Text to Remote Sensing Image Generation with Structured Generative Adversarial Networks

Rui Zhao and Zhenwei Shi*, *Member IEEE*

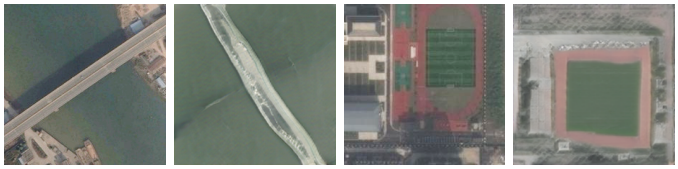
***Abstract*—Synthesizing high-resolution remote sensing images based on the given text descriptions has great potential in expanding the image data set to release the power of deep learning in the remote sensing image processing field. However, there has been no efficient research carried out on this formidable task yet. Given a remote sensing image, the structural rationality of ground objects is critical to judge it whether real or fake, e.g., real bridges are always straight while a sinuous one can be easily judged as fake. Inspired by this, we propose a multi-stage structured generative adversarial network (StrucGAN) to synthesize remote sensing images in a structured way given the text descriptions. StrucGAN utilizes structural information extracted by an unsupervised segmentation module to enable the discriminators to distinguish the image in a structured way. The generators of StrucGAN are thus forced to synthesize structural reasonable image contents which could enhance the image authenticity. The multi-stage framework enables the StrucGAN to generate remote sensing images with increasing resolution stage by stage. The quantitative and qualitative experiments results show that the proposed StrucGAN achieves better performance compared with the baseline, and it could synthesize high resolution, realistic, structural reasonable remote sensing images which are semantically consistent with the given text descriptions.**

***Index Terms*—Remote Sensing Image Synthesize, Text Description, Generative Adversarial Networks, Structural Rationality.**

I. INTRODUCTION

DEEP learning technology greatly drives research progress in remote sensing image processing. Massive data is the cornerstone of high-performance deep learning algorithms, while the high-cost imaging platforms (e.g., airborne, space-borne) impose restrictions on the scale of the remote sensing data set. This limits the deep learning technology to exert its full potential in the remote sensing field.

Recently, Generative Adversarial Networks (GANs) [1] have drawn great attention in a variety of research fields. The interesting but challenging task which needs to generate image according to the given natural language descriptions, namely text to image generation, is active one of them. The success of GANs in this task shed light on the possibility of controllably generating images that can be passed for genuine ones. If GANs can generate sufficiently realistic remote sensing images, then we can construct large-scale remote sensing image data sets in a controllable and low-cost manner. This will unlock the potential of deep learning in remote sensing image processing tasks.



(a+) (a-) (b+) (b-)

(a) A bridge is built over the river.
(b) A playground is surrounded by some buildings.

Fig. 1. Two cases of fake images generated by AttnGAN given the text descriptions and real images that correspond to the same text descriptions. The image (a+) is the real image that corresponds to the caption (a), while the image (a-) is the generated fake image given the caption (a). Since bridges are always straight or of low curvature, the sinuous bridge in (a-) can be easily told the fake. The synthetic playground in (b-), which is in the shape of a square, is also easily judged as fake. These two cases suggest that the structure of the synthesized object is an important feature that affects whether the synthesis is realistic.

Great progress has been achieved in the text to natural image generation. Reed *et al.* [2] proposed a deep architecture and GAN formulation to effectively synthesis plausible images given the text descriptions. Their follow-up work [3] synthesizes images conditioned on more specific instructions (e.g., object locations). Zhang *et al.* [4], [5] proposed Stacked Generative Adversarial Networks (StackGANs) which stacked several different scale GANs to generate photo-realistic images given text descriptions. Xu *et al.* [6] proposed an attentional generative network (AttnGAN) to pay attention to the relevant words in descriptions and the image sub-regions when synthesizing the image. Qiao *et al.* [7] proposed a semantic-preserving text-to-image-to-text framework to guarantee semantic consistency between the text description and visual content.

