Filmy Cloud Removal on Satellite Imagery with Multispectral Conditional Generative Adversarial Nets

Kenji Enomoto[1*] Ken Sakurada[1] Weimin Wang[1] Hiroshi Fukui[2] Masashi Matsuoka[3]
Ryosuke Nakamura[4] Nobuo Kawaguchi[1]
[1]Nagoya University [2]Chubu University [3]Tokyo Institute of Technology
[4]Advanced Industrial Science and Technology
{enoken, weimin}@ucl.nuee.nagoya-u.ac.jp, {sakurada, kawaguti}@nagoya-u.jp
fhiro@vision.cs.chubu.ac.jp, matsuoka.m.ab@m.titech.ac.jp, r.nakamura@aist.go.jp

Abstract

In this paper, we propose a method for cloud removal from visible light RGB satellite images by extending the conditional Generative Adversarial Networks (cGANs) from RGB images to multispectral images. Satellite images have been widely utilized for various purposes, such as natural environment monitoring (pollution, forest or rivers), transportation improvement and prompt emergency response to disasters. However, the obscurity caused by clouds makes it unstable to monitor the situation on the ground with the visible light camera. Images captured by a longer wavelength are introduced to reduce the effects of clouds. Synthetic Aperture Radar (SAR) is such an example that improves visibility even the clouds exist. On the other hand, the spatial resolution decreases as the wavelength increases. Furthermore, the images captured by long wavelengths differs considerably from those captured by visible light in terms of their appearance. Therefore, we propose a network that can remove clouds and generate visible light images from the multispectral images taken as inputs. This is achieved by extending the input channels of cGANs to be compatible with multispectral images. The networks are trained to output images that are close to the ground truth using the images synthesized with clouds over the ground truth as inputs. In the available dataset, the proportion of images of the forest or the sea is very high, which will introduce bias in the training dataset if uniformly sampled from the original dataset. Thus, we utilize the t-Distributed Stochastic Neighbor Embedding (t-SNE) to improve the problem of bias in the training dataset. Finally, we confirm the feasibility of the proposed network on the dataset of four bands images, which include three visible light bands and one near-infrared (NIR) band.

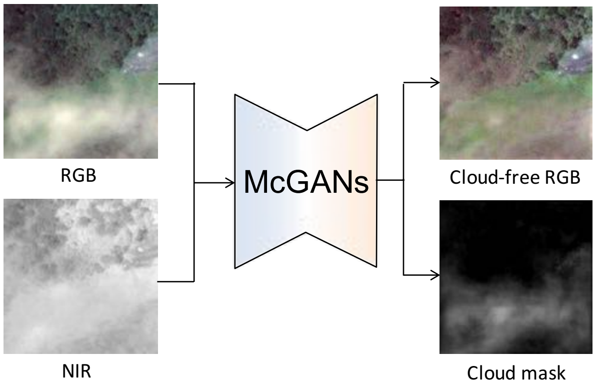

Figure 1: McGANs for cloud removal

1. Introduction

Satellite images have been widely utilized in various of fields such as remote sensing, computer vision, environmental science and meteorology. With the help of satellite images, we can observe the situation on the ground for natural environment monitoring (pollution, forest or rivers), transportation improvement and prompt emergency response to disasters. There are many research area dealing with satellite images, e.g., object recognition from the satellite images, change detection for ground usage or disaster situation analysis.

However, the obscurity caused by the cloud makes it unstable to monitor the situation on the ground with a visible light camera. To be unaffected by the cloud, images captured by longer wavelengths are introduced. Synthetic Aperture Radar (SAR) [6] is such an example, which improves visibility even in the presence of clouds. On the other hand, the spatial resolution decreases as the wavelength increases. Furthermore, the image captured by a long wavelength differs considerably in appearance from the one captured by visible light. This affects the visibility for observation.

In this paper, we propose Multispectral conditional Generative Adversarial Networks (McGANs) based on conditional Generative Adversarial Networks (cGANs), for cloud removal from visible light RGB satellite images with multispectral images as inputs. See Fig.1 for illustration. Compared with cGANs, the input channels of McGANs are expended for multispectral images. For the input of RGB images obscured by clouds and the registered NIR images, McGANs is trained to output the RGB images that are close to the ground truth. However, it is impractical to capture the cloud-free and the cloud obscured images of the completely same scene at the same time. Hence, we synthesize images with the simulated clouds over the ground truth RGB images to generate the training data. Furthermore, the prediction accuracy is expected to be improved by training the networks to detect the region of cloud simultaneously. Both the synthesized and the ground truth RGB images are color corrected to eliminate the affection of color tone caused by variety of imaging conditions such as weather, lighting and the processing method of the image sensor.

In the available dataset, the ratio of images of the forest or the sea is very high, which will introduce bias in the training dataset if uniformly sampled from the original dataset. Thus, we utilize the t-Distributed Stochastic Neighbor Embedding (t-SNE) [13] to reduce the bias problem of the training dataset. Finally, we confirm the feasibility of the proposed networks on the dataset of four bands images, which includes three visible light bands and one near-infrared (NIR) band.

