Generative Adversarial Training for Weakly Supervised Cloud Matting

Supplementary Material

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

1 Training details

As is suggested by I. Goodfellow *et al.*[4], instead of training G to minimize $\log(1 - D(G(x, y)))$, in practice, we try to maximize $\log D(G(x, y))$. This is because in early stage of learning, $\log(1 - D(G(x, y)))$ tends to saturate. This revision on objective provides much stronger gradients early in learning.

We also consider the other two variants of the adversarial objectives in recent works, i.e. WGAN [2] and LSGAN [7], to stabilize our training. Particularly, for the WGAN based objectives, we train the G to minimize $-D(G(x, y))$, and train the D to maximize $D(y) - D(G(x, y))$. For the LSGAN based objectives, we train the G to minimize $(D(G(x, y)) - 1)^2$, and train the D to minimize $(D(y) - 1)^2 + D(G(x, y))^2$. In these two cases, the sigmoid function at the last layer of D is removed so that to produce logits rather than probabilities.

2 Dataset

The statistics of our dataset are given in Table 1.

Image Info.	# total imgs	1,209
	source	GaoFen-1 PMS and WFV
training set	# imgs	681
	# thin cloud imgs	81
	# thick cloud imgs	404
	# background imgs	196
testing set	# imgs	528
	# thin cloud imgs	81
	# thick cloud imgs	391
	# background imgs	56

Table 1: A summary of our experimental dataset.

3 Implementation details of our baseline model

In our ablation studies, we compare our method with a baseline model that is trained without any help of the adversarial training. As there is no ground truth value for cloud reflectance and attenuation, we *manually*

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

synthesize a set of images and corresponding ground truth references.

Specifically, the thick clouds images (where $\alpha \approx 1$) and background images (with no clouds, where $\alpha \approx 0$) in our training set are used to generate the synthesized images and their ground truth maps. We use the image regions that completely covered by thick clouds as the "ground truth" cloud reflectance of synthesized images. Then, the synthetic image can be generated by simply performing a linear combination between the clouds and a clear background image, where a random alpha value is used as the combination weights.

4 Configurations of our Networks

Table 2 lists the detailed configurations of our cloud generator G and our cloud discriminator D. Table 3 lists the detailed configurations of our cloud matting network F.

The column "Filters" gives the configuration of the convolutional filters, where $n \times n/m$ corresponds to the size ($n \times n$) and number of filters (m). "C(2)_P" represents two "stacked convolution layers" followed by a pooling layer. "U" represents an up-sampling layer with bi-linear interpolation. "DC" represents a fractional-strided convolution layer [14] (a.k.a. the transposed convolution) for up-sampling the feature maps. "|" represents feature fusion by concating two feature maps. "+" represents feature fusion by element-wise summation.

	Layer	**Input**	**Stride**	**Filters**
Generator	C(2)_P_1	image	2	3x3 / 64
	C(2)_P_2	C(2)_P_1	2	3x3 / 64
	C(2)_P_3	C(2)_P_2	2	3x3 / 64
	C(2)_P_4	C(2)_P_3	2	3x3 / 64
	C(2)_U_5	C(2)_P_4	1/2	5x5 / 64
	C(2)_U_6	C(2)_U_5 + C(2)_P_3	1/2	5x5 / 64
	C(2)_U_7	C(2)_U_6 + C(2)_P_2	1/2	5x5 / 64
	C(2)_U_8	C(2)_U_7 + C(2)_P_1	1/2	5x5 / 64
	C(1)_9	C(2)_U_8	1	5x5 / 3
Discriminator	C(2)_P_1	image	2	3x3 / 128
	C(2)_P_2	C(2)_P_1	2	3x3 / 256
	C(2)_P_3	C(2)_P_2	2	3x3 / 256
	C(2)_P_4	C(2)_P_3	2	3x3 / 256
	C(2)_P_5	C(2)_P_4	2	3x3 / 256
	FC_1	C(2)_P_6	-	- / 512
	FC_2	FC_1	-	- / 1

Table 2: Detailed configuration of our cloud generator G and our cloud discriminator D.

