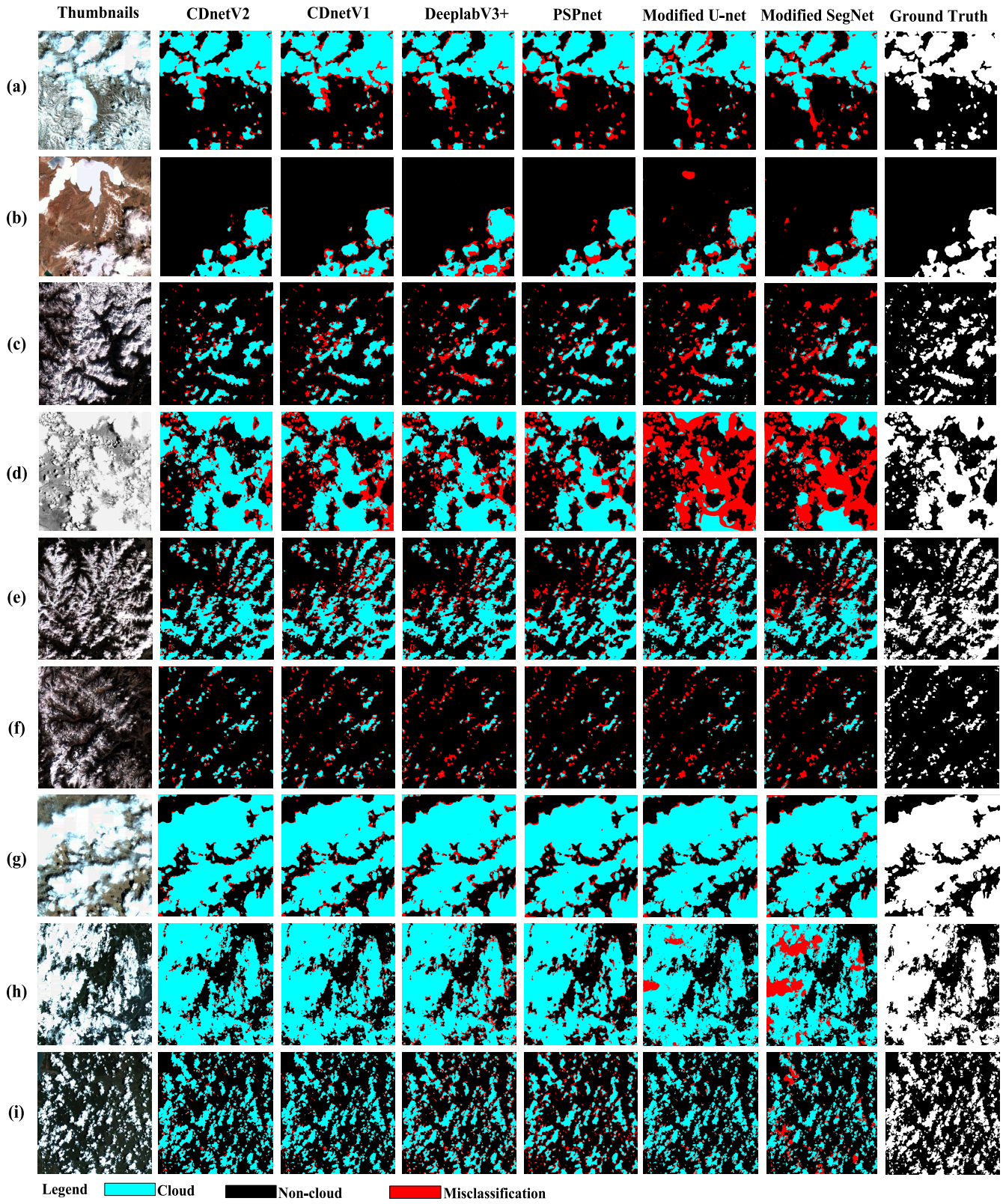

Fig. 7. Comparison of cloud extraction results on different cases. (a)–(f) Six cloud–snow coexistence extreme cases. (g)–(i) Three cloud-only cases.

V. Experimental Results

A. Ablation Study

We implement four ablation experiments on our ZY-3 satellite thumbnails data set to demonstrate the advancement and effectiveness of our designed network.

