Attention Mechanism Cloud Detection With Modified FCN for Infrared Remote Sensing Images

LIYUAN LI, XIAOYAN LI, XIN LIU, WENWEN HUANG, ZHUOYUE HU, AND FANSHENG CHEN

1 Key Laboratory of Intelligent Infrared Perception, Chinese Academy of Sciences, Shanghai 200083, China
2 CAS Key Laboratory of Infrared System Detection and Imaging Technology, Shanghai Institute of Technical Physics, Shanghai 200083, China
3 University of Chinese Academy of Sciences, Beijing 100049, China
4 Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China

Corresponding authors: Fansheng Chen (cfs@mail.sitp.ac.cn), Xiaoyan Li (xiaoyanlee2018@163.com), and Zhuoyue Hu (uestchu@163.com)

This work was supported in part by the CASEarth Minisatellite Thermal Infrared Spectrometer Project of the Shanghai Institute of Technical Physics, Chinese Academy of Sciences under Grant XDA19010102; and in part by the National Natural Science Foundation of China under Grant 61975222.

ABSTRACT Semantic segmentation (SS) has been widely applied for cloud detection (CD) in remote sensing images (RSIs) with high spatial and spectral resolution because of its effective pixel-level feature extraction structure. However, the typical model of lightweight SS, namely the fully convolutional network (FCN) with only seven layers, has difficulty in extracting high-level features, and the heavy pyramid scene parsing network (PSPNet) with complicated calculations is not practical in real-time CD, let alone on-orbit CD. So, in view of the problems above, we propose a compact attention mechanism cloud detection network (AM-CDN) based on the modified FCN to refine and fuse the multi-scale features for on-orbit CD. Specifically, taking the FCN as the baseline, our model increases the numbers of hidden layers and adds the residual connections between the input and output to eliminate the network degradation and extract the advanced context feature maps effectively. To expand the receptive field without losing the spatial information, the ordinary convolutions in FCN are replaced by the dilated convolution in AM-CDN. And inspired by the selective kernels of human vision, we introduce the convolutional attention mechanism (AM) into the encoder to adaptively adjust the receptive field to highlight the key texture features. According to experimental results using Landsat-8 infrared RSIs, the accuracy of the proposed CD method is 95.31%, which is 10.17% higher than that of FCN. And the calculation complexity of AM-CDN is only 7.63% of that of PSPNet.

INDEX TERMS Cloud detection (CD), semantic segmentation (SS), fully convolutional network (FCN), attention mechanism (AM), remote sensing images (RSIs).

I. INTRODUCTION

In recent years, remote sensing images (RSIs) have been used in many fields, including geological mapping, environmental monitoring, disaster relief, small target identification and real-time battlefield monitoring [1]–[3]. However, most of the earth surface is covered by the cloud every day, which leads to the obscure of critical information in RSIs. Moreover, the harsh weather containing heavy rain, hurricanes and tornadoes, and the changes of local or global climate can be tracked by detecting the cloud distribution. Therefore, cloud detection (CD) is crucial in weather forecast and microclimate research.

Generally, even adopting the reference defined by the World Meteorological Organization, the data of cloud coverage and types observed by empirical observers are not objective [4]. In rain or fog, visible RSIs are rough, and they are unavailable at night due to the dependence on sunlight. Due to the strong transmission of thermal infrared (TIR) in the atmosphere, the TIR camera can work through smoke,
fog, rain, and snow in the day and night. The short-wave infrared (SWIR) RSIs contain target details extracted from deep shadows, and the contrast of TIR RSIs is lower than its. Therefore, SWIR RSIs is an important supplement to TIR RSIs in target recognition. Taking the above considerations into account, the thermal infrared RSIs [5] in two channels and short-wave infrared RSIs of single channel are combined for automatic, all-day, and all-weather CD in this paper. Compared with visible images, the number of the existing infrared RSIs is extremely small, and some high-frequency components of infrared RSIs are easily missed during the sampling or scanning, which makes it more challenging to obtain the detection result effectively.

In stable atmosphere, although traditional threshold methods [6], time differential methods [7] and classic machine learning methods [8], are simple and fast in calculation, they are susceptible to the background and heavily depend on the camera specifications and the parameter settings. In addition, under circumstance of the low temperature of surface or the low cloud height, the contrast between objectives and background in infrared images is so weak that it is necessary to extract more advanced texture information for accurate detection.

With the rapid development of GPU, artificial intelligent methods based on deep learning particularly convolution neural networks (CNNs) are extensively used to learn the multi-level features in RSIs to improve the accuracy and robustness of CD algorithms even a small part of dataset is mislabeled. In particular, when targets are mixed with the heavy atmospheric aerosols with dust or smoke, artificial intelligent methods are more effective than traditional methods.

