CD-CTFM: A Lightweight CNN-Transformer Network for Remote Sensing Cloud Detection Fusing Multiscale Features

Wenxuan Ge, Xubing Yang, Rui Jiang, Wei Shao, and Li Zhang

***Abstract*—Clouds in remote sensing images inevitably affect information extraction, which hinders the following analysis of satellite images. Hence, cloud detection is a necessary preprocessing procedure. However, most existing methods have numerous calculations and parameters. In this article, a lightweight convolutional neural network (CNN)-Transformer network, CD-CTFM, is proposed to solve the problem, which is based on encoder–decoder architecture and incorporates the attention mechanism. In the encoder part, we utilize a lightweight network combing CNN and Transformer as backbone, which is conducive to extracting local and global features simultaneously. The backbone of CD-CTFM also incorporates attention gate based on dark channel extraction module. Moreover, a lightweight feature pyramid module is designed to fuse multiscale features with contextual information. In the decoder part, a lightweight channel-spatial attention module is integrated into each skip connection between encoder and decoder to extract low-level features while suppressing irrelevant information without introducing many parameters. Finally, the proposed model is evaluated on two cloud datasets, 38-Cloud and MODIS. The results demonstrate that CD-CTFM achieves comparable accuracy as the state-of-art methods and outperforms in terms of efficiency.**

***Index Terms*—Attention mechanism, cloud detection, deep learning, lightweight network, vision transformer (ViT).**

I. Introduction

SINCE the 1960s, with the rapid development of satellite remote sensing technology, the United States and China have launched a variety of satellites, such as LandSat series and GaoFen series [1], [2], [3]. Remote sensing images collected from satellites are widely used in various fields, including land cover mapping, weather forecasting, marine pollution monitoring, and other fields [4], [5]. However, optical remote sensing images are inevitably affected by cloud appearing on them, resulting in attenuation or loss of image information. Therefore, to improve the utilization of images with cloud, it is necessary to detect cloud before analyzing images.

Over the years, researchers have proposed a multitude of approaches from different perspectives. Traditional cloud detection methods can be broadly divided into two categories: threshold-based and machine-learning-based methods [6]. Although they are lightweight, their performance is not so satisfactory due to limitations of algorithms. Threshold-based methods are implemented using a threshold of different image parameters [7]. Among threshold-based algorithms, the Function of mask (Fmask) is a widely used cloud detection approach [8]. In Fmask, several rules are designed by expert knowledge of spectral characteristics to distinguish clouds from noncloud. Fmask identifies cloud by calculating cloud temperature probability, which is based on the assumption that cloud and corresponding shadows share similar shape and shadows follow the project geometry. However, threshold-based methods have poor universality in that thresholds vary as per the location. Machine-learning-based methods, such as decision tree [9], support vector machine [10], and Bayesian classification [11], identify cloud by learning from training data, which improves the performance of cloud detection. In [9], a classification tree were used to build the decision tree, which was designed based on empirical studies and simulations. With a great deal of repeating scenes coming from the same area, cloud pixels can be replaced by real surface types. According to Ishida et al. [10], discriminant analysis was incorporated into support vector machine (SVM), which made it possible to subjectively determine the definition of typical cloudy and clear sky. Moreover, if incorrect results occurred, feature space used for cloud detection would be improved to adjust the classifier. The Bayesian method calculated a probability of cloud for each image pixel, based on the satellite observations and prior probability [11]. However, machine-learning-based methods are highly affected by the handcraft features, which is highly dependent on expert knowledge and experience.

With the advance in deep learning, methods based on convolutional neural networks (CNNs) have achieved great success in the field of computer vision (CV), overwhelmingly surpassing

Manuscript received 31 October 2023; revised 2 January 2024; accepted 27 January 2024. Date of publication 5 February 2024; date of current version 15 February 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61802193, in part by the Innovation and Entrepreneurship Training Program for College Students in Jiangsu Province under Grant 202210298120Y, and in part by the Shenzhen Science and Technology Program under Grant JCYJ20220530172403008. *(Corresponding authors: Wei Shao; Li Zhang.)*

Wenxuan Ge, Xubing Yang, and Li Zhang are with the College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China (e-mail: gwx@njfu.edu.cn; xbyang@njfu.edu.cn; lizhang@njfu.edu.cn).

Rui Jiang is with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: j_ray@njupt.edu.cn).

Wei Shao is with the Shenzhen Research Institute, Nanjing University of Aeronautics and Astronautics, Shenzhen 518038, China (e-mail: shaowei20022005@nuaa.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3361933

traditional algorithms. Scholars have applied deep learning to optical satellite imagery, including cloud detection. Li et al. [12] proposed graph-feature-enhanced selective assignment network to mine meaningful features and effectively fuse multisource remote sensing data. Focusing on subtle traits extraction, Zhang et al. [13] designed a spatial-logical aggregation network with morphological transformation for tree species classification. Hang et al. [14] presented a multiscale progressive segmentation network to obtain satisfactory segmentation results on small or large objects. In addtion, he designed an edge-enhanced multiscale convolutional network to identify oceanic eddy [15]. Francis et al. [16] proposed CloudFCN, which is based on fully convolutional networks, to extract multiscale spectral features of remote sensing images. Luotamo et al. [17] designed an architecture of two cascaded CNN model components successively processing undersampled and full-resolution images. Inspired by cognitive psychology and neuroscience, scholars have applied visual attention mechanism to the field of CV. Visual attention mechanism selectively concentrates the significant parts within remote sensing images and obtains discriminative features to improve accuracy, for which some cloud detection approaches incorporate the attention mechanism to further improve accuracy. In LSCNet, large kernel sparse ConvNet was weighted by multifrequency attention for remote sensing scene understanding [18]. Guo et al. [19] incorporated attention mechanism into their U-Net architecture-based cloud detection model, CloudAttU, to achieve the fusion of high-level and low-level features. CAA-UNet was a cloud detection network based on asymmetric encoder–decoder architecture, where attention gate (AG) was improved and integrated into each skip connection, making the cloud detection model distinguish the cloud and noncloud more accurately [20]. Zhang et al. [21] designed multiscale global attention (MGA) module in their CRSNet in order to strengthen the channel and spatial information. MGA was made up of three submodules: hierarchical multiscale convolution module, global spatial attention module (SAM), and global channel attention module (CAM). However, the above methods have a large number of parameters and computations, which can be tens of millions. To alleviate this problem, Yao et al. [22] designed a lightweight CAM in CD-AttDLV3+, which was based on DeepLabV3+ architecture, to strengthen the learning of significant channels. Also, to further reduce parameters and computations, CD-AttDLV3+ used MobileNetV2 as its backbone. However, there are still two problems in these cloud detection methods. On the one hand, a majority of existing methods only utilize the stack of convolutional layers to obtain local spatial features, ignoring global semantic information of image. On the other hand, existing models with attention mechanism still have much room for improvement in both performance and efficiency.

In this work, we propose a lightweight encoder–decoder architecture CD-CTFM for cloud detection. CD-CTFM utilizes a lightweight network, on basis of CNN and Transformer, as backbone. CNN has an excellent ability of local features extraction, while Transformer is good at extracting global contextual information. Therefore, the combination of CNN and Transformer is conducive to fusing both local and global features, thus improving accuracy of cloud detection. Considering dark channel plays an important role in haze removal according to [23], and cloud and haze have a strong visual similarity, dark channel prior must help in cloud detection. Dark channel prior of image can reflect the cloud distribution to a large extent, so the backbone of CD-CTFM incorporates AG, which is based on dark channel extraction module (DCEM). A novel lightweight feature pyramid module (LWFPM), which consists of five paths and three of them are shared and dilation (SD) blocks, is proposed to fuse multiscale features. SD block is a residual structure made up of pointwise (PW) convolution, shared convolution (SC), and dilated convolution (DC). Besides, hierarchical feature fusion (HFF), which is parameter-free, is used to solve the gridding effect caused by DC. CD-CTFM incorporates a lightweight channel-spatial attention module (LWCSAM) to suppress invalid information and highlight discriminative features. To alleviate the problem of information attenuation or loss caused by pooling, LWCSAM introduces mixed pooling technique, which is based on stochastic model. CD-CTFM has achieved satisfactory results on public datasets 38-Cloud and MODIS. To summarize, the main contributions of this work are listed as follows.

1) CD-CTFM utilizes a lightweight CNN-Transformer network as backbone and the backbone incorporates dark channel-based AG.
2) A novel LWFPM based on SD block is proposed to fuse multiscale features, where HFF technique is introduced to solve the gridding effect caused by DC.
3) CD-CTFM incorporates a LWCSAM to suppress invalid information and highlight discriminative features.
4) We conduct extensive experiments on 38-Cloud and MODIS datasets. The results demonstrate that CD-CTFM can obtain the best tradeoff in terms of performance and efficiency.

The rest of this article is organized as follows. In Section II, related works, such as visual transformer, attention mechanism, dilated covolution, and lightweight model, are introduced. The proposed CD-CTFM, LWFPM, and LWCSAM are detailed in Section III. Section IV illustrates experimental settings and results. Section V discusses some limitations of the proposed method. Finaly, Section VI concludes this article.

II. Related Work

A. Visual Transformer

Transformer is an excellent deep learning model that has been widely used in many fields [24], such as natural language processing, CV, and speech processing [25]. In the field of CV, Transformer has been adopted for various tasks, e.g., image classification [26], object detection [27], image generation [28], and video understanding [29], [30], [31].

