# AERIAL IMAGE AND MAP SYNTHESIS USING GENERATIVE ADVERSARIAL NETWORKS

*Jun Gu[1,2,3], Yue Zhang[1,2], Wenkai Zhang[1,2], Hongfeng Yu[1,2], Siyue Wang[4], Yaoling Wang[1,2,3], Lei Wang[1,2]*

[1]Institute of Electronics, Chinese Academy of Sciences, Beijing, China
[2]Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing, China
[3]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China
[4]Northeastern University Department of Electrical and Computer Engineering

## ABSTRACT

Accurate automatic conversion between aerial images and maps is a valuable and challenging task in computer vision and computer graphics. Deep convolutional neural networks (CNN) have achieved promising results on this task but the results accuracy is not ideal. In this paper, we propose a solution to improve the precision and quality of the transforming results. The core learning method is based on generative adversarial networks (GANs). A novel generator and a multi-scale discriminator are introduced in our network. The generator operates at the progressive method to gurantee the spatial consistency between the inputs and outputs, and our multi-scale discriminator focuses on increasing the capacity of the network and guides the generator to generate better results. In particular, our architecture can also be used as a general neural network for style translation. Analytic experiments on the aerial-to-map dataset show that our network outperforms the existing method, advancing both accuracy and visual appearance.

***Index Terms***— Generative adversarial network, image-to-image translation, aerial photos

## 1. INTRODUCTION

Maps and aerial photos are essential public resources that are widely used in our regular travel. It's hard for ordinary people to draw a map without continuous training when given a correlated aerial image, and it's also a time-consuming and laborious task even for professional people. Therefore, reducing the time and labor cost of drawing is very meaningful. Meanwhile, aerial images resource can better show the landscape of the city with the aerial view which allow people to obtain a stronger sense of 3D. Therefore, specially designed techniques are essential for achieving a high-quality automatic conversion of aerial images and maps. Fig. 1 illustrates the corresponding images in these two domains.


**Fig. 1**. Example images in aerial photos (left) and maps (right).

Recently, domain transfer methods based on Generative Adversarial Networks (GAN) [1] have drawn considerable attention. Isola et al. [2] proposed a general purpose framework, which is based on the conditional GAN, to pixel-to-pixel image synthesis problems. However, this basic method requires paired image sets for the training process. Zhu et al. [3] introduced a cyclic manner into GAN framework to perform image translation withn unpaired training data.

Although significant success between the domains translation task of aerial images and maps has been achieved with GAN, there still remains some room to produce images with more acceptable quality and precision. Analyzing the results of CycleGAN, we find that the position of the objects in the images generated by the network is not aligned well with the input images, and the edges of objects such as houses and roads are also not clear.

In this paper, a novel GAN-based network is proposed to improve the conversion results precision and quality between the domains of aerial images and maps. The generator in our network is divided into two sub-networks which progressively synthesis more satisfactory results. Meanwhile, a multi-scale discriminator is proposed to enhance the network capability by integrating different scales information and improve its ability to guide the generator. Additional, our architecture can also be treated as general neural network for image-to-image translation with unpaired datasets.

This work was supported in part by the National Natural Science Foundation of China under Grants 41801349.

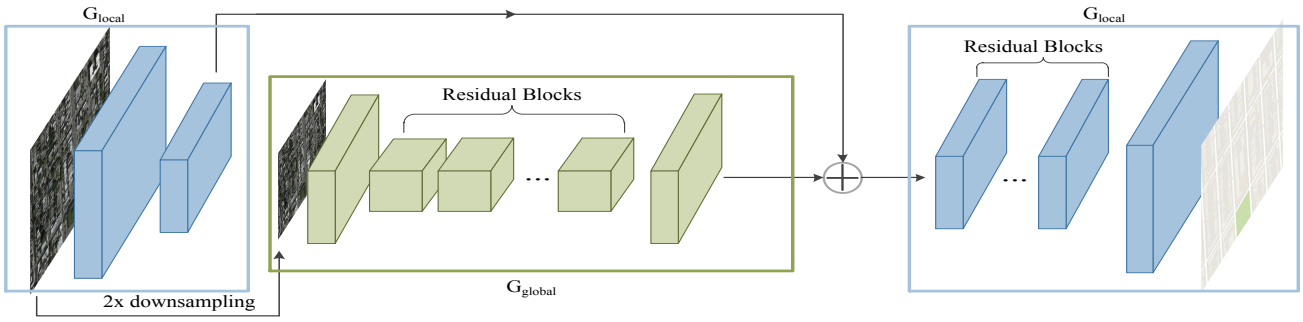


**Fig. 2**. Network architecture of our novel generator.

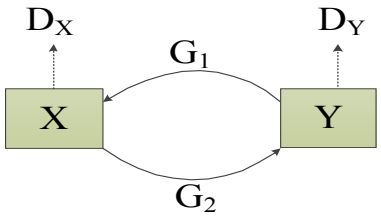


**Fig. 3**. Two mapping functions $G_1$ and $G_2$ in our model, and associated adversarial discriminators $D_X$ and $D_Y$.