Instead of using the finite context information, semantic segmentation (SS) uses the spatial and spectral information from the whole image of all channels to segment the cloud at pixel-level. By adding differentiable interpolation layers in CNNs, the first end-to-end SS, namely the fully convolutional network (FCN) [9], utilize input information efficiently to output classification maps even for small datasets [10]. Deep learning semantic segmentation networks are mostly based on FCN model, and the details will be discussed in Related Work.

In the cognitive science, due to the limited capacity of brain information processing, humans selectively focus on a portion of the available information and ignore useless visible information. Originating from the human visual system, the soft attention mechanism (AM) [11] is used to locate the most salient features of the targets and eliminate the redundancy of computer vision tasks. The structure of soft attention mechanism is divided into three: spatial attention, channel attention, and mixed module. In spatial attention module, the spatial information in images is transformed to generate masks and then scored to extract key information. However, the spatial attention module utilizes image features in each channel equally, so it ignores the channel information and is confined to the stage of original image features extracting. In the channel attention module, signals on each channel are given a weight to represent the relevance between the channel and the key information. The higher the weight, the higher the correlation. For example, squeeze-and-congestion networks (SENet) [12] adaptively recalibrate the characteristic responses of the channels by explicitly modeling the interdependencies between the channels. And SENet won championship in the ImageNet Large Scale Visual Recognition Challenge 2017 (ILSVRC2017). In the mixed domain, both channel and spatial attention were scored to combine these two ideas. For example, the convolutional block attention module (CBAM) [13] generates attention maps in both channel and spatial dimensions to enhance the constraints on input features. However, more spatial information of different scales needs to be captured to enrich the feature map. SKNet [14] is an enhanced version of SENet. In SKNet, different convolution kernel weights are used for multi-scale images to dynamically extract multi-scale spatial information. Aiming at solving the existing problems, we propose a submodule of Modified ResNet, namely SKConv, which contains three convolution kernels. The cloud detection for remote sensing images is greatly improved by SKConv module.

This paper proposes a high-precision and compact attention mechanism cloud detection (AM-CDN) based on the modified fully convolutional network using infrared remote sensing images for all-day real-time CD.

The main contributions of this study as follows:

1. The preprocessing of datasets. Aiming for all-day, and all-weather CD, the dataset we collected consists of the TIR and SWIR images. TIR RSIs can also record targets information during the night, and the SWIR RSIs contain target details extracted from deep shadows. To enrich the datasets, the digital numbers of TIR and SWIR RSIs are converted into the top of atmosphere (TOA) brightness temperature and the TOA reflectance, respectively.

2. The improved model architecture. Taking the FCN as the baseline, we propose a compact network for CD. Frist, our model increases the numbers of hidden layers and adds the residual connections between the input and output to eliminate the network degradation and extract the advanced context feature maps effectively. Then, the dilated convolutions are properly utilized to expand the receptive field without losing the spatial information. Third, and more important, SK-Conv containing channel AM is also added in the encoder to locate the key targets by comparing the importance of features to enhance or suppress different features weights. The comparisons and analyzations of the advantages and disadvantages between the new and old architectures for cloud identification are described in detail in Section III.

3. The setting of loss function. To achieve better results, we choose the optimized multiplier factors to fuse both Binary Cross Entropy (BCE) and Dice Loss to solve the class imbalance problem to improve the stability of the model during training.

Owe to the efficient extraction of the high-level texture and the low-level features maps, containing size and position...
information, AM-CDN, with low cost of the floating point of operations (FLOPs), is more sensitive to the subtle changes in infrared features of the cloud than traditional methods including Kmeans and OTSU, and classical SS, containing FCN, U-Net, SegNet, DeepLabV3+, PSPNet, and the proposed method has been validated on embed system with the master chip as XCZUEG. The overview of cloud detection method, and the demonstration of porting the proposed model to hardware platforms are shown in figure 1.

The remainder of this paper is organized as follows: Section II presents an overview of the related work. Section III introduces preprocessing of datasets, and the architecture of the proposed neural network. Section IV describes the setting of loss function, the evaluation metrics, and the performance of different neural networks. In Section V, we discuss some of the problems that arose in the experiment and suggest future research directions. The final section concludes the entire article.

![Figure 1](image-url)

**FIGURE 1.** The overview of cloud detection method, and the demonstration of porting the proposed model to hardware platforms. To facilitate hardware porting, we converted the PyTorch model to the tensorflow model by using a cross-framework model.