As for semantic segmentation task, which requires modeling rich interactions between pixels, Transformer also has a wide range of applications. Some recent works have designed encoder–decoder architectures based on transformer. Zheng et al. [32] proposed a vision transformer (ViT) encoder and three different types of decoders, which are based on naive upsampling, progressive upsampling, and multilevel feature aggregation, respectively, in their segmentation transformer. Xie et al. [33] designed SegFormer, of which the encoder is a hierarchical pyramid ViT while the decoder is based on MLP and simple upsampling operation. In Segmenter, Strudel et al. [34] also empolyed a ViT encoder to extract feature maps, and used a mask Transformer as the decoder, predicting sementation masks.

As a type of semantic segmentation application, cloud detection task tends to get good results on the above models. In addition, scholars have also proposed Transformer-based methods specifically for cloud detection. Zhang et al. [35] adopted MobileViT as the backbone in their CloudViT. Singh et al. [36] proposed a novel spatial-spectral attention transformer for cloud detection with a spatial-spectral attention module that generates an enhanced feature map to replace convolution by using the image patches directly.

In CD-CTFM, we also utilize the visual transformer as the backbone network. Uniqueness of the backbone lies in two aspects. On one hand, we amalgamate CNN and Transformer, allowing network to effectively capture both global and local features. On the other hand, dark-channel-based AG is integrated into backbone, leveraging the prior knowledge of the dark channel to enhance the model's performance.

B. Attention Mechanism

Motivated by the fact that humans can effectively find salient regions in complex scenes, attention mechanism, a dynamic weight adjustment process based on feature maps, was introduced into the field of CV. Attention mechanisms can be divided into many types according to approach [37], such as channel attention, spatial attention, temporal attention, and branch attention. In the field of semantic segmentation, including cloud detection task, the most commonly used attention mechanisms are channel attention and spatial attention.

Channel attention adaptively recalibrates the weight of each channel, because different channels in different feature maps usually contain information of different degrees of importance. SENet proposed by Hu et al. [38] used SE modules to obtain the weight of feature maps of different channels. Yao et al. [22] introduced the CAM in their cloud detection method, CD-AttDLV3+, to strengthen the learning of important channels.

Spatial attention can be interpreted as an adaptive spatial region selection mechanism, which foucses on where to pay attention. Oktay et al. [39] proposed a simple and yet effective mechanism, the AG, to focus on targeted regions while suppressing feature activations in irrelevant regions. In cloud detection model CAA-UNet, Zhang et al. [20] modified AG and integrated into each skip connection to hightlight salient features.

Channel-spatial attention combines the advantages of channel attention and spatial attention. Woo et al. [40] designed a novel concolutional block attention module, which stacked channel attention and spatial attention in series. Guo et al. [41] proposed a channel-spatial attention-based module, adaptive feature fusing model (AFFM), to fuse multilevel feature maps. AFFM was composed of three submodules: channel attention fusion model, spatial attention fusion module, and channel attention refinement model.

The proposed LWCSAM falls under the category of channel-spatial attention, formed by the cascade of a CAM and a SAM. Its primary innovation revolves around the modeling of cross-dimension interaction information between channel dimension and either vertical spatial dimension or horizontal spatial dimension.

C. Dilated Convolution

In a CNN, if the sizes of convolutional kernals are too small, the network tends to be limited to local information, decreasing the accuracy of network. Conversely, if the kernals are too large, the number of parameters and calculations will be too high, which makes it hard to deploy model on mobile devices. This problem can be solved by combining DC with different expansion rates, according to atrous spatial pyramid pooling module in DeepLab [42]. CD-AttDLV3+, proposed by Yao et al. [22] was based on DeeplabV3+ architecture, which was able to robustly segment objects at multiple scales. He et al. [43] proposed a deformable context feature pyramid module in their DABNet to improve the adaptive modeling capability of multiscale features, which is also based on ASPP.

The difference between LWFPM and other methods lies in its introduction of the parameter-free HFF technique to address the grid effect, and employing the SD block instead of regular DC, achieving improved accuracy at a low cost.

D. Lightweight Model

Due to the limited performace of mobile devices, deploying models with huge parameters is impracticable. Therefore, it is necessary to carry out lightweight design of the network. Recently, a quantity of lightweight CNN have been proposed. MobileNetV2 proposed by Google achieved great success in compressing model [44], which was able to separate the network expressiveness from its capacity thanks to inverted residuals and linear bottlenecks. Han et al. [45] designed a novel Ghost module to generate more feature map from cheaper operations.

As for cloud detection task, researchers have also designed a multitude of lightweight models. Hu et al. [46] proposed two novel lightweight modules in their LCDNet, one of which was lightweight bottleneck, which had the ability to quickly capture multiscale features. The other lightweight module proposed in LCDNet was lightweight self attention module, which could quickly establish the spatial location information of remote sensing image (RSI). The abovementioned CloudViT was also a lightweight cloud detection method. It was composed of a multiscale dark channel extractor, used to guide the network based on dark channel priors, and an attention-based context aggregation module, utilized to make cloud detection results more accurate.

III. Methodology

A. Overview of CD-CTFM

The CD-CTFM exploits the encoder–decoder architecture, as shown in Fig. 1. The encoder part captures local spatial

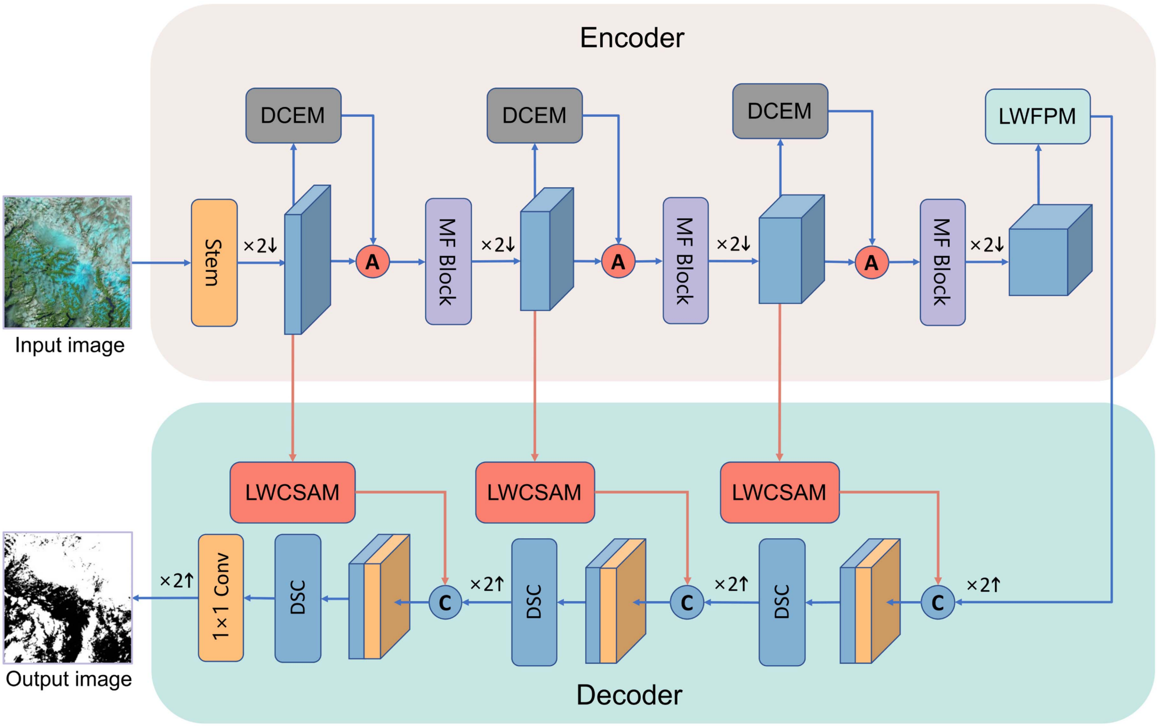

Fig. 1. Overall framework of CD-CTFM. The encoder contains a lightweight backbone and LWFPM while the decoder is based on DSC. LWCSAM filters information propagated through the skip connections.

features and global context information simultaneously, making the results of semantic segmentation more reliable. However, processing large amounts of semantic features increases parameters and computations, which is contradictory to the aim of lightweight cloud detection network. To ease this problem, CD-CTFM combines dark channel prior [23] with attention mechanism in the AG. With the guidance of the AG, the encoder is able to highlight discriminant feature regions while suppressing irrelevant regions, improving the performance of network. Besides, local spatial features are further fused with global multiscale features, for which the encoder is able to gain stronger semantic information. To compensate for the attenuated or lost low-level features, the decoder part fuses low-level information with deep features through skip connectinos step by step, which is based on LWCSAM.

In the encoder part, we propose a lightweight network based on mobile-former (MF) block and DCEM as backbone, combing CNN, and Transformer. In the backbone, a stem layer composed of a convolution layer and a bottleneck layer is first adopted to adjust the number of channels and downsample the feature maps. Bottleneck is a structure that compresses and amplifies information, which is able to remove high-frequency noise and reduce parameters. Given a multichannel image F_I as input, the extracted feature maps of the stem layer can be stated as

$$F_{\text{stem}} = \text{PW}(\text{DW}(\text{PW}(\text{Conv}(F_I)))) \tag{1}$$

where $\text{Conv}(\cdot)$ is convolution, $\text{PW}(\cdot)$ is PW convolution, and $\text{DW}(\cdot)$ is depthwise convolution.