1) Baseline: We use the ResNet-50 [46] as the backbone architecture, which is pretrained on ImageNet data set [60]. The backbone has the output feature maps that is 1/16th of the input image. The output segmentation results are directly upsampled to the same size as the input images. We list

TABLE I
CLOUD EXTRACTION ACCURACY (%) OF DIFFERENT ABLATION NETWORKS ON ZY-3 SATELLITE THUMBNAILS. ALL RESULTS ARE THE AVERAGING RESULTS ON ALL THE TESTING THUMBNAILS

Baseline	√	√	√	√	√
U-shape	×	√	√	√	√
AFFM	×	×	√	×	√
HSIGFs	×	×	×	√	√
OA	90.39	91.63	92.17	93.33	**95.76**
MIoU	77.59	80.76	81.52	82.26	**86.62**
Kappa	71.32	74.67	77.44	78.13	**82.51**
PA	68.67	71.71	77.51	80.12	**87.75**
UA	89.78	**91.32**	89.51	88.27	88.58

evaluation performance of the backbone network as shown in Table I.

2) Ablation for U-Shape: We design an encoder–decoder network with U-shape structure, similar to [33] and [34], to leverage features at different layers to generate detailed boundaries and recover results spatial resolution. It directly fuses multilevel feature maps by using the elementwise summation operation. The results of the U-shape network, named, "Baseline + U-shape," is shown in Table I. It can be seen that the U-shape network improves the performance from 77.59% to 80.76% in terms of MIoU, and from 71.32% to 74.67% in terms of Kappa.

3) Ablation for AFFM: Based on U-shape network structure, we further introduce an adaptive multilevel feature fusion module (AFFM) to ensure that feature maps at different levels are seamlessly merged as shown in Fig. 2. Table I shows the U-shape network structure equipped with AFFMs, named, "Baseline + U-shape + AFFM," further improve the performance.

4) Ablation for HSIGFs: Since high-level semantic information features will be gradually diluted when they are transmitted through the decoder layers, we introduce a series of HSIGFs to explicitly make layers at decoder be aware of the locations of the cloud objects. We name this network "Baseline + U-shape + HSIGFs." As shown in Table I, the U-shape network structure equipped with HSIGFs also improves its performance.

5) Ablation for HSIGFs + AFFM: U-shape network equipped with HSIFs and AFFM modules, named "Baseline + U-shape + HSIGFs + AFFM," is the proposed CDnetV2. Its experimental evaluation results are also listed in Table I. It can be seen that CDnetV2 achieves the highest performance in terms of OA, MIoU, Kappa, and PA compared with other ablation networks. The experimental results demonstrate that U-shape, HSIGFs, and AFFM are beneficial for cloud detection accuracy. Fusing these models together makes CDnetV2 achieve its promising performance.

B. Cloud Detection Results on ZY-3 Data Set

1) Quantitative Results: Table II lists quantitative results in terms of OA, MIoU, Kappa, PA, and UA. CDnetV2 achieves the best quantitative results among all competing methods. To be specific, compared with CDnetV1 [1], CDnetV2 achieves performance gain by more than 2.11%, 3.28%, 3.30%, 5.38%, and 1.86%. In addition, our method consistently outperforms other two CNN-based cloud detection models, MSegNet [37] and MU-Net [34]. It also outperforms other two generic images segmentation networks, PSPnet [47] and DeeplabV3+ [43]. The quantitative results demonstrate that the proposed CDnetV2 are able to achieve promising cloud detection performance.

TABLE II
CLOUD EXTRACTION ACCURACY (%) OF DIFFERENT CNN-BASED METHODS ON ZY-3 SATELLITE THUMBNAILS

Method	OA	MIoU	Kappa	PA	UA
MSegNet [37]	90.86	81.20	75.57	73.78	86.13
MUnet [34]	91.62	82.51	76.70	74.44	87.39
PSPnet [47]	90.58	81.63	75.36	76.02	87.52
DeeplabV3+ [43]	91.80	82.62	77.65	75.30	87.76
CDnetV1 [1]	93.15	82.80	79.21	82.37	86.72
CDnetV2	**95.76**	**86.62**	**82.51**	**87.75**	**88.58**