II. RELATED WORK

Different from CNN, arbitrary size of RSIs with high spatial and spectral resolution can be effectively segmented by the FCN [15]. FCN is completely a data-driven algorithm instead of relying on the experience. But the segmentation boundary is fuzzy due to the partial loss of spatial information during up-sampling. Thus, lots of derivatives of FCN [16], [17] are proposed for the processing of RSIs, including SegNet, U-Net, DeepLab, and PSPNet proposed in chronological order.

SegNet [18] uses the first 13 layers convolutional network of VGG16 in the encoder, and the outputs in the decoder are classified by the soft-max to independently generate category probability for each pixel. Symmetric semantic segmentation model, namely U-Net [19], is mainly composed of a contraction path and a symmetric expansion path to obtain context information. DeepLabv1 [20] proposes a new extended convolution semantic segmentation model, which can increase the receptive field without increasing the number. DeepLabv3+ [21] adds encoder-decoder module to recover the output to the same size of original pixel information, and uses depthwise separable convolution to improve the speed. To introduce more contextual information, the pyramid pooling module are used to output different sizes of feature maps through four parallel pooling in pyramid scene parsing network (PSPNet) [22]. Based on FCN and PSPNet, by adding four scale filters during the pooling, the low-level spatial information and the high-level semantic information are output by four cascading convolution layers, and then the multi-level features are integrated to a global feature map in MF-CNN [16]. However, there is a lot of redundant information in the different feature maps extracted from semantic segmentation, resulting in the key features being submerged in them.

In recent years, the combination of semantic segmentation and attention mechanism has become a hot topic. Wu et al. [23] add the spatial AM in the decoder to select and reconstruct color feature maps carefully. But the boundary features are ignored without considering the information between channels, so the accuracy is only improved by 0.66%. To extract more features, the AM are introduced in the U-Net between copy and crop connection, and Gabor filters are also utilized for images segmentation [24]. By using a DeepLabv3+ on spectral image and a random forest (RF) classifier on hand-crafted features, Du et al. [25] propose a method to obtains two initial probabilistic labeling predictions. However, the above methods are composed of multiple modules, the results are affected by the outputs of these independent modules.

To reduce the interference of invalid features during sharing feature maps, more attention modes are introduced for CD. Mou et al. [26] plug the spatial and channel relation module to learn about global relationships between any two spatial positions in the existing fully convolutional network (FCN) framework. To refine the multi-feature maps captured by the network Li et al. [27] take the ResNet-34 as the encoder and substitute the plain skip connections within U-Net into attention blocks at multiple stages. Focusing on removing the cloud shadows, Guo et al. [28] propose a novel FCN based on the multi-scale aggregation and channel AM. Taking SegNet as the baseline and ResNet50 as the backbone, SCAttNet [29] adds the combination module of channel and spatial attention to the last layer of backbone. The models based on PSPNet [30] embedding with the channel [31] and the spatial AM [32] are applicable even for thumbnail CD datasets [33]. But these methods are limited by the large number of parameters and the high consumption of time and storage.

III. DATASET AND ARCHITECTURE

A. EXPERIMENTAL DATASET

The datasets we used is the public Landsat-8 Biome Cloud Validation Masks [4] provided by the United States Geological Survey (USGS), containing different types of clouds and eight ground object scenes. We select RSIs with 30m resolution and bands of $2.11 \sim 2.29 \mu m$, $10 \sim 11.19 \mu m$, and $11.5 \sim 12.51 \mu m$. Since the black filled pixels are meaningless, the original images are rotated and cropped.
to 4480 × 4480. The surface thermal radiation intensity is roughly represented by digital numbers (DN) in infrared images. As the surface thermal radiation increasing, the DN is higher. Intensity value with physical significance, such as top of atmosphere (TOA) brightness temperature, is the total amount of energy received by the sensor, including both surface and atmospheric radiation.

\[
\text{TOA}_A = M_L \cdot \text{DN} + A_L \\
\text{TOA}_{BT} = \frac{K_2}{\ln(\frac{K_1}{\text{TOA}_A} + 1)} \\
\text{TOA}_R = K \cdot \text{DN} + c
\]

where \(M_L = 3.342 \times 10^{-4}\) and \(A_L = 0.1\) are the multiplicative and additive resetting factor of the thermal infrared images respectively. We choose the thermal conversion constant \(K_1 = 774.89, K_2 = 1321.08\) in 10 ~ 11.19\(\mu\)m, and \(K_1 = 480.89, K_2 = 1201.14\) in 11.5 ~ 12.51\(\mu\)m. In short-wave infrared images, \(k = 2.0 \times 10^{-5}\) and \(c = -0.1\) are the gain and offset.