Following the stem layer is DCEM. For F_{stem}, its dark channel F_{dark} is given by

$$F_{\text{dark}}(x) = \min_{y\in\Omega(x)}\left(\min_{c\in C} F_{\text{stem}}[c](y)\right) \tag{2}$$

where $\Omega(x)$ is a local patch centered at *x*, C is the channel set of F_{stem}, c is an element of C, and $F_{\text{stem}}[c]$ is the c channel of F_{stem}. As Fig. 2 shows, the dark channel of RSI is the reulst of two commutative minimum operators, of which one is performed on all channel values of RSI at pixel level and the other is a minimum filter.

After extraction, F_{stem} and its corresponding dark channel F_{dark} are forwarded to the AG. As shown in Fig. 3, the two inputs are elementwise merged after convolution and batch (CB) nomalization blocks, respectively. After a rectified linear unit (ReLU) activation function, CB block and sigmoid activation function, we can get attention coefficient α. Finally, the output of AG F_a is obtained by multiplying F_{stem} by attention coefficient A. The above process is formulated as

$$F_a = BN(\text{Conv}(F_{\text{stem}})) + BN(\text{Conv}(F_{\text{dark}})) \tag{3}$$

$$A = \text{Sigmoid}(BN(\text{Conv}(\text{ReLU}(F_a)))) \tag{4}$$

$$F_A = F_{\text{stem}} \odot A \tag{5}$$

where $\odot$ denotes elementwise product.

Then, feature maps guided by dark channel and global tokens (learnable parameters) are forwarded to a sequnece of MF block (see Fig. 4). In particular, the lengths of MF block sequences are set to 2, 2, and 4, respectively. Different from ViT [47], there

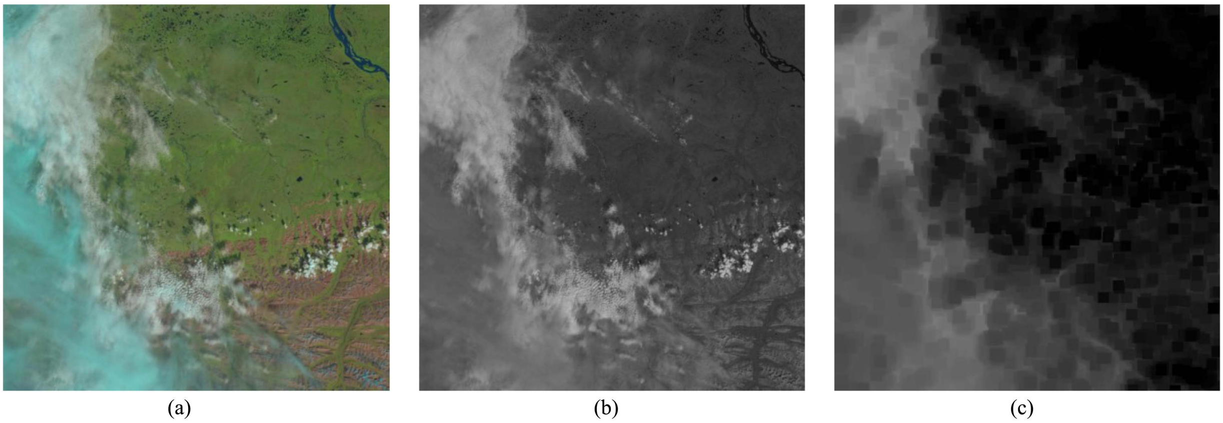

Fig. 2. Process of DCEM. (a) Remote sensing image from 38-Cloud dataset. (b) Minimum of the four channel values of the image at pixel level. (c) Dark of the image. It is the result of performing minimum filter on (b), where the kernel size is 60×60.

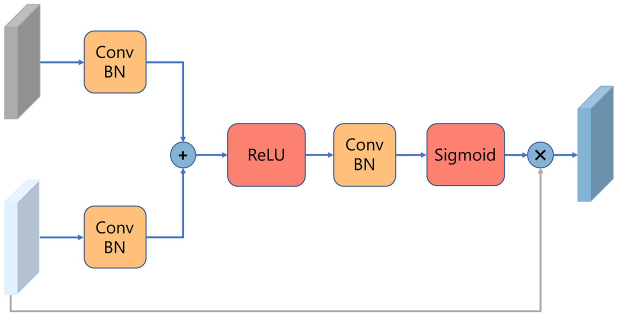

Fig. 3. Structure of the AG used in backbone.

are only 6 global tokens in CD-CTFM, which are much fewer than that of ViT. Global tokens are initialized from learnable embeddings rather than linear projection of image patches. As a result, fewer tokens can not only contain enough global prior knowledge of RSI, but also contribute to reducing parameters and computations.

In MF blocks, local spatial features and global context information are extracted simultaneously and fused bidirectionally, resulting in more representation power. Each MF block is composed of four subblocks. Mobile subblock is essentially a bottleneck, which contains two PW convolutions and a depthwise convolution. Former subblock is made up of a multihead attention and a feedforward network. The other two parts are mobile-former subblock and former-mobile subblock, which are used to fuse information between mobile subblock and former subblock based on attention mechanism.

After the performance of three stages of DCEM, AG, and MF sequences, we obtain the output of the backbone. Followed by the backbone is the LWFPM to fuse high-level multiscale features with global interaction information.

In the decoder part, deep features coming from LWFPM are up-sampled layer by layer. To compensate for the attenuated or lost low-level features, a LWCSAM is introduced into each skip connection, which suppress irrelevant low-level information and highlight discriminative features.

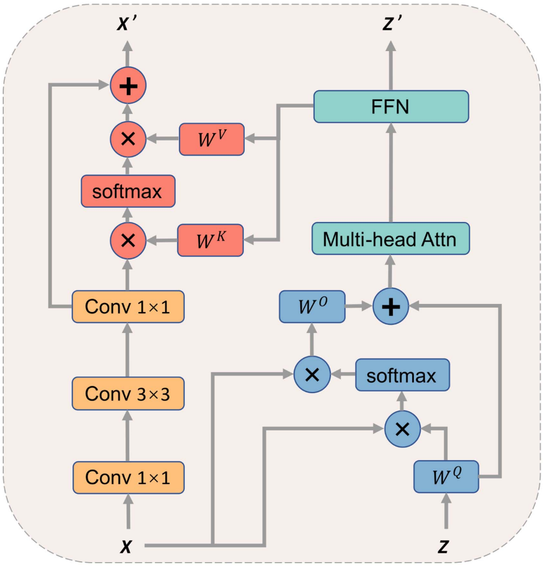

Fig. 4. Details of MF block. Yellow blocks belong to Mobile sub-block, green blocks belong to Former subblock, blue blocks belong to mobile-former subblock, and red blocks belong to former-mobile subblock.

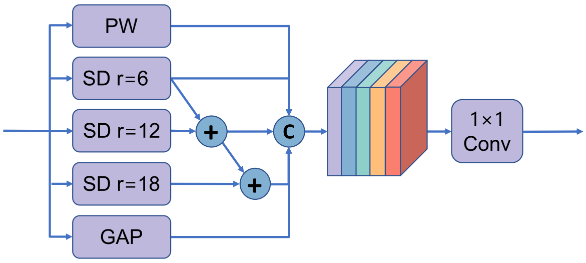

Fig. 5. Structure of LWFPM.

B. Lightweight Feature Pyramid Module

To further extract features propagated through backbone, we design LWFPM, which consists of five parallel paths (see Fig. 5). Except for one global average pooling (GAP) path and one

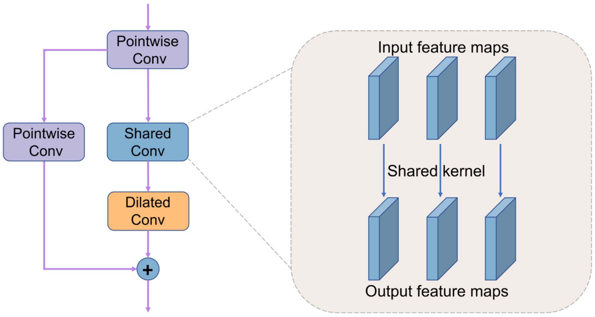

Fig. 6. Details of SD block.

PW convolution path, the rest of the paths consist of three novel SD blocks, of which the dilated rate is 6, 12, and 18, respectively. As shown in Fig. 6, residual structure-based SD block contains two PW blocks, a SC, and a DC. The effect of the PW is dimensionality reduction while the purpose of SC is deep features extraction, both of which are conducive to decreasing parameters and computations. To fuse high-level multiscale features with context information, DC with different dilated rates are added. Inspired by ESPNet [48], LWFPM deploys HFF to solve the gridding effect caused by DC without introducing extra parameters. Given that the input of LWFPM is X, the output Y can be obtained by the following equations:

$$X_1 = \text{PW}(X) \tag{6}$$

$$X_2 = \text{DC}_6(\text{SC}(\text{PW}(X))) + X \tag{7}$$

$$X_3 = \text{DC}_{12}(\text{SC}(\text{PW}(X))) + X \tag{8}$$

$$X_4 = \text{DC}_{18}(\text{SC}(\text{PW}(X))) + X \tag{9}$$

$$X_5 = \text{GAP}(X) \tag{10}$$

$$Y = \text{PW}(\text{concat}(X_1, X_2, X_3, X_4, X_5)) \tag{11}$$

where PW is PW convolution, DC_r is DC with a dilated rate equal to r, SC is shared convolution, and GAP is the global average pooling.

C. Lightweight Channel-Spatial Attention Module

With the help of the skip connections, CD-CTFM can get more low-level information to detect the cloud. However, some feature maps contain too much invalid or irrelevant information, affecting the accuracy of the model. To alleviate this problem, a LWCSAM is proposed, as shown in Fig. 7. LWCSAM is composed of CAM and SAM, which are cascaded directly.

