Generative Adversarial Training for Weakly Supervised Cloud Matting

Zhengxia Zou*
University of Michigan, Ann Arbor

Wenyuan Li
Beihang University

Tianyang Shi
NetEase Fuxi AI Lab

Zhenwei Shi
Beihang University

Jieping Ye
Didi Chuxing & University of Michigan, Ann Arbor

Abstract

The detection and removal of cloud in remote sensing images are essential for earth observation applications. Most previous methods consider cloud detection as a pixel-wise semantic segmentation process (cloud v.s. background), which inevitably leads to a category-ambiguity problem when dealing with semi-transparent clouds. We re-examine the cloud detection under a totally different point of view, i.e. to formulate it as a mixed energy separation process between foreground and background images, which can be equivalently implemented under an image matting paradigm with a clear physical significance. We further propose a generative adversarial framework where the training of our model neither requires any pixel-wise ground truth reference nor any additional user interactions. Our model consists of three networks, a cloud generator G, a cloud discriminator D, and a cloud matting network F, where G and D aim to generate realistic and physically meaningful cloud images by adversarial training, and F learns to predict the cloud reflectance and attenuation. Experimental results on a global set of satellite images demonstrate that our method, without ever using any pixel-wise ground truth during training, achieves comparable and even higher accuracy over other fully supervised methods, including some recent popular cloud detectors and some well-known semantic segmentation frameworks.

1. Introduction

The rapid development of remote sensing technology has opened a door for people to better understand the earth. Remote sensing satellites fly around the earth several times a day, providing up-to-date information for human activities in all walks of life, such as disaster relief, land monitoring, and military reconnaissance. Despite its wide applications, as reported by C. Stubenrauch et al [37], on average of more than half of the earth's surface is covered by clouds every day, which has greatly limited the data accessibility and has increased difficulty in data analysis. The research on cloud detection and removal thus has received great attention in recent years.

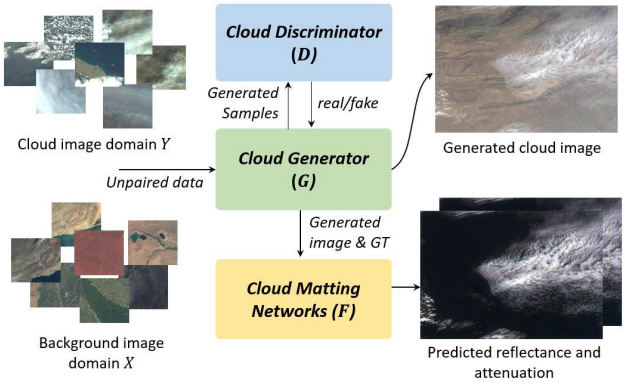

Figure 1. An overview of the proposed method. Our model consists of a cloud generator G, a cloud discriminator D, and a cloud matting network F. On one hand, G and D contradict each other to generate realistic and physically meaningful cloud images. On the other hand, F predicts the cloud reflectance and attenuation so that the background under the cloud can be recovered.

Clouds in an image may visually present various transparency, where in most cases, the energy received by an imaging sensor can be approximated by a linear combination of the reflected energy of the clouds and the ground objects [27, 38]. In recent years, most of the cloud detection methods frame the detection as a pixel-wise classification process (cloud vs background), i.e. to generate binary masks of the predicted foreground (cloud) and background regions. Some commonly used methods include the band grouping/thresholding methods [13, 15, 20, 45, 48, 49], and the semantic segmentation based methods [1, 18, 39, 41, 43, 44]. As most of these methods are borrowed from the computer vision community without considering the mechanism behind the imaging process, the pixel-wise classification based paradigm will inevitably lead to a category-ambiguity in terms of detecting semi-transparent clouds

*Corresponding author: Zhengxia Zou (zzhengxi@umich.edu)

(thin clouds). In addition, current cloud detection and cloud removal methods [8, 21, 24, 28, 34, 40] are separately investigated despite the high correlation between them.

In this paper, we reformulate cloud detection and removal as a mixed energy separation between foreground and background images. This idea can be equivalently implemented under an image matting framework [17, 31, 42, 50] by predicting of multiple outputs, including the "foreground cloud map" and the "alpha matte" (attenuation). Most of the recent image matting methods consider the learning and prediction of the alpha matte under a regression paradigm in a fully supervised way [3, 5, 35, 42]. Although it proves to be effective for traditional matting problems, for a cloud image, it is difficult to obtain the ground truth of foreground map and alpha matte as it involves quantitatively determining some important physical parameters such as cloud reflectance and atmospheric attenuation. To this end, we further propose a generative adversarial training framework to achieve weakly supervised matting of a cloud image by incorporating the physics behind it. Particularly, the training of our framework does not require any pixel-wise ground truth references.

Our model consists of three networks: a cloud generator G, a cloud discriminator D and a cloud matting network F, as shown in Fig. 1. On one hand, the G takes in a pair of cloud and background images, and generates a new cloud image and its "ground truth". D takes in the generated image to discriminate it is real or fake and feeds this information back to G to further make the generated images indistinguishable. On the other hand, the cloud matting network F takes in the cloud image and produce matting outputs: the predicted cloud reflectance and attenuation maps. The learning of F is instructed by the "ground truth" generated by G so that this process can be easily implemented under a standard regression paradigm. The three networks can be jointly trained in an end-to-end fashion with a clear physical significance.

Our contributions are summarized as follows:

1) Current cloud detection methods frame the detection as a pixel-wise classification problem, which inevitably leads to the defect of category ambiguity when dealing with the semi-transparent clouds. This paper reformulates both of the cloud detection and cloud removal as a foreground-background energy separation process, which can be equivalently implemented under an image matting framework.