	Layer	Input	Stride	Filters
Cloud Matting Network	C(2)_1	image	1	3x3 / 64
	P_C(2)_2	C(2)_1	2	3x3 / 128
	P_C(2)_3	P_C(2)_2	2	3x3 / 256
	P_C(2)_4	P_C(2)_3	2	3x3 / 512
	P_C(2)_5	P_C(2)_4	2	3x3 / 1024
	DC_1	P_C(2)_5	1/2	3x3 / 512
	C(2)_6	DC_1 | P_C(2)_4	1	3x3 / 512
	DC_2	C(2)_6	1/2	3x3 / 256
	C(2)_7	DC_2 | P_C(2)_3	1	3x3 / 256
	DC_3	C(2)_7	1/2	3x3 / 128
	C(2)_8	DC_3 | P_C(2)_2	1	3x3 / 128
	DC_4	C(2)_8	1/2	3x3 / 64
	C(2)_9	DC_4 | P_C(2)_1	1	3x3 / 64
	C(1)_10	C(2)_9	1	3x3 / 3

Table 3: A detailed configuration of our cloud matting network F.

5 Is the cloud matting network F necessary?

As our cloud generator G and our matting networks are all built based on the same physical imaging model, a natural question would be, is the cloud matting network F necessary? or can we replace the F with the G? The answer is “no”. We cannot remove F or replace F with G. This is because the G does not simply play a “copy-and-paste” role for clouds in our method. Instead, it may also modify the cloud’s shape and transparency based on its “own will”. Fig. 1 gives an example of why G-only cannot be used for cloud detection.

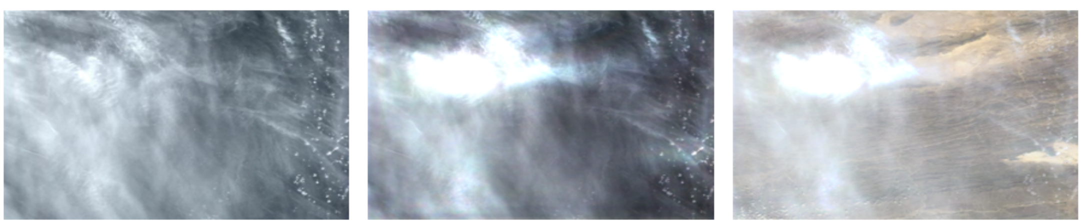

Figure 1: Left to right: G’s input, generated cloud reflectance, and re-composed cloud image. Note that the cloud has been modified by G and thus cannot be used as the final detection output.

6 Limitations

one of our limitations is when dealing with the snow-covered regions, especially when the snow presents high reflectance. Fig. 2 shows a failure case of our method.

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.

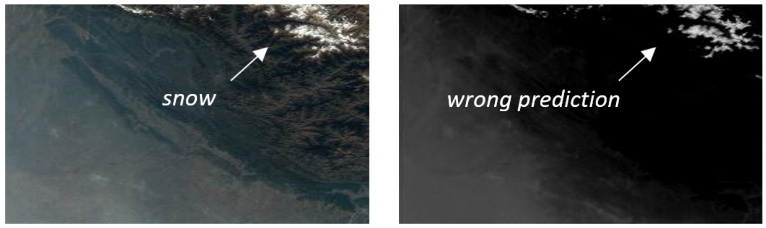

Figure 2: A failure case of our method: cloud detection in a snow-covered area. Left: input; Right: predicted cloud reflectance.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Volodymyr Mnih. *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.

[9] 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.

[10] 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.

[11] 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.

[12] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

[14] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[15] 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.

[16] 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.