CAM is made up of two parallel branches, aiming at modeling channel attention based on cross-dimension interaction information between the channel dimension C and either the vertical spatial dimension H or the horizontal spatial dimension W. In CAM, the input feature maps X are first rotated around either W-axis or H-axis, getting X_W and X_H, respectively. Then we need to compress the 0th dimension of feature maps in order to construct the importance of each point on the C × W (or C × H) plane. Since single maximum pooling or average pooling tends to result in loss of much valuable information, the stochastic mixed pooling technique is used [49], which is able to preserve rich information with high probability while keeping computation lightweight. The stochastic mixed pooling can be formulated as

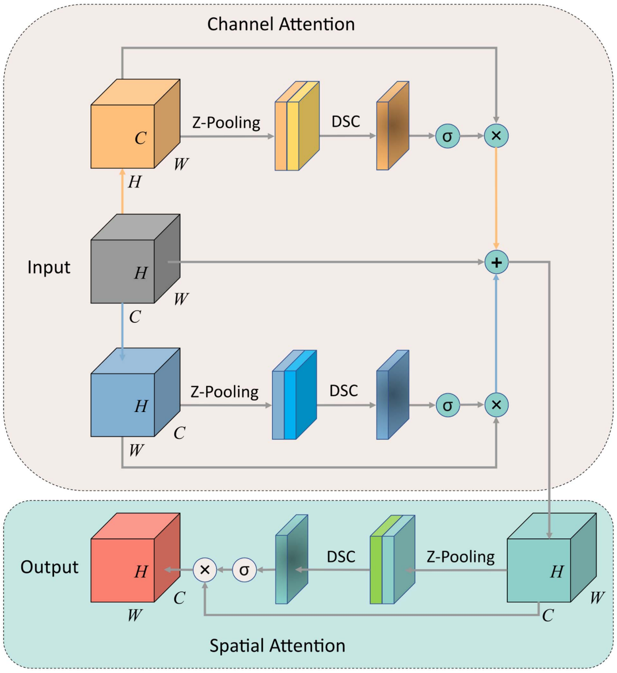

Fig. 7. Details of LWCSAM. Yellow arrows represent rotating feature maps around W-axis while blue arrows represent rotating feature maps around H-axis. In addition, σ is a sigmoid activation function.

$$y_c = \lambda \cdot \max_{(i,j)\in\mathcal{P}} x_{ijc} + (1-\lambda) \cdot \frac{1}{|\mathcal{P}|} \cdot \sum_{(i,j)\in\mathcal{P}} x_{ijc} \tag{12}$$

where x and y are pixel values before and after pooling, c represents channel, (i, j) denotes pixel of RSI, $\mathcal{P}$ is the set of pixels, and λ is a random value between 0.0 and 1.0. After pooling and subsequent depthwise separable convolutions (DSC) and sigmoid layers, attention maps are multiplied by corresponding feature maps, getting X'_W and X'_H, respectively. At last, the output of CAM, which is denoted as X', can be obtained by adding X'_W, X'_H, and the feature maps that is not rotated.

SAM is employed to build spatial attention, of which the structure is similar to that of CAM. X' is first passed through the stochastic mixed pooling layer, DSC and sigmoid layer, After which the shape of feature map is reduced to $1 \times H \times W$. Then, the result of sigmoid activation function and X' are aggregated by elementwise multiplication. The output of SAM can be stated as

$$X'' = \text{Sigmoid}(\text{DSC}(\text{MP}(X'))) \odot X' \tag{13}$$

where MP is mixed pooling.

D. Loss Function

By reason of the class imbalance problem manifested in cloud detection tasks, CD-CTFM combine the dice loss and the binary cross entropy loss as its loss function [50], which is defined as follows:

$$L(y_i, \hat{y}_i, \Theta) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)) + \lambda\left(1 - \frac{2\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n}(y_i^2 + \hat{y}_i{}^2)}\right) + \frac{\mu}{2}\|\Theta\|^2 \quad (14)$$

where the coefficient $\lambda \geq 0$ specifies a relative importance of the dice loss versus the binary cross entropy loss, $\frac{\mu}{2}$ indicates the relative significance of regularization, y_i is the ground truth, $\hat{y}_i$ is the prediction result and n is the number of pixels in each image.

IV. Experiments and Results

A. Experimental Settings

1) Datasets: To train and test the CD-CTFM, we conduct experiments on two cloud detection datasets: 38-Cloud and MODIS datasets.

The 38-Cloud dataset is first released in [51]. The dataset contains 18 LandSat-8 images for training and 20 images for testing. Due to the large size of images, it is difficult to directly use these images as inputs. Therefore, each image is cropped into 384 × 384 nonoverlapping patches. After cropping, the training set contains 8400 patches while the test set contains 9201 patches. Each patch has four corresponding spectral channels: red (band 4), green (band 3), blue (band 2), and near infrared (band 5).

MODIS dataset contains 1422 remote sensing images, which are separated into 1272 training images and 150 test images [52]. After cropping them into 512 × 512 patches, the training set and test set contain 19 080 and 2250 nonoverlapping patches, respectively. Each patch consists of ten spectral channels: band 1, 3, 4, 18, 20, 23, 28, 29, 31, and 32.

2) Evaluation Metrics: In the experiment, the ground truths and prediction results are divided into cloud and non-cloud classes at pixel level. The performance of CD-CTFM is evaluated by five quantitative metrics, including mean intersection-over-union (mIoU), precision, recall, F1-score, and overall accuracy (OA) [53]. These metrics are defined as follows:

$$\text{mIoU} = \frac{\text{TP}}{\text{TP + FN + FP}} \quad (15)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \quad (16)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \quad (17)$$

$$\text{F1-score} = \frac{2 \times \text{TP}}{2 \times \text{TP + FP + FN}} \quad (18)$$

$$\text{OA} = \frac{\text{TP + TN}}{\text{TP + TN + FP + FN}} \quad (19)$$

where TP, TN, FP, and FN represent the total number of true positive, true negative, false positive, and false negative pixels, respectively. Besides, we utilize two metrics, including giga floating-point operations per second (GFLOPS) and the number of model parameters, to measure the efficiency of models.

3) Implementation Details: Our experiments are based on the Pytorch framework and Ubuntu 20.04 equipped with an NVIDIA 3090 24 G GPU. All models are optimized by the stochastic gradient descent algorithm and the learning rate decays from 0.001 to 0. Besides, the batch sizes, momentum, and epochs are 32, 0.9, and 50, respectively.

B. Ablation Studies

In this section, we utilize 38-Cloud and MODIS datasets to verify the effectiveness of backbone, LWFPM and LWCSAM in CD-CTFM.

1) Effectiveness of Backbone: To begin with, the backbone of CD-CTFM is replaced by that of MobileNetV2 in order to evaluate the performance and efficiency of our proposed backbone. As we can see in Table I, although GFLOPS and parameters of CD-CTFM is only 28% and 91% of the model based on MobileNetV2, CD-CTFM still has a 1.98% increase in mIoU in 38-Cloud dataset. Subsequently, we conducted experiments on CD-CTFM without DCEM so that we can evaluate the impact of dark channel on cloud detection. The result is also given in Table I. With guidance of AG based on dark channel, CD-CTFM gains a 0.15% increase in mIoU and a 0.12% increase in F1-Score when it comes to the MODIS dataset while the parameters and GFLOPS are almost the same.

2) Effectiveness of LWFPM: To verify the effectiveness of LWFPM, it is replaced by a simple PW convolution and ASPP, respectively. Table II gives the performance and efficiency of CD-CTFM with different feature fusion modules. In terms of performance, LWFPM achieves higher scores than ASPP in all metrics in MODIS dataset. As for 38-Cloud dataset, OA of LWFPM is only 0.14% lower than ASPP, recall is 0.96% lower than ASPP, and other evaluation metrics are all higer than that of ASPP. With regard to efficiency, the params of LWFPM are fewer than that of ASPP while GFLOPS is more both in 38-Cloud and MODIS datasets.

3) Effectiveness of LWCSAM: In this part, LWCSAM is first removed to evaluate its performance and efficiency. As given in Table III, LWCSAM is almost a parameter-free module and its GFLOPS is quite small. On the other hand, LWCSAM improves performance in all metrics, especially mIoU. Then, LWCSAM is compared with SE module, of which the result is also exhibited in Table III. While SE module slightly outperforms LWCSAM on some performance metrics, such as recall, LWCSAM is lighter than SE.