2) We propose a weakly supervised method for cloud image matting based on generative adversarial training. The training of our method does not require on any pixel-wise ground truth reference. The proposed method is able to generate realistic cloud images with a clear physical significance.

2. Related work

2.1. Image matting

Image matting refers to a group of the methods that aim to extract the foreground from an image [17, 31, 42, 50], which is important in image and video editing. The matting task usually produces an "alpha matte" that can be used to separate foreground from the background in a given image, which naturally corresponds to the cloud detection and removal process. Traditional image matting methods can be divided into two groups: 1) sampling-based methods [9, 11, 33] and 2) propagation-based methods [6, 17, 46], where the former produces the alpha matte by a predefined metric given a set of the foreground and background sampling regions, while the latter formulates the prediction as the propagation of the foreground and background regions. As the matting is an ill-posed problem, some methods also take in the user interactions (e.g. trimap [31, 42] or scribbles [17]) as additional inputs which specify the predefined foreground, background, and unknown regions to produce more accurate predictions. In recent years, the deep learning techniques have greatly promoted the image matting research progress [3, 5, 35, 42] and most of these methods are built under a regression paradigm. Different from all the above approaches, our method takes advantage of the recent popular adversarial training, neither relying on any pixel-wise ground truth reference, nor any additional user interactions.

2.2. Generative adversarial networks

The Generative Adversarial Network (GAN) [10] has received great attention in recent years, and has achieved impressive results in various tasks such as image generation [7, 30], image style transfer [14, 47] and image super-resolution [16]. A typical GAN consists of two neural networks: a generator network and a discriminator network, where the former learns to map from a latent space to a particular data distribution of interest, while the latter aims to discriminate between instances from the true data distribution and those generated. The key to GAN's success is the idea of an adversarial training framework under which the two networks will contest with each other in a minimax two-player game and forces the generated data to be, in principle, indistinguishable from real ones. Very recently, S. Lutz *et al*. has adopted GAN to improve image matting [25]. In their method, the generator network is trained to improve the prediction of alpha matte by considering the adversarial loss from the discriminator to distinguish well-composited images. However, this method still requires the pixel-wise ground truth and additional user interaction for training, which is not suitable for our tasks because our ground truth cannot be directly obtained.

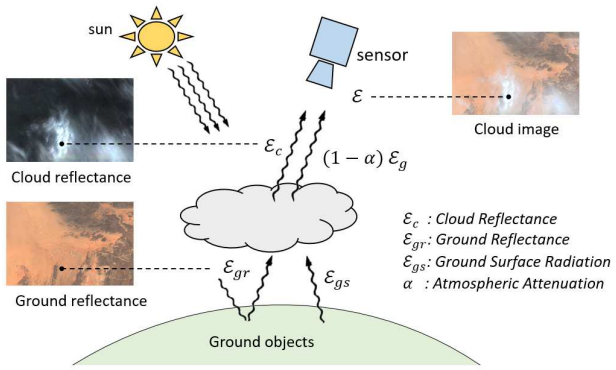

Figure 2. An illustration of the imaging model of cloud images [27, 38]. The energy $\mathcal{E}$ received by a sensor per unit time can be approximated by a linear combination of $\mathcal{E}_c$: the reflectance energy of the cloud, and $\mathcal{E}_g$: the radiation of ground objects.

3. Imaging model

When a satellite or an aircraft flies over the clouds, the onboard imaging sensor receives the energy of ground objects and clouds at the same time. The amount of energy $\mathcal{E}$ received per unit of time can be approximately considered as a linear combination of three components [27, 38], 1) the reflected energy of the clouds $\mathcal{E}_c$, 2) the reflected energy of the ground objects E_{gr}, and 3) the radiation of ground objects E_{gs}:

$$\begin{aligned}\mathcal{E}_{sensor} &= \mathcal{E}_c + (1-\alpha)(\mathcal{E}_{gr} + \mathcal{E}_{gs}) \\ &= \mathcal{E}_c + (1-\alpha)\mathcal{E}_g\end{aligned} \tag{1}$$

where α is defined as an atmospheric attenuation factor ($\alpha \in [0,1]$): the larger the α is, the thicker the cloud will be: $\alpha = 0$ indicates there is no cloud, while $\alpha = 1$ indicates the ground objects are completely occluded. Here we refer $r_c = \mathcal{E}_c$ and $r_g = \mathcal{E}_g$ the "reflectance", as we assume the solar radiation is set to a constant and thus can be neglected. In this way, a cloud image y can be expressed as a linear combination of a cloud reflectance map r_c and a background image r_g:

$$y = r_c + (1-\alpha)r_g. \tag{2}$$

Notice that this equation is quite similar with the well-known "matting function" [17], and therefore, we can simply consider r_c as the "foreground image", α as the "alpha matte", and r_g as the clear "background image" to be recovered. According to the above imaging model, we are able to deal with cloud detection and cloud removal in a unified image matting framework:

I. Cloud detection. As r_c and α correspond to how much energy is reflected and attenuated by clouds, either of them can be used as an indicator of how much clouds is covering on the ground objects. The cloud detection task thus can be considered as a prediction of r_c and α given an input image y. When α is set to 1, the prediction will degenerate to a traditional binary classification based cloud detection method.

II. Cloud removal. Cloud removal is essentially a background recovery problem. According to Eq. 2, the background image can be easily derived by:

$$r_g = (y - r_c)/(1-\alpha), \;\; 0 \le \alpha < 1. \tag{3}$$

This means once we have obtained r_c and α, the cloud can be easily removed and background images can be thus recovered. Notice that when α is close to 1, the ground reflectance is completely lost and thus cannot be recovered.