To increase spectral information, GDAL software is used to fuse infrared three-channels images, as shown in figure 2 (d). With the increase of the input, the size of the deep CNNs increases dramatically. Limited by GPU memory, the RSIs with large scale cannot be processed at one time. Although image integrity is compromised, images should be compressed or cropped during datasets making. The numbers of images in Landsat-8 Biome Cloud Validation Masks we used is 96, and the size of each image is 4480 × 4480. So, these images are divided into 224 × 224 blocks. Then, by random sampling, 38400 blocks are divided into training, verification and test set in a ratio of 8:1:1. The training datasets include 30720 images, and the verification and test datasets include 3840 images, respectively.

**B. AM-CDN ARCHITECTURE**

The overall architecture of AM-CDN is shown in figure 3. Taking FCN as baseline, we increase the number of hidden layers with residual connections and introduce the convolutional AM, meanwhile replacing the standard convolution with dilated convolution [34]. By increasing the number of convolution layers with the residual connections in the FCN’s encoder, our model extract more advanced context feature maps from the large perceptual domain without network degradation.

In the Modified ResNet, the feature maps are convoluted by SK-Conv’s different cores to extract and fuse the multi-level features. Then, the multi-scale information is refined by the Modified FCN. Combining the channel AM in SK-Conv, the middle outputs split by three convolution kernels sizes are aggregated from multiple paths based on selection weights to dynamically select the size of receptive field for different input scales and extract global and local context information. In section C, D, and E, the key modules including Modified FCN, Modified ResNet, and SK-Conv are analyzed in detail.

**C. MODIFIED FCN MODULE**

In ResNet, the input is convoluted to 1/8, 1/16 and 1/32 of the original input. In Modified FCN, F1, F2, and F3 are output by Block2, Block3 and Block4, respectively. After the pooling, F1’ and F2’ are preserved. Then, to enlarge the receptive field, F4 and F5 are upsamples by dilated deconvolution to be restored to the same size as F2’ and F1’. Finally, to increase the recovery information, the heat map F4 are added with F2’ to iterate forward, and F5 combining with F1’ are deconvolved to upsample the prediction to the same resolution as the input.

Since the pooling result in the resolution reduction, the dilated convolution is used to replace the standard convolution and pooling operation in the FCN. Without the spatial dimension losing, our model increases the corresponding receptive field index to improve the detection precision.

**D. MODIFIED RESNET MODULE**

With the layers of deep CNNs increasing, more and more features can be extracted from the input, but accompanied with the problem of gradient dispersion or explosion during training. To avoid the above problems, in the Modified ResNet module, we add residual connections [35] in deep CNNs. Therefore, the features are learned from residual mapping instead of the identity mapping to limit the loss during the information transmission. By passing the input directly to the output, the integrity of information is protected well.

As shown in figure 4, the feature maps (F1, F2 and F3) outputted by Modified ResNet, are the input of Modified FCN module (1/8, 1/16 and 1/32 of the initial images). To succinctly express the diagram, the solid and dashed lines represent skip connections, namely residual connections. Unlike solid lines that connect inputs and outputs of the same size, dashed lines are used to connect the increased dimension input-output maps. And \(\times n\) represents that the module is reused \(n\) times. Excluding the output, the Modified ResNet contains a total of 35 layers.
E. SUBMODULE SKCONV

When we look at the objects from different distances, the size of the neuron’s receptive fields in the visual cortex will be adjusted according to the stimulus [14]. Corresponding to the CNNs, instead of being determined by a specific task or a model, the convolution kernel size of the receiving domain can be selected adaptively by the SK-Conv module according to multiple scales of the input. SK-Conv, including Split, Fuse and Selection, is added to the deep residual network module to highlight the importance of cloud texture among various features.