C. Comparison With Other Methods

To further validate the effectiveness of our proposed model, we compare CD-CTFM with five non-lightweight models and four lightweight methods on MODIS and 38-Cloud datasets. Five nonlightweight models contain two semantic segmentation model (UNet [54] and DeepLabV3+ [55]) and three cloud detection models (CloudFCN [16], CloudAttU [19], and CRSNet [21]). Four lightweight methods contain two traditional

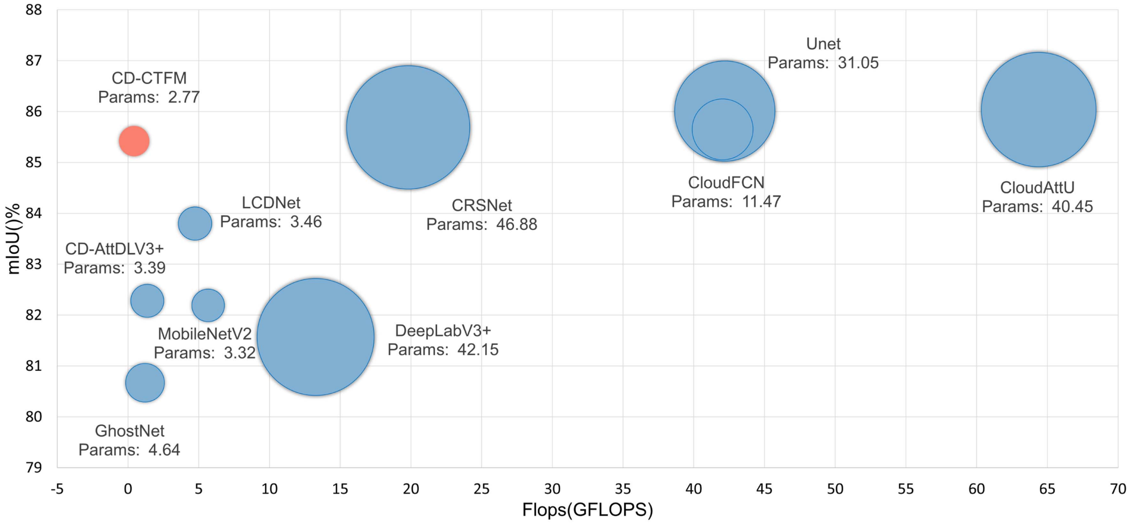

Fig. 8. MIoU performace, GFLOPS, and parameters of cloud detection models on 38-Cloud dataset. *x*-axis is the GFLOPS, *y*-axis is the MIoU, and the size of bubble represents the parameters. Here, CD-CTFM is compared with UNet, DeepLabV3+, CloudFCN, CloudAttU, CRSNet, MobileNetV2, GhostNet, CD-AttDLV3+, and LCDNet.

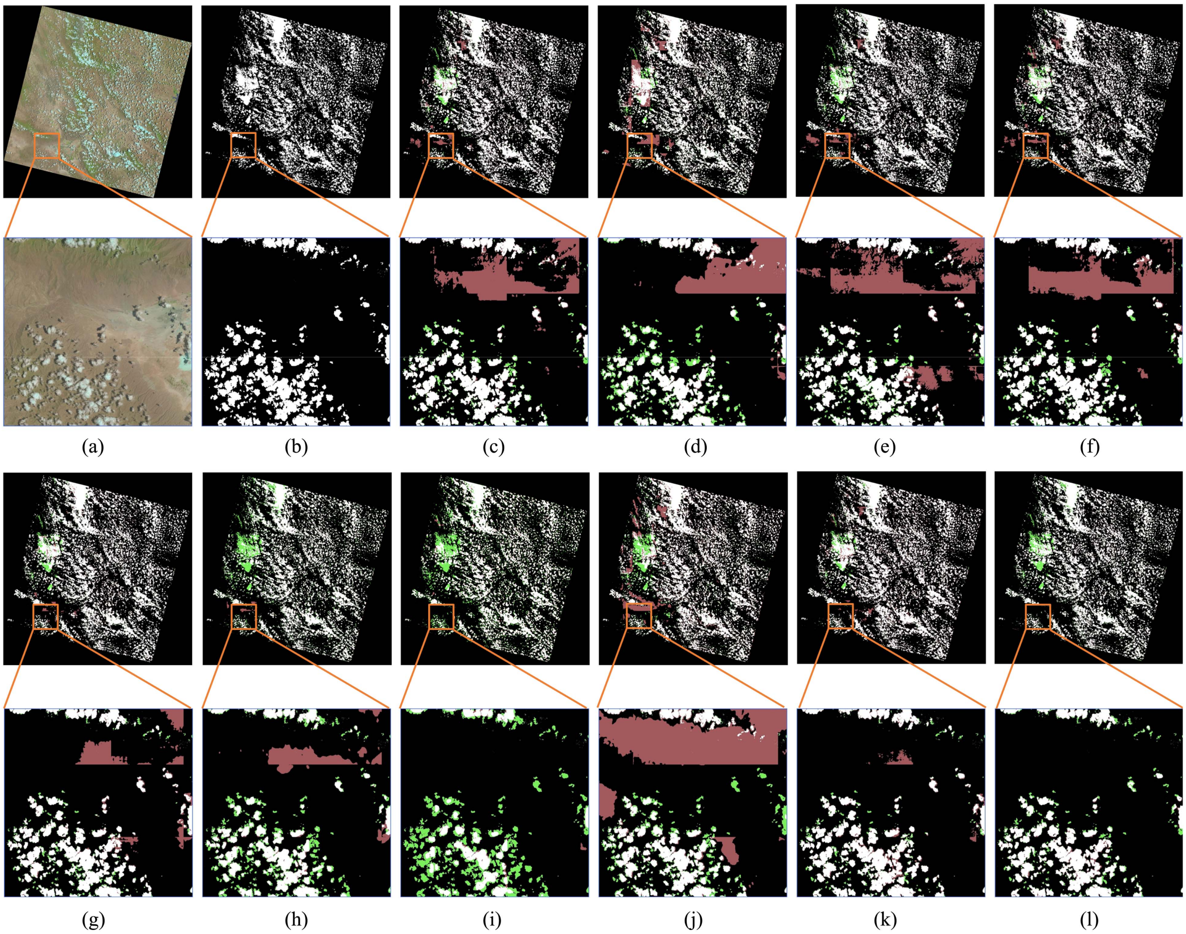

Fig. 9. Comparison between the results of different methods in 38-Cloud dataset. White area represents cloud, black area represents noncloud, red area represents false-positive detection and green area represents false-negative detection. (a) False-color RSI. (b) UNet. (c) DeepLabV3+. (d) CloudFCN. (e) CloudAttU. (f) CRSNet. (g) Ground truth. (h) MobileNetV2. (i) GhostNet. (j) CDAttDLV3+. (k) LCDNet. (l) CD-CTFM (ours).

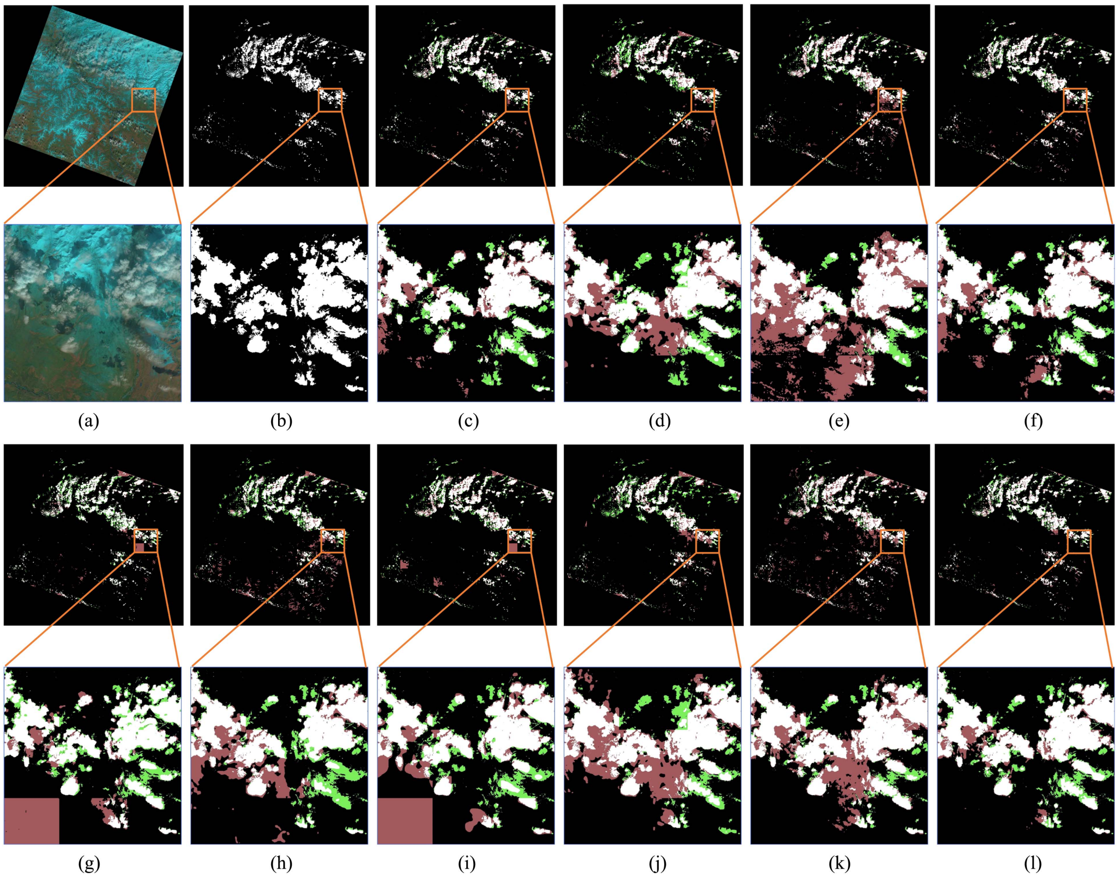

Fig. 10. Comparison between the results of different methods in 38-Cloud dataset. White area represents cloud, black area represents noncloud, red area represents false-positive detection and green area represents false-negative detection. (a) False-color RSI. (b) UNet. (c) DeepLabV3+. (d) CloudFCN. (e) CloudAttU. (f) CRSNet. (g) Ground truth. (h) MobileNetV2. (i) GhostNet. (j) CDAttDLV3+. (k) LCDNet. (l) CD-CTFM (ours).

