Deep Matting for Cloud Detection in Remote Sensing Images

Wenyuan Li, Zhengxia Zou*, and Zhenwei Shi* *Member, IEEE*

***Abstract*—Cloud detection, as an important pre-processing operation for remote sensing images analysis, has received increasing attention in recent years. Most of the previous cloud detection methods consider the detection as a pixel-wise image classification problem (cloud vs background), which inevitably leads to a category-ambiguity when dealing with the detection of thin clouds. In this paper, starting from the remote sensing imaging mechanism on cloud images, we re-examine the cloud detection under a totally different point of view, i.e. to formulate cloud detection as a mixed energy separation between foreground and background images. This process can be further equivalently implemented under a deep learning based image matting framework with a clear physical significance. More importantly, the proposed method is capable to deal with three different but related tasks, i.e. "cloud detection", "cloud removal", and "cloud cover assessment", under a unified framework. Experimental results on three satellite image datasets demonstrate the effectiveness of our method, especially for those hard but common examples in remote sensing images, such as the thin and wispy cloud.**

***Index Terms*—Remote sensing image, Cloud detection, Image matting, Deep learning, Convolutional Neural Networks**

I. Introduction

THE rapid development of remote sensing technology in recent years has opened a door for people to better understand the earth. Optical remote sensing, as a large family of remote sensing imaging technologies, has been extensively applied in many areas in recent years, such as land monitoring, disaster relief, military reconnaissance, etc. Despite its wide applications, the fact that the ground objects in an optical remote sensing image usually being covered by clouds has greatly limited the usage of optical images and increased the difficulty of image analysis. As reported by C. Stubenrauch et al [1], on average, more than 50% of the earth's surface is covered by clouds every day. The research on cloud detection is of great importance and has received increasing attention in recent years.

The work was supported by the National Key R&D Program of China under the Grant 2017YFC1405605, the National Natural Science Foundation of China under the Grant 61671037, the Beijing Natural Science Foundation under the Grant 4192034 and the National Defense Science and Technology Innovation Special Zone Project. *(Corresponding author: Zhengxia Zou (e-mail: zzhengxi@umich.edu) and Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn))*

Wenyuan Li and Zhenwei Shi are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China.

Zhengxia Zou is with the Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

Most of the previous cloud detection methods frame the cloud detection as a semantic segmentation process, i.e. to generate a binary mask for the foreground (cloud) and background image regions under a pixel-wise classification (cloud vs background) paradigm. Some commonly used methods include the band grouping/thresholding methods [2–7], the traditional image segmentation methods [8–10], and the recent popular deep learning based segmentation methods [11–17]. As these methods are mostly borrowed from the computer vision community without considering the mechanism behind the remote sensing imaging, the pixel-wise classification will inevitably lead to a category-ambiguity in terms of detecting thin clouds. Therefore, a common defect of these methods lies that they are not able to deal with thin clouds properly.

The cloud in an image usually presents in a mixed form of visual appearance with the cloud itself and the ground objects underneath. Cloud may have various thicknesses and present in various transparency. As the energy received by an imaging sensor can be approximated by a linear combination of the reflectance of the clouds and the ground objects, a remote sensing image can be considered as a superimposition of a "clouds layer" and a "background layer". Therefore, cloud detection is naturally a mixed image separation problem.

In the image processing field, image matting [18–20] refers to a group of methods that aim to extract foreground objects from an image, which have very similar idea compared to the above descriptions. Matting is an important task in image and video editing. Some related works can be traced back to the 1990s [18]. Traditional image matting methods can be divided into two large groups of families: 1) sampling-based methods [21–23] and 2) propagation-based methods [19, 24, 25], where the former one produces the matte by a predefined metric given a set of the foreground and background sampling regions, and the latter one reformulates the prediction as the propagation of the foreground and background regions. The matting task usually produces a "matte" that can be used to separate the foreground from the background in a given image, which naturally corresponds to the cloud detection process. To this end, starting from the "mixed energy imaging model (to be introduced)" of cloud images, we propose a brand new cloud detection paradigm called "Deep Cloud Matting", which reformulates the cloud detection as a mixed energy separation between foreground and background images and can be equivalently implemented under an image matting framework. As the recent advances in deep learning technology have greatly promoted the progress of image matting [20, 26], we take advantage of the deep Convolutional Neural Network (CNN) and the multi-task learning framework so that the matting-based

cloud detection can be implemented by learning to predict multiple outputs, including the "cloud reflectance map" and the "cloud opacity map", under a unified deep convolutional architecture. The proposed framework is scalable, flexible, has clear physical significance, and can be jointly trained in an end-to-end fashion. In particular, we consider the traditional cloud detection model as a special case of our method.

To improve the prediction of some hard examples such as thin and wispy clouds, the attention mechanism is further integrated into our method. Attention was originally introduced in machine translation to improve the performance of an Encoder-Decoder RNN model by taking into account the input from several time steps to make one prediction [27]. In a CNN-based model, the introduction of mechanism attention helps investigate the spatial correlations of different feature locations, and now has been widely used in many computer vision tasks such as object detection [28], optical character recognition [29], image captioning [30, 31], etc. In the proposed method, by introducing the cloud foreground map (indicating where there is cloud) as the pixel-wise attention weights to the loss functions of the other two tasks, the learning of the cloud reflectance and opacity can be well instructed. The above design has two advantages. The first one is it makes the training process focus more on those hard examples. The second one is, it is helpful to reduce the correlation between the predictions of cloud reflectance and opacity. Our method predicts not only the detailed cloud region but also the accurate cloud reflectance and opacity which can be further used for cloud removal and cloud cover assessment. Experimental results on three satellite image datasets have demonstrated the effectiveness of our method.

The contributions of this paper are summarized as follows:

1) In most previous cloud detection literature, the detection is framed as a pixel-wise classification process without considering the nature of the remote sensing imaging. This pixel-wise classification paradigm results in an inherent defect when dealing with the detection of thin and wispy clouds. This paper proposes a brand new cloud detection framework, which reformulates cloud detection as a foreground-background energy separation process. This idea can be further implemented under a classical image matting framework which is derived from a mixed energy imaging model of cloud images.
2) To improve the detection of some "hard examples" such as the thin and wispy cloud, the attention mechanism is introduced in our method to reduce coupling between tasks and make the learning process focus more on those hard examples.
3) In previous remote sensing literatures, cloud detection [2–17], cloud cover assessment [32–34] and cloud removal [35–42] are investigated separately despite the high correlation between them. This inhibits joint optimization and makes the implementation of the methods highly complicated. Instead of designing individual algorithms, the proposed method deals with the three tasks under the same framework and is trained in an end-to-end fashion.

The rest of this paper is organized as follows. In section II we introduce the mixed energy imaging model of cloud images and how cloud detection is formulated under an image matting framework. In section III, we give a detailed introduction to our proposed method, including network configuration, multi-task loss function, and implementation details. In section IV, we introduce the dataset used in our experiments. Some experimental results are given in section V. In section VI, we discuss the drawbacks and limitations of our method and the conclusions are drawn in section VII.

II. Mixed Energy Imaging Model of Cloud Image

When a satellite or an aircraft flies over a cloud-covered area, the onboard imaging sensors receive the reflectance energy of ground objects and the cloud at the same time. The amount of energy received by the sensors per unit of time can be approximately considered as a linear combination of the three terms [43, 44], 1) the reflectance energy of the clouds, 2) reflectance energy of ground objects without the interference of the clouds, and 3) the radiation of ground objects, as shown in Fig. 1. This process can be described as follows:

$$\begin{aligned} E &= E_{CR} + (1-\alpha)(E_{BR} + E_{BE}) \\ &\approx E_{CR} + (1-\alpha)E_{BR}, \end{aligned} \tag{1}$$

where E represents the total energy received by the sensor, E_{CR} represents the energy reflected by the clouds, E_{BR} represents the reflectance energy of the ground objects without the occlusion of clouds, and E_{BE} represents the radiation of ground objects, which can be usually neglected for visible bands. α is an attenuation factor of the ground reflectance due to the occlusion of clouds ($0 \leq \alpha \leq 1$). The larger the α is, the thicker the cloud will be: $\alpha = 0$ means there is no cloud while $\alpha = 1$ means the ground object is completely occluded by the cloud. We refer to the above model as the "Mixed Energy Imaging Model" of cloud images.

To this end, a remote sensing image $I(x)$ can be generally expressed as a linear combination of a cloud reflectance map $R_c(x)$ and a background reflectance map $R_b(x)$:

$$I(x) = R_c(x) + [1-\alpha(x)]R_b(x) \tag{2}$$

where x represents the pixel locations in an image.