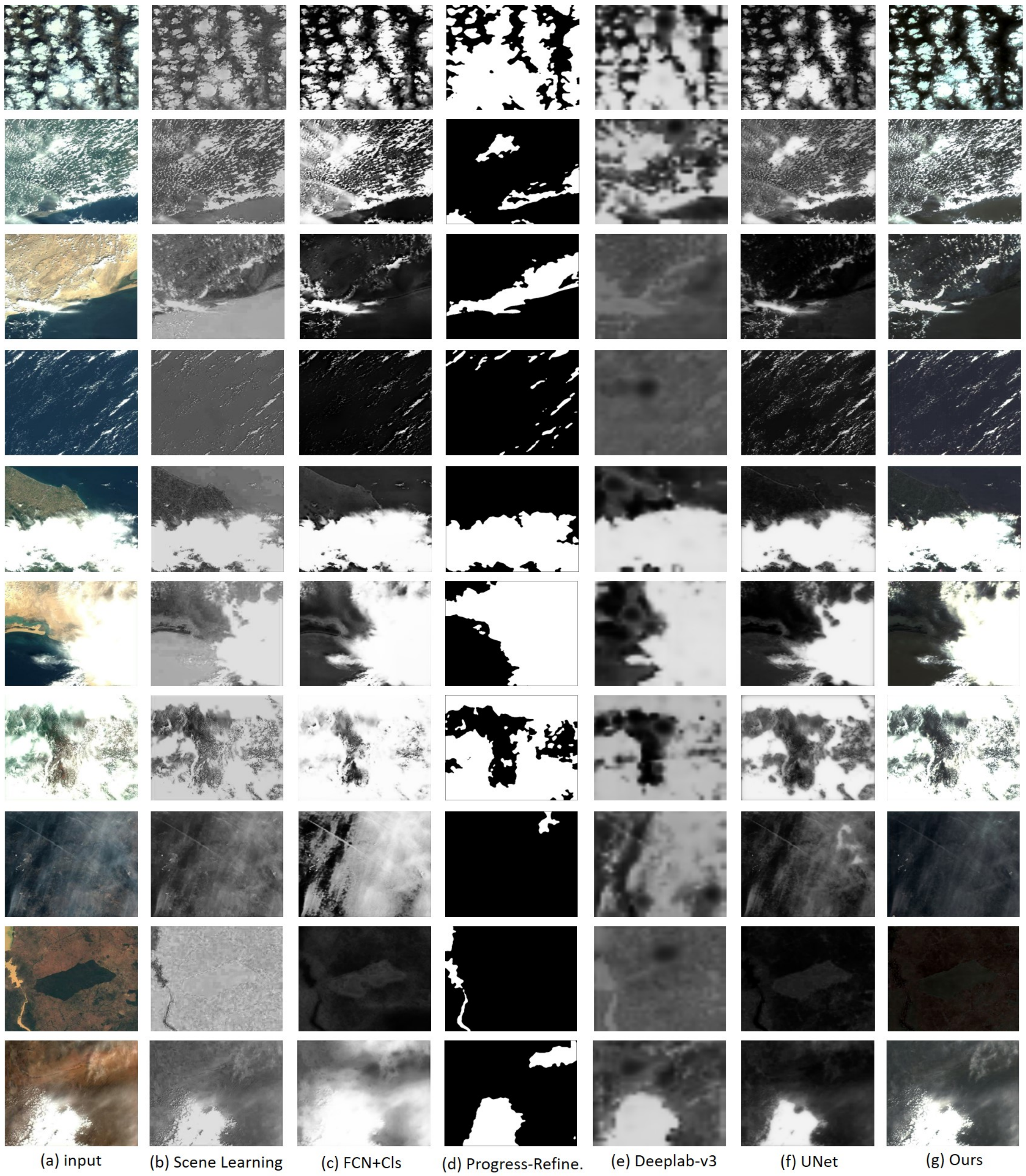

Figure 3: Some examples of the cloud detection results of different methods, where Scene Learning [1], FCN+Cls [15], and Progress-Refine [16] are recent published cloud detection methods. Deeplab-v3 [3] and UNet [11] are well-known semantic segmentation methods.