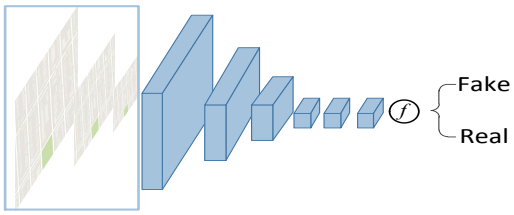


**Fig. 4**. Network architecture of our multi-scale discriminator.

## 2. THE PROPOSED METHOD

### 2.1. Framework

The framework has been used for addressing the image-to-image translation [2] problem which converts an input image from one domain $X$ to the output image in another domain $Y$. In our application, the domain $X$ represents the aerial photos while the domain $Y$ contains images with the desired characteristics of maps. Our model contains a forward mapping $G_1 : X \rightarrow Y$and a backward mapping $G_2 : Y \rightarrow X$. In addition, two adversarial discriminators $Dx$ and $Dy$ are introduced in the whole structure, where $Dx$ aims to distinguish between images $\{x\}$ and translated images $\{G_1\{y\}\}$; in the same way, $Dy$ aims to discriminate between $\{y\}$ and $\{G_2\{x\}\}$. The whole framework is shown in Fig. 3.

### 2.2. Loss Function

The full obeject of loss function are formulates as a combination of adversarial and cycle consistency loss:

$$\begin{aligned} L(G_1, G_2, D_X, D_Y) =& L_{GAN}(G_1, D_Y, X, Y) \\ &+ L_{GAN}(G_2, D_X, Y, X) \\ &+ L_{CYC}(G_1, G_2) \end{aligned} \quad (1)$$

The original CycleGAN uses simple GAN objective as the adversarial losses to both mapping functions. Recently provides an alternative way of using least square GAN [4] which is more stable and generates higher quality results. We use the least square adversarial loss as the critic function. The objective function $L_{GAN}(G, D)$ is calculated as follows:

$$L_{GAN}(G) = \frac{1}{2} E_{x \sim p_{data}(x)}[(D(G(x)))^2] \quad (2)$$

$$\begin{aligned} L_{GAN}(D) =& \frac{1}{2} E_{y \sim p_{data}(y)}[(D(y) - 1)^2] \\ &+ \frac{1}{2} E_{x \sim p_x(x)}[(D(G(x)) + 1)^2] \end{aligned} \quad (3)$$

The cycle consistency loss is the same as utilized in CycleGAN.

$$\begin{aligned} L_{CYC}(G_1, G_2) =& E_{x \sim p_{data}(x)}[||G_2(G_1(x)) - x||_1] \\ &+ E_{y \sim p_{data}(y)}[||G_1(G_2(y)) - y||_1] \end{aligned} \quad (4)$$

where we denote the data distribution as $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$.

### 2.3. Generator Architecture

The generator is decomposed into two sub-networks as shown in Fig. 2: $G_{global}$ and $G_{local}$. $G_{global}$ is designed as a global generator operator at a low resolution and $G_{local}$ outputs an image with a high resolution. The $G_{global}$ is proposed as the main part to generator the basic structure of the images and the $G_{local}$ can be regarded as a local enhancer network to capture fine structure of the inputs.


**Fig. 5**. Comparisions for mapping aerial photos to maps on Google Maps. From left to right: input, ground truth, CycleGAN, our generator and our model.

The CNN architecture of $G_{global}$ is similar to one proposed by Johnson et al. [5], which has been proven successful for style transfer task. It contains two stride convolution blocks with stride $\frac{1}{2}$, nine residual blocks and two transposed convolution blocks. Each residual block consists of a convolutional layer, instance normalization layer and ReLU activation.

The $G_{local}$ consists of two convolutional blocks, a set of residual blocks and a transposed convolutional blocks. According to Fig. 2, the input to the residual blocks in $G_{local}$ is the element-wise sum of the output feature map of the two convolutional layers in $G_{local}$, and the feature map of $G_{global}$.