According to the above model, we deal with three different problems, i.e. cloud detection, cloud cover assessment, and cloud removal, under a unified framework.

- **Task 1: Cloud Detection.**

As the cloud reflectance map $R_c(x)$ defines how much energy is reflected by clouds per unit time, the cloud detection task can be thus considered as the learning of a mapping from the input cloud image to the cloud reflectance map: $I(x) \mapsto R_c(x)$. When $\alpha(x) = 1$ and $R_c(x)$ is set to binary values (i.e, 0 and 1), the prediction will degenerate to a traditional cloud detection method, which simply ignores the reflectance.

- **Task 2: Cloud Cover assessment**

As $\alpha(x)$ corresponds to how much of the reflectance energy of ground objects have been suppressed due to the occlusion of clouds, we define it as the "thickness" of the cloud. Therefore, the cloud cover assessment tasks can be considered as the

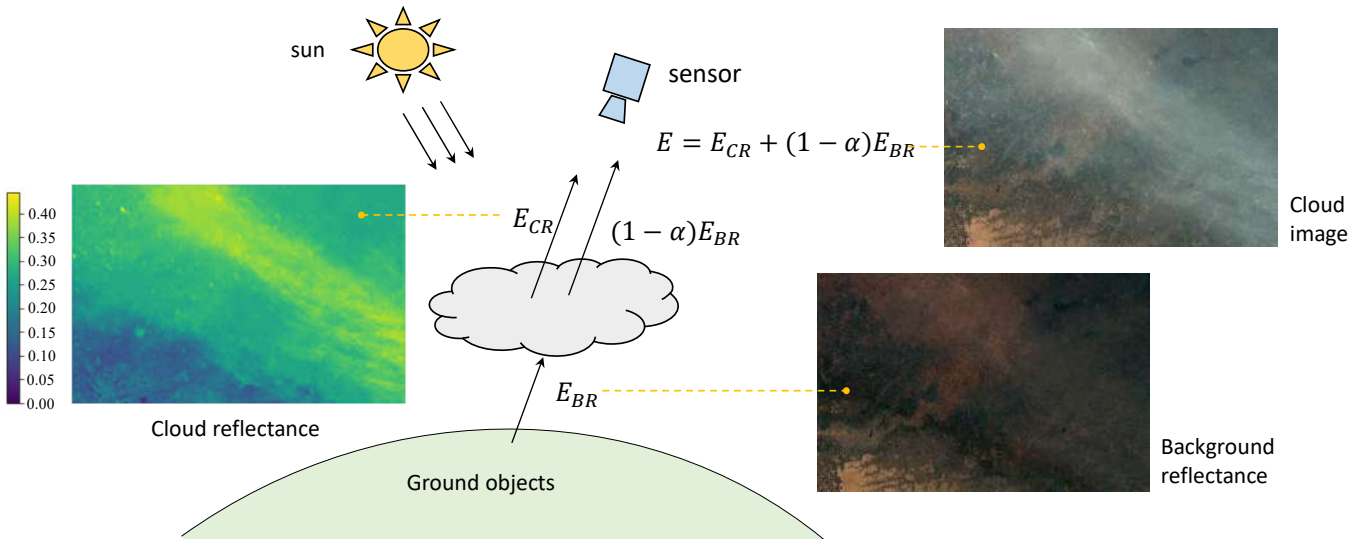


Fig. 1. An illustration of the "Mixed Energy Imaging Model" of cloud images [43, 44]. The energy received by a sensor per unit time can be approximately considered as a linear combination of the reflectance energy of the cloud E_{CR} and the ground objects E_{BR}.

learning of a mapping from the input to the cloud opacity map: $I(x) \mapsto \alpha(x)$.

- **Task 3: Cloud Removal**

Cloud removal is essentially a background image recovery problem. According to (2), it is easy to obtain the background reflectance image as follows:

$$R_b(x) = \frac{I(x) - R_c(x)}{1 - \alpha(x)}, \quad \alpha(x) \neq 1. \tag{3}$$

This means once we have obtained $R_c(x)$ and $\alpha(x)$, the cloud can be easily removed and background images can be thus recovered. It should be noticed that when $\alpha(x) = 1$, the ground is completely covered by clouds thus can not be recovered.

III. Deep Cloud Matting

In recent years, deep Convolutional Neural Network (CNN) has played a central role in many computer vision tasks such as image classification [45, 46], object detection [47–49], etc. CNN has also been extensively applied in a variety of remote sensing tasks such as object detection [50], scene labeling [51], remote sensing image captioning [52], and cloud detection [12], etc. A CNN model learns high-level image representations with better discrimination and robustness by constructing multiple layers of neural networks as opposed to those in traditional methods, where image features are designed manually. In this paper, we build our network based on a CNN architecture.

A. Networks Architecture

We formulate the learning and prediction of the cloud reflectance map $R_c(x)$ and the cloud opacity map $\alpha(x)$ under a multitask learning framework. The networks consist of an encoder and multiple decoders, where the encoder aims to learn high-level feature representations of the input image, and the decoders aim to predict multiple desired outputs throughout multiple heads, as shown in Fig. 2.

To improve the prediction of hard examples, e.g. thin cloud, we further integrate the attention mechanism to our model by introducing an additional attention branch to the decoders. The decoder, therefore, consists of three output branches: the first one aims to predict cloud reflectance map $R_c(x)$, the second one aims to predict the cloud opacity map $\alpha(x)$, and the last one, i.e. the attention branch, aims to generate binary cloud mask for cloud pixels, meanwhile, to instruct the learning of the previous two branches and help them concentrate more on difficult regions. Specifically, the attention branch takes in the foreground mask $A(x)$ (where 1 is the cloud-covered pixels and 0 is the no cloud pixels) as its ground truth reference. The predicted attention scores are employed as the pixel-wise weights of the loss functions of the other two tasks (to be introduced in the next subsection).

As a CNN model consists of a series of convolutional and pooling layers, features in deeper layers will have stronger invariance but less equivariance. Although this could be beneficial to category recognition, it usually suffers the loss of details such as the object's edge and boundary. To improve the learning of features with both high-level semantics and local details, the feature fusion is employed in our method by introducing the skip-connection across different layers from the encoder to the three decoders, as shown in Fig. 2. The above networks can be end-to-end trained with the help of a multi-task loss function.

B. Multi-task Loss Function

Our loss function L consists of three parts, 1) the loss of attention branch $L(A(x))$, 2) the loss of cloud reflectance prediction $L(R_c(x))$, and 3) the loss of cloud opacity prediction $L(\alpha(x))$:

$$L = \sum_x (\gamma_1 L(A(x)) + \gamma_2 L(R_c(x)) + \gamma_3 L(\alpha(x))), \tag{4}$$

where γ_1, γ_2, and γ_3 are three positive coefficients for balancing the learning weights between three tasks. x is the pixel

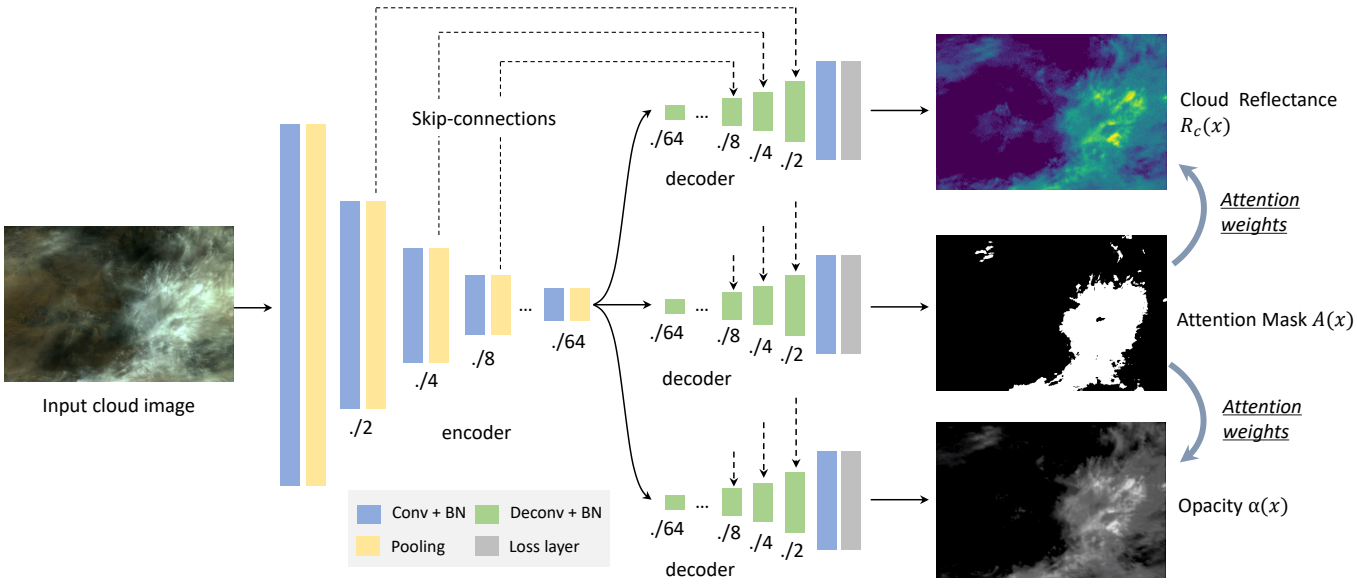


Fig. 2. An overview of the architecture of the proposed networks. The networks consist of an encoder for learning feature representation and three decoders for predicting multiple outputs including cloud reflectance, cloud opacity, and the cloud mask (attention map).

locations. The details of the three loss functions are described as follows.