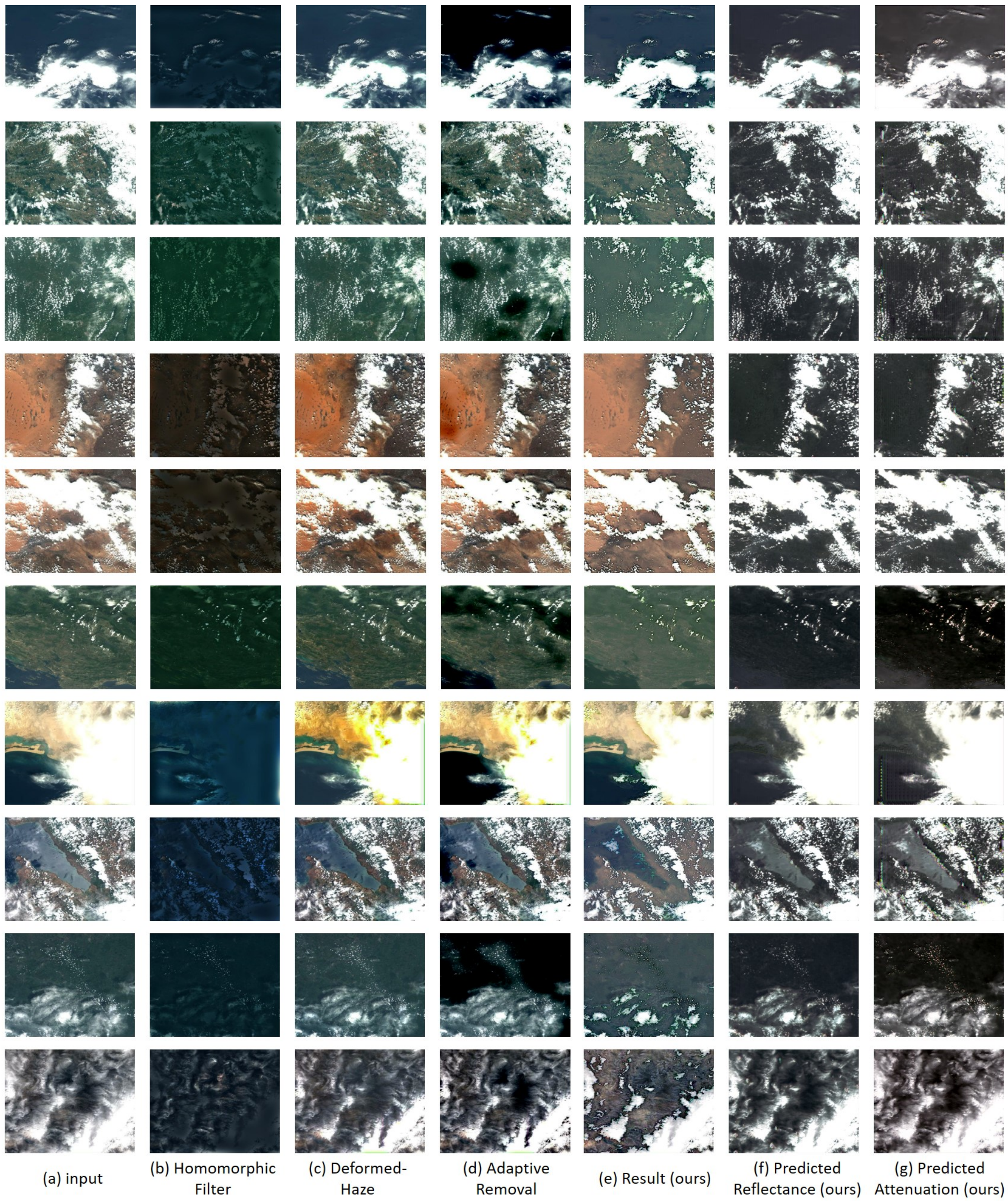

Figure 4: Some example results of the thin cloud removal, where Homomorphic Filter [6] is a classical cloud removal method, Deformed-Haze [9] and Adaptive Removal [13] are two recent proposed cloud removal methods.