TABLE I
ABLATION STUDY OF PROPOSED BACKBONE

Dataset	Backbone	Performance					Efficiency	
		mIoU	Precision	Recall	F_1-Score	OA	params (M)	GFLOPS
38-Cloud	MobileNetV2	82.79±1.26	89.22±0.99	**89.95±1.58**	89.58±1.27	95.41±1.84	3.03	1.48
	CD-CTFM without DCEM	84.54±0.88	91.62±0.26	88.82±0.66	90.20±0.24	**95.88±1.53**	**2.77**	**0.41**
	CD-CTFM	**84.77±0.10**	**92.44±0.41**	89.08±0.52	**90.73±0.46**	95.57±0.90	2.77	0.42
MODIS	MobileNetV2	84.11±0.62	91.67±0.71	90.14±1.88	90.40±1.21	93.79±0.32	3.03	1.50
	CD-CTFM without DCEM	85.27±0.81	**92.71±0.79**	90.47±0.87	91.20±0.73	94.47±0.53	**2.77**	**0.42**
	CD-CTFM	**85.42±0.10**	92.46±0.08	**90.77±0.22**	**91.32±0.12**	**94.52±0.63**	2.77	0.43

The bold entities represent the best-performing algorithms.

TABLE II
ABLATION STUDY OF LWFPM

Dataset	Module	Performance					Efficiency	
		mIoU	Precision	Recall	F_1-Score	OA	params (M)	GFLOPS
38-Cloud	None	82.51±1.13	90.72±1.14	87.44±0.86	89.05±0.98	95.04±0.80	**2.31**	**0.33**
	ASPP	83.46±0.90	88.77±0.67	**90.04±0.48**	89.40±0.34	**95.71±0.55**	2.83	0.39
	LWFPM	**84.77±0.10**	**92.44±0.41**	89.08±0.52	**90.73±0.46**	95.57±0.90	2.77	0.42
MODIS	None	84.14±0.73	91.82±1.01	90.07±1.29	90.40±1.01	93.86±0.87	**2.31**	**0.34**
	ASPP	84.61±0.44	91.94±0.67	90.41±0.75	90.75±0.26	94.09±0.90	2.83	0.40
	LWFPM	**85.42±0.10**	**92.46±0.08**	**90.77±0.22**	**91.32±0.12**	**94.52±0.63**	2.77	0.43

The bold entities represent the best-performing algorithms.

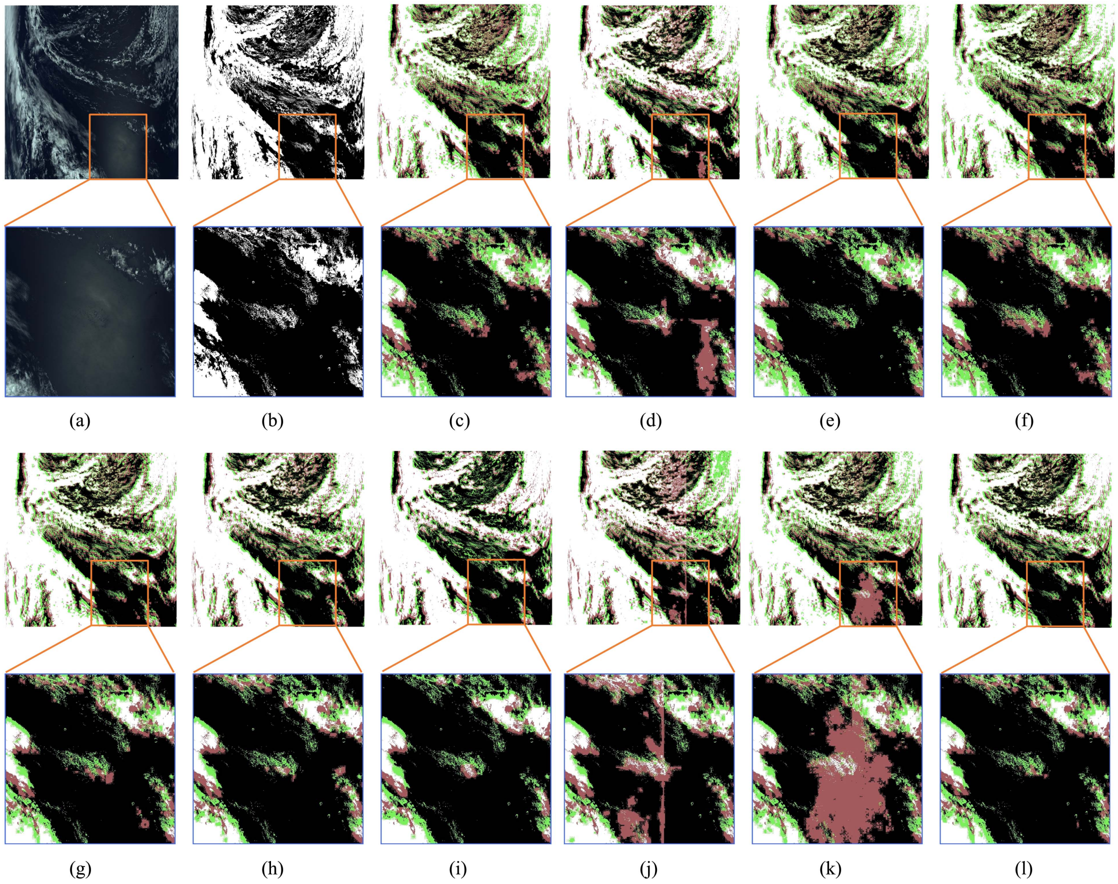

Fig. 11. Comparison between the results of different methods in MODIS dataset. White area represents cloud, black area represents noncloud, red area represents false-positive detection and green area represents false-negative detection. (a) False-color RSI. (b) UNet. (c) DeepLabV3+. (d) CloudFCN. (e) CloudAttU. (f) CRSNet. (g) Ground truth. (h) MobileNetV2. (i) GhostNet. (j) CDAttDLV3+. (k) LCDNet. (l) CD-CTFM (ours).

TABLE III
Ablation Study of LWCSAM

Dataset	Module	Performance					Efficiency	
		mIoU	Precision	Recall	F_1-Score	OA	params (M)	GFLOPS
38-Cloud	None	83.37±1.28	91.64±0.86	88.26±0.46	89.92±0.63	94.87±0.40	**2.77**	**0.41**
	SE	**84.86±1.01**	90.91±0.49	**89.91±0.70**	90.41±0.13	**95.82±0.40**	2.78	0.42
	LWCSAM	84.77±0.10	**92.44±0.41**	89.08±0.52	**90.73±0.46**	95.57±0.90	2.77	0.42
MODIS	None	84.57±1.05	91.68±0.81	90.60±0.75	90.76±0.66	94.04±0.87	**2.77**	**0.42**
	SE	**85.74±0.83**	**92.80±0.82**	**90.86±0.74**	**91.54±0.36**	**94.63±0.72**	2.78	0.43
	LWCSAM	85.42±0.10	92.46±0.08	90.77±0.22	91.32±0.12	94.52±0.63	2.77	0.43

The bold entities represent the best-performing algorithms.

lightweight models (MobileNetV2 [44], and GhostNet [45]) and two cloud detection models (CD-AttDLV3+ [22], and LCD-Net [46]).

1) Results on 38-Cloud Dataset: Table IV gives the performance and efficiency of various cloud detection models on 38-Cloud dataset. On the one hand, CD-CTFM outperforms four lightweight models on performance and efficiency. Compared with other top-performing lightweight models, mIoU of CD-CTFM is 1.91% higher than LCDNet, while the parameters and GFLOPS are both significantly fewer than that of LCDNet. On the other hand, compared to five nonlightweight models, CD-CTFM achieves similar or better results on performance metrics, but obviously improves efficiency. For example, the mIoU index of CD-CTFM is 0.6% lower than CloudAttU, which has the highest value in mIoU, but the parameters and computations of CloudAttU are 40.45 (M) and 64.21 GFLOPS, which are 13.6 and 159.5 times larger than CD-CTFM, respectively. Even though CRSNet has the highest recall value, its precision value is lower than that of CD-CTFM. As for efficiency, the parameters of CRSNet is 15 times larger than parameters of our proposed CD-CTFM. Fig. 8 visually illustrates the comparison between different models on 38-Cloud dataset.