4. Method

In this paper, we frame the prediction of cloud reflectance r_c and attenuation α under a deep learning based regression paradigm. As it is difficult to obtain the ground truth references, we propose a new adversarial training framework for weakly supervised matting by generating cloud images and the corresponding "ground truth".

4.1. Adversarial training

Our model consists of three networks: a cloud generator G, a cloud discriminator D and a cloud matting network F, as is shown in Fig. 1. Suppose X represents the cloud image domain, Y represents background image domain, and $x_i \in X$ and $y_j \in Y$ are their training samples.

Instead of leaning a mapping directly from X to Y as is suggested by previous GAN-based image translation methods [14, 47], we learn their intermediate states r_c and α with our generator. Fig. 3 shows the processing flow of our method. Specifically, G takes in two images: a clear background image x and a cloud image y, and "creates" a new cloud image $\hat{y} = G(x,y)$ according to (2):

$$\begin{aligned}G(x,y) &= \hat{r}_c + (1-\hat{\alpha})x \\ &= g_1(x,y) + (1 - g_2(x,y))x,\end{aligned} \tag{4}$$

where $g_1(x,y) = \hat{r}_c$ and $g_2(x,y) = \hat{\alpha}$ represents the mappings from the input image pair (x,y) to the generated cloud reflectance $\hat{r}_c$ and attenuation $\hat{\alpha}$. After we obtain the synthesized image $\hat{y}$, the discriminator D is introduced to distinguish between the synthesized images $\hat{y}$ and the real ones y. We express the objective function as follows:

$$\begin{aligned}\mathcal{L}_{adv}(G,D) = &E_{y\sim p(y)}\{\log D(y)\} \\ &+ E_{x,y\sim p(x,y)}\{\log(1 - D(G(x,y)))\},\end{aligned} \tag{5}$$

where the generator G is trained to capture the distribution of real cloud images and make $\hat{y}$ cannot be distinguished from real images. Meanwhile, the discriminator D is trained to do as well as possible at detecting the fake

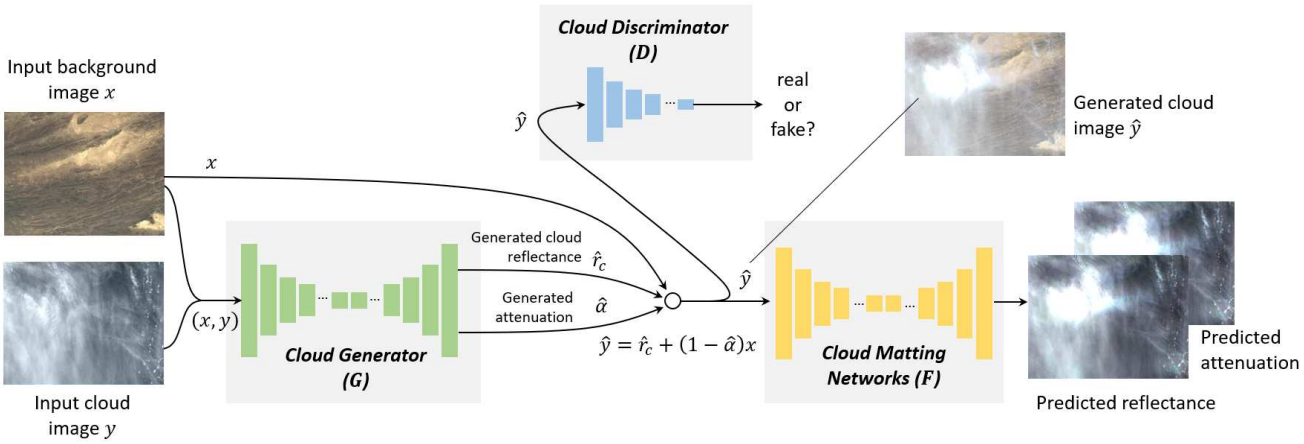

Figure 3. In our method, the cloud generator G takes in a clear background image y and a cloud image x, and generates a cloud reflectance map $\hat{r}_c$ and an attenuation map $\hat{\alpha}$, which are then used to synthesize a new cloud image $\hat{y}$ based on the imaging model $\hat{y} = \hat{r}_c + (1 - \hat{\alpha})x$. A cloud discriminator D is trained to discriminate whether its input (y or $\hat{y}$) is real or fake, and the generator, G, learns to fool the discriminator. The cloud matting network, F, takes in $\hat{y}$ as its input, and produce multiple outputs: the predicted cloud reflectance r_c and attenuation α, where we use $\hat{r}_c$ and $\hat{\alpha}$ as its "ground truth" so that F can be trained under a standard regression paradigm. The three networks can be jointly trained in an end-to-end fashion with a clear physical significance.

ones. The adversarial training process of G and D can be essentially considered as a minimax optimization process, where G tries to minimize this objective while D tries to maximize it: $G^\star = \arg\min_G \max_D \mathcal{L}_{adv}(G, D)$.

4.2. Cloud matting

We consider the learning of our cloud matting network F as a standard regression problem. As the adversarial training progresses, the generated cloud images $\hat{y}$ are fed to the cloud matting network, meanwhile, the generated cloud reflectance $\hat{r}_c$ and the attenuation $\hat{\alpha}$ are used as its "ground truth".

Suppose r_c and α are the predicted cloud reflectance and attenuation maps and F represents their mapping functions: $(r_c, \alpha) = F(\hat{y})$. The regression loss L_{matt} for the matting network therefore can be expressed as the summary of two terms: 1) the loss for cloud reflectance prediction $\mathcal{L}_r$ and 2) the loss for attenuation $\mathcal{L}_\alpha$ prediction:

$$\begin{aligned}\mathcal{L}_{matt}(F) &= \mathcal{L}_r + \mathcal{L}_\alpha \\ &= E_{x,y\sim p(x,y)}\{\|r_c - \hat{r}_c\|_1 + \|\alpha - \hat{\alpha}\|_1\}.\end{aligned} \tag{6}$$

We use $l1$ distance rather than $l2$ as the regression loss since the $l1$ encourages less blurring.