2. Related Work

In the field of remote sensing, microwave is usually utilized since it is unaffected by the cloud cover. Synthetic Aperture Radar (SAR) is mounted on airplanes and satellites to overcome the shortage of low spatial resolution of the microwave. Nonetheless, the resolution of SAR images is still much lower than that of the images captured by visible light. Besides, it is difficult to understand the SAR images directly. To improve the visibility of SAR images, there also exists the work about coloring these SAR images [6].

In the field of computer vision, many dehazing methods have been proposed for RGB images only [8, 2] or for both RGB and NIR images [19, 5, 20]. The pre-knowledge or assumption about the color information of the hazed imaged is necessary in the former method. In the latter, NIR images, which possess higher penetrability through fog than the visible light, are used as the guide to dehaze the RGB images.

Generative Adversarial Networks (GANs) [7] is the most relevant to our work. GANs is consisted of two types of networks, Generator and Discriminator. Generator is trained to generate images that cannot be discriminated by Discriminator with the ground truth, while Discriminator is trained to discriminate between the generated images by the Generator and the ground truth. The conditional version of GANs was also proposed in [14]. However, learning by GANs is unstable. To increase the stability, convolutional networks and Batch Normalization are introduced to Deep Convolutional Generative Adversarial Networks (DCGANs) [17] is proposed.

Research about image generation based on cGANs and DCGANS has been widely applied for image restoration or the removal of certain objects such as rain and snow [15, 21]. In particular, the method in [10] can generate general and high-quality images by combing Generator of U-Net [18] and Discriminator of PatchGAN [12]. The Generator of U-Net spreads the missing spatial features in the convolution layers of Encoder to each layer of Decoder by introducing the skip connection between layers of Encoder and Decoder. PatchGAN is able to model the high frequencies for sharp details by training the Discriminator on the image patches. Generally, these cGANs-based methods predict the obscured regions of the image with the surrounding unobscured information only from the input RGB images.

Based on the aforementioned research, we propose the cloud removal networks by taking the advantage of the color information from visible light images and the high penetrability from images captured by longer wave. The proposed networks predict the obscured region from not only the RGB images but also images captured by longer wavelengths that can partly or completely penetrate the cloud. Our final purpose is to implement the networks that can merge SAR images captured by the cloud-penetrating microwave. As the first step, we construct and evaluate the networks for cloud removal with the visible light RGB images and the near-infrared spectrum NIR images in this work (the region of NIR wavelength is the closest to visible light).

3. Dataset Generation for Cloud Removal

In this work, images captured by the WorldView-2 earth observation satellite are used. Both visible light images and the NIR images possess the resolution of $20,000 \times 20,000$ with the spatial resolution of 0.5 [m/pixel]. We chose eight comparatively cloudless images, which mainly captured urban areas, for actual learning. In total, 37,000 images with a resolution of 256×256 are extracted for training McGANs.

3.1. Synthesis of cloud-obscured images

Both cloud obscured images and cloud-free images are indispensable to train the networks for cloud removal, as they form the training and ground truth data respectively.

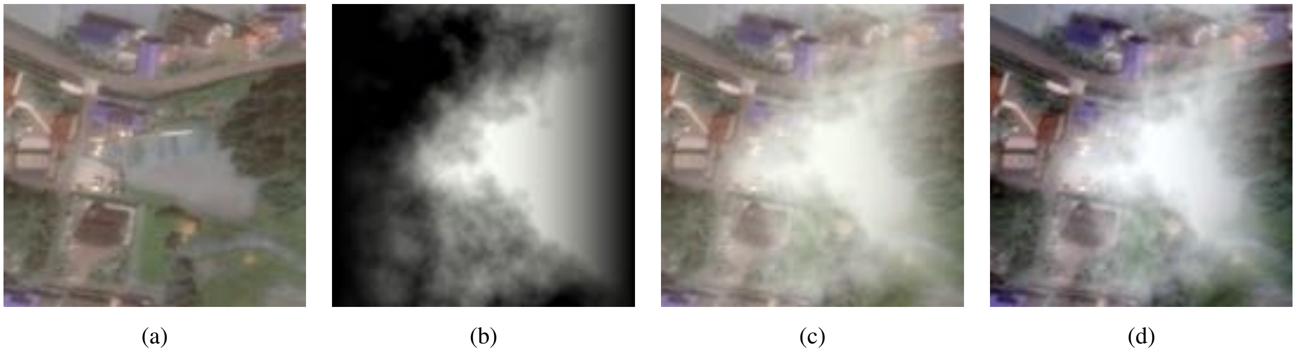

Figure 2: Synthesis of cloud obscured images. a: Original RGB image. b: Simulated cloud using Perlin noise. c: Merged image with the cloud by alpha blending. d: Final result after color correction

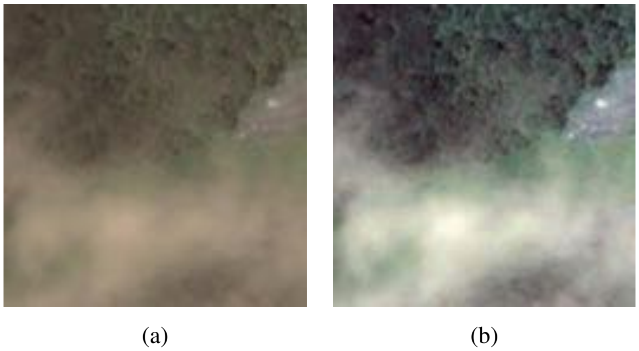

Figure 3: Example of color correction for real RGB image with cloud. a: Original RGB image obscured by the cloud. b: Color corrected result.