Despite the recent success in the text to natural image generation, the text to high-resolution remote sensing image generation remains challenging. Bejiga *et al.* proposed the first work [8] that dealt with the text to remote sensing image generation. In which, a conditional GAN is applied to generate very low spatial resolution grayscale remote sensing images from ancient text descriptions of geographical areas. In their following works [9], [10], Bejiga *et al.* improved the text encoding by using a pre-trained Doc2Vec encoder [11], which could utilize different levels of information available from the input text. However, these works generated grayscale remote sensing images with a very low spatial resolution which missed many details. Zheng *et al.* [12] proposed a reranking audio-

The work was supported by the National Key R&D Program of China under the Grant 2019YFC1510905, the National Natural Science Foundation of China under the Grant 61671037 and the Beijing Natural Science Foundation under the Grant 4192034. *(Corresponding author: Zhenwei Shi.)*

Rui Zhao (e-mail: ruizhaoipc@buaa.edu.cn) and Zhenwei Shi (Corresponding author, e-mail: shizhenwei@buaa.edu.cn) are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China.

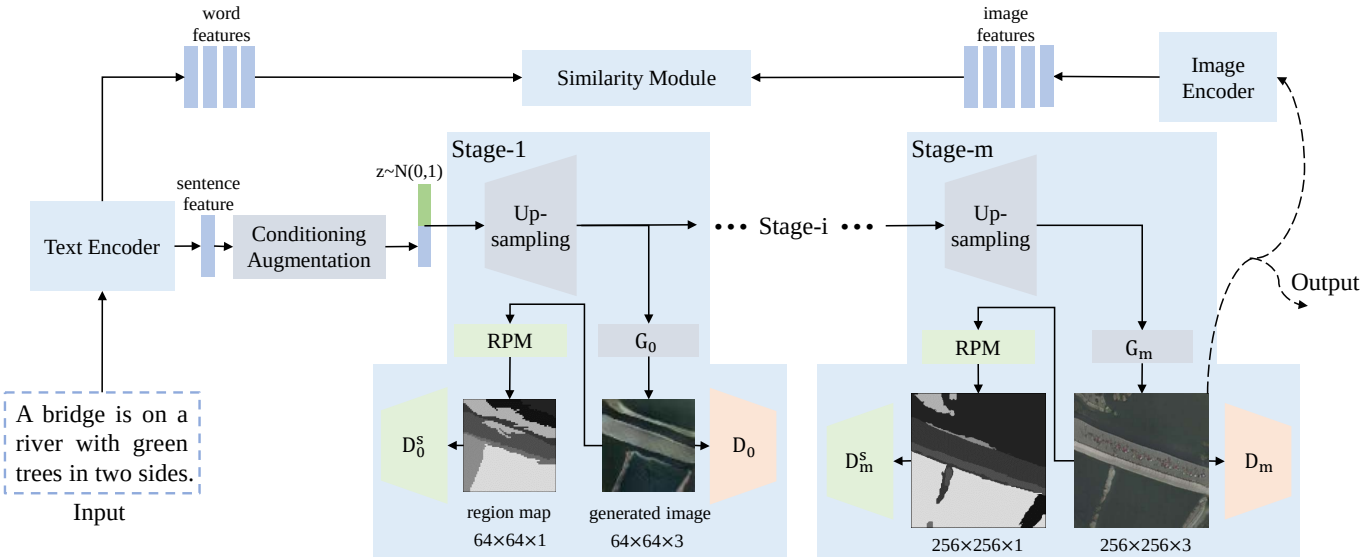


Fig. 2. The overview of the proposed structured generative adversarial network for text to remote sensing image generation.

image translation method to retrieved remote sensing images given the audio descriptions. In this work, remote sensing images were real data retrieved from the existing database, and the input was audio description rather than text description. Other researches focus on the inverse task, namely image caption generation [13], which generates text descriptions based on the input remote sensing images.

The main challenge in the text to high-resolution remote sensing image generation task is that the contents of remote sensing images have strong structure characteristics (e.g., bridge, playground). The unnatural structure of the synthetic contents will make people spot the fake. For example, since bridges always are straight or have small curvature, the synthetic sinuous bridges will be easily judged as fake. Another example is that playgrounds are always elliptical, while the synthetic square one would be judged fake, as shown in Fig. 1.

To address the above challenge, we propose a structured generative adversarial network (StrucGAN) to generate remote sensing images in a structured way given the text descriptions. The proposed StrucGAN uses AttnGAN as the backbone to achieve the multi-stage refinement text to image generation. Novelly, to synthesize structural reasonable image content, StrucGAN utilizes an unsupervised segmentation module to extract structured information of the remote sensing image contents and construct structured discriminators to distinguish authenticity based on the structured information. Since the discriminators can distinguish images in a structured way, the generators are forced to generate structural reasonable image content. The experiments on the RSICD dataset [14] show that the proposed StrucGAN can generate more realistic remote sensing images compared with the baseline.