During the training process, the three networks G, D and F can be alternatively updated under a unified objective. Our final objective function is defined as follows:

$$\mathcal{L}(D, G, F) = \mathcal{L}_{adv}(D, G) + \beta\mathcal{L}_{matt}(F), \tag{7}$$

where $\beta > 0$ controls the balance between the adversarial training and the cloud matting. We aim to solve:

$$G^\star, F^\star = \arg\min_{G,F}\max_D \mathcal{L}(D, G, F). \tag{8}$$

By taking advantage of the adversarial training and the physics behind the imaging process, the learning of the matting network can be well instructed even there is no pixel-wise ground truth available. This makes the training extremely efficient because the manual labeling of the data is no longer required. As the training of our model only requires image-level annotations (i.e., an image belongs to X or Y), we refer to our method as a "weakly supervised" cloud matting method.

4.3. Implementation details

1) Architecture of the networks. Our cloud generator G consists of an eight-layer encoder and an eight-layer decoder. We add skip connections between all channels at layer i and layer $n - i$, following the general configuration of the "U-Net" [32] for building both of the high-level semantic features and low-level details. Our cloud discriminator D is a standard convolutional network with 10 convolutional layers followed by 2 fully-connected layers. Our cloud matting network F takes similar configurations with G, but only different in terms of the number of layers and the number of filters.

2) Saturation penalty. As the clouds are mostly in white color, the r-g-b channels of the generated cloud reflectance map should be close to each other. To improve the training stability, we add an additional saturation penalty term to the objective of G, i.e. to compute the saturation value of each pixel and penalize those pixels with large saturation values:

$$\mu(s) = \gamma\|s\|_2^2, \tag{9}$$

where $s = (\max(r, g, b) - \min(r, g, b))/\max(r, g, b)$. To test the importance of this additional constraint, we also

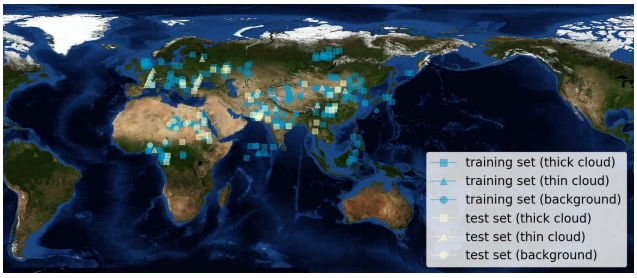

Figure 4. (Better viewed in color) A global distribution of our experimental data. The dataset covers the most types of ground features over a worldwide distribution, such as city, ocean, plains, plateaus, glacier, desert, gobi, and etc.

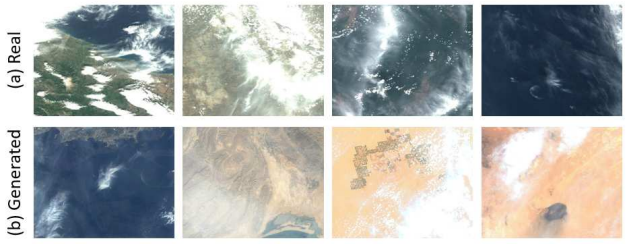

Figure 5. The first row shows some real cloud images of GF-1 satellite. The second row shows our generated images. Samples are fair random draws, not cherry-picked.

compare to an un-penalized variant in which the generator does not takes the saturation prior into consideration (see Section 5.4 for more details).

3) Training details. The Batch-normalization and ReLU activation function are embedded in G, D and F after all convolution layers, except for their outputs. For the last layer of G and D, we use the sigmoid function to convert the output logits to probabilities. To increase the diversity of the generated images, the background images X and cloud images Y are randomly rotated, flipped and cropped during training. We set the weight for saturation penalty as $\gamma = 1.0$. We use the Adam optimizer for training, with batch_size $= 4$. We use Xavier initialization for all networks. For the first 10 training epochs, we set $\beta = 0$. The learning rates are set to 10^{-5} for G and 10^{-6} for D. For the next 80 epochs, we set $\beta = 1.0$. The learning rate of F is set to 10^{-4} and we reduce the learning rates of (G, D) to their 1/10. All images are resized to 512×512 for training and evaluation.

To verify the stability of our training framework, we also take the recent two improvements of GAN into consideration, i.e. WGAN [2] and LSGAN [26].

5. Experimental results and analysis

5.1. Dataset and Metrics

Our experimental dataset consists of 1,209 remote sensing images that are captured by two cameras on Gaofen-1 satellite: the panchromatic and multi-spectral (PMS) sensor with the image size of about 4500×4500 pixels, and the wide field-of-view (WFV) sensors with the image size of about 12000×13000 pixels. There are 681 images in our training set and 528 in our testing set, where each of them is further split into three subsets: a "thick cloud set", a "thin cloud set" and a "background set". Since the raw images have four bands (blue, green, red, and infrared) and are in 16-bit depth, all images are converted to 8-bit RGB images before fed into the networks. Apart from that, we did not perform any other pre-processing operations. The dataset covers the most types of ground features over a worldwide distribution, such as city, ocean, plains, plateaus, glacier, desert, gobi, and etc, as is shown in Fig. 4. It should be noticed that although there are some publicly available cloud detection datasets [19, 20], we do not make evaluations on them because they are too small (only about 100 images) to obtain statistically significant results.