TABLE III
STATISTICAL RESULTS OF CLOUDAGE ESTIMATION ERROR IN TERMS OF THE MAD AND ITS VARIANCE

Methods	Mean value (μ)	Standard Deviation (σ^2)
CDnetV2	0.0241	0.0220
CDnetV1 [1]	0.0357	0.0288
DeeplabV3+ [43]	0.0456	0.0301
PSPnet [47]	0.0487	0.0380
MUnet [34]	0.0544	0.0583
MSegNet [37]	0.0572	0.0591

Meanwhile, we further evaluate the cloud cover estimation accuracy of our method compared with competing approaches. Fig. 8 plots the normalized estimated cloudages against the groundtruth. Ideally, results for perfect cloudage estimation would lies on the solid line with a slope of one for various cloudage. Either overestimation or underestimation would cause deviation against the benchmark line. Fig. 8 shows that the results estimated by the proposed CDnetV2 distributed more closer to the benchmark than those estimated by competing approaches, which is also verified by their R-square and root mean square error (RMSE) results. Table III reports statistical results of cloudage estimation error in terms of the mean absolute deviation (MAD) and its variance. Both MAD and variance of our estimation results are smaller than those of competing methods, which demonstrates that CDnetV2 is more accurate and robust in cloudage estimation than competing methods.

2) Qualitative Results: Fig. 6 shows cloud detection visual results on nine typical thumbnails covered with snow. For visual inspection, we mark the correctly detected cloud pixels, noncloud pixels, and misclassified pixels in bright cyan, black, and red, respectively. Results in Fig. 6 show that CDnetV2 is far more accurate than the competing methods. The results by MSegNet [37] and MUNet [34] present a large number of misclassified pixels. PSPnet [47] and DeeplabV3+ [43] developed for generic image segmentation tasks. Using them

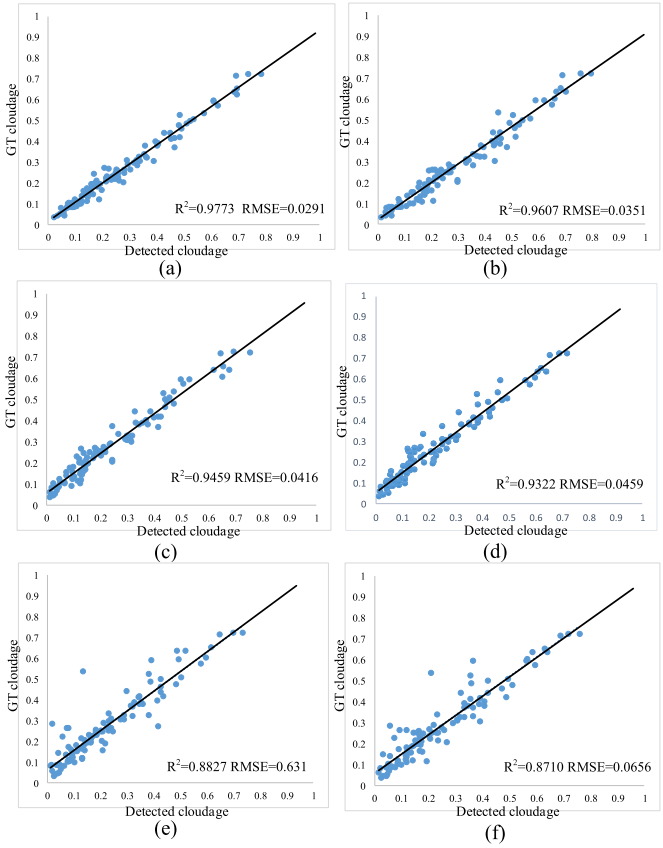

Fig. 8. Comparison of cloud cover estimation accuracy between (a) CDnetV2, (b) CDnetV1, (c) DeeplabV3+, (d) PSPnet, (e) MUnet, and (f) MSegNet.

in RS images leads to suboptimal cloud detection results as the cloud objects in RS images have dramatic spatial variability. CDnetV1 is able to achieve promising performance on most of areas, but it not explicitly addresses the coexistence of cloud and snow in cloud detection. In contrast, CDnetV2 is able to effectively detect cloud from thumbnails covered with snow thanks to the adaptive feature maps fusion strategy and high-level semantic segmentation information compensation. In addition, we further present more visual results on challenge cases in Fig. 7. In Fig. 7(a), the image is almost covered by half snow and half cloud; in Fig. 7(c), the high mountains are covered by snow, half of which are further overlaid by cloud; in Fig. 7(f), only a small amount of cloud float on the snow covered mountains; while in Fig. 7(g)–(h), almost all the grounds are covered by cloud. The results show that our method is able to handle all these challenging cases, which justifies its effectiveness and robustness.