- **Attention Branch**

The prediction of the attention map is essentially a pixel-wise binary classification process. We use the binary cross-entropy loss as its loss function:

$$L(A(x)) = -\widehat{A}(x)log(A(x)) - (1-\widehat{A}(x))log(1-A(x)), \tag{5}$$

where $A(x)$ and $\widehat{A}(x)$ are the predictions (probabilities) and binary labels of the attention branch.

- **Cloud Reflectance Branch**

As the reflectance of a cloud pixel is a continuous value, we formulate the prediction of cloud reflectance as a regression problem. To obtain more robust prediction, especially for the the outliers, e.g, some thin clouds with low reflectance, we use the $L1$ (absolute value) function as the loss of this branch:

$$L(R_c(x)) = A(x)|R_c(x) - \widehat{R_c}(x)|, \tag{6}$$

where $R_c(x)$ and $\widehat{R_c}(x)$ represent the predicted and true cloud reflectance value respectively. $A(x)$ is used as pixel-wise weights of the loss function to make the learning focus on cloud regions.

- **Cloud Opacity Branch**

Cloud opacity can be also learned through a regression process. The loss function is defined as follows:

$$L(\alpha(x)) = A(x)|\alpha(x) - \widehat{\alpha}(x)| \tag{7}$$

where $\alpha(x)$ and $\widehat{\alpha}(x)$ represent the predicted and true cloud opacity values. $A(x)$ is used as pixel-wise weights of the loss function to make the learning focus on cloud regions.

TABLE I
A DETAILED CONFIGURATION OF OUR NETWORKS.

	Layer	Input	Ker	Stride	#Ker	$\sigma(\cdot)$
Encoder	conv_pool1	image	3x3	2	64	ReLu
	conv_pool2	conv_pool1	3x3	2	128	ReLu
	conv_pool3	conv_pool2	3x3	2	256	ReLu
	conv_pool4	conv_pool3	3x3	2	256	ReLu
	conv_pool5	conv_pool4	3x3	2	512	ReLu
	conv_pool6	conv_pool5	3x3	2	512	ReLu
	conv7	pool6	3x3	1	512	None
Decoder: 1~3	deconv1	conv7	3x3	2	512	ReLu
	deconv2	deconv1 + conv6	3x3	2	512	ReLu
	deconv3	deconv2 + conv5	3x3	2	256	ReLu
	deconv4	deconv3 + conv4	3x3	2	128	ReLu
	deconv5	deconv4 + conv3	3x3	2	128	ReLu
	deconv6	deconv5 + conv2	3x3	2	64	ReLu
	conv_output	deconv6 + conv1	3x3	1	1	None

C. Implementation details

Default Setting. We build a seven-layer convolutional network as our encoder and another seven-layer convolutional network as our decoder. The detailed configuration of our networks is shown in Table I. The columns "Ker" means the size of the convolutional kernel, "Stride" means the convolutional or pooling stride, "#Ker" means the number of filters, "$\sigma(\cdot)$" means the kind of the nonlinear activation layers. The rows "conv" means convolution operation, "deconv" means deconvolution operation [53] (a.k.a. the transposed convolution), which is used for up-sampling the feature map. Apart from the output layer, Batch Normalization [54] is embedded in all convolution and deconvolution layers to speed up training. The decoders of the three tasks have the same architecture. Since the output map of our attention branch does not require a high accuracy, a rough guide is enough for the training. Therefore, we set $\gamma_1 = 1$, $\gamma_2 = \gamma_3 = 10$. We used the Adam optimizer [55] with the learning rate of 10^{-4} for training. We train at

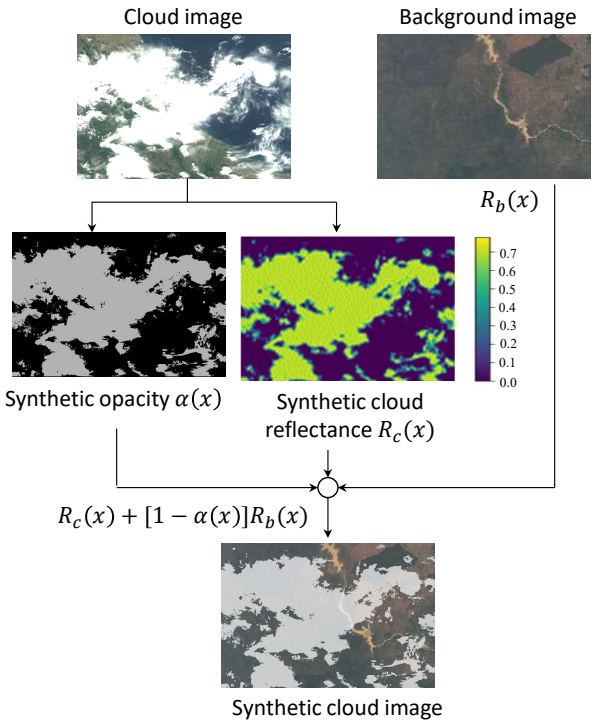


Fig. 3. An illustration of the synthetic training data generation process.

batch size 3 for 16 epochs.

Data Augmentation. In order to increase the diversity of training data and reduce the gap between real data and synthetic data, extensive data augmentation methods are adopted. For an image of a size 512×512, we first rotate it with the angle randomly selected from [0°, 90°, 180°, 270°]. Then a patch with a size of 410×410 is randomly cropped from the rotated image and resize to 512×512. Finally, we randomly flip the augmented image.

IV. Experimental Dataset

Our experimental dataset consists of 328 remote sensing images that are captured by the Gaofen-1 satellite. The original images have two resolutions: 8m/pixel from the panchromatic and multi-spectral (PMS) sensor with the image size of about 4500×4500 pixels, and 16m/pixel from the wide field-of-view (WFV) sensors with the image size of about 12000×13000 pixels. The statistics of our dataset are given in Table II. There are 72 images in our training set and 256 images in our testing set. Each image has been down-sampled to a fixed size, 512×512 pixels, for training. Since the raw pixel of the original GF-1 data has four bands (blue, green, red, and infrared) and is 16-bit depth, all images are converted to 8-bit RGB images before they are fed into the networks. Apart from that, we do not perform any other pre-processing operations. The dataset covers the most type of ground features, such as city, ocean, plains, plateaus, glacier, desert, gobi, etc. Each image in our dataset has been manually labeled with a pixel-wise binary cloud mask as the ground truth of the attention branch. During the labeling process, if the background details can be clearly observed through the clouds, we consider the clouds as "thin clouds", otherwise we consider them as "thick clouds". Besides, if more than half of the clouds pixels in an image are thin clouds, this image is considered as a thin cloud image, otherwise, it is considered as a thick cloud image.

TABLE II
A Summary of our GF-1 experimental dataset.