For the cloud detection task, the Precision-Recall (PR) curve and the "Average Precision (AP)" score are used as our evaluation metrics. All images in our dataset have been manually labeled with pixel-wise binary cloud masks as their ground truth in spite of the fact that we did not use this information for training.

For the cloud removal task, three different scores are evaluated, including the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). These metrics are defined as follows: MAE $= \frac{1}{N}\|y - \hat{y}\|_1$, MSE $= \frac{1}{N}\|y - \hat{y}\|_2^2$, MAPE $= \frac{1}{N}\|(y - \hat{y})/\hat{y}\|_1$, where y is the predicted output, $\hat{y}$ is the ground truth, and N is the total number of pixels. Besides, since the backgrounds under thick cloud regions cannot be fully recovered, the above metrics are only evaluated on those pixels whose ground truth attention value is smaller than 0.5.

5.2. Cloud detection results

We compare our method with some recent popular cloud detection methods, including Scene Learning [1], Fully Convolutional Networks based pixel-wise Classification (FCN+Cls) [44], and Progressive Refinement Detection [45] on our test set. As these methods are all essentially performing binary classification on each image pixel, we also compare with some well-known semantic segmentation frameworks, including Deeplab-v3 [4], and UNet [32]. Fig. 6 shows some cloud detection examples, where the different columns correspond to the input image and the outputs of different detection methods.

Table 1 shows their corresponding AP scores on "thick cloud set" and "thin cloud set". Fig. 7 shows the comparison of their PR curves. As the Progressive Refinement [45] is a thresholding-based method and only produce binary masks, the PR curve becomes a single point thus we can not com-

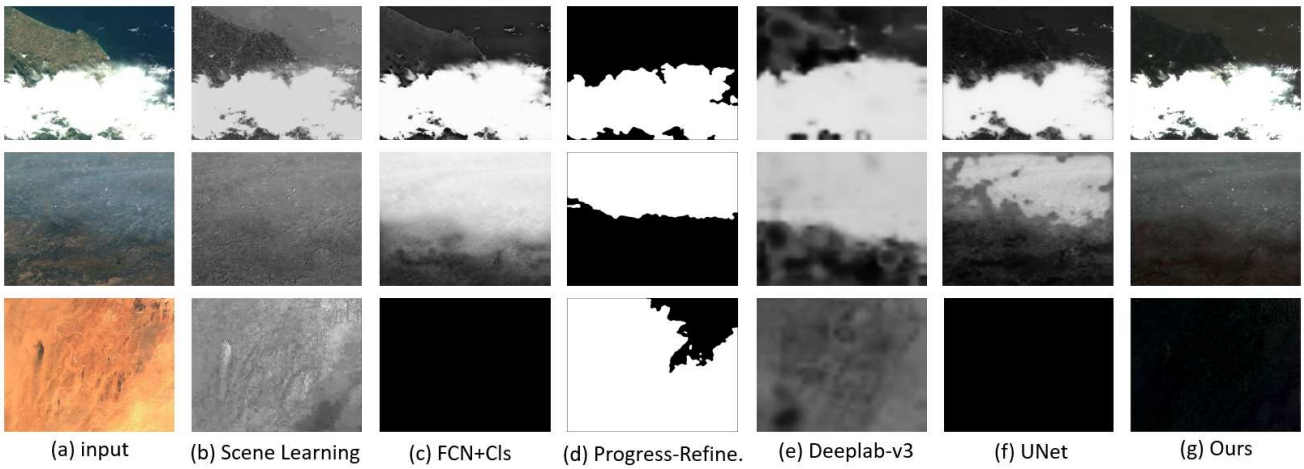

Figure 6. Some examples of the cloud detection results of different methods, where Scene Learning [1], FCN+Cls [44], and Progress-Refine [45] are recent published cloud detection methods. Deeplab-v3 [4] and UNet [32] are two well-known semantic segmentation methods.

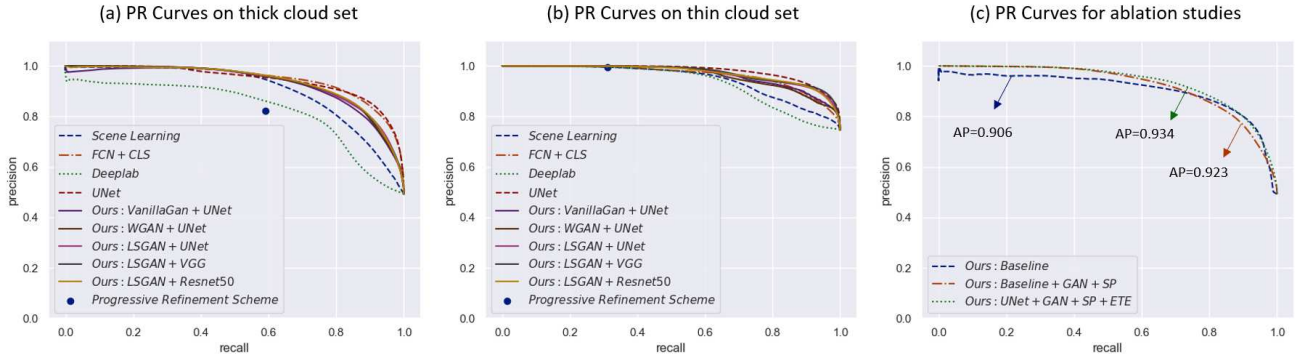

Figure 7. (Better viewed in color) The precision-recall curves of different cloud detection methods on (a) thick-cloud set, (b) thin-cloud set and (c) ablation studies. A higher curve or a higher AP score suggests a better detection result.

pute AP score based on its result. In this experiment, we also compare our methods on two variants of adversarial objective functions: WGAN [2] and LSGAN [26]. Notice that Deeplab-v3 are built on larger backbone networks, for a fair comparison, we further replace the encoder of our cloud matting network F with VGG [36] and Resnet50 [12] while keeping other parameter configurations unchanged.