In a word, both the quantitative results in Table II and qualitative results in Figs. 6 and 7 demonstrate that CDnetV2 is a promising cloud detection method for satellite thumbnails with cloud–snow coexistence.

VI. Discussion and Analysis

A. Analysis on the AFFM Module

In the proposed AFFM module, the two AFFMs, CAFM and SAFM, efficiently fuse feature maps from different levels seamlessly by learning attentional information along channel and spatial dimensions, respectively. To understand how AFFM works, we visualize the channel-attention and spatial-attention weights in CAFM and SAFM modules, respectively.

TABLE IV
Standard Deviation (σ^2) of Channel Attention Weights in Different Stages

Standard Deviation	stage 1	stage 2	stage 3	stage 4
average σ^2	0.0281	0.0432	0.0397	0.0573

1) Channel-Attention Weights in CAFM: Taking a thumbnail from the testing set as an example, Fig. 9 shows weights for each channel in low-level and high-level feature maps from stage 1 to stage 4. Here, we roughly refer to the feature maps of the encoder layers as the low-level feature map, and those of the decoder layers as the high-level ones. It can be seen that the weight value of each channel is not the same when they fuse together. With the expansion of difference in two branches feature maps, weight values have a great change. As we can see from Fig. 9(b) and (c), the weight values are close to 0.5 in stage 1, while most of them are far away from 0.5 in stages 2 and 3, and this phenomenon is most obvious in stage 4. In order to further observe changes of the channel attention weights in different stages, we calculate the average standard deviation (σ^2) of channel attention weights in different stages over the whole testing set, as shown in Table IV. The standard deviations progressively increase as the difference between of the two feature maps increase from stage 1 to stage 4. In a word, the learned channel attention weight values in CAFM help adaptive fusion multilevel/scale features.

2) Spatial-Attention Weights in SAFM: Being complementary to the channel-attention fusion, the SAFM focuses on the spatial attention aspect to fuse multilevel/multiscale feature maps. In Fig. 10, we show the spatial attention weight maps at different stages for a two branches case. The two spatial attention weight maps have different spatial variations. In particular, at stage 4, the spatial distribution of weights is highly correlated with the shapes of ground objects. For better visual results, we enlarge the two spatial attention weighting maps in stage 4, as shown in Fig. 10(b) and (c). Most of the spatial attention weight values in snow regions in Fig. 10(c) are higher than those in Fig. 10(b), which is in accordance with our expectation that the semantic segmentation information for distinguishing snow and cloud pixels is mostly provided by high-level features at stage 4. In addition, from the enlarged subimage results we can find that the spatial attention weight map of low-level feature map has higher weight values at cloud objects boundary than that of high-level feature map, which is in accordance with our expectation that the spatial information is mostly provided by the low-level feature maps. In a word, multilevel feature maps fusion based on spatial attention way makes the network leverage promising spatial information for further improving cloud detection accuracy.

B. Computational Complexity Analysis

In Table V, we evaluate computational complexity of these networks with three evaluation criterions, that is, floating