Our work mainly has the following two contributions:

- We shed light on the possibility of improving the structural rationality of contents to synthesize realistic remote sensing images.
- The structured generative adversarial network is proposed to synthesize realistic high-resolution remote sensing images that are semantically consistent with the given text description.

II. Methodology

In the text to image generation task, the existing architectures [2], [3], [8]–[10] are essentially conditional GANs, the architectures [4], [5] are stacked conditional GANs, and the architecture AttnGAN [6] are stacked conditional GANs with attention mechanism. We reimplement the AttnGAN as the backbone and add novel branches based on our proposed structured mechanism. Each branch consists of a region proposal module and a structured discriminator.

The overview of the proposed StrucGAN is shown in Fig. 2. We briefly review the structure of the backbone in section II-A and detailedly introduce the proposed structured mechanism in section II-B. Then we introduce the modified loss functions in section II-C.

A. Overall Structure

A bi-directional Long Short-Term Memory (LSTM) [15] is used as the text encoder to extract semantic features from the text description. The output word features matrix is indicated by $w \in \mathrm{R}^{M \times N_w}$, where M is the dimension of the word feature vector and N_w is the number of words. The sentence feature vector, $s \in \mathrm{R}^{M}$, is the concatenated last hidden states of the bi-directional LSTM. The conditioning augmentation module [4] converts the sentence vector s to the conditioning vector s, which is a latent variable randomly sampled from an independent Gaussian distribution $N(\ (s)); \ (s))$.

The huge gap between semantics and image contents makes it difficult to generate high-resolution images based on the text descriptions in one step. To tackle this challenge, the stacked generative adversarial networks are used to gradually generate

images of small-to-large scales. The stage-i generator G_i, takes the hidden state h_i as input and generate image $\hat{x}_i$, namely

$$\hat{x}_i = G_i(h_i). \tag{1}$$

The hidden state h_i is generated by the up-sampling module F_i, which is defined as follows.

$$h_i = \begin{cases} F_i(z, s), & i = 1, \\ F_i(h_{i-1}, F_i^{att}(w, h_{i-1})), & i = 2, 3, ..., m, \end{cases} \tag{2}$$

where z is a vector sampled from a standard normal distribution. The up-sampling module F_i increases the spatial size of the hidden state by twice. Namely, the length and width of the image generated in each stage are twice that of the image generated in the previous stage. F_i^{att} is the attention module that uses two fully connected layers to map word features w and previous hidden state h_{i-1} into the same space, takes their product, and normalize the product through the softmax function as attention weight which is then used to weight word features to generates the word-context vector. Finally, the previous hidden state and the corresponding word-context features are added together to input the up-sampling module.

For each generator G_i, one pixel-level discriminator D_i is constructed using downsampling blocks and fully connected layers. The pixel-level discriminator takes the generated image and the sentence feature vector as input and produces the decision score.

A convolutional neural network (CNN) with two followed perceptron layers is used to map the image to semantic vector $v \in \mathrm{R}^{M \times N_p}$ and global semantic vector $v \in \mathrm{R}^{M}$, where N_p denotes the spatial size of the feature map extracted by the last convolutional layer. The similarity module, namely the deep attentional multimodal similarity model (DAMSM) proposed in [6], measures similarity between the semantic vectors and the word features through the local and global matching scores. The local matching score is defined as follows,

$$R(x, y) = log(\sum_{i=1}^{T} exp(\frac{(\sum_{j} \alpha_{i,j} v_j)^T w_i}{k \sum_{j} \alpha_{i,j} v_j k k w_i k})^{\frac{1}{\gamma}}), \tag{3}$$

where x denotes the text description, y denotes the image, w_i denotes the i^{th} word feature and $\alpha = softmax(w^T v)$ denotes the attention weights. The γ is a smoothing factor set to 5. The global matching score is defined as

$$R'(x, y) = \frac{v^T s}{kvkksk}. \tag{4}$$

B. Structured Mechanism

For each stage of generating images of the small-to-large scale, besides the pixel-level discriminator, we further construct a branch based on the proposed structured mechanism to force the generator to produce structural reasonable images. Each branch consists of a region proposal module and a structured discriminator.