Figs. 9 and 10 display the qualitative results on 38-Cloud dataset. White pixel represents detecting cloud for cloud and

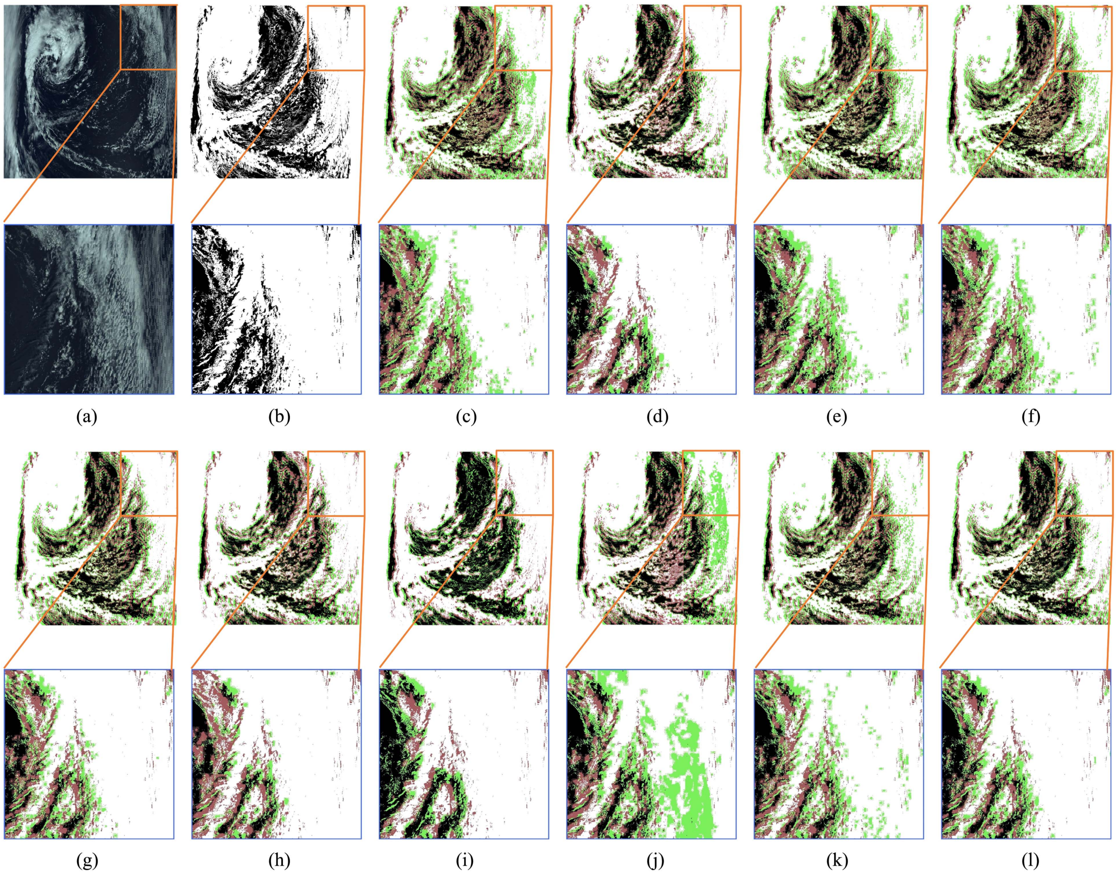

Fig. 12. Comparison between the results of different methods in MODIS dataset. White area represents cloud, black area represents non-cloud, red area represents false-positive detection and green area represents false-negative detection. (a) False-color RSI. (b) UNet. (c) DeepLabV3+. (d) CloudFCN. (e) CloudAttU. (f) CRSNet. (g) Ground truth. (h) MobileNetV2. (i) GhostNet. (j) CDAttDLV3+. (k) LCDNet. (l) CD-CTFM (ours).

TABLE IV
PERFORMANCE AND EFFICIENCY OF VARIOUS CLOUD DETECTION MODELS ON 38-CLOUD DATASET

Category	Method	Performance					Efficiency	
		mIoU	Precision	Recall	F_1-Score	OA	params (M)	GFLOPS
Non-Lightweight Models	UNet	83.82±0.43	89.68±0.41	89.85±0.38	89.76±0.10	95.75±1.47	31.05	42.02
	DeepLabv3+	81.64±0.62	87.72±0.62	88.36±0.66	88.04±0.63	95.03±0.74	41.04	**13.24**
	CloudFCN	83.31±0.45	88.81±0.62	89.61±1.32	89.21±0.39	95.66±0.90	**11.44**	40.53
	CloudAttU	84.73±0.86	90.62±0.52	89.95±1.27	90.28±0.40	95.92±0.98	40.45	64.21
	CRSNet	**85.76±1.35**	**92.34±0.52**	**89.97±1.70**	**91.14±1.13**	**96.10±0.76**	46.80	18.64
Lightweight Models	MobileNetV2	82.37±0.43	90.97±1.02	87.31±0.05	89.10±0.46	94.90±0.78	3.31	5.65
	GhostNet	82.39±0.42	91.98±0.99	87.72±0.90	89.80±0.94	95.16±0.95	4.64	1.19
	CD-AttDLV3+	81.24±0.41	88.85±0.47	87.58±0.09	88.21±0.19	94.49±0.95	3.54	1.33
	LCDNet	82.86±0.47	90.15±0.82	89.08±0.57	89.61±0.18	94.34±1.53	3.47	4.78
	CD-CTFM	**84.77±0.10**	**92.44±0.41**	**89.08±0.52**	**90.73±0.46**	**95.57±0.90**	**2.77**	**0.42**

The bold entities represent the best-performing algorithms.

black pixle represents detecting background for background correctly. Red pixel represents mistaking background for cloud and green pixel represents mistaking cloud for background. Fig. 9 contains flat lands with complicated texture and broken clouds. Thanks to DCEM incorporated in backbone, CD-CTFM cannot only clearly detect correct cloud regions, but also exclude surface with complex texture, which are like clouds in terms of color and shape. Fig. 10 demonstrates the results on remote sensing image with cloud snow coexistence. CD-CTFM is able to identify ice and snow surfaces with complex textures, benefiting from the combination of CNN with Transformer and the fusion of multiscale features.

2) Results on MODIS Dataset: Quantitative results of various cloud detection models on MODIS are displayed in Table V. For lightweight models, CD-CTFM achieves the best quantitative results among all competing methods. To be specific, compared with CD-AttDLV3+, CD-CTFM achieves performance gain by 3.14%, 1.58%, 1.96%, 2.17%, and 1.59%

TABLE V
PERFORMANCE AND EFFICIENCY OF VARIOUS CLOUD DETECTION MODELS ON MODIS DATASET

Category	Method	Performance					Efficiency	
		mIoU	Precision	Recall	F_1-Score	OA	params (M)	GFLOPS
Non-Lightweight Models	UNet	86.01±0.90	92.82±0.67	91.44±0.39	91.65±0.53	94.67±0.38	31.05	42.19
	DeepLabv3+	81.57±0.48	90.13±0.44	88.62±0.68	88.67±0.14	92.53±0.62	42.15	**13.26**
	CloudFCN	85.65±0.52	92.04±0.19	**91.79±0.10**	91.45±0.12	94.36±0.46	**11.47**	42.04
	CloudAttU	**86.04±0.54**	**92.93±0.58**	91.42±0.24	**91.66±0.36**	**94.71±0.82**	40.45	64.39
	CRSNet	85.69±0.64	92.67±0.94	91.02±0.41	91.47±0.63	94.57±0.47	46.88	19.81
Lightweight Models	MobileNetV2	82.19±0.96	90.59±1.37	88.75±0.43	89.14±0.88	92.87±0.47	3.32	5.67
	GhostNet	80.67±0.19	90.00±0.82	87.10±1.05	87.99±0.93	92.28±0.47	4.64	1.20
	CD-AttDLV3+	82.28±1.03	90.87±1.39	88.80±0.61	89.14±0.98	92.93±0.79	3.39	1.36
	LCDNet	83.80±0.93	90.85±1.44	90.73±0.18	90.16±0.76	93.50±1.03	3.46	4.74
	CD-CTFM	**85.42±0.10**	**92.46±0.08**	**90.77±0.22**	**91.32±0.12**	**94.52±0.63**	**2.77**	**0.43**

The bold entities represent the best-performing algorithms.

on mIoU, precision, recall, F1-score, and OA, respectively. The quantitative results demonstrate that the proposed CD-CTFM are able to achieve promising cloud detection performance. For nonlightweight models, CD-CTFM also achieves similar results on performance metrics but better on efficiency metrics. For example, mIoU of CD-CTFM is 0.27% lower than CRSNet, OA of CD-CTFM is 0.05% lower than CRSNet, and the parameters and GFLOPS of CRSNet are 15 and 46 times larger than that of CD-CTFM, respectively.

Qualitative results on MODIS dataset are shown in Figs. 11 and 12. Clouds in Fig. 11 have quite different visibility due to various factors, such as altitude. Most models do not perform well when testing on this RSI, especially DeepLabV3+, CDAttDLV3+, and LCDNet, but performance of CD-CTFM is satisfactory. Fig. 12 is almost full of large clouds. CDAttDLV3+ is able to achieve promising performance on most of areas, but it mistakes much broken clouds for background. In contrast, CD-CTFM can effectively detect cloud from RSI covered with broken clouds thanks to the fusion of local and global features and dark channel prior.

V. Discussion

The proposed approach for cloud detection has some advantages compared with other methods.

On the one hand, the integration of CNN and Transformer empowers CD-CTFM with the capacity to recognize ice and snow. When focusing solely on local color information, models tend to misclassify ice and snow as clouds due to the white appearance ice and snow exhibit. For instance, models like CDAttDLV3+, LCDNet, and CloudFCN easily misclassify ice and snow as clouds. Most of their network modules are primarily focused on extracting local information. The MF block in CD-CTFM effectively endows the model with the capability to fuse both local and global information, enabling the model to distinguish between ice and snow from clouds more effectively.

On the other hand, the incorporation of attention mechanism enables an accuracy improvement with a limited parameter utilization. In LWCSAM, the utilization of CAM, which is based on cross dimensional interaction information, is advantageous for channel attention modeling, while the subsequent SAM is able to proficiently model spatial attention. In addition, within the backbone network, the use of AG effectively integrates the dark channel prior into the feature maps. Compared with models, such as GhostNet and DeepLabv3+, that do not employ attention mechanisms, CD-CTFM demonstrates superior performance in key metrics, such as mIoU and F1-score.

However, the proposed method still exhibits certain limitations that need to be addressed in future research. To be specific, CD-CTFM still lags slightly behind in performance metrics, such as mIoU and OA, when compared with certain traditional nonlightweight cloud detection models, such as CRSNet. To address the aforementioned limitations, we can investigate the subsequent avenues for improvement.