It can be seen from Table 1 and Fig. 7 that our method achieves a higher cloud detection accuracy especially for those thin cloud images. As for their overall performances, our method achieves comparable performance with FCN+Cls [44] and UNet [32], and outperforms other cloud detection methods and semantic segmentation methods. We also notice that deeplab-v3 [4] has a relatively low detection accuracy. This is simply because it down-samples the input x8, which produces very coarse outputs. The advantage of our method is not only suggested by the metrics but also in terms of the interpretation of some causal factors (reflectance and attenuation) of the imaging process. Although our model is trained *without the help of any pixel-wise labels*, the experimental result still demonstrates that it *achieves comparable and even higher accuracy with other popular cloud detection methods, which are trained in a fully supervised manner*.

5.3. Cloud removal results

Once we have obtained the cloud reflectance and attenuation maps, the background can be easily recovered by using Eq. 3. In this experiment, we compare our method with a classical cloud removal method: Homomorphic Filter [24], and two recent proposed methods: Deformed-Haze [28] and Adaptive Removal [40]. We further compare our method with a sota image-to-image translation method, CycleGAN [47]. Fig. 8 shows some example results. It can be seen that the thin cloud has nicely been removed by our method and the ground object has been recovered. As the CycleGAN is essentially performing "style transfer" rather than cloud removal, it may introduce unexpected

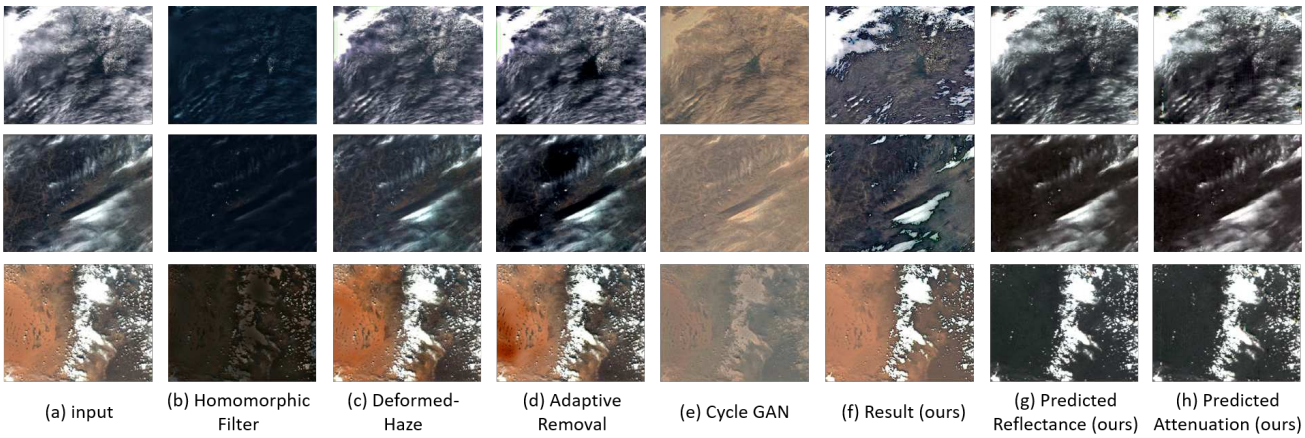

Figure 8. Some example results of the thin cloud removal by different methods. As the Homomorphic Filter (HF) [24] is designed to simply remove the low-frequency components of an image, it wrongly removes all backgrounds while removing the clouds. Deformed-Haze [28] and Adaptive Removal [40] performs better than HF but still suffers from a color distortion problem. For our method, the thin cloud has been removed and the ground object has been nicely recovered.

Method	AP_{thick}	AP_{thin}	AP_{avg}
Scene Learning [1]	0.9027	0.9457	0.9242
FCN+Cls [44]	**0.9463**	0.9643	**0.9553**
Progressive Refine. [45]	–	–	–
Deeplab-v3 [4]	0.8287	0.9320	0.8804
UNet [32]	**0.9434**	**0.9791**	**0.9613**
Ours: VanillaGan+UNet	0.9262	0.9662	0.9462
Ours: WGan+UNet	0.9327	0.9629	0.9478
Ours: LSGAN+UNet	0.9336	0.9709	0.9523
Ours: LSGAN+VGG	**0.9344**	**0.9732**	**0.9538**
Ours: LSGAN+Resnet50	0.9341	**0.9711**	0.9526

Table 1. A Comparison of cloud detection results of different methods. A higher score suggests a better result. The top-3 best results in each entry are marked as bold. As the Progressive Refine [45] is a thresholding-based method and only produces binary outputs, we can not compute AP based on its detection results.

"color shift". Another disadvantage of the CycleGAN is it cannot generate the cloud reflectance/attenuation as it ignores the physics behind.

As there is no ground truth for the cloud removal task, to make a quantitative comparison of different methods, we run cloud generator G on our testing set to randomly synthesize 1,390 cloud images and use the original background image as their "ground truth" for evaluation. Table 2 shows the quantitative comparison results of the proposed method and other three cloud removal methods, in which we can see our method performs better than the other three methods under all metrics. We do not further compare with other recent image matting methods [3, 5, 35, 42] because the training of these methods requires the ground truth of alpha matte or the user interactions.