To increase the inter-channel information, the convolution kernel is $3 \times 3$ with 1, 2, and 3 dilations rate respectively, so the receptive field sizes are expanded to $3 \times 3$, $5 \times 5$ and $7 \times 7$. As shown in figure 5, first, the original input is split by three dilated convolutions to generate three feature maps. Then, added by U1, U2 and U3, U is pooled by the global average pooling function to output the $C \times 1 \times 1$ vector $S$. Through a fully connected layer function, $d \times 1$ vector $Z$ is output, as shown in formula (4).

$$z = \delta(B(WS)) \quad (4)$$

where $\delta$ represents the activation function ReLU, $B$ represents batch normalization, $W$ represents the $d \times c$ vector, and $d = \max(C/r, L)$, $C$ represents input channels number, $r = 16$ and $L = 32$ represents the reduced ratio and lower bound.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad (5)$$

where $e$ is the base of the natural logarithm, $K = 3, j = 1, 2, 3, \sigma(z)_1, \sigma(z)_2, \text{ and } \sigma(z)_3$ are multiplied by U1, U2 and
U3, respectively to generate A1, A2 and A3. Finally, the information of A is added by three softmax (A1, A2 and A3) from different branch fused. As shown in formula (5), the sum of the $\sigma(z_j)$ is equal to 1, so setting the weights of the feature maps in the branches enable our model automatically choose the appropriate convolution kernel size.

IV. EXPERIMENTAL SETTING AND RESULTS
A. LOSS FUNCTION AND EVALUATION METRICS

By calculating the intersection of segmentation mask and ground truth (GT), Binary Cross Entropy (BCE) [36] is the most commonly used loss function to evaluate the binary classification model. Dice Loss [36] is the intersection of mask and GT divided by the total pixel. When calculating the intersection, pixels of a category are considered as a whole, so Dice Loss is not affected by a large number of mainstream pixels.

To achieve better results, we use both BCE and Dice Loss to solve the class imbalance problem to improve the stability of the model during training, as shown in formula (6-9).

$$L_{Bce} = -\sum_i T_i \ln (P_i) + (1 - T_i) \ln (1 - P_i)$$

$$L_{Dice} = 1 - \frac{2|P \cap T|}{|P| + |T|}$$

$$L = W_{Bce}L_{Bce} + W_{Dice}L_{Dice}$$

$$W_{Bce} + W_{Dice} = 1$$

where $T$ is the prediction, and $P$ is the GT. We assign the multiplier factors $W_{Bce} = 0.95$ and $W_{Dice} = 0.05$ for $L_{Bce}$ and $L_{Dice}$, respectively.

In most cases, a certain index is not comprehensive, so the metrics including the time of training and testing. As shown in formula (10-12), Pixel Accuracy (Pixel Acc.), Mean Accuracy (Mean Acc.), Mean IoU and GFLOPs [23] are used for evaluation metrics to fully evaluate the precision and complexity of the proposed method, where $G = 10^3$.

$$Pixel\_Accuracy = \frac{T}{M} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Mean\_Accuracy = \frac{1}{2} \left( \frac{TP}{TP + FN} + \frac{TN}{FP + TN} \right)$$

TABLE 1. Confusion matrix.

| Ground truth | Predicted class |
|--------------|-----------------|
| Cloud        | TP              | FN              |
| Non-Cloud    | FP              | TN              |

TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.
TABLE 2. The experimental results comparison of different models.

| Models      | CONV   | T_train($) | T_infer($) | Pixel Acc. (%) | Mean Acc. (%) | Mean IoU (%) | GFLOPs |
|-------------|--------|------------|------------|----------------|---------------|--------------|--------|
| Kmeans      | -      | -          | -          | 68.31          | 70.52         | 56.29        | -      |
| OTSU        | -      | -          | -          | 85.83          | 83.33         | 70.29        | -      |
| FCN         | Standard Conv | 538.80    | 4.39       | 85.14          | 90.22         | 69.19        | 39.07  |
| U-Net       | Standard Conv | 979.32    | 6.08       | 87.67          | 92.27         | 71.35        | 85.68  |
| SegNet      | Standard Conv | 876.61    | 5.35       | 84.15          | 85.49         | 64.09        | 51.30  |
| DeepLabV3+  | Depthwise Separable Conv | 2880.45  | 8.22       | 91.75          | 94.23         | 75.87        | 37.40  |
| PSPNet      | Dilated Conv | 4921.02   | 12.37      | 95.29          | 95.18         | 79.94        | 97.75  |
| ResNet-FCN  | Dilated Conv | 534.90    | 3.39       | 92.72          | 94.38         | 75.40        | 7.45   |
| SK-CDN      | Dilated Conv | 673.22    | 4.42       | 94.01          | 93.66         | 72.51        | 7.49   |
| AM-CDN      | Dilated Conv | 306.10    | 4.01       | 95.31          | 95.26         | 79.43        | 7.46   |

The bold font is the optimal and suboptimal results; \( T_{\text{train}} \) is training time of 100 epoch divided by 100; \( T_{\text{infer}} \) is the inference time to generate 4480×4480 images.

\[
\text{Mean IoU} = \frac{1}{2} \left( \frac{TP}{TP + FN + FP} + \frac{TN}{FP + FN + TN} \right)
\]

\((12)\)

As shown in TABLE 1, added by the \( TP \) and the \( TN \), \( T \) is the number of the pixels classified correctly. \( M \) is the total number of pixels in the image (excluding filling pixels). In most cases, the high accuracy is accompanied by high FLOPs, and FLOPs is positively correlated with time. Therefore, using a certain metrics is not comprehensive to evaluate the model. With the high accuracy maintained, the lower the complexity of CD method, the better the effectiveness.