However, the appearance varies greatly as the imaging conditions, such as lighting and status on the ground, changes with time even for the same location. Therefore, we create the dataset for learning by synthesizing the simulated cloud on the cloudless or cloud-free ground truth images. Furthermore, to compensate for the difference in color tone between the cloud synthesized images and the original images, color correction [9, 4] is performed on both images.

In this work, the clouds are simulated by Perlin noise [16] firstly. Then the simulated clouds are combined with the RGB images by alpha blending to generate obscured images. Fig.2 shows an example of the image synthesis process. The RGB image (Fig.2a) is overlaid by a Perlin noise simulated cloud (Fig.2b) with the alpha blending method to synthesize the image (Fig.2c). Then generated image is further processed by color correction (Fig. 2d). To show the necessity of color correction, we take another image (Fig.3) for comparison. Fig.3a is the original RGB image of a different location from that in Fig.2. The color corrected result is shown in Fig.3b. By comparing the two groups of images, we can observe that the variety of color tone is greatly improved with the process of color correction.

3.2. Uniformization of the dataset with t-SNE

Since most of the earth is covered with seas and forests, the contents of the satellite images used in the work are also mainly of these two types. If we randomly sample the images for training, the learning result is prone to overfitting in certain categories due to the bias of the training data. Hence, we utilized t-SNE to sample the images by categories to avoid this problem.

First, we extract a feature vector of 4096 dimensions from each image with the AlexNet [11]. The extracted feature vectors are mapped to the 2D space with t-SNE. Then, we uniformly sample 2000 images from the 2D feature space to create the training dataset.

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [3] and the land use image dataset UC Merced land use dataset [1] (21 classes and 100 images for each class) are used for training the AlexNet. The processed results of the features from the two datasets after using t-SNE are shown in Fig.4. Fig.4a shows the distribution of the training images mapped with features from ImageNet dataset, and Fig.4b shows the result with features from UC Merced land use dataset. In the Fig.4a, images of the urban areas are clustered in the upper region, forest images are clustered at the right, images of the sea are clustered in the lower region and images of farmlands are clustered at the left. We can see that the images are well clustered by their categories. We also can see a similar result in Fig. 4b except that some images from the same category are distributed separately, e.g., images of forests are divided into the left and the lower parts. This is probably caused by the differences between the images used in this work and the

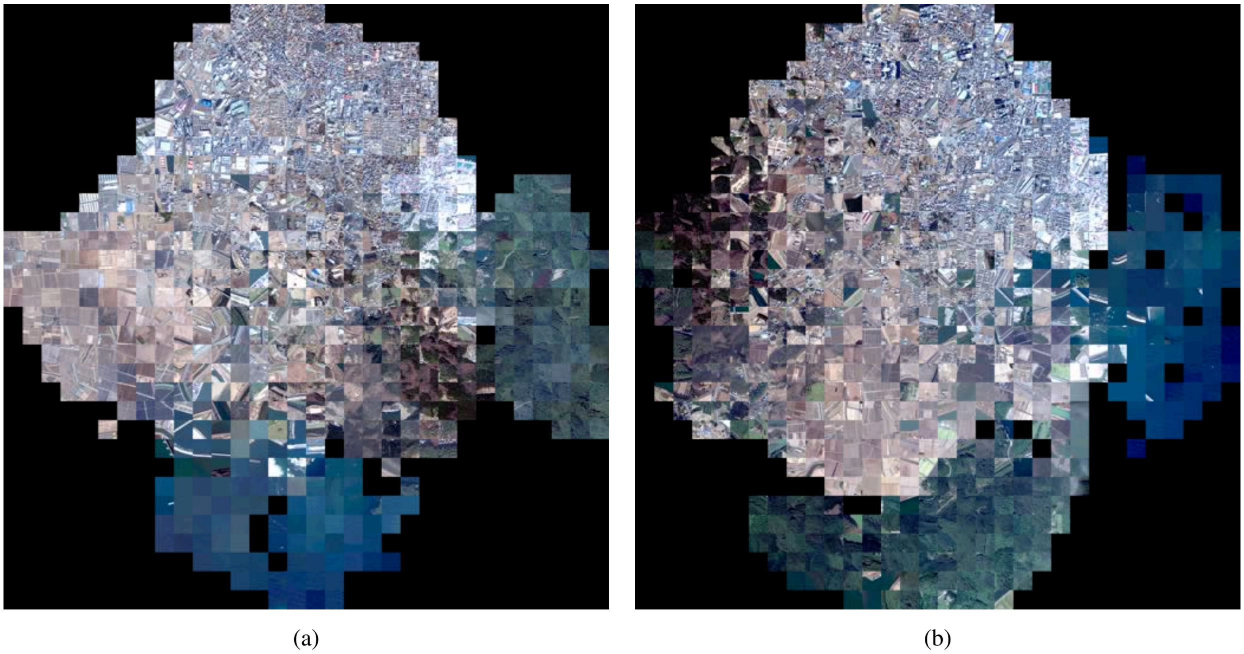

Figure 4: Visualization by t-SNE. a:ImageNet [3]. Images of urban areas are clustered in the upper region, forests images are clustered at the right, images of the sea are clustered in the lower region, ant the farmlands are clustered at the left. b:UC Merced Land Use Dataset [1]. Some images from the same category are distributed separately, for example images of forests are divided into the left and the lower parts.

images in UC dataset, in addition to the insufficiency of the images in the UC dataset. Therefore, we adopt the features extracted by the AlexNet from ImageNet for t-SNE.