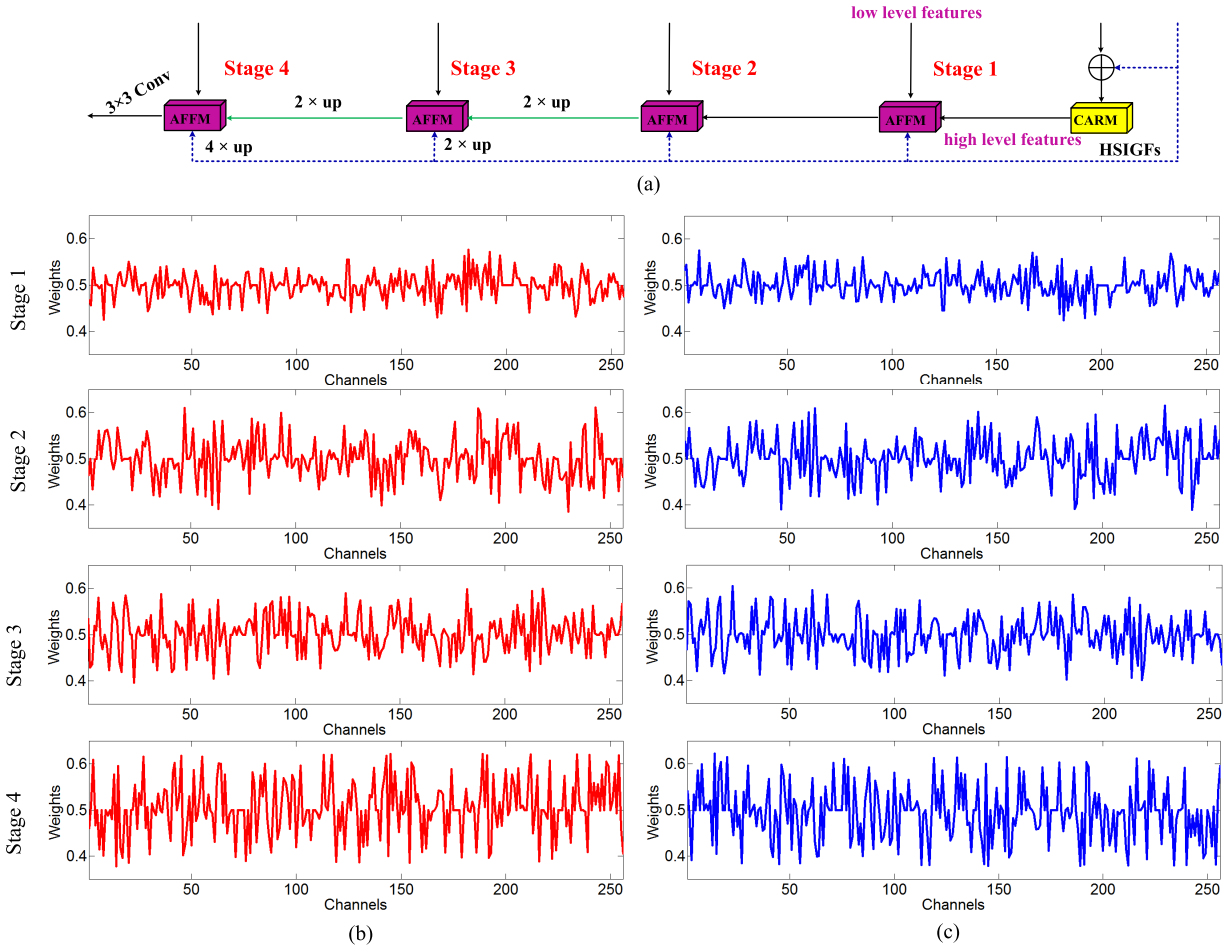

Fig. 9. Visualization of channel attention weights values at different fusion stages. (a) Different feature maps fusion stages. (b) Low level channel attention weights. (c) High level channel attention weights.

TABLE V
COMPUTATIONAL COMPLEXITY ANALYSIS OF DIFFERENT CNN-BASED METHODS

Methods	GFLOPs (224×224)	Trainable params	Running time (s) (1k×1k)
CDnetV2	31.5	65.9 M	1.31
CDnetV1 [1]	48.5	64.8 M	1.26
DeeplabV3+ [43]	31.8	40.3 M	1.14
PSPnet [47]	19.3	46.6 M	1.05
MUnet [34]	25.2	8.6 M	1.09
MSegNet [37]	90.2	29.7 M	1.28

[1] 1 GFLOPs = 1×10^9 FLOPs
[2] 1 M = 1×10^6

point operations (FLOPs), number of trainable parameters, and running time. FLOPs are calculated from input data of size 224×224 for easy comparison with reported results in the literature [61]. Running time is calculated from input data of size $1\ \text{k} \times 1\ \text{k}$. It can be seen that the FLOPs value of CDnetV2 is lower than that of CDnetV1 and MSegNet, but is higher than that of DeeplabV3+, PSPnet, MUnet, and MSegNet. The reason is that CDnetV2 has more training parameters than other competing methods. However, the running times are close for all the compared methods as the running time not only depends on number of operations, but also relate to the efficiency of memory access [62]. In CDnetV2, we use HSIGF for multiple times to prevent the dilution of high-level semantic information at decoder side, which increases the memory access and hence slightly higher running time. We will further improve the network structure of CDnetV2 to reduce its computational complexity in our future work.