B. RESULTS

To avoid the overfitting caused by too many iterations, 200 epochs are selected during training. Due to the limitation of the computing power, the batch size is set to 16. The initial learning rate was set to 0.001 and optimized by the Adam optimizer.

With only seven convolution layers, the last rough outputs obtained from the down sampling are restored to the same size of the input to calculate the softmax loss pixel by pixel during the classification of FCN. To be more insensitive to the details, the encoder of FCN is replaced by the modified ResNet with the following advantages: (a) The learning capacity of encoder is improved by increasing layers of our model. (b) Meanwhile, more skip connections are added to combine background semantic information and solve the problem of loss non-convergence. (c) The data volume of the model is reduced for rapid convergence. (d) The dilated convolutions are added to expand the size of receptive field to enhance context information during extracting the feature maps. (e) The SK-Conv module, containing three convolution cores, is introduced to integrate the multi-receptive field information effectively during outputting the feature maps.

We compare the performance of segmentation networks with encoder-decoder, as shown in figure 6. AM-CDN are proved the superiority especially for Urban, Shrubland, Grass/Crops and Water. Our model has a significant improvement of Accuracy (95.87%), but a low Mean IOU (49.91%) in Snow/Ice due to the spectral similarity. The possible reason is that Accuracy is the proportion of the true cloud pixels among all judged as cloud pixels, while Mean IOU is the intersection over union ratio of detection result and GT. In conclusion, our model has a better performance in returning the attribute for relevant instances.

3SK-Conv with convolutional AM added at the 9th, 17th and 29th layers of ResNet-34, and Modified ResNet with automatic selection of convolution added only at the 9th layer are tested to design a more effective module for CD, as shown in figure 7. We quantitatively test two options, as shown in TABLE 2. We speculate that 3SK-Conv as submodule of SK-CDN with more parameters even compact the necessary feature maps in spatial and channel dimensions. So, we should improve the accuracy scientifically instead of increasing complexity.

As showed in TABLE 2, compared with the original FCN, the layers of ResNet-FCN are appropriately increased, and the accuracy are improved greatly, which demonstrates the superiority of Modified ResNet module. Different from ResNet-FCN, SKConv modules are added in SK-CDN, and the experimental results shows the effectiveness of attention mechanism applied in remote sensing image processing. For rational layout of attention mechanism, we also compared the results of SK-CDN and AM-CDN.

By comprehensively evaluating, our model (AM-CDN) is superior to traditional algorithms including Kmeans.
and OTSU, classical SS containing FCN, U-Net, SegNet, DeepLabV3+, PSPNet, and modified model (ResNet-FCN and SK-CDN). Modified by FCN, the Accuracy of our model reaches 95.31%, which is comparable to that of PSPNet. And the FLOPs computation required by AM-CDN is only 7.46G, which is 7.63% of that of PSPNet.

By quantitatively comparing the precision, computational speed, and complexity of the models, the positive effect of the Modified FCN, Modified ResNet and especially SK-Conv modules are verified on CD. The memory capacity needed by AM-CDN is only 85.3M to be easily ported to the hardware platforms, so that our model has been verified on the embedded system.

V. DISCUSSION AND FUTURE WORK

Our method has the following advantages. First, the encoder-decoder structure is helpful for extracting high-level features, and the multi-scale information can be fully utilized. The encoder is used to extract multiple features, and the decoder is used to explain abstract information. Then, with the network layers increasing, the residual connections are added between the input and output to prevent the loss of important semantic information. Next, without adding extra parameters or reducing the spatial resolution, the ordinary convolutions are replaced by the dilated convolutions to expand the corresponding receptive field index in our model. Moreover, the convolutional AM is used to filter the spatial and channel information respectively to effectively enhance the key features during the network training.

Although our model has a strong ability to return relevant instances on many scenes, the intersection between detection result and GT is lower than their union value resulting in the dim boundary detection.