The number of images in each cluster is shown in a heat map in Fig.5. From Fig.5 we can see that the images are uniformly distributed except in some regions of the grids. Images are uniformly sampled by the grid to improve the overfitting caused by the bias of in the training data.

4. Multispectral conditional Generative Adversarial Networks (McGANs)

In this paper, we propose Multispectral conditional Generative Adversarial Networks (McGANs), which extends the input of cGANs to multispectral images in order to be capable of merging input visible light images and images of longer wavelengths to remove clouds from the visible light images. The detailed architecture of McGANs are shown in Fig. 6 and Tab.1.

We extend the input of the cGANs model proposed in [10] to four channels RGB-NIR images [1]. Furthermore, the output is also extended to a total number of four channels, including the predicted RGB image after cloud removal and

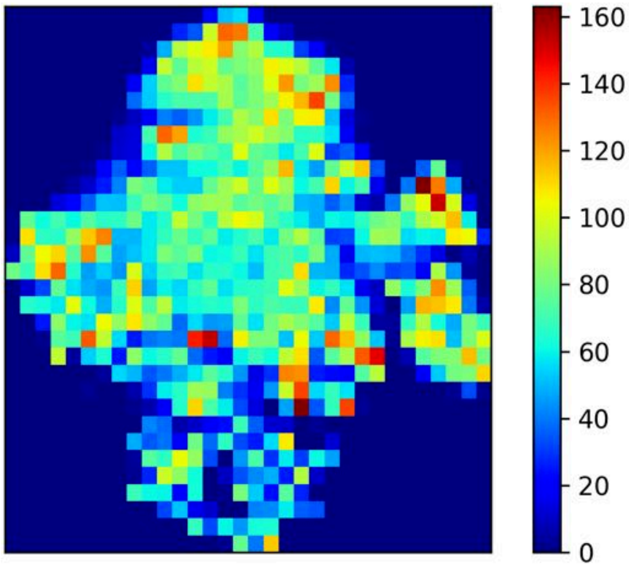

Figure 5: Heat map of image distribution mapped by t-SNE. The colors indicate the number of images in the corresponding 2D feature space.

[1] By adding images captured using other wavelength, such as far infrared rays and microwaves, the input will be further extended

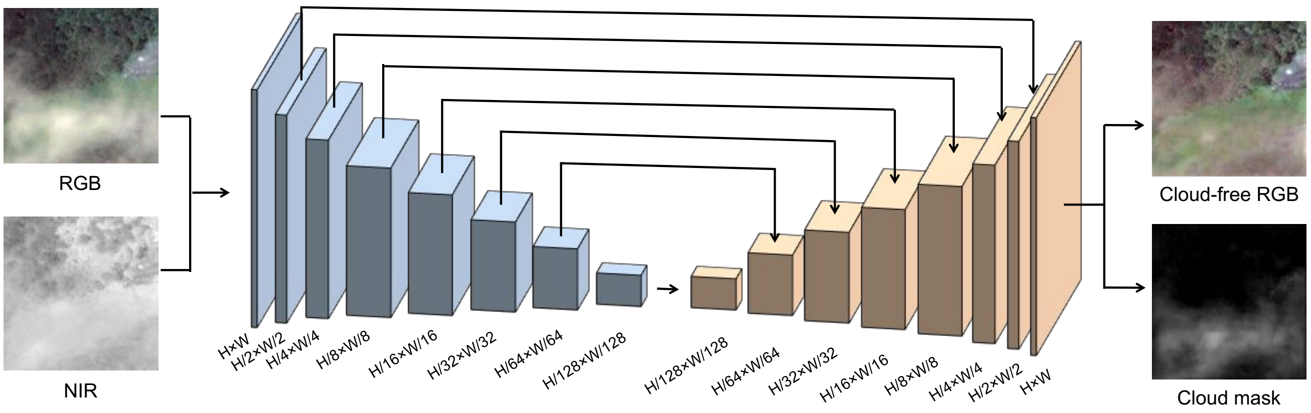

Figure 6: Network Architecture of Generator

Table 1: Network Architecture of McGANs

Encoder	Decoder	Discriminator
CR (64, 3, 1)	CBRD (512, 4, 2)	CBR (64, 4, 2)
CBR (128, 4, 2)	CBRD (512, 4, 2)	CBR (128, 4, 2)
CBR (256, 4, 2)	CBRD (512, 4, 2)	CBR (256, 4, 2)
CBR (512, 4, 2)	CBR (512, 4, 2)	CBR (512, 4, 2)
CBR (512, 4, 2)	CBR (256, 4, 2)	C (1, 3, 1)
CBR (512, 4, 2)	CBR (128, 4, 2)	
CBR (512, 4, 2)	CBR (64, 4, 2)	
CBR (512, 4, 2)	C (4, 3, 1)	

the grayscale mask image, which is estimated simultaneously to improve the prediction accuracy. The input RGB-NIR image, the output RGB image and the cloud mask image are normalized to $[-1, 1]$ at each channel and then transferred to the network.