Image Info.	# total imgs	328
	img size (pxl)	4,500×4,500, 12,000×13,400
	resolution	8m/pxl, 16m/pxl
	source	GaoFen-1 PMS and WFV
training set	# imgs	72
	# thin cloud imgs	0
	# thick cloud imgs	38
	# no cloud imgs	34
testing set	# imgs	256
	# thin cloud imgs	99
	# thick cloud imgs	134
	# no cloud imgs	23

Note that it is impracticable to manually annotate their accurate ground truth values. This is because the cloud reflectance and opacity are both in continuous values. In image matting, a current solution to this problem is to use synthetic data [20, 26]. We have followed this idea to generate a set of synthetic images with their "ground truth" labels for cloud reflectance maps, opacity maps, and attention maps. The thick clouds images (where $\alpha(x) \approx 1$) and background images (with no clouds, where $\alpha(x) \approx 0$) in our training set are used to generate the synthetic images and their ground truth maps. The synthetic data generation process is shown in Fig. 3. We use the image regions that completely covered by thick clouds as the ground truth cloud reflectance of synthetic images. We use the images without any clouds as the ground truth background reflectance of the synthetic images. Then, the synthetic image can be generated by performing a linear combination of the clouds and backgrounds based on the Eq. (2), where a random opacity value is generated as the combination weights. To increase the diversity of the synthetic images, the background images and cloud reflectance maps are randomly rotated, flipped and cropped. We select 38 background images and 34 thick clouds images from the training set to synthesize 10405 images for training the cloud matting networks. We select 23 background images and 43 thick clouds from the testing set to synthesize 5934 images for evaluating cloud detection and cloud removal accuracy. Since we train our networks with synthetic images but test on real ones. To obtain more convincing results, we need more real images for the test phase. Therefore, we leave more images in the testing set.

In addition to our GF-1 dataset, we also test on two public cloud detection datasets, the GF1_WHU dataset [5], and the Landsat-8 dataset [56]. The GF1_WHU dataset [5] consists of 108 images and the Landsat-8 dataset [56] consists of 96 images. Since we do not use images of the two datasets to train our model. We treat all images in these two datasets as the test ones. And all images in the above two datasets are converted to 8-bit images and down-sampled to 512×512 pixels. Besides, because the networks in our method are

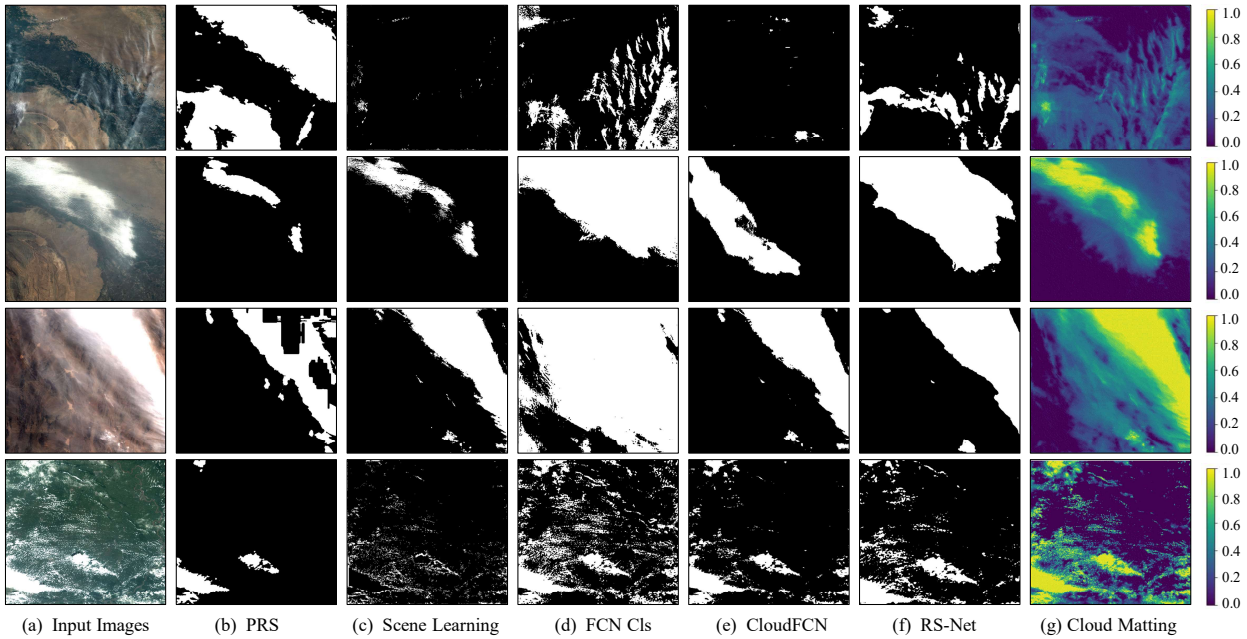


Fig. 4. (Better viewed in color) Some examples of cloud detection results of comparison methods. The first column shows the input cloud images. The second to sixth columns are the results of the comparison methods. The last column is the predicted cloud reflectance of our method (Cloud Matting + Proposed Networks).

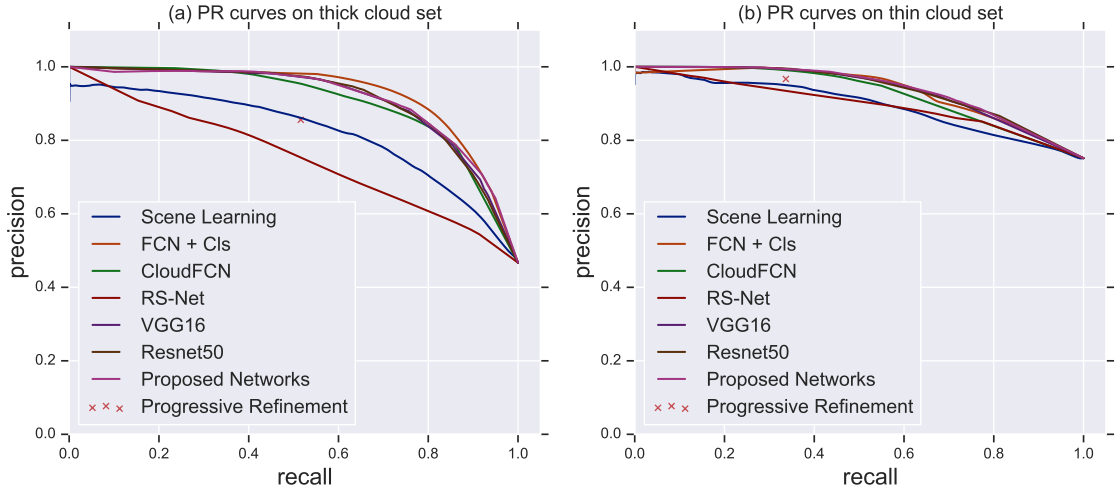


Fig. 5. (Better viewed in color) The precision-recall curves of different cloud detection methods. (a) Results on thick-cloud images. (b) Results on thin-cloud images.

trained on RGB images, when testing on these two datasets, we have also selected these three bands accordingly.

V. Experimental Results and Analysis

For the cloud detection task, the Precision-Recall (PR) curve and the "Average Precision (AP)" score are used as our evaluation metrics. The PR curve shows the relationship between the detection precision and recall rate by changing different thresholds on a detection output:

$$\text{Precision} = \frac{N_c}{N_c + N_f}, \quad \text{Recall} = \frac{N_c}{N_t}, \tag{8}$$

where N_t is the total number of cloud pixels in groundtruth, N_c is the number of correct detection of cloud pixels, and N_f is the number of false alarm pixels.

For cloud reflectance and opacity prediction tasks, three different metrics are used including the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE), which are defined as follows:

$$\begin{aligned} \text{MAE} &= \sum_i |y_i - \widehat{y}_i|/N, \\ \text{MSE} &= \sum_i (y_i - \widehat{y}_i)^2/N, \\ \text{MAPE} &= \sum_i |\frac{y_i - \widehat{y}_i}{\widehat{y}_i}|/N, \end{aligned} \tag{9}$$