The artificial intelligent methods can extract the deeper features of the data through continuous convolution, so it has high robustness. It worth nothing that although the artificial intelligent methods present more accurate and efficient than the conventional CD algorithms, the overfitting of networks, misjudgment between smooth texture terrain and clouds, imprecise recognition for thin clouds, and the losing details caused by compression or cropping are still the significant challenges needed to be improved further.

VI. CONCLUSION

In this paper, the AM-CDN is proposed to extract multi-scale feature maps combined by the residual connections between input and output layers. And the dilated convolutions are used to expand the feature selection receptive field. Moreover, the SK-Conv combined with three channels information is conducive to adapting reception field and selecting key information from a large number of redundant feature maps. Instead of adjusting parameters manually, our model with high-level accuracy and strong generalization has the capacity of all-day and all-weather for cloud detection. The proposed method is suitable for n-channel images in many kinds of scenes. It can be easily transplanted to the hardware due to the low coat of time and calculation. The further work will focus on filtering the valid information from large scale images with dim texture.

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] T. Shi, Q. Xu, Z. Zou, and Z. Shi, “Automatic raft labeling for remote sensing images via dual-scale homogeneous convolutional neural network,” Remote Sens. Environ., vol. 10, no. 7, p. 1130, Jul. 2018.
[3] Shi, Qi, Liu, Ni, and Zhang, “Urban land use and land cover classification using multisource remote sensing images and social media data,” Remote Sens., vol. 11, no. 22, p. 2719, Nov. 2019.
[4] Y. Zhang, W. Rossow, and A. Lacas, “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., vol. 109, pp. 1121–1125, Oct. 2004.
[5] S. Foga, P. L. Scaramuzza, S. Guo, Z. Zhu, R. D. Dilley, Jr., T. Beckmann, G. L. Schmidt, J. L. Dwyer, M. J. Hughes, and B. Laue, “Cloud detection algorithm comparison and validation for operational Landsat data products,” Remote Sens. Environ., vol. 194, pp. 379–390, Jun. 2017.
[6] Z. Li, H. Shen, H. Li, G. Xia, and L. Zhang, “Multi-feature combined cloud and shadow detection in GaoFen-1 wide field of view imagery,” Remote Sens. Environ., vol. 191, pp. 342–358, Mar. 2017.
[7] 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.
[8] Z. Yue, X. Fengying, and Z. Jiang, “A cloud detection method for Landsat 8 images based on PCANet,” Remote Sensing Pattern Recogniton., vol. 10, no. 6, p. 877, Jun. 2018.
[9] F. Li, Z. Long, P. He, P. Feng, X. Guo, X. Ren, B. Wei, M. Zhao, and B. Tang, “Fully convolutional pyramidal networks for semantic segmentation,” IEEE Access, vol. 8, pp. 229132–229140, 2020.
[10] W. Li, Z. Zou, and Z. Shi, “Deep matting for cloud detection in remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 12, pp. 8490–8502, Dec. 2020.
[11] L. Li, M. Xu, H. Liu, Y. Li, X. Wang, L. Jiang, Z. Wang, X. Fan, and N. Wang, “A large-scale database and a CNN model for attention-based glaucoma detection,” IEEE Trans. Med. Imag., vol. 39, no. 2, pp. 413–424, Feb. 2020.
[12] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogniton., Jun. 2018, pp. 7132–7141.
[13] S. Woo, J. Park, J. Lee, and I. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), vol. 11211, Oct. 2018.
[14] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in Proc. Comput. Vis. Pattern Recogniton. (CVPR), Mar. 2019.
[15] J. An, J. Wu, and J. Ma, “Automatic cloud segmentation based on fused fully convolutional neural networks,” J. Signal Process., pp. 530–1003, Jul. 2019, doi: 10.1007/978-3-030-26763-6_50.
[16] 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.
[17] L. Li, X. Li, L. Jiang, X. Su, and F. Chen, “A review on deep learning techniques for cloud detection methodologies and challenges,” Signal, Image Video Process., vol. 15, no. 7, pp. 1527–1535, Oct. 2021.
[18] 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.
[19] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Assist. Intervent., 2015, pp. 234–241.
[20] L. C. Chen et al., “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” Comput. Sci., vol. 4, pp. 357–361, 2014.
[21] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, Encoder-Decoder With Atrous Separable Convolution for Semantic Image Segmentation. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-030-01234-2_40.
[22] 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. 6230–6239.

[23] Z. Wu, J. Li, Y. Wang, Z. Hu, and M. Molnár, “Self-attentive generative adversarial network for cloud detection in high resolution remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 17, no. 10, pp. 1792–1796, Oct. 2020.