Network Architecture

Details of the network structure about McGANs used in this work are shown in Tab.1. The layer of Convolution, Batch Normalization, and ReLU are represented by C, B, R respectively. D indicates that the Dropout is applied. Numbers in parentheses indicate the number, size, stride of the convolution filters sequentially. In addition, Leaky ReLU is used in all ReLU layers of Encoder and Discriminator.

The objective of a conditional GAN can be expressed as

$$\begin{aligned}\mathcal{L}_{cGAN}(G, D) =& \mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x, y)] + \\ & \mathbb{E}_{x\sim p_{data}(x), z\sim p_z(z)}[\log(1 - D(x, G(x, z)))],\end{aligned} \tag{1}$$

where Generator G tries to minimize the objective against an adversarial Discriminator D that tries to maximize it. To encourage less blurring, L1 loss can be added to the objective as follows [10]

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \mathcal{L}_{L1}(G). \tag{2}$$

Let I_M be the input multispectral image and I_T be the target RGB image with a total of four channels, including RGB and the grayscale mask image of the cloud. The L1 Loss function (denoted as $\mathcal{L}_{L1}$) of the Generator is defined in Eq.3. λ_c represents the weight of each channel for the loss calculation [2], and $\phi(I_M)$ represents the predicted result from the input image I_M from the trained networks.

$$\mathcal{L}_{L1}(G) = \frac{1}{4HW}\sum_{c=1}^{4}\sum_{v=1}^{H}\sum_{u=1}^{W}\lambda_c|I_T^{(u,v,c)} - \phi(I_M)^{(u,v,c)}|_1 \tag{3}$$

5. Evaluation Results

To evaluate our proposed method, experimental results are listed and discussed in this Section. From the experimental results, we expect to show that the proposed McGANs are able to improve visibility by cloud removal with RGB and NIR satellite images.

As explained earlier, the satellite images captured at different times (even though they might be of the same area), vary greatly in their appearance as imaging conditions, such as lighting and the situation on the ground, change. This makes it difficult to acquire the ground truth of the area blocked by the cloud. We use 2,000 groups of images as described in Sec.3 to train the network. Each group includes an image of the area not obscured by the cloud, a mask image of the simulated clouds using Perlin noise, a synthesized image and an NIR image. All images are processed with color correction. The number of minibatch is set to 1 and the number of epochs is 500.

[2] λ_c is set to 1 in this work.