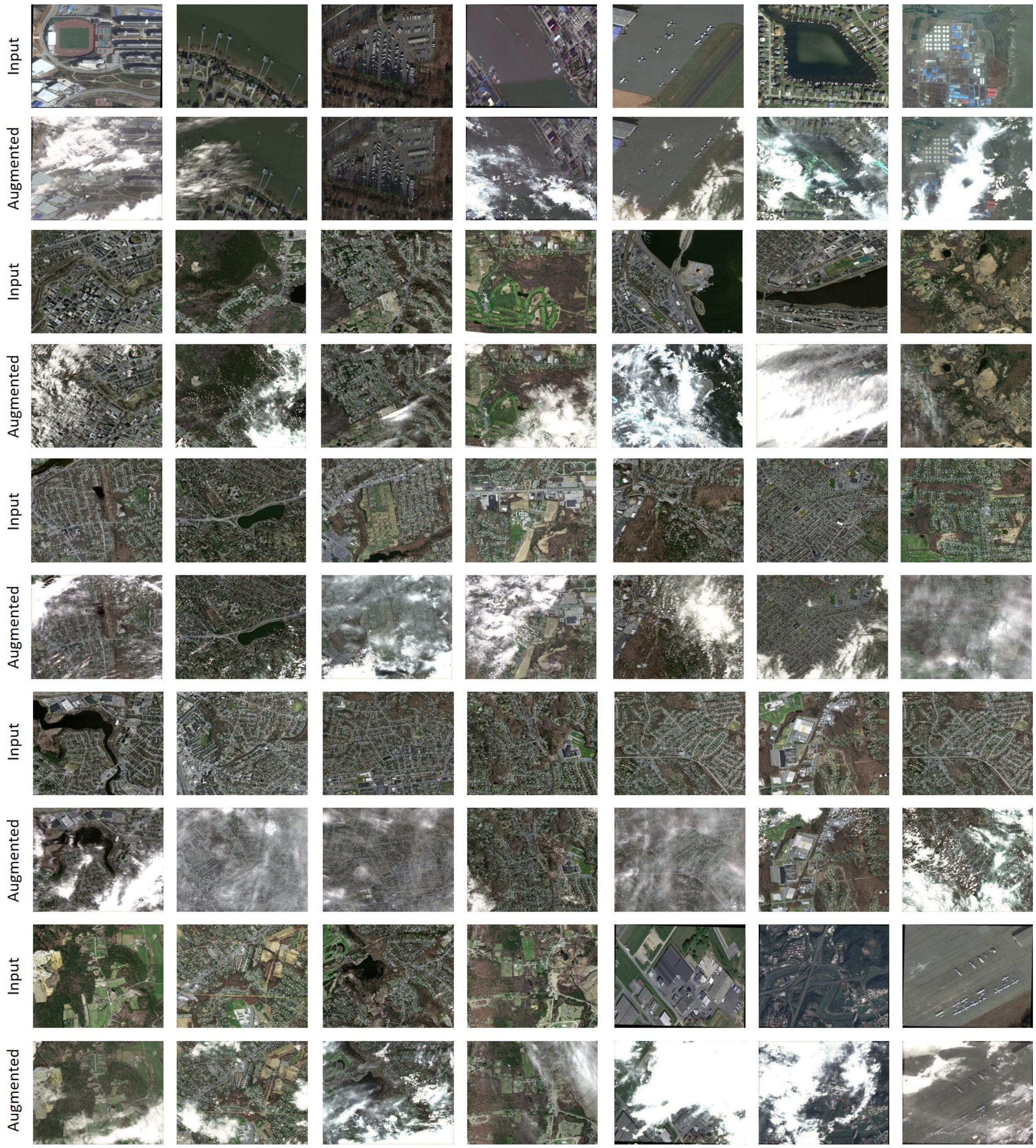

Figure 5: Some examples of “cloud augmentation” on high-resolution Google Earth images. Input images are from DOTA dataset [12] and Massachusetts Roads/Building Dataset [8].