C. Limitations

Although CDnetV2 achieves satisfactory cloud detection results for most tough cases, there are still some errors in thin cloud and fine particle size cloud masks. The thin cloud objects may have only minor differences with underground objects. Since our proposed network uses downsampling operations including pooling operations and convolutional operations with stride greater than one, it is difficult to retain enough pixels to represent the existence of spotwise cloud areas. Fig. 11 provides examples of cloud detection errors in areas covered by thin cloud and fine particle size cloud. It can be seen that there are many omission pixels in these segmentation

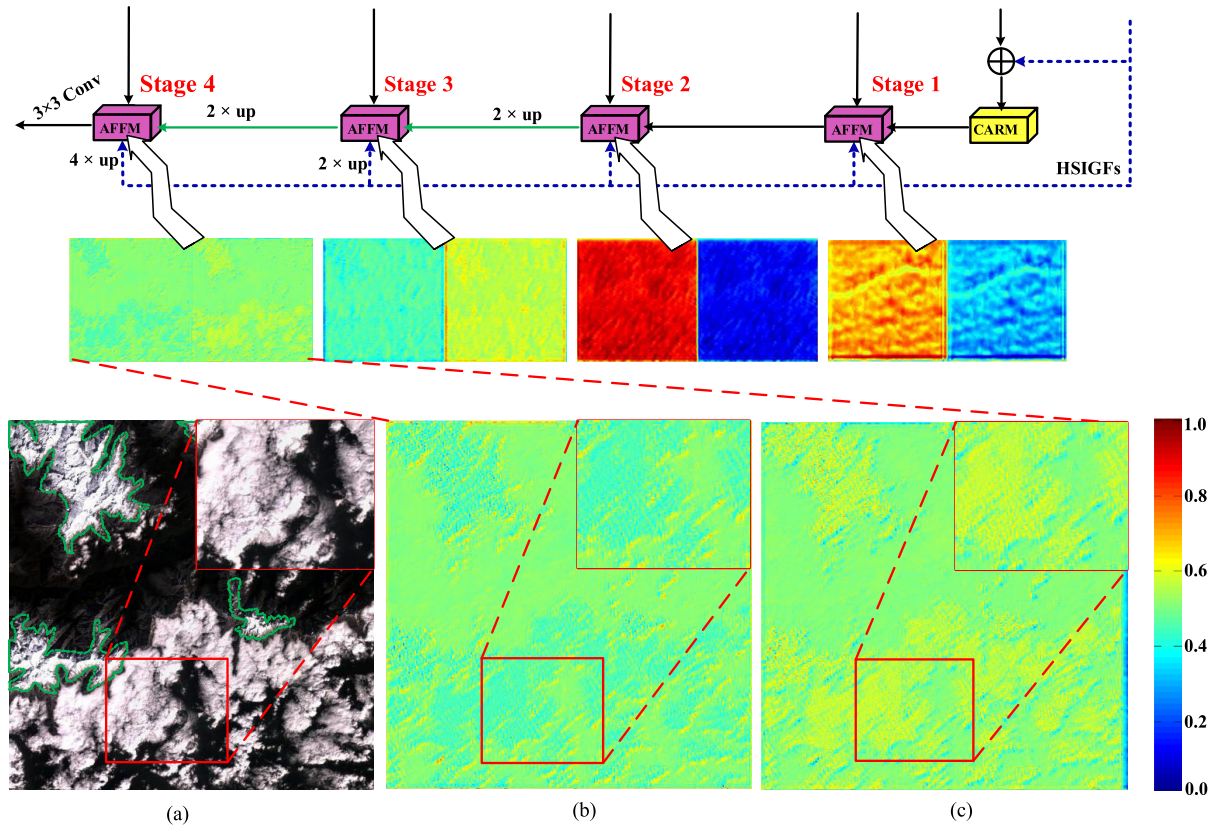

Fig. 10. Visualization of spatial attention weight maps at different fusion stages. The snow regions are sketched by green dotted lines in the input thumbnail. (a) Input ZY-3 satellite thumbnail. (b) Low-level features spatial weight map. (c) High-level features spatial weight map.