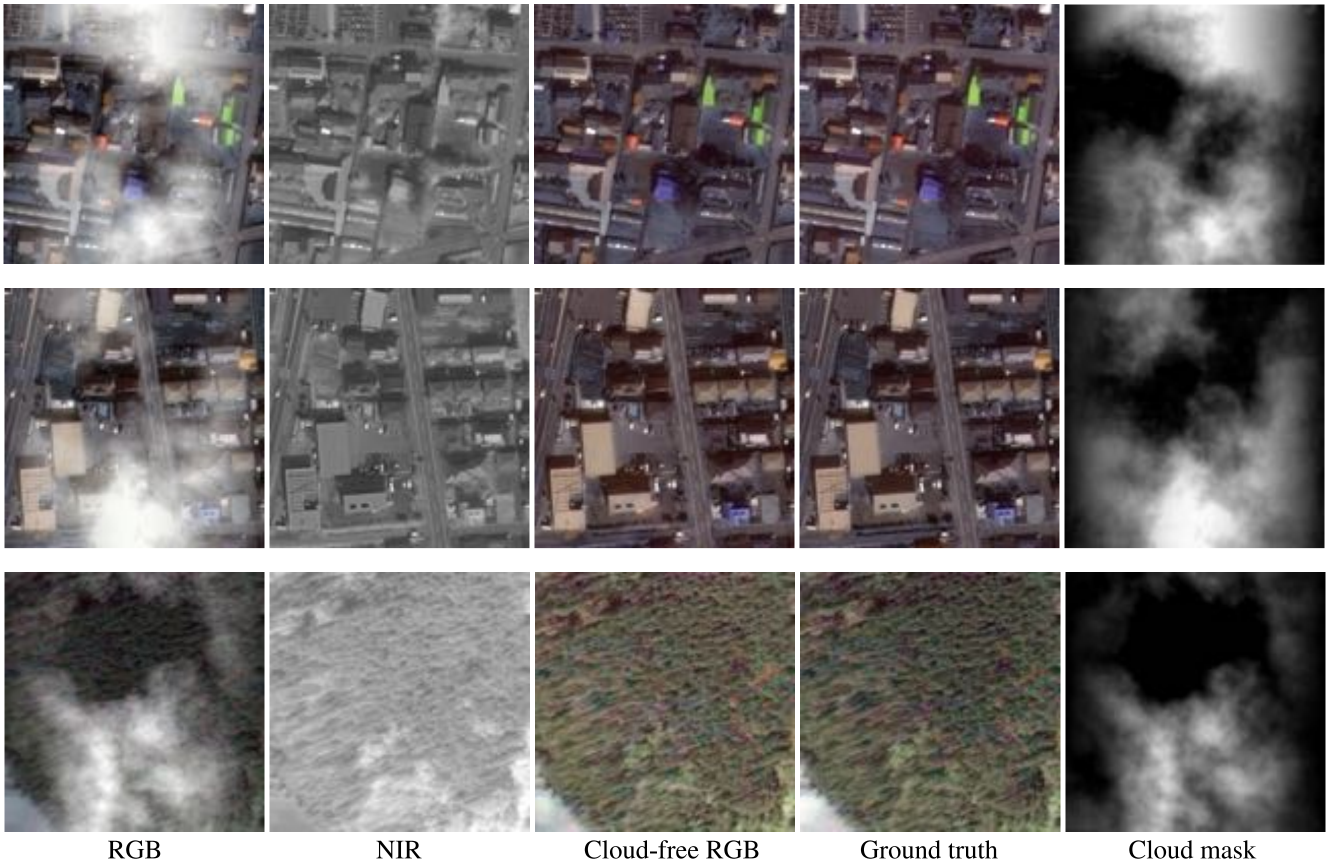

Figure 7: Prediction results by McGANs with the synthesized cloud images

To verify the advantage of using multispectral images for cloud removal, we also compare them against the RGB images generated by the networks (NIR-cGANs) with only NIR images as input. NIR images are used as input and images that are not obscured by the cloud are used as ground truth. The same dataset (as in McGANs) is used for training NIR-cGANs. The number of minibatch and epochs is also the same.

Sample results of the synthesized cloud obscured images are shown in Fig.7. The columns represent the synthesized cloud obscured RGB images, NIR images, RGB images predicted by McGANs, the ground truth and the mask images of the clouds predicted by McGANs, from left to right.

Sample results of real cloud obscured images are shown in Fig.8, Fig.9 and Fig.10. The columns represents RGB images obscured by the cloud, NIR images, RGB images predicted by McGANs, RGB images predicted by NIR-GANs and the mask images of the clouds predicted by McGANs, from left to right. From Fig.8, we can observe that although the images, which are generated only with NIR images, look like visible light RGB images, their colors differ from the ground truth. While the clouds are well removed in the predicted results by McGANs except for the region obscured by the cloud that infrared can not penetrate. Even for these regions in the predicted images, the color appears similarly to the very light color in the input images. This also proves that McGANs dose not predict color from the information only from the NIR images.

On the other hand, in Fig.9, we can see that the white object is erroneously recognized as the cloud from the output mask image. This indicates that it is difficult to separate the cloud from the white object with only the visible light images and NIR images, when they are overlapped. In addition, as seen in Fig.10, clouds are not removed when they are too thick to be penetrated by NIR. The purpose of this research is to observe the real situation on the ground. Thus, the regions blocked by clouds in the NIR image will not be predicted, which is different from [15]. When predicting the area blocked by clouds in both visible light and NIR images, it is necessary to model the cloud penetration of NIR based on the visible light images, process the simulated cloud with the penetration model and then synthesize