Method	MAE	MSE	MAPE
Homomorphic Filter [24]	0.2374	0.0731	0.4312
Deformed-Haze [28]	0.1820	0.0441	0.3645
Adaptive Removal [40]	0.1290	0.0230	0.2774
Cycle GAN [47]	0.0904	0.0153	0.2404
Ours: VanillaGan+UNet	**0.0720**	**0.0088**	**0.1735**
Ours: WGAN+UNet	**0.0706**	**0.0086**	**0.1704**
Ours: LSGAN+UNet	**0.0753**	**0.0095**	**0.1791**
Ours: LSGAN+VGG	0.0759	0.0096	0.1791
Ours: LSGAN+Resnet50	0.0753	0.0095	0.1790

Table 2. Comparison of different methods on cloud removal task. Lower scores indicate better. The top-3 best results in each entry are marked as bold.

Ablations				**Accuracy**		
Adv	S-p	E2E	Bg	AP_{thick}	AP_{thin}	AP_{avg}
×	×	×	✓	0.9061	0.9608	0.9335
✓	×	×	✓	–	–	–
✓	✓	×	✓	0.9230	0.9630	0.9430
✓	✓	✓	×	0.9307	0.9658	0.9483
✓	✓	✓	✓	**0.9336**	**0.9709**	**0.9523**

Table 3. Ablation studies of four technical components of our method 1) adversarial training (Adv), 2) saturation penalty (S-p), 3) end-to-end training (E2E), and 4) background input (BG) on cloud detection task. Higher scores indicate better. We observe when do not apply S-p, the training does not converge and the outputs of G collapse to a single nonsensical image.

5.4. Ablation analyses

The ablation analyses are conducted on the cloud detection task to analyze the importance of each component of

Figure 9. (a) Input high-resolution images from Google Earth. (b) augmented cloud images with our method.

the proposed framework, including 1) adversarial training, 2) saturation penalty, 3) end-to-end training, and 4) background input. The baseline methods are first evaluated, then we gradually integrate these techniques. All evaluations of this experiment are performed on basis of our "LSGAN + UNet" implementation.

1) w/o adversarial training (Adv). This ablation setting corresponds to our baseline method, where we train our cloud matting network F without any help of adversarial training. Since there is no ground truth for the matting task, we *manually* synthesized a set of images and corresponding ground truth for training.

2) w/o saturation penalty (S-p). To test the importance of "saturation penalty" of our method, we simply remove the term of Eq. 9 in our objective function while keeping other settings unchanged and test its performance.

3) w/o end-to-end training (E2E). We also compare with another variant of our method, in which the (G, D) and F are separately trained based on their own objectives. We first train (G, D) for the first 10 epochs, then, we freeze their weights and train F for the next 80 epochs.

4) w/o background input (BG). While it may be possible for us to remove the BG from the G, we found the joint input helps improve the results and can be considered as an integration of geographical domain knowledge. This is because the cloud and its BG are not completely independent of each other (e.g., deserts tend to have less thick clouds than other areas).

Table 3 shows their evaluation accuracy. As we can see, the integration of the "adversarial training" and "end-to-end training" yields noticeable improvements of the detection accuracy. We also notice that when we do not apply the "saturation penalty", the training does not converge and the outputs of G collapse to a single nonsensical image. We further evaluate our method w/o the help of a BG input, and we observe the cloud detection accuracy drops 0.4% compared with our full implementation.

5.5. Cloud montage

The proposed framework can be further applied to generate cloud on a given background image, or to "transplant" the cloud in one image to another. We refer to this process as "cloud montage". This can be considered as a new way of performing data augmentation and may have great potential for improving the performance of many applications such as occluded object detection, scene recognition, and image segmentation. Fig. 9 shows some examples of our image augmentation results, where the clouds are generated on some very high-resolution aerial images from Google Earth.

Object detectors	w/o augm.	w/ augm.
SSD [23] (VGG)	78.9%	**81.7%**
RetinaNet [22] (Resnet-50)	83.3%	**87.3%**

Table 4. Comparison of the object detection results (VOC2012-AP) on the occluded target detection dataset [29]. There are noticeable improvements over the two baseline detectors by performing "cloud augmentation" based on the proposed method.

5.6. Improving occluded target detection

In this experiment, we choose airplane detection as an example to evaluate the effectiveness of the above data augmentation on occluded target detection. Specifically, we train two well-known object detector, SSD [23] and retinaNet [22] as our baseline detectors. We use VGG [36] and Resnet-50 [12] as their backbones. The baseline detectors are trained on LEVIR dataset [51] (consists of 22,000 images and 7,749 annotated targets), and then evaluated on a publicly available occluded target detection dataset [29] (consists of 47 images and 184 annotated targets where 96 are occluded). We compare our baseline detectors with their enhanced versions, in which the detectors are trained with the augmented training configuration. Table 4 shows their average precision scores. We observe noticeable improvements over the baseline detectors with the help of "cloud augmentation".

6. Conclusion

We propose a weakly supervised method for the detection and removal of clouds in remote sensing images based on adversarial training. The proposed method inherently incorporates the cloud imaging mechanism and considers our task as a cloud-background energy separation problem. Our experimental results demonstrate that without ever using any pixel-wise ground truth references during training, our method achieves comparable or even better performance over other methods, which are trained in a fully supervised manner. In addition, the proposed framework can be used for generating cloud images of various styles on any given backgrounds. This can be viewed as a new way of performing data augmentation and has great potential for improving occluded object detection and recognition.

References

[1] Zhenyu An and Zhenwei Shi. Scene learning for cloud detection on remote-sensing images. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 8(8):4206–4222, 2015.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[3] Guanying Chen, Kai Han, and Kwan-Yee K Wong. Tom-net: Learning transparent object matting from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9233–9241, 2018.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[5] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 618–626. ACM, 2018.