### 2.4. Discriminator Architecture

Due to hardware limitation and large detection distance, scene distribution of aerial photos is more complicated than ordinary image. This characteristic is a challenge to the GAN discriminator to differentiate the real and synthesized images. To address this issue, a multi-scale discriminators are proposed in this paper, as shown in Fig. 4. Three discriminators with the same network structure operate at different image scales. The generated images are downsampled by a factor of 2 and 4 to create an image pyramid of 3 scales. Then we trained the three discriminators to distinguish real and synthesized images at the three different scales, respectively. The discriminator operating at the finest scale is specialized in guiding the generator to produce finer outputs. We use PatchGAN [6] for our discriminator. With these discriminators, the minimax game of GAN can be described as follows:

$$min_G max_{D_1,D_2,D_3} \sum_{i=1,2,3} L_{GAN}(G, D_i) \qquad (5)$$

## 3. EXPERIMENT AND RESULT

### 3.1. Dataset

The dataset, which first used in [2], is from Google Maps. This dataset is composed of six folders: two for training, two for validation, and others for test. There are 1096 aerial photos with a resolution of $256 \times 256$ and the same numbers of paired maps in the training folders. Images are shuffled during our training to ensure that the data is not paired.

### 3.2. Implementation Details

We implement the proposed models via Pytorch framework and train them using NVIDIA Tesla P100 GPUs. Our models are optimized with Adam by setting $\beta 1 = 0.9$, $\beta 2$ = 0.999 and $\epsilon = 10^{-8}$. We set the initial learning rate to $0.0002$ and halve it decreased every 50 epochs. The models were trained with a batch size $= 1$, which showed empirically better results on validation. We use the peak signal-to-noise ratio (PSNR)

[dB] and structural similarity index measure (SSIM) [7] as one criterion to evaluate the performance.

**Table 1**. PSNR and SSIM scores for CycleGAN, our improve generator and our network, evaluated on the validation dataset.

| Model | CycleGAN | Our Generator | Ours Model |
|---|---|---|---|
| Aerial2Maps | 27.9527 | 28.7608 | **29.1910** |
| PSNR/SSIM | 0.6939 | 0.7513 | **0.7606** |
| Maps2Aerial | 15.7559 | 15.7682 | **16.0755** |
| PSNR/SSIM | 0.2326 | 0.2415 | **0.2439** |

### 3.3. Comparision with the state-of-the-art

Our model is compared with the stat-of-the-art method CycleGAN. For a fair and convincing comparison, we slightly adjust the method and retrain the network under our experimental dataset to obtain their best performance. Fig. 5 illustrates some translation results of our method and CycleGAN. Comparing the first and last columns of the result shown in Fig. 5, the translation performance is slightly better when the objects in the input images are neatly arranged. However, in the case of intricate roads, buildings and water bodies, our approach has shown a certain advantage. On the whole, our approach is able to generate more clearly edge of the objects and keep the spatial consistency. Table 1 presents the ultimate mean PSNR and SSIM over the validation images of the two methods, and our proposed method outperforms the CycleGAN.

### 3.4. Model Analysis

To demonstrate the effects of the generator and discriminator in our model and verify the usefulness of these two components. We carry out a set of ablation experiments. It can be concluded from the indicators and the visual appearance of results that the network with our novel generator and the multi-scale discriminator achieves the best performance.

In the mapping task of aerial photos to maps, the two scales generator can differentiate the position of the objects well. When translating the maps to aerial images domain, the improved generator can capture the edges of the objects better than the original generator in CycleGAN. The results as shown in Fig. 5 confirms that our generator works better for the task of complex domain data conversion and can capture the distribution of the target domain well.

The multi-scale discriminator in our network is inspired by the method of multi-scale pipelines used in computer vision [8]. The capacity of the discriminator is improved with the multi-scale information and our discriminator can direct the generator to produce more accurate images. As shown in Table 1, the indicators and the visual performance have been improved with this multi-scale discriminator further.

## 4. CONCLUSION

In this paper, a novel generator and a multi-scale discriminator are proposed into the Cycle Adversarial Network for the domain translation of maps and aerial photos. Our generator synthesizes the image to the original scale with a progressive strategy, and the discriminator adopts a multi-scale approach to utilize the information from a different level to enhance the network power. Experiment results on the benchmark dataset demonstrate that the proposed method can guarantee accurate results and outperform the existing state-of-the-art unpaired image translation method of CycleGAN. In particular, our network can also be used in other image style translation tasks.

## 5. REFERENCES


[1] 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*, 2014, pp. 2672–2680.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

[4] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2813–2821.

[5] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[6] Chuan Li and Michael Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 702–716.

[7] Zhou Wang, A. C Bovik, H. R Sheikh, and E. P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans Image Process*, vol. 13, no. 4, pp. 600–612, 2004.

[8] Peter Burt and Edward Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on communications*, vol. 31, no. 4, pp. 532–540, 1983.