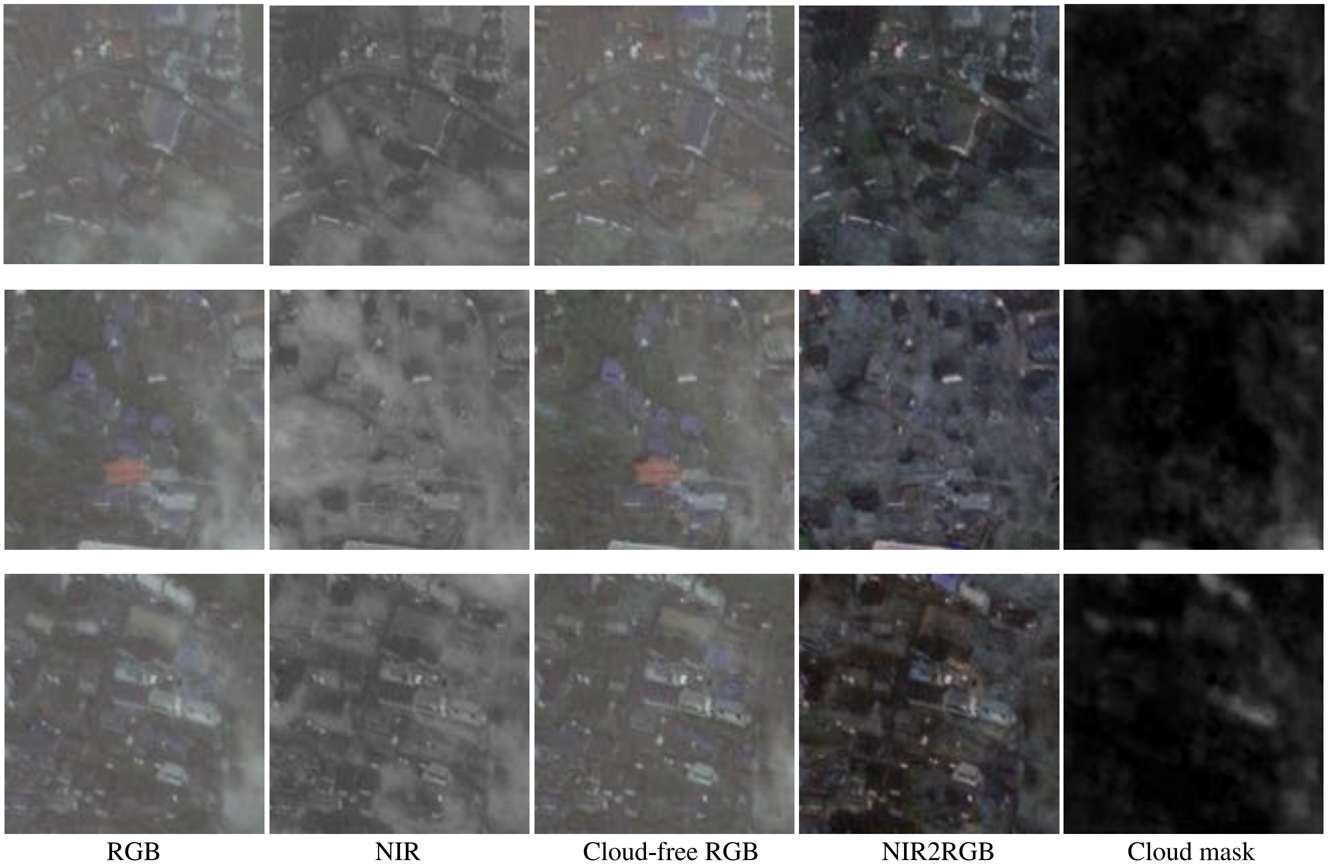

Figure 8: Prediction results by McGANs with real cloud images

the modeled cloud on the NIR images. To verify the necessity of the NIR images, we also compare the results generated by our proposed method with that generated from only a RGB image as the input. For thin clouds that can be partly penetrated by the visible light, the results dose not differ much. However, for clouds that can be only penetrated by NIR light, the result with the presence of NIR appears more natural as shown in Fig.11. We can see some line contours of the roads on the ground from the upper left part of the NIR image in Fig.11, while these contours are occluded in the RGB image. This can be considered as the reason why the result generated with both the NIR and RGB images looks more natural that generated with the RGB image.

From the above results, we have confirmed that the proposed McGANs can remove clouds and predict the color properly when the cloud is thin enough to be penetrated by the NIR.

6. Conclusion

In this paper, we have proposed a method to remove thin clouds from satellite images formed using visible light by extending cGANs to multispectral images. The dataset for training networks is constructed by synthesizing simulated clouds with Perlin noise over images without clouds, which makes it possible to generate cloud obscured training images and ground truth of the same area. In addition, to avoid overfitting to certain categories caused by biased datasets, we introduce t-SNE to sample images uniformly in each category. Finally, the experimental results evaluated on the constructed data prove that the clouds in the visible light images can be removed if they are penetrated in NIR images.

In the future, we will extend McGANs to far infrared (FIR) images and SAR images which captured by longer wavelengths and build the networks that can remove all the clouds in the visible light images. The findings obtained by analyzing the filters of McGANs in this work can also be applied to establish the model of cloud penetration for waves at each wavelength region or to the physical model of SAR. In addition, the simulated clouds with Perlin noise used in this work are somewhat different from real clouds in visible light images. Therefore, statistical analysis of actual cloud images is necessary to improve the reality of the simulated clouds for training data. Furthermore, we aim to improve the prediction accuracy for different areas by increasing the number and variety of images.