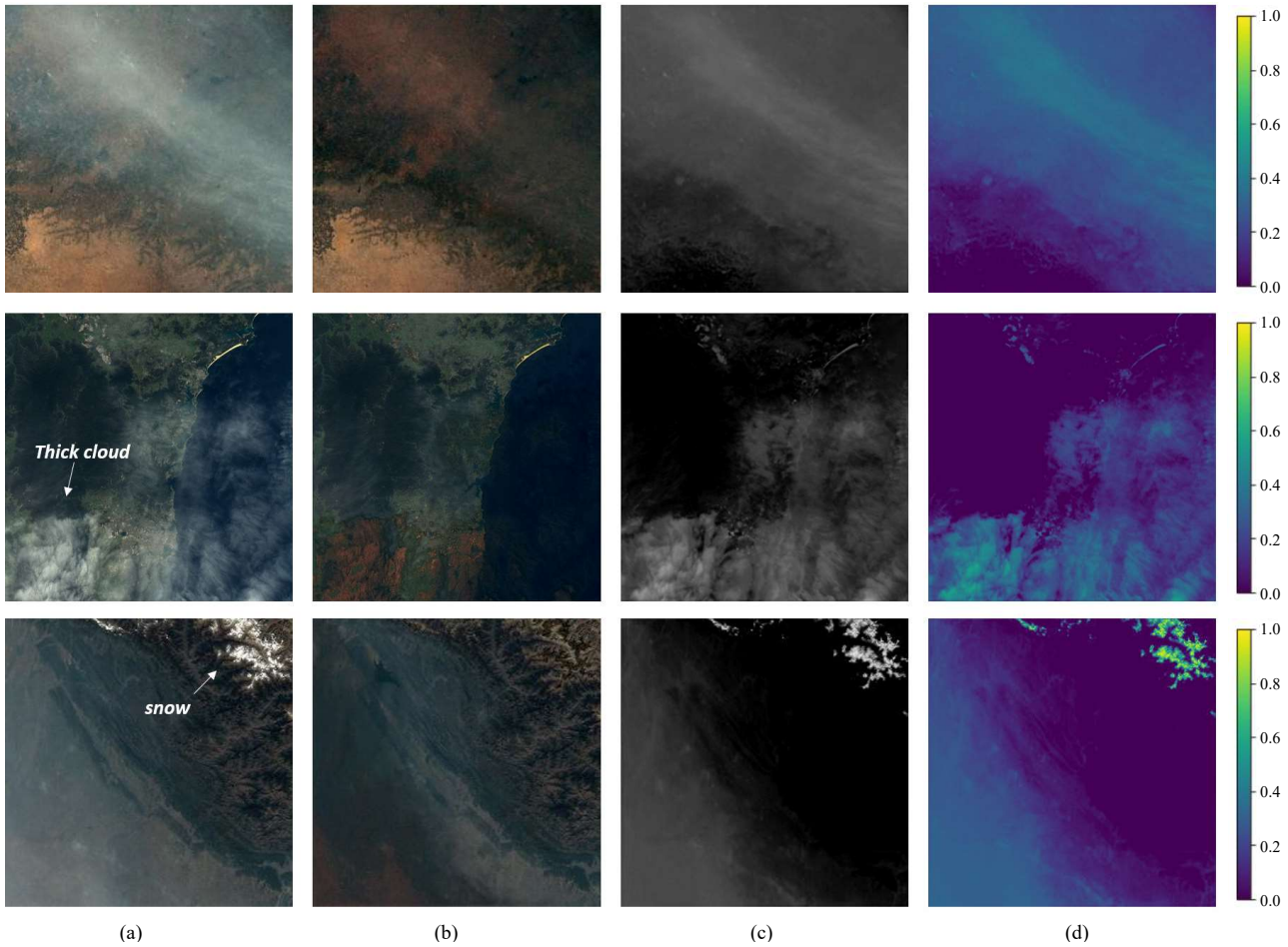


Fig. 6. (Better viewed in color) Some examples of the thin cloud removal result of our method. (a) Input image. (b) Cloud removal result. (c) Predicted opacity (cloud cover assessment result). (d) Predicted cloud reflectance. The last row shows a failure case of our method, there is no cloud in this image but our model has wrongly predicted the snow as the cloud. This may be because our training set does not contain any snow images.

where y and $\widehat{\mathbf{y}}$ are the predicted output and the groundtruth reference, respectively. i is the pixel id and N is the total number of pixels in an image.

A. Detection Results

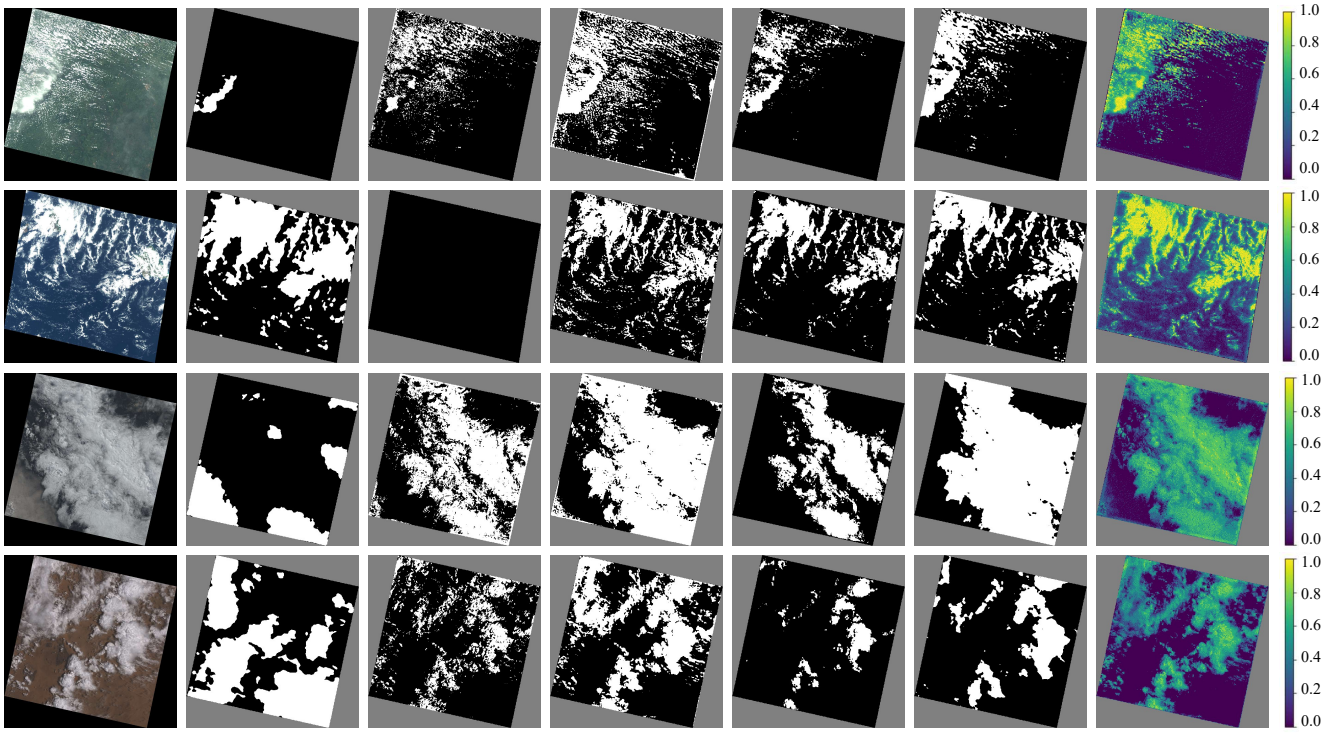
We compare our method with some recent cloud detection methods, including the Progressive Refinement [6], Scene Learning [8], Fully Convolutional Networks + pixel-wise Classification (FCN+CLS) [14], CloudFCN [15] and RS-Net [16], on our test set. We also replace the encoder of the cloud matting network with VGG16 [57] and Resnet50 [46] to evaluate the performance of our framework. Fig. 4 shows some cloud detection examples. The first column shows the input cloud images. The second to sixth columns are the results of the comparison methods. The last column is the predicted cloud reflectance of our method.

Since there are great differences between thin clouds and thick clouds in visual appearance and detection difficulty, we evaluate our method based on the results of thin clouds and thick clouds separately. It can be seen from Table. III and Fig. 5 that our method has a higher cloud detection accuracy especially for those thin cloud images, regardless of encoder's structure. For thick cloud images, our method has comparable detection results on high recall areas with FCN+CLS [14], and their curves are crossed with each other. As the Progressive Refinement [6] only produces binary output masks, we can not compute its AP in Table III and can only mark their precision and recall as a single point in Fig. 5 for comparison. The advantage of our method is not only suggested by the metrics we currently use, but also in terms of the physical mechanism of the cloud images. Although our method is trained solely with synthetic data, the experimental result demonstrates that it still achieves comparable accuracy with other popular cloud detection methods on real data.

Table. IV and Fig. 7 show the cloud detection results on the GF1_WHU dataset [5] and Landsat-8 dataset [56]. It can be seen that our method can not only obtain comparable AP scores with other cloud detection methods on the above two datasets but also extract the cloud reflectance accurately from the image. This suggests that our method can be applied to a variety of satellite platforms.