The region proposal module (RPM) takes the image as input and produces the region map.

$$r_i = RPM(x_i). \tag{5}$$

Specifically, we first apply the guided image filter [16] to smooth the generated image while keeping its edges and structures. Then we use a pre-defined method "Selective Search" [17] as the region proposal module to segment the input image to a set of class-agnostic segmentation proposals based on the color and texture features. Then the region map is generated by replacing the pixel value in image x_i with the mean value of the image pixels in each corresponding proposal. To make the selective search method adapt to the remote sensing images, we specifically tune three key parameters to make sure that segmentation proposals are not too fragmented while retaining as many detailed regions as possible. These parameters include a smooth parameter σ of the Gaussian filter, a parameter m_{size} which controls the minimum bounding box size of the proposals, and a scale parameter s_{scale} which controls the initial segmentation scales. These parameters are set as $\sigma = 0.8$, $m_{size} = 100$, and $s_{scale} = 100$.

The structured discriminator is constructed using down-sampling blocks and fully connected layers. Besides the sentence feature, the structured discriminator takes the region map generated by the region proposal module as input. The stage-i structured discriminator D_i^s computes the decision score as follows,

$$D_i^s(r_i, s) = F^d([F_i^r(r_i), F^s(s)]), \tag{6}$$

where F_i^r is the down-sampling module, F^s is a fully connected layer followed by a spatially replicate operation, and F^d is the decision score computing module constructed by a 1×1 convolutional layer followed by a fully connected layer. The square brackets denote channel dimension wise concatenation.

Since the structural information is extracted by the selective search module and transferred to the structured discriminator, the structured discriminator can distinguish the generated structural unreasonable image from the real one. This forces the generator to produce structural reasonable images.

C. Loss Functions

We proposed the structural loss to train the generators and discriminators in a structured way. The total loss functions of the proposed method contain adversarial loss, structural loss, and image-text matching loss. During the training, minimizing adversarial loss forces the model to generate realistic images, minimizing structural loss forces the model to generate structural reasonable images, and minimizing image-text matching loss forces the model to generate images that are semantically consistent with the input text description.

The adversarial loss function for each generator G_i is defined as

$$L_{G_i} = -\mathrm{E}_{\hat{x}_i \sim p_{G_i}}[\log D_i(\hat{x}_i)] - \mathrm{E}_{\hat{x}_i \sim p_{G_i}}[\log D_i(\hat{x}_i, s)], \tag{7}$$

while the adversarial loss function for each discriminator D_i is defined as

$$\begin{aligned} L_{D_i} = & -\mathrm{E}_{x_i \sim p_{data_i}}[\log D_i(x_i)] - \mathrm{E}_{\hat{x}_i \sim p_{G_i}}[\log(1 - D_i(\hat{x}_i))] \\ & - \mathrm{E}_{x_i \sim p_{data_i}}[\log D_i(x_i, s)] - \mathrm{E}_{\hat{x}_i \sim p_{G_i}}[\log(1 - D_i(\hat{x}_i, s))]. \end{aligned} \tag{8}$$

The structural loss function for each generator G_i is defined as

$$L_{G_i^s} = \mathbb{E}_{\hat{r}_i \sim p_{G_i}}[\log D_i^s(\hat{r}_i)] \quad \mathbb{E}_{\hat{r}_i \sim p_{G_i}}[\log D_i^s(\hat{r}_i; s)]; \quad (9)$$

while the structural loss function for each discriminator D_i is defined as

$$L_{D_i^s} = \mathbb{E}_{r_i \sim p_{data_i}}[\log D_i^s(r_i)] \quad \mathbb{E}_{\hat{r}_i \sim p_{G_i}}[\log(1 \quad D_i^s(\hat{r}_i))] \quad \mathbb{E}_{r_i \sim p_{data_i}}[\log D_i^s(r_i; s)] \quad \mathbb{E}_{\hat{r}_i \sim p_{G_i}}[\log(1 \quad D_i^s(\hat{r}_i; s))]; \quad (10)$$

where r_i denotes the region map corresponding to the real image x_i, while $\hat{r}_i$ denotes the region map corresponding to the generated image $\hat{x}_i$