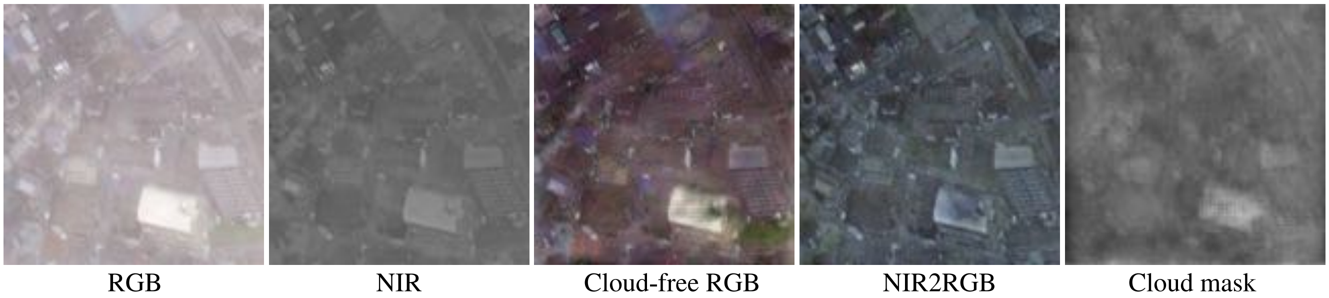

Figure 9: Failure case due to a white object

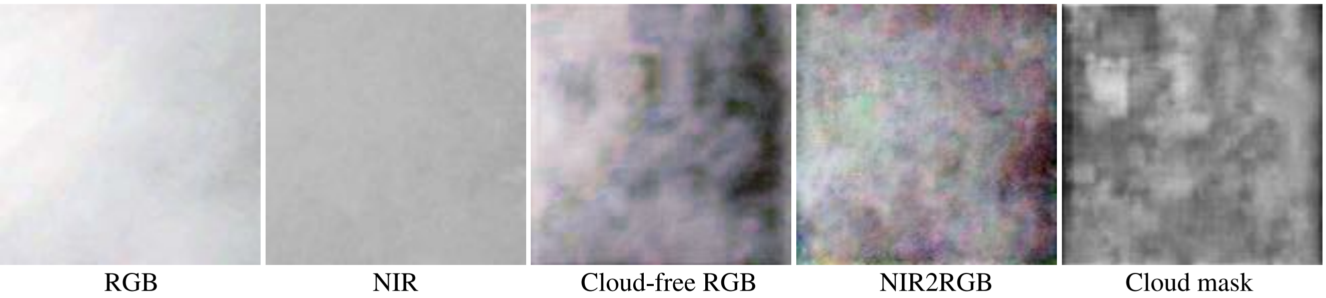

Figure 10: Thick cloud case

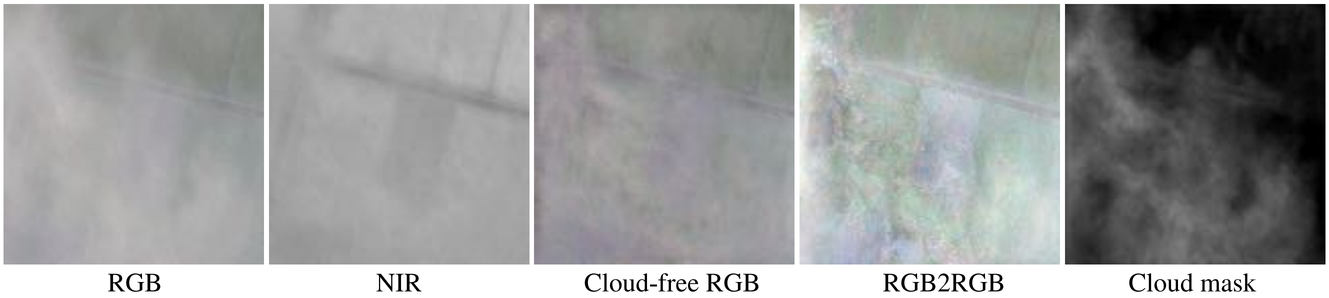

Figure 11: A prediction result with cGANs with only a RGB image

References

[1] Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification.

[2] D. Berman, T. Treibitz, and S. Avidan. Non-local image dehazing. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1674–1682, June 2016.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[4] H. S. Faridul, T. Pouli, C. Chamaret, J. Stauder, A. Tremeau, and E. Reinhard. A Survey of Color Mapping and its Applications. In *Eurographics 2014 - State of the Art Reports*. The Eurographics Association, 2014.

[5] C. Feng, S. Zhuo, X. Zhang, L. Shen, and S. Susstrunk. NEAR-INFRARED GUIDED COLOR IMAGE DEHAZING. In *IEEE International Conference on Image Processing (ICIP)*, pages 2363–2367, 2013.

[6] R. Furuta. Synthetic Aperture Radar (SAR) Utilization for Disaster Management. In *Technological Seminar on Environmental Monitoring*, 2014.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in*

Neural Information Processing Systems (NIPS), pages 2672–2680, 2014.

[8] K. He, J. Sun, and X. Tang. Single Image Haze Removal Using Dark Channel Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(12):2341–2353, 2011.

[9] R. W. G. Hunt. *The Reproduction of Colour*. John Wiley & Sons, 2005.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv preprint arXiv:1611.07004*, 2016.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[12] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision (ECCV)*, pages 702–716, 2016.

[13] L. v. d. Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(Nov):2579–2605, 2008.

[14] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014.

[15] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context Encoders: Feature Learning by Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016.

[16] K. Perlin. Improving noise. 21(3):681–682, 2002.

[17] A. Radford, L. Metz, and S. Chintala. UNSUPERVISED REPRESENTATION LEARNING WITH DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORKS. *arXiv preprint arXiv:1511.06434*, 2015.

[18] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[19] L. Schaul, C. Fredembach, and S. Süsstrunk. Color Image Dehazing using the Near-Infrared. In *IEEE International Conference on Image Processing (ICIP)*, pages 1629–1632, 2009.

[20] T. Shibata, M. Tanaka, and M. Okutomi. Unified Image Fusion based on Application-Adaptive Importance Measure. In *IEEE International Conference on Image Processing (ICIP)*, pages 1–5, 2015.

[21] H. Zhang, V. Sindagi, and V. M. Patel. Image Deraining Using a Conditional Generative Adversarial Network. *arXiv preprint arXiv:1701.05957*, 2017.