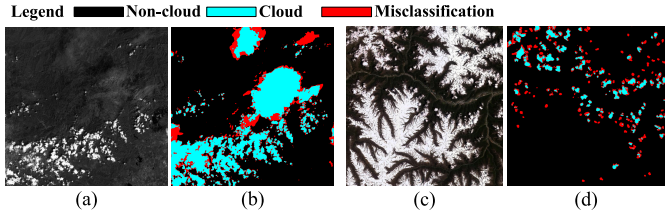

Fig. 11. Cloud detection results with considerable amount of misclassification on two challenging cases. (a) Thumbnail with thin cloud. (b) Detection result of (a). (c) Thumbnail with spotwise cloud and detection result of (c).

results. In future work, we will focus on these points to further improve this article.

VII. Conclusion

In this article, we propose an encoder–decoder network, named CDnetV2, for cloud detection from thumbnails with cloud–snow coexistence. The proposed network focuses on adaptively fusing multilevel/multiscale feature maps and remedying high-level semantic information diluted at decoder layers to improve cloud detection accuracy. The experimental results show that CDnetV2 achieves state-of-the-art cloud detection performance on ZY-3 satellite thumbnails. It shows great promise for practical application. In the future, we will extend the proposed method to other satellite thumbnails and make it widely used for cloud detection and cloudage estimation. In our future study, we will attempt to simultaneously detect cloud and snow on thumbnails. In addition, we will explore lightweight network by using knowledge distillation strategy [63] to further reduce the computational expense.

Acknowledgment

The authors thank Land Satellite Remote Sensing Application Center, Ministry of Natural Resources of China for providing ZY-3 satellite thumbnails. They also thank the editors and reviewers for their valuable suggestions.

References

[1] 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.

[2] J. Long, Z. Shi, W. Tang, and C. Zhang, "Single remote sensing image dehazing," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 59–63, Jan. 2014.

[3] A. J. Feijt, *Quantitative Cloud Analysis Using Meteorological Satellites=[Kwantitatieve Analyse Van Wolken Met Meteorologische Satellieten]*. Wageningen, The Netherlands: Wageningen Univ., 2000.

[4] 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.

[5] 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, Jan. 1991.

[6] 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. 1989.

[7] International Society for Optics and Photonics, "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.

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

[9] 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.

[10] 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.

[11] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," 2019, *arXiv:1904.09569*. [Online]. Available: http://arxiv.org/abs/1904.09569

[12] 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.

[13] 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, Mar. 2015.

[14] 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.

[15] Y. Oishi, H. Ishida, and R. Nakamura, "A new Landsat 8 cloud discrimination algorithm using thresholding tests," *Int. J. Remote Sens.*, vol. 39, no. 23, pp. 9113–9133, Dec. 2018.

[16] C. Latry, C. Panem, and P. Dejean, "Cloud detection with SVM technique," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jun. 2007, pp. 448–451.

[17] M. Hughes and D. 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, 2014.

[18] 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.

[19] K. Wohlfarth *et al.*, "Dense cloud classification on multispectral satellite imagery," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens. (PRRS)*, Aug. 2018, pp. 1–6.

[20] D. Varshney, P. K. Gupta, C. Persello, and B. R. Nikam, "Snow and cloud discrimination using convolutional neural networks," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 5, pp. 59–63, Nov. 2018.

[21] M. Xia, W. Liu, B. Shi, L. Weng, and J. Liu, "Cloud/snow recognition for multispectral satellite imagery based on a multidimensional deep residual network," *Int. J. Remote Sens.*, vol. 40, no. 1, pp. 156–170, Jan. 2019.

[22] 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.

[23] G. Morales, S. G. Huamán, and J. Telles, "Cloud detection in high-resolution multispectral satellite imagery using deep learning," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 2018, pp. 280–288.

[24] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[25] Y. Zi, F. Xie, and Z. Jiang, "A cloud detection method for Landsat 8 images based on PCANet," *Remote Sens.*, vol. 10, no. 6, p. 877, 2018.

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

[27] X. Zeng, J. Yang, and X. Deng, "Cloud segmentation of remote sensing images on Landsat-8 by deep learning," in *Proc. 2nd Int. Conf. Big Data Res. (ICBDR)*, 2018, pp. 174–177.

[28] 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.* Berlin, Germany: Springer, 2015, pp. 234–241.

[29] 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.

[30] S. Mohajerani, T. A. Krammer, and P. Saeedi, "Cloud detection algorithm for remote sensing images using fully convolutional neural networks," 2018, *arXiv:1810.05782*. [Online]. Available: http://arxiv.org/abs/1810.05782

[31] J. Drönner *et al.*, "Fast cloud segmentation using convolutional neural networks," *Remote Sens.*, vol. 10, no. 11, p. 1782, 2018.