B. Cloud Removal Evaluation

According to Equation 3, once we have obtained the reflectance image and opacity of the cloud, the background can be easily recovered. In this way, we can evaluate the



(a) Input Images (b) PRS (c) Scene Learning (d) FCN Cls (e) CloudFCN (f) RS-Net (g) Ours

Fig. 7. (Better viewed in color) Some examples of the cloud detection results of comparison methods on the GF1_WHU dataset [5] (the first two rows) and Landsat-8 dataset [56] (the last two rows). The first column shows the input cloud images. The second to sixth columns are the results of the comparison methods. The last column is the predicted cloud reflectance of our method (Cloud Matting + Proposed Networks). The parts marked in gray correspond to the "black areas" in the input image, and the parts marked in black and white correspond background and cloud separately.

TABLE III
A COMPARISON OF CLOUD DETECTION RESULTS OF DIFFERENT METHODS. A HIGHER SCORE SUGGESTS A BETTER RESULT.

Method	AP_{thick}	AP_{thin}
Scene Learning [8]	0.8200	0.8944
FCN+Cls [14]	**0.9253**	0.9332
Progressive Refinement [6]	–	–
CloudFCN [15]	0.9031	0.9261
RS-Net [16]	0.7518	0.8976
Ours (Cloud Matting+VGG16)	0.9110	0.9356
Ours (Cloud Matting+Resnet50)	0.9109	0.9370
Ours (Cloud Matting+Proposed Networks)	0.9139	**0.9382**

TABLE IV
COMPARISON RESULTS OF DIFFERENT METHODS ON THE GF1_WHU DATASET [5] AND THE LANDSAT-8 DATASET [56]. THE "AVERAGE PRECISION (AP)" IS USED AS AN EVALUATION METRIC. A HIGHER SCORE SUGGESTS A BETTER RESULT.

Method	GF1_WHU [5]	Landsat-8 [56]
Scene Learning [8]	0.4237	0.5282
FCN+Cls [14]	0.7594	0.7712
Progressive Refinement [6]	–	–
CloudFCN [15]	0.8105	0.6821
RS-Net [16]	0.5690	0.5724
Ours (Cloud Matting+VGG16)	0.7891	0.6896
Ours (Cloud Matting+Resnet50)	0.8035	0.6981
Ours (Cloud Matting+Proposed Networks)	0.7549	0.7074

TABLE V
A COMPARISON OF DIFFERENT METHODS ON THE CLOUD REMOVAL TASK. LOWER SCORES INDICATE BETTER. THE BEST RESULTS IN EACH ENTRY ARE MARKED AS BOLD.

Method	MAE	MSE	MAPE
Homomorphic Filter [38]	0.3400	0.1444	0.4312
Deformed-Haze [39]	0.0997	0.0169	0.2100
Adaptive Removal [40]	0.0655	0.0089	0.1449
SM-DCP [41]	0.1763	0.0485	0.3652
Ours (Cloud Matting+VGG16)	0.0599	0.0080	0.1180
Ours (Cloud Matting+Resnet50)	0.0648	0.0087	0.1306
Ours (Cloud Matting+Proposed Networks)	**0.0570**	**0.0068**	**0.1140**

performance of cloud removal by calculating the MAE, MSE, and MAPE of the recovered images and their "ground truth". We compare our method with four classical cloud removal methods: Homomorphic Filter [38], Deformed-Haze [39] , Adaptive Removal [40] and sphere model improved DCP (SM-DCP) [41]. We use our synthetic data sets to quantitatively assess the effects of cloud removal. From Table V we can see our method achieves the best cloud removal result.

Fig. 6 shows some examples of cloud removal results for real images, where the column (a) shows the input image, column (b) shows the cloud removal result, column (c) shows the predicted opacity, and column (d) shows the predicted cloud reflectance. It can be seen that the thin cloud has been removed and the ground object has been nicely recovered. In

TABLE VI
ABLATION STUDIES OF CLOUD REFLECTANCE AND OPACITY PREDICTION. ABLATIONS ARE PERFORMED ON 1) ATTENTION MECHANISM, 2) FEATURE FUSION, AND 3) BATCH NORMALIZATION.

Ablations			Cloud Reflectance Prediction			Cloud Opacity Prediction		
Attention	Feature Fusion	Batch-Norm	ε_{MAE}	ε_{MSE}	ε_{MAPE}	ε_{MAE}	ε_{MSE}	ε_{MAPE}
×	×	×	7.44%	2.70%	86.3%	9.86%	4.47%	79.4%
×	×	✓	7.33%	2.28%	85.0%	9.84%	3.83%	78.23%
×	✓	✓	4.90%	0.717%	72.5%	9.55%	2.25%	**68.3%**
✓	✓	✓	**3.04%**	**0.302%**	**71.6%**	**6.56%**	**2.10%**	73.7%

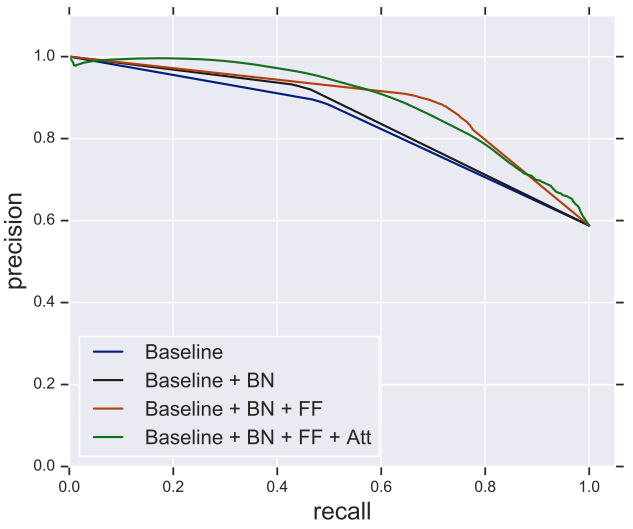


Fig. 8. (Better viewed in color) Ablation studies of cloud detection accuracy on testing data. Ablations are performed on 1) attention mechanism (Att), 2) feature fusion (FF), and 3) batch normalization (BN).

the 4th row of this figure, as the cloud opacity is close to 1, the background under the thick cloud region is hard to recover and thus leading to slight color distortion. The last row of this figure shows a failure case of our method, where there is actually no cloud in this image but our model has wrongly identified the snow as the cloud. This is mainly because of our training set does not contain any snow images, and this can be easily improved by adding more snow images for training. Since this is not the focus of this paper, we do not make any further evaluations on distinguishing cloud and snow pixels.

C. Ablation Analyses

In this experiment, the ablation analyses are made to analyze the importance of each technical component of our method, including 1) attention mechanism, 2) feature fusion and 3) Batch Normalization (BN). The baseline methods are first evaluated, then we gradually integrate these techniques. Table VI shows their comparison results of the reflectance and opacity on the sub-testing set. Fig. 8 shows their comparison of the cloud detection results on the whole testing set. The integration of the first two techniques yields a noticeable improvement of the prediction accuracy of cloud reflectance and opacity, while the BN is trivial for the improvement of accuracy. Nevertheless, the model trained with BN still benefits from a faster convergence speed. The reason behind the improvement is twofold: on one hand, as there is a strong correlation between the cloud reflectance and opacity, the attention map helps to eliminate the coupling between these two tasks, on the other hand, feature fusion is beneficial for predicting a more detailed output for some areas such as some small pieces of clouds and the clouds with sharp edges. Fig. 9 shows the effectiveness of the attention mechanism. We can see that, in the area covered by clouds, the values of its attention map are larger, no matter where it is covered by thin clouds or thick clouds. Therefore, the attention map can be used to guide the training of the network, and make the networks concentrate on the area covered by clouds and obtain better cloud detection results. Meanwhile, predicted values without attention are smaller than those with attention.