The image-text matching loss is defined as

$$L_m = \sum_i^K log(\quad fR(\hat{x}_i; y_j)g_{j=1}^K) \quad \sum_i^K log(\quad fR(\hat{x}_j; y_i)g_{j=1}^K) \quad \sum_i^K log(\quad fR^0(\hat{x}_i; y_j)g_{j=1}^K) \quad \sum_i^K log(\quad fR^0(\hat{x}_j; y_i)g_{j=1}^K); \quad (11)$$

where is the softmax function and K is the size of a batch of generated image and text description pairs. Taking one of the softmax items, for example, it is defined as

$$fR(\hat{x}_i; y_j)g_{j=1}^K = \frac{exp(R(\hat{x}_i; y_i))}{\sum_j^K exp(R(\hat{x}_i; y_j))}: \quad (12)$$

The total loss function of generator is defined as

$$L_G = \sum_i^M (L_{G_i} + L_{G_i^s}) + \quad L_m; \quad (13)$$

where denotes a weight factor, which is set to 5.

III. Experiments

In the experiments, we implemented the model with three stages of generators, pixel-level discriminators, and structured discriminators. These three generators synthesize three-channel remote sensing images with the spatial size of 64 64 pixels, 128 128 pixels, 256 256 pixels, respectively.

A. Dataset and Metrics

Experiments are conducted on the remote sensing captioning dataset named RSICD which is constructed by Lu et al. [14]. It contains a total of 10921 high-resolution remote sensing images, of which the training set contains 8004 images, the validation set, and the test set contains 2187 images. Each image is labeled with 5 description sentences and there are 3323 different label words in the label file altogether.

We use the inception score [18] and R-precision [6] as the quantitative evaluation metrics. The inception score is defined as

$$\text{Inception Score} = exp(\mathbb{E}_x D_{KL}(p(y|x)||p(y))); \quad (14)$$

where x denotes the generated image, and y is the class label predicted by the inception model. Inception score is based on the intuition that a good model should generate diverse and meaningful images. That is, the KL divergence between the marginal distribution $p(y)$ and the conditional distribution $p(y|x)$ should be large.

TABLE I
Evaluation scores of different methods on the RSICD dataset [14].

Method / Data	Inception Score		R-precision(%) k=1	k=3	k=5	k=10
real data	7.32	.10	2.47	6.22	9.97	16.83
AttnGAN [6]	5.33	.14	1.85	4.17	7.31	14.81
StrucGAN(ours)	**5.84**	**.04**	**2.50**	**6.20**	**8.15**	**16.20**

Since the inception score can not measure whether the generated images are semantically consistent with the input text description, we further use the R-precision to evaluate in this respect. Specifically, for each generated image, we use it to query the corresponding text description from a candidate description set consist of one ground truth t_i and 99 randomly selected mismatching descriptions. Using the DAMSM to measure the similarity between the image and candidate descriptions, we rank the retrieval results and select top k results $Y_i^k = \{y_1; y_2; ...; y_k\}$. The R-precision is defined as follows,

$$\text{R-precision}_k = \frac{1}{n} \sum_n I(t_i \ 2 \ Y_i^k); \quad (15)$$

where I is the indicator function and $k = 1; 3; 5; 10$ in pur experiments.

The higher inception score means that the images generated by the model are more meaningful and diverse, while the higher R-precision means that the generated images have stronger semantic consistency with the text descriptions.

B. Quantitative Results

We compare our StrucGAN with the previous state-of-the-art AttnGAN model, which is borrowed from the field of natural image processing field, for text-to-image generation on the RSICD test set. Table. I shows the comparison results on quantitative evaluation metrics including the inception score and R-precision. The first row shows the scores of real data, which means we directly compute the metric scores using real images in the test set and the labeled text descriptions. The metric scores of real data can reflect the diversity of images in the data set and the difficulty of synthesizing these images.

Compared with AttnGAN as the baseline, the proposed StucGAN achieves state-of-the-art performance. StucGAN has a higher mean (5.84 compared with 5.33) and lower variance (0.04 compared with 0.14) on the inception score, which shows that it could generate more diverse and meaningful images. The R-precision scores improvements show that, compared with AttnGAN, StucGAN can generate images that have stronger semantic consistency with the input text descriptions.

C. Qualitative results

Fig. 3 shows the generated images byAttnGAN (second column) and StrucGAN (third column) based on the 5 different