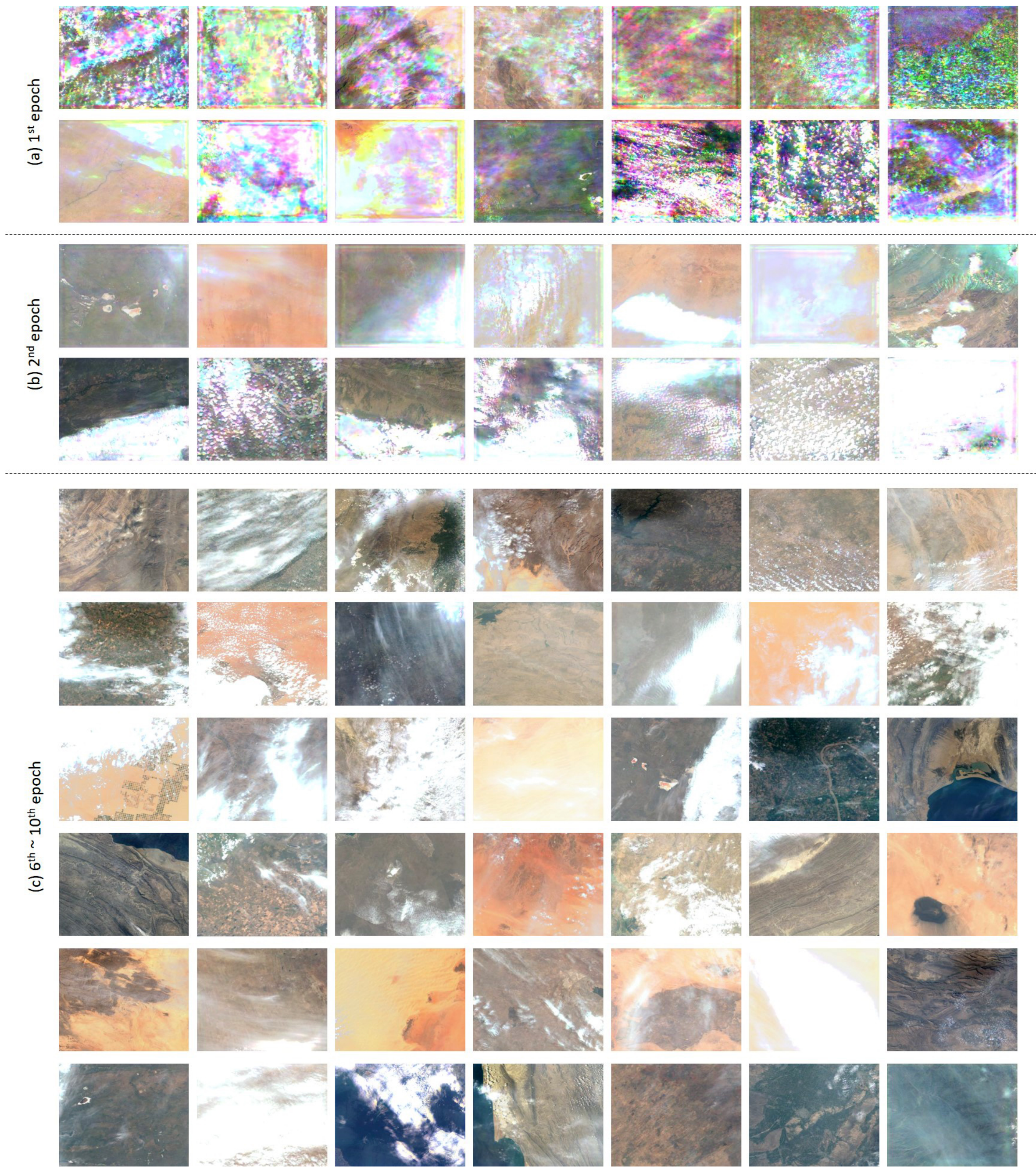

Figure 6: Some examples of generated cloud images by our method: (a) 1st training epoch, (b) 2nd training epoch, (c) 6th-10th training epoch. Samples are fair random draws, not cherry-picked.

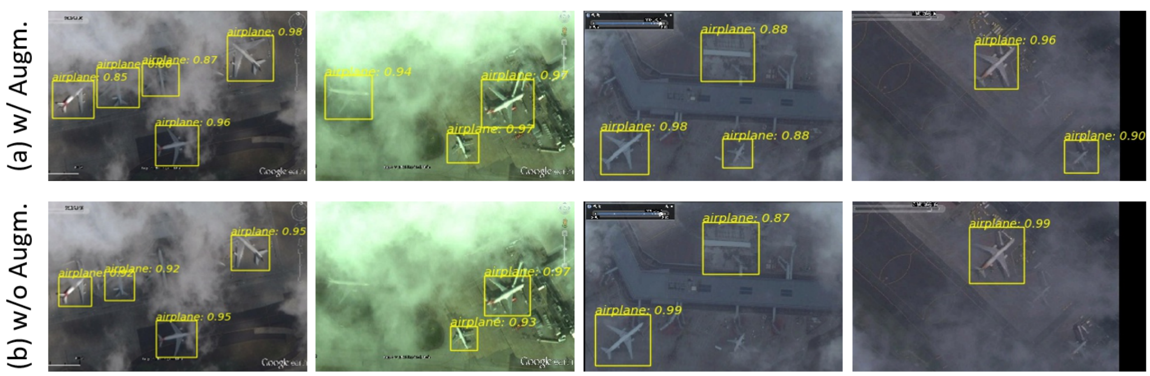

Figure 7: The object detection results on the occluded target detection dataset [10] with RetinaNet detector [5]: (a) detection results trained with cloud augmentation and (b) without cloud augmentation.