[24] J. Zhang, Q. Zhou, J. Wu, Y. Wang, and H. Wang, “A cloud detection method using convolutional neural network based on Gabor transform and attention mechanism with dark channel SubNet for remote sensing image,” Remote Sens., vol. 12, no. 19, p. 3261, 2020.

[25] S. Du, S. Du, B. Liu, and X. Zhang, “Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images,” Int. J. Digit. Earth, vol. 14, no. 3, pp. 357–378, Mar. 2021.

[26] L. Mou, Y. Hua, and X. X. Zhu, “Relation matters: Relational contextual-aware fully convolutional network for semantic segmentation of high-resolution aerial images,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[27] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, “Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images,” IEEE Geosci. Remote Sens. Lett., early access, Mar. 15, 2021, doi: 10.1109/LGRS.2021.3063381.

[28] Y. Guo, X. Cao, B. Liu, and M. Gao, “Cloud detection for satellite imagery using attention-based U-Net convolutional neural network,” Symmetry, vol. 12, no. 6, p. 1056, Jun. 2020.

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

[30] S. Ji, P. Dui, M. Lu, and Y. Zhang, “Simultaneous cloud detection and removal from bitemporal remote sensing images using cascade convolutional neural networks,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 1, pp. 732–748, Jan. 2021.

[31] Z. Yao, J. Jia, and Y. Qian, “MCNet: Multi-scale feature extraction and content-aware reassembly cloud detection model for remote sensing images,” Symmetry, vol. 13, no. 1, p. 28, Dec. 2020.

[32] J. Yu, Y. Li, X. Zheng, Y. Zhong, and P. He, “An effective cloud detection method for Gaofen-5 images via deep learning,” Remote Sens., vol. 12, no. 13, p. 2106, Jul. 2020.

[33] J. Guo, J. Yang, H. Yue, H. Tan, C. Hou, and K. Li, “CDnetV2: CNN-based cloud detection for remote sensing imagery with cloud-snow coexistence,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 1, pp. 700–713, Jan. 2021.

[34] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in Proc. Int. Conf. Learn. Represent. (ICLR), Apr. 2016.

[35] 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.

[36] J. Zhang, X. Li, L. Li, P. Sun, X. Su, T. Hu, and F. Chen, “Lightweight U-Net for cloud detection of visible and thermal infrared remote sensing images,” Opt. Quantum Electron., vol. 52, no. 9, p. 397, Sep. 2020.

XIAOYAN LI received the B.S. degree in mechanism design, manufacturing, and automation from Northwest A&F University, Xi’an, China, in 2016. He is currently pursuing the Ph.D. degree in electronic circuit and systems with the Shanghai Institute of Technical Physics of the Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China. His current research interests include on-orbit accurate navigation and geometric calibration of remote sensing satellites.

XIN LIU received the B.S. degree from the Wuhan University of Science and Technology, Wuhan, China, in 2017. He is currently pursuing the Ph.D. degree in electronic circuit and systems with the Shanghai Institute of Technical Physics of the Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China. His current research interests include image processing, machine learning, and high-performance computing.

WENWEN HUANG received the B.S. degree in electronic information science and technology from Jilin University, Changchun, China, in 2020. She is currently pursuing the M.D. degree in electrical and information engineering with the Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China. Her current research interests include deep learning and computer vision.

ZHUOYUE HU received the B.S. degree in information display and opto-electronic technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2016. She is currently pursuing the Ph.D. degree in physical electronics with the Shanghai Institute of Technical Physics of the Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China. Her current research interest includes radiometric calibration of remote sensing satellites.

FANSHENG CHEN received the B.S. degree in optoelectronic information engineering from Shandong University, Jinan, China, in 2002, and the Ph.D. degree in physical electronics from the Shanghai Institute of Technical Physics of the Chinese Academy of Sciences, Shanghai, China, in 2007. Since 2013, he has been a Professor with the Shanghai Institute of Technical Physics of the Chinese Academy of Sciences. He has been committed to the research and development of the space infrared staring detection instruments, the high spatial and temporal resolution photodetector payloads, and the application of infrared multispectral information acquisition technology in artificial intelligence, target recognition, and other relative aspects. His research interests include the design of spatial high resolution remote sensing and detection payloads, high-speed and low noise information acquisition technology, and infrared dim small target detection technology.

LIYUAN LI received the B.S. degree in optoelectronics information science and engineering from the Dalian University of Technology, Dalian, China, in 2018. She is currently pursuing the Ph.D. degree in physical electronics with the Shanghai Institute of Technical Physics of the Chinese Academy of Sciences, Shanghai, China. Her current research interest includes dim and small targets detection of IR through machine learning.