1) Designing refined structures for the backbone network to better utilize dark channel prior and modifying the AG between encoder and decoder to better fuse low-level and high-level information.
2) Exploring techniques to improve model generalization ability, such as leveraging large vision language models, in a lightweight way [56].
3) Introducing knowledge distillation and model compression techniques to further reduce computations of our model.

In future research, we will endeavor to address these limitations and propose improved model to further enhance the performance and efficiency of cloud detection model.

VI. Conclusion

In this article, we propose a lightweight network, CD-CTFM, to detect clouds efficiently. A lightweight CNN-Transformer network is utilized as the backbone to extract local and global feature simultaneously, where dark channel prior is introduced to improve performance. Then, we design a LWFPM to fuse multiscale features with global context information. A LWCSAM is integrated between the encoder and the decoder through skip connection. Experimental results on 38-Cloud and MODIS datasets demonstrate that CD-CTFM achieves better or similar performance while decreasing parameters and computations, compared with state-of-art methods.

References

[1] D. Roy et al., "Landsat-8: Science and product vision for terrestrial global change research," *Remote Sens. Environ.*, vol. 145, pp. 154–172, 2014.

[2] F. Li, L. Xin, Y. Guo, D. Gao, X. Kong, and X. Jia, "Super-resolution for GAOFEN-4 remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 1, pp. 28–32, Jan. 2018.

[3] S. N. Goward, J. G. Masek, D. L. Williams, J. R. Irons, and R. Thompson, "The landsat 7 mission: Terrestrial research and applications for the 21st century," *Remote Sens. Environ.*, vol. 78, no. 1/2, pp. 3–12, 2001.

[4] K. Anderson, B. Ryan, W. Sonntag, A. Kavvada, and L. Friedl, "Earth observation in service of the 2030 agenda for sustainable development," *Geo-Spatial Inf. Sci.*, vol. 20, no. 2, pp. 77–96, 2017.

[5] X. Du and H. Wu, "Feature-aware aggregation network for remote sensing image cloud detection," *Int. J. Remote Sens.*, vol. 44, no. 6, pp. 1872–1899, 2023.

[6] J. Song, Z. Yan, Y. Niu, L. Zou, and X. Lin, "Cloud detection method based on clear sky background under multiple weather conditions," *Sol. Energy*, vol. 255, pp. 1–11, 2023.

[7] S. Mahajan and B. Fataniya, "Cloud detection methodologies: Variants and development–A review," *Complex Intell. Syst.*, vol. 6, no. 2, pp. 251–261, 2020.

[8] S. Qiu, Z. Zhu, and B. He, "Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery," *Remote Sens. Environ.*, vol. 231, 2019, Art. no. 111205.

[9] L. Xu, S. Fang, R. Niu, and J. Li, "Cloud detection based on decision tree over tibetan plateau with modis data," *Int. Arch. Photogramm. Remote. Sens. Spatial Inf. Sci*, vol. 39, pp. 535–538, 2012.

[10] H. Ishida, Y. Oishi, K. Morita, K. Moriwaki, and T. Y. Nakajima, "Development of a support vector machine based cloud detection method for modis with the adjustability to various conditions," *Remote Sens. Environ.*, vol. 205, pp. 390–407, 2018.

[11] C. E. Bulgin, J. P. D. Mittaz, O. Embury, S. Eastwood, and C. J. Merchant, "Bayesian cloud detection for 37 years of advanced very high resolution radiometer (AVHRR) global area coverage (GAC) data," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 97.

[12] W. Li, J. Wang, Y. Gao, M. Zhang, R. Tao, and B. Zhang, "Graph-feature-enhanced selective assignment network for hyperspectral and multispectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 11, 2022, Art. no. 5526914.

[13] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, "Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 3, 2023, Art. no. 5501212.

[14] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 16, 2022, Art. no. 5412012.

[15] R. Hang, G. Li, M. Xue, C. Dong, and J. Wei, "Identifying oceanic eddy with an edge-enhanced multiscale convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9198–9207, Oct. 19, 2022.

[16] A. Francis, P. Sidiropoulos, and J.-P. Muller, "CloudFCN: Accurate and robust cloud detection for satellite imagery with deep learning," *Remote Sens.*, vol. 11, no. 19, 2019, Art. no. 2312.

[17] M. Luotamo, S. Metsämäki, and A. Klami, "Multiscale cloud detection in remote sensing images using a dual convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4972–4983, Jun. 2021.

[18] J. Wang, W. Li, M. Zhang, and J. Chanussot, "Large kernel sparse convnet weighted by multi-frequency attention for remote sensing scene understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 16, 2023, Art. no. 5626112.

[19] 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, 2020, Art. no. 1056.

[20] L. Zhang, J. Sun, X. Yang, R. Jiang, and Q. Ye, "Improving deep learning-based cloud detection for satellite images with attention mechanism," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Dec. 10, 2021, Art. no. 6005505.

[21] C. Zhang, L. Weng, L. Ding, M. Xia, and H. Lin, "CRSNet: Cloud and cloud shadow refinement segmentation networks for remote sensing imagery," *Remote Sens.*, vol. 15, no. 6, 2023, Art. no. 1664.

[22] X. Yao, Q. Guo, and A. Li, "Light-weight cloud detection network for optical remote sensing images with attention-based DeepLabv3 architecture," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3617.

[23] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[24] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[25] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, 2022.

[26] M. Chen et al., "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[28] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two pure transformers can make one strong gan, and that can scale up," in *Proc. Neural Inf. Process. Syst.*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235421596

[29] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7464–7473.

[30] L. Liu et al., "Scotch and soda: A transformer video shadow detection framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10449–10458.

[31] R. Karim, H. Zhao, R. P. Wildes, and M. Siam, "MED-VT: Multiscale encoder-decoder video transformer with application to object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6323–6333.

[32] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[33] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12077–12090, 2021.

[34] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, " Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7262–7272.

[35] B. Zhang, Y. Zhang, Y. Li, Y. Wan, and Y. Yao, "CloudViT: A lightweight vision transformer network for remote sensing cloud detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Dec. 29, 2022, Art. no. 5000405.

[36] R. Singh, M. Biswas, and M. Pal, "A transformer-based cloud detection approach using sentinel 2 imageries," *Int. J. Remote Sens.*, vol. 44, no. 10, pp. 3194–3208, 2023.

[37] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.

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

[39] O. Oktay et al., "Attention u-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[40] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[41] 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.

[42] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[43] Q. He, X. Sun, Z. Yan, and K. Fu, "DABNet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 5, 2021, Art. no. 5601216.

[44] 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.*, 2018, pp. 4510–4520.

[45] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1580–1589.

[46] K. Hu, D. Zhang, M. Xia, M. Qian, and B. Chen, "LCDNet: Light-weighted cloud detection network for high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4809–4823, Jun. 10, 2022.

[47] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[48] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 552–568.

[49] D. Yu, H. Wang, P. Chen, and Z. Wei, "Mixed pooling for convolutional neural networks," in *Proc. 9th Int. Conf. Rough Sets Knowl. Technol.*, 2014, pp. 364–375.

[50] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[51] S. Mohajerani, T. A. Krammer, and P. Saeedi, "A cloud detection algorithm for remote sensing images using fully convolutional neural networks," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process.*, 2018, pp. 1–5.

[52] X. Li, X. Yang, X. Li, S. Lu, Y. Ye, and Y. Ban, "GCDB-UNet: A novel robust cloud detection approach for remote sensing images," *Knowl.-Based Syst.*, vol. 238, 2022, Art. no. 107890.

[53] D. Ma, R. Wu, D. Xiao, and B. Sui, "Cloud removal from satellite images using a deep learning model with the cloud-matting method," *Remote Sens.*, vol. 15, no. 4, 2023, Art. no. 904.

[54] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*2015, pp. 234–241.

[55] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[56] Y. Hu, J. Yuan, C. Wen, X. Lu, and X. Li, "RSGPT: A remote sensing vision language model and benchmark," 2023, *arXiv:2307.15266*.

Wenxuan Ge is currently working toward the B.S. degree in computer science from Nanjing Forestry University, Nanjing, China.

His research interests include computer vision and remote sensing image processing.

Xubing Yang received the B.S. degree in mathematics from Anhui University, Hefei, China, in 1997, and the M.S. and Ph.D. degrees in computer applications from the Nanjing University of Aeronautics & Astronautics, Nanjing, China, in 2004 and 2008, respectively.

Since 2008, he joined Nanjing Forestry University where he is currently working as an Associate Professor with Computer Science and Engineering Department. His research interests include pattern recognition, machine learning, and neural computing.

Rui Jiang was born in Jiangsu, China in 1985. He received the B.S. and Ph.D. degrees in electronic engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2007 and 2013, respectively.

In 2013, he joined the Department of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, where he is now an Associate Professor. His research interests include machine learning and Internet of Things.

Wei Shao received the B.Sc. and M.Sc. degrees in information and computing science from the Nanjing University of Technology, Nanjing, China, in 2009 and 2012, respectively, and the Ph.D. degree in software engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, in 2018.

He is currently an Associate Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. His research interests include machine learning, computer vision, and bioinformatics.

Li Zhang received the B.S. degree in computer science from the Changsha University of Science and Technology, Changsha, China, and the M.S. and Ph.D degrees in computer science from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2007, 2010, and 2015, respectively.

He is currently an Associate Professor with the College of Information Science and Technology, Nanjing Forestry University, Nanjing. His research interests include machine learning, remote sensing, and medical imaging analysis.