D. Cloud Image Montage

The proposed framework can be also used for another important application: cloud image montage, i.e. to transplant the cloud in one image to another background image. This can be simply implemented by the following transformation:

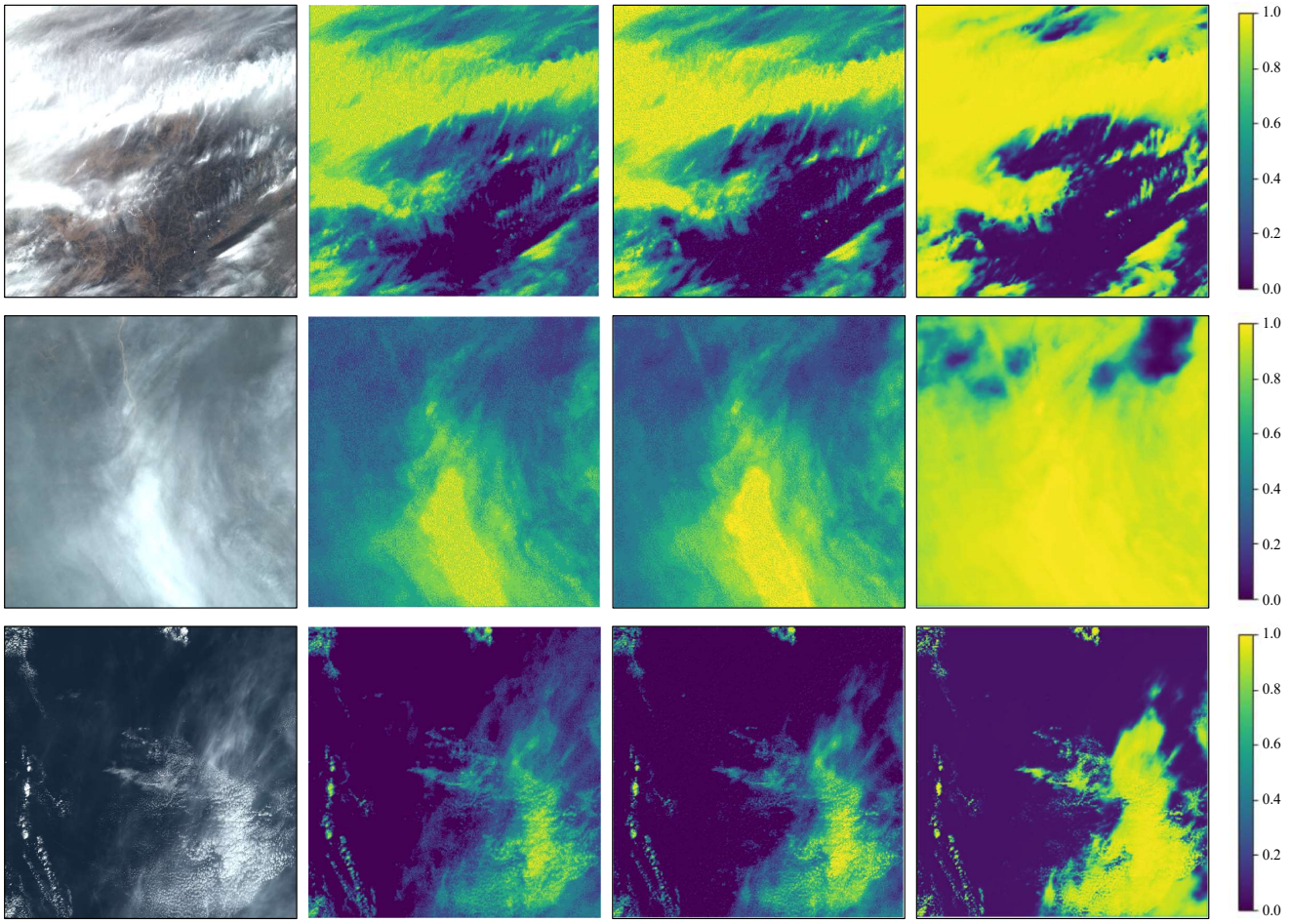
$$I'(x) = R_c^B(x) + [1 - \alpha^B(x)]I^A(x), \tag{10}$$

where $I^A(x)$ is an image from a background image set A, $R_c^B(x)$ and $\alpha^B(x)$ are the predicted cloud reflectance and opacity of an image from a cloud image set B, $I'(x)$ is the generated cloud montage output. Fig. 10 shows some examples of our cloud montage generation results. In this figure, some very high-resolution aerial images from Google Earth are used as the background images, and the clouds from GaoFen-1 images from our test set are used as the foreground cloud styles.

The above process can be considered as a new way of data augmentation, which helps us to generate some hard examples. It may have great potential for improving the performance of some remote sensing applications, such as occluded target detection, scene recognition, and image segmentation.

E. Computational Complexity, Parameters, and Speed

We use three different metrics to compare the computational complexity, parameters, and speed of our method with other CNN-based cloud detection methods. In Table VII, we record the number of model parameters (Params), the number of floating-point operations (FLOPs), and the inference time of different models. We use images of 512×512 pixels to compute FLOPs and inference time, and test on an Nvidia



(a) Input Image (b) Reflectance without Attention (c) Reflectance with Attention (d) Attention Map

Fig. 9. (Better viewed in color) The effectiveness of the proposed attention mechanism. The first column shows the input images. The second and the third columns show the predicted reflectance maps w/o and w/ the help of attention loss. The last column shows the attention maps.

TABLE VII
A COMPARISON ON MODEL PARAMETERS, FLOPS, AND INFERENCE TIME.

Method	**Params**	**FLOPs**	**Inference time**
FCN+Cls [14]	18.5M	12.4G	0.103s
CloudFCN [15]	23.3M	155M	0.054s
RS-Net [16]	7.9M	39M	0.288s
Ours (Cloud Matting+VGG16)	74.0M	25.8G	0.224s
Ours (Cloud Matting+Resnet50)	139M	17.3G	0.410s
Ours (Cloud Matting+Proposed Networks)	79.5M	22.1G	0.210s

GeForce RTX 2080 Ti graphics card. Compared with other methods, our method has more parameters but has comparable inference time with RS-Net [16]. This is because our method uses different branches to predict the reflectance, the opacity and the attention map, which requires more parameters and memories.

VI. DISCUSSION

Although the experimental results on three satellite image datasets demonstrate the effectiveness of our method, it still has some limitations:

1) We do not take the shadow detection into consideration. Since shadows often appear with clouds, it's also important to consider the shadow detection in our framework. In fact, the proposed model formulated in equation 2 can be simplified to $I(x) = [1 - \alpha(x)]R_b(x)$ by setting $R_c(x) = 0$. Then our framework can be naturally extended to shadow detection tasks.
2) As illustrated in the last row of Fig. 6, our method may produce a wrong detection result when there is snow in the image. The reason behind this could be the limited snow samples in our training set. To improve the detection performance on snow covered regions, we may simply add more snow images to our training set.
3) We use a simple data synthesis method to support the training of our model. Despite our primary verification and the promising results obtained, there is still much room for improvement in our method, especially for data synthesis. In our future research, we may design a more sophisticated data synthesis process (e.g. nonuniform opacity by using adversarial training, where we have already obtained some promising results [58]) to generate more realistic cloud images.

Fig. 10. Some examples of the cloud image montage results. The first row shows the input images from image set A. The second row shows the input images from image set B. The third row shows the generation outputs by combining the cloud style of the image set A and backgrounds of the image set B.

VII. Conclusion

We propose a brand new method for cloud detection in remote sensing images which inherently incorporates the cloud imaging mechanism and jointly deals with three different but related problems, cloud detection, cloud cover assessment, and cloud removal, under the same framework. Different from previous methods that consider cloud detection as a pixel-wise binary classification problem, we re-formulate cloud detection as a mixed energy separation problem. Experimental results on three satellite image datasets demonstrate the effectiveness of our method. Besides, the proposed framework can also be used to synthesize cloud images of specific style, which can be considered as a new way of data augmentation and may have great potential for improving the performance of many remote sensing applications such as occluded object detection and recognition.

References

[1] C. Stubenrauch, W. Rossow, S. Kinne, S. Ackerman, G. Cesana, H. Chepfer, L. Di Girolamo, B. Getzewich, A. Guignard, A. Heidinger *et al.*, "Assessment of global cloud datasets from satellites: Project and database initiated by the gewex radiation panel," *Bulletin of the American Meteorological Society*, vol. 94, no. 7, pp. 1031–1049, 2013.

[2] R. R. Irish, J. L. Barker, S. N. Goward, and T. Arvidson, "Characterization of the landsat-7 etm+ automated cloud-cover assessment (acca) algorithm," *Photogrammetric engineering & remote sensing*, vol. 72, no. 10, pp. 1179–1188, 2006.

[3] Z. Zhu and C. E. Woodcock, "Object-based cloud and cloud shadow detection in landsat imagery," *Remote sensing of environment*, vol. 118, pp. 83–94, 2012.

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

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

[6] Q. Zhang and C. Xiao, "Cloud detection of rgb color aerial photographs by progressive refinement scheme," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 11, pp. 7264–7275, 2014.

[7] G. J. Jedlovec, S. L. Haines, and F. J. LaFontaine, "Spatial and temporal varying thresholds for cloud detection in goes imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 6, pp. 1705–1717, 2008.