[32] 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.

[33] S. Dev, S. Manandhar, Y. Hui Lee, and S. Winkler, "Multi-label cloud segmentation using a deep network," 2019, *arXiv:1903.06562*. [Online]. Available: http://arxiv.org/abs/1903.06562

[34] 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.

[35] Z. Zhang, A. Iwasaki, G. Xu, and J. Song, "Cloud detection on small satellites based on lightweight U-net and image compression," *J. Appl. Remote Sens.*, vol. 13, no. 2, Apr. 2019, Art. no. 026502.

[36] S. Ghassemi and E. Magli, "Convolutional neural networks for on-board cloud screening," *Remote Sens.*, vol. 11, no. 12, p. 1417, 2019.

[37] 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.

[38] J. Lopez, S. Santos, C. Atzberger, and D. Torres, "Convolutional neural networks for semantic segmentation of multispectral remote sensing images," in *Proc. IEEE 10th Latin-American Conf. Commun. (LATINCOM)*, Nov. 2018, pp. 1–5.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[40] S. Ozkan, M. Efendioglu, and C. Demirpolat, "Cloud detection from RGB color remote sensing images with deep pyramid networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 6939–6942.

[41] 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.

[42] 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.

[43] L. 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)*, pp. 801–818, 2018.

[44] G. Morales, S. G. Huamán, and J. Telles, "Cloud detection in high-resolution multispectral satellite imagery using deep learning," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 2018, pp. 280–288.

[45] 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.

[46] K. He, X. Zhang, S. Ren, and S. Jian, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

[47] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[48] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," 2019, *arXiv:1903.06586*. [Online]. Available: https://arxiv.org/pdf/1903.06586

[49] S. Woo, J. Park, J. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[50] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[51] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[53] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[54] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. ICML*, vol. 2, 2016, p. 7.

[55] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.
[56] 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, 2017.
[57] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*. [Online]. Available: http://arxiv.org/abs/1712.04621
[58] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.
[59] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.
[60] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[61] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," 2016, *arXiv:1611.06440*. [Online]. Available: http://arxiv.org/abs/1611.06440
[62] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[63] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531

Jianhua Guo (Student Member, IEEE) 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 pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China.

He was a Visiting Student at the Land Satellite Remote Sensing Application Center, Ministry of Natural Resources of China from 2015 to 2017. His main research focuses on remote sensing image matching, enhancement, classification, and segmentation.

Jingyu Yang (Senior Member, IEEE) 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, since 2009, where he is 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 the Signal Processing Laboratory, EPFL, 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 VCIP 2016 and the Area Chair in ICIP 2017. He was selected into 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.

Huanjing Yue (Member, IEEE) 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 2012, and from 2013 to 2015. She visited the Video Processing Laboratory, University of California at San Diego, San Diego, CA, USA, from 2016 to 2017. She is an Associate 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.

Hai Tan received the B.E. degree in surveying and mapping engineering from Liaoning Technical University, Fuxin, China, in 1998, the M.A. degree in cartography and GIS from the Chinese Academy of Surveying and Mapping, Beijing, China, in 2001, and the Ph.D. degree in cartography and GIS from Information Engineering University, Zhengzhou, China, in 2014.

He is an Associate Research Fellow with the Land Satellite Remote Sensing Application Center, MNR, Beijing. His main research focuses on remote sensing image quality inspection and evaluation.

Chunping Hou received the M.Eng. and Ph.D. degrees in electronic engineering from Tianjin University, Tianjin, China, in 1986 and 1998, respectively.

She was a Post-Doctoral Researcher with the Beijing University of Posts and Telecommunications, Beijing, China, from 1999 to 2001. Since 1986, she has been with the faculty of Tianjin University, where she is a Professor and the Director of the Broadband Wireless Communications and 3-D Imaging Institute. Her research interests include 3-D image processing and communication systems.

Kun Li (Member, IEEE) 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 Ecole Polytechnique de Federale Lausanne, Lausanne, Switzerland, in 2012 and from 2014 to 2015. She is 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 received the Platinum Best Paper Award in the IEEE ICME 2017. She was selected for the Peiyang Scholar Program of Tianjin University in 2016.