[6] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013.

[7] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

[8] Kenji Enomoto, Ken Sakurada, Weimin Wang, Hiroshi Fukui, Masashi Matsuoka, Ryosuke Nakamura, and Nobuo Kawaguchi. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. *arXiv preprint arXiv:1710.04835*, 2017.

[9] Eduardo SL Gastal and Manuel M Oliveira. Shared sampling for real-time alpha matting. In *Computer Graphics Forum*, volume 29, pages 575–584. Wiley Online Library, 2010.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *CVPR 2011*, pages 2049–2056. IEEE, 2011.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Richard R Irish, John L Barker, Samuel N Goward, and Terry Arvidson. Characterization of the landsat-7 etm+ automated cloud-cover assessment (acca) algorithm. *Photogrammetric engineering & remote sensing*, 72(10):1179–1188, 2006.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[15] Gary J Jedlovec, Stephanie L Haines, and Frank J LaFontaine. Spatial and temporal varying thresholds for cloud detection in goes imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6):1705–1717, 2008.

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[17] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2008.

[18] Pengfei Li, Limin Dong, Huachao Xiao, and Mingliang Xu. A cloud image detection method based on svm vector machine. *Neurocomputing*, 169:34–42, 2015.

[19] Zhiwei Li, Huanfeng Shen, Qing Cheng, Yuhao Liu, Shucheng You, and Zongyi He. Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150:197–212, 2019.

[20] Zhiwei Li, Huanfeng Shen, Huifang Li, Guisong Xia, Paolo Gamba, and Liangpei Zhang. Multi-feature combined cloud and cloud shadow detection in gaofen-1 wide field of view imagery. *Remote sensing of environment*, 191:342–358, 2017.

[21] Chao-Hung Lin, Po-Hung Tsai, Kang-Hua Lai, and Jyun-Yuan Chen. Cloud removal from multitemporal satellite images using information cloning. *IEEE transactions on geoscience and remote sensing*, 51(1):232–241, 2013.

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[24] ZK Liu and Bobby R Hunt. A new approach to removing cloud cover from satellite imagery. *Computer vision, graphics, and image processing*, 25(2):252–256, 1984.

[25] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. *arXiv preprint arXiv:1807.10088*, 2018.

[26] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

[27] OR Mitchell, Edward J Delp, and Peilu L Chen. Filtering to remove cloud cover in satellite imagery. *IEEE Transactions on Geoscience Electronics*, 15(3):137–141, 1977.

[28] Xiaoxi Pan, Fengying Xie, Zhiguo Jiang, and Jihao Yin. Haze removal for a single remote sensing image based on deformed haze imaging model. *IEEE Signal Processing Letters*, 22(10):1806–1810, 2015.

[29] Shaohua Qiu, Gongjian Wen, and Yaxiang Fan. Occluded object detection in high-resolution remote sensing images

using partial configuration object model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(5):1909–1925, 2017.

[30] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[31] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1826–1833. IEEE, 2009.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[33] Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen. Improving image matting using comprehensive sampling sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 636–643, 2013.

[34] Huanfeng Shen, Huifang Li, Yan Qian, Liangpei Zhang, and Qiangqiang Yuan. An effective thin cloud removal procedure for visible remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 96:224–235, 2014.

[35] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *European Conference on Computer Vision*, pages 92–107. Springer, 2016.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[37] CJ Stubenrauch, WB Rossow, Stefan 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*, 94(7):1031–1049, 2013.

[38] CT Swift and DJ Cavalieri. Passive microwave remote sensing for sea ice research. *Eos, Transactions American Geophysical Union*, 66(49):1210–1212, 1985.

[39] Xi Wu and Zhenwei Shi. Utilizing multilevel features for cloud detection on satellite imagery. *Remote Sensing*, 10(11):1853, 2018.

[40] Fengying Xie, Jiajie Chen, Xiaoxi Pan, and Zhiguo Jiang. Adaptive haze removal for single remote sensing image. *IEEE Access*, 6:67982–67991, 2018.

[41] Fengying Xie, Mengyun Shi, Zhenwei Shi, Jihao Yin, and Danpei Zhao. Multilevel cloud detection in remote sensing images based on deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(8):3631–3640, 2017.

[42] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[43] Zhiyuan Yan, Menglong Yan, Hao Sun, Kun Fu, Jun Hong, Jun Sun, Yi Zhang, and Xian Sun. Cloud and cloud shadow detection using multilevel feature fused segmentation network. *IEEE Geoscience and Remote Sensing Letters*, (99):1–5, 2018.

[44] Yongjie Zhan, Jian Wang, Jianping Shi, Guangliang Cheng, Lele Yao, and Weidong Sun. Distinguishing cloud and snow in satellite images via deep convolutional network. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1785–1789, 2017.

[45] Qing Zhang and Chunxia Xiao. Cloud detection of rgb color aerial photographs by progressive refinement scheme. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):7264–7275, 2014.

[46] Yuanjie Zheng and Chandra Kambhamettu. Learning based digital matting. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 889–896. IEEE, 2009.

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[48] Zhe Zhu, Shixiong Wang, and Curtis 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*, 159:269–277, 2015.

[49] Zhe Zhu and Curtis E Woodcock. Object-based cloud and cloud shadow detection in landsat imagery. *Remote sensing of environment*, 118:83–94, 2012.

[50] Douglas E Zongker, Dawn M Werner, Brian Curless, and David H Salesin. Environment matting and compositing. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 205–214. ACM Press/Addison-Wesley Publishing Co., 1999.

[51] Zhengxia Zou and Zhenwei Shi. Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images. *IEEE Transactions on Image Processing*, 27(3):1100–1111, 2018.