[8] Z. An and Z. Shi, "Scene learning for cloud detection on remote-sensing images," *IEEE Journal of selected topics in applied earth observations and remote sensing*, vol. 8, no. 8, pp. 4206–4222, 2015.

[9] P. Li, L. Dong, H. Xiao, and M. Xu, "A cloud image detection

method based on svm vector machine," *Neurocomputing*, vol. 169, pp. 34–42, 2015.
[10] N. Greeshma, M. Baburaj, and S. N. George, "Reconstruction of cloud-contaminated satellite remote sensing images using kernel pca-based image modelling," *Arabian Journal of Geosciences*, vol. 9, no. 3, p. 239, 2016.
[11] F. Xie, M. Shi, Z. Shi, J. Yin, and D. Zhao, "Multilevel cloud detection in remote sensing images based on deep learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 8, pp. 3631–3640, 2017.
[12] X. Wu and Z. Shi, "Utilizing multilevel features for cloud detection on satellite imagery," *Remote Sensing*, vol. 10, no. 11, p. 1853, 2018.
[13] Z. Yan, M. Yan, H. Sun, K. Fu, J. Hong, J. Sun, Y. Zhang, and X. Sun, "Cloud and cloud shadow detection using multilevel feature fused segmentation network," *IEEE Geoscience and Remote Sensing Letters*, no. 99, pp. 1–5, 2018.
[14] Y. Zhan, J. Wang, J. Shi, G. Cheng, L. Yao, and W. Sun, "Distinguishing cloud and snow in satellite images via deep convolutional network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1785–1789, 2017.
[15] A. Francis, P. Sidiropoulos, and J.-P. Muller, "Cloudfcn: Accurate and robust cloud detection for satellite imagery with deep learning," *Remote Sensing*, vol. 11, no. 19, p. 2312, 2019.
[16] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sensing of Environment*, vol. 229, pp. 247–259, 2019.
[17] K. Xu, K. Guan, J. Peng, Y. Luo, and S. Wang, "Deepmask: an algorithm for cloud and cloud shadow detection in optical satellite remote sensing images using deep residual network," *arXiv preprint arXiv:1911.03607*, 2019.
[18] D. E. Zongker, D. M. Werner, B. Curless, and D. H. Salesin, "Environment matting and compositing," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1999, pp. 205–214.
[19] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 228–242, 2008.
[20] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.
[21] E. S. Gastal and M. M. Oliveira, "Shared sampling for real-time alpha matting," in *Computer Graphics Forum*, vol. 29, no. 2. Wiley Online Library, 2010, pp. 575–584.
[22] K. He, C. Rhemann, C. Rother, X. Tang, and J. Sun, "A global sampling method for alpha matting," 2011.
[23] E. Shahrian, D. Rajan, B. Price, and S. Cohen, "Improving image matting using comprehensive sampling sets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 636–643.
[24] Y. Zheng and C. Kambhamettu, "Learning based digital matting," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 889–896.
[25] Q. Chen, D. Li, and C.-K. Tang, "Knn matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 9, pp. 2175–2188, 2013.
[26] G. Chen, K. Han, and K.-Y. K. Wong, "Tom-net: Learning transparent object matting from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9233–9241.
[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
[28] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6995–7003.
[29] Z. Wojna, A. Gorban, D.-S. Lee, K. Murphy, Q. Yu, Y. Li, and J. Ibarz, "Attention-based extraction of structured information from street view imagery," *arXiv preprint arXiv:1704.03549*, 2017.
[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.
[31] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and vqa," *arXiv preprint arXiv:1707.07998*, 2017.
[32] A. Cazorla, F. Olmo, and L. Alados-Arboledas, "Development of a sky imager for cloud cover assessment," *JOSA A*, vol. 25, no. 1, pp. 29–39, 2008.
[33] M. Souza-Echer, E. Pereira, L. Bins, and M. Andrade, "A simple method for the assessment of the cloud cover state in high-latitude regions by a ground-based digital camera," *Journal of Atmospheric and Oceanic Technology*, vol. 23, no. 3, pp. 437–447, 2006.
[34] I. D. R. Eberhardt, B. Schultz, R. Rizzi, I. D. Sanches, A. R. Formaggio, C. Atzberger, M. P. Mello, M. Immitzer, K. Trabaquini, W. Foschiera *et al.*, "Cloud cover assessment for operational crop monitoring systems in tropical areas," *Remote Sensing*, vol. 8, no. 3, p. 219, 2016.
[35] C.-H. Lin, P.-H. Tsai, K.-H. Lai, and J.-Y. Chen, "Cloud removal from multitemporal satellite images using information cloning," *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 1, pp. 232–241, 2013.
[36] H. Shen, H. Li, Y. Qian, L. Zhang, and Q. Yuan, "An effective thin cloud removal procedure for visible remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 96, pp. 224–235, 2014.
[37] K. Enomoto, K. Sakurada, W. Wang, H. Fukui, M. Matsuoka, R. Nakamura, and N. Kawaguchi, "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," *arXiv preprint arXiv:1710.04835*, 2017.
[38] Z. Liu and B. R. Hunt, "A new approach to removing cloud cover from satellite imagery," *Computer vision, graphics, and image processing*, vol. 25, no. 2, pp. 252–256, 1984.
[39] X. Pan, F. Xie, Z. Jiang, and J. Yin, "Haze removal for a single remote sensing image based on deformed haze imaging model," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1806–1810, 2015.
[40] F. Xie, J. Chen, X. Pan, and Z. Jiang, "Adaptive haze removal for single remote sensing image," *IEEE Access*, vol. 6, pp. 67 982–67 991, 2018.
[41] J. Li, Q. Hu, and M. Ai, "Haze and thin cloud removal via sphere model improved dark channel prior," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 3, pp. 472–476, 2018.
[42] X. Li, L. Wang, Q. Cheng, P. Wu, W. Gan, and L. Fang, "Cloud removal in remote sensing images using nonnegative matrix factorization and error correction," *ISPRS journal of photogrammetry and remote sensing*, vol. 148, pp. 103–113, 2019.
[43] K. Steffen, J. Key, D. J. Cavalieri, J. Comiso, P. Gloersen, K. S. Germain, and I. Rubinstein, "The estimation of geophysical parameters using passive microwave algorithms," *Microwave remote sensing of sea ice*, vol. 68, pp. 201–231, 1992.
[44] C. Swift and D. Cavalieri, "Passive microwave remote sensing for sea ice research," *Eos, Transactions American Geophysical Union*, vol. 66, no. 49, pp. 1210–1212, 1985.
[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[47] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years:

A survey," *arXiv preprint arXiv:1905.05055*, 2019.
[48] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
[49] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
[50] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1100–1111, 2018.
[51] H. Lin, Z. Shi, and Z. Zou, "Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1665–1669, 2017.
[52] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3623–3634, 2017.
[53] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
[54] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[56] 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 sensing of environment*, vol. 194, pp. 379–390, 2017.
[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[58] Z. Zou, W. Li, T. Shi, Z. Shi, and J. Ye, "Generative adversarial training for weakly supervised cloud matting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 201–210.

Wenyuan Li received his B.S. degree from North China Electric Power University, Beijing, China in 2017. He is currently working toward his doctorate degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include deep learning, image processing, and pattern recognition.

Zhengxia Zou received his B.S. degree and his Ph.D. degree from the Image Processing Center, School of Astronautics, Beihang University in 2013 and 2018. He is now working at the Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, as a postdoc research fellow. His research interests include computer vision, pattern recognition, and remote sensing image analysis. He serves as the PC member/reviewer for several top conferences and top journals, including the NeurIPS, CVPR, AAAI, TIP, SPM, TGRS, etc. He was selected as one of the 2017 best reviewers for the Infrared Physics and Technology. His personal website is http://www-personal.umich.edu/~zzhengxi/.

Zhenwei Shi (M'13) received his Ph.D. degree in mathematics from Dalian University of Technology, Dalian, China, in 2005. He was a Postdoctoral Researcher in the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was Visiting Scholar in the Department of Electrical Engineering and Computer Science, Northwestern University, U.S.A., from 2013 to 2014. He is currently a professor and the dean of the Image Processing Center, School of Astronautics, Beihang University. His current research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi serves as an Associate Editor for the *Infrared Physics and Technology*. He has authored or co-authored over 100 scientific papers in refereed journals and proceedings, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Neural Networks, the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Geoscience and Remote Sensing Letters and the IEEE Conference on Computer Vision and Pattern Recognition. His personal website is http://levir.buaa.edu.cn/.