# 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 multistage 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 multistage 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 that are semantically consistent with the given text descriptions.**

***Index Terms*— Generative adversarial networks (GANs), remote sensing image synthesize, structural rationality, text description.**

## I. Introduction

DEEP learning technology greatly drives research progress in remote sensing image processing. Massive data are the cornerstone of high-performance deep learning algorithms, while the high-cost imaging platforms (e.g., airborne and spaceborne) 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 that needs to generate images 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.

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 GANs (StackGANs) that stacked several different scale GANs to generate photorealistic 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 subregions 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.* [8] proposed the first work that dealt with the text-to-remote-sensing-image generation, in which a conditional GAN is applied to generate very low spatial resolution gray-scale remote sensing images from ancient text descriptions of geographical areas. In their following works, Bejiga *et al.* [9], [10] improved the text encoding by using a pretrained Doc2Vec encoder [11], which could utilize different levels of information available from the input text. However, these works generated gray-scale remote sensing images with a very low spatial resolution that missed many details. Zheng *et al.* [12] proposed a reranking audio–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 studies 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 and playground). The unnatural structure of the

Manuscript received December 30, 2020; revised February 26, 2021; accepted March 21, 2021. Date of publication March 30, 2021; date of current version January 5, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFC1510905, in part by the National Natural Science Foundation of China under Grant 61671037, and in part by the Beijing Natural Science Foundation under Grant 4192034. *(Corresponding author: Zhenwei Shi.)*

The authors are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, also 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 (e-mail: ruizhaoipc@buaa.edu.cn; shizhenwei@buaa.edu.cn).

Digital Object Identifier 10.1109/LGRS.2021.3068391

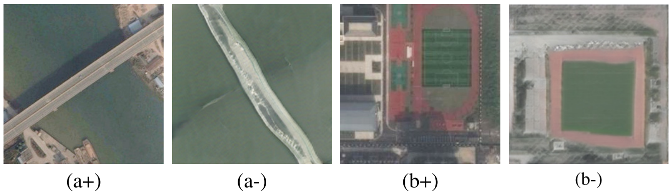

Fig. 1. Two cases of fake images generated by AttnGAN given the text descriptions and real images that correspond to the same text descriptions. (a) Bridge is built over the river. (b) Playground is surrounded by some buildings. (a+) Real image that corresponds to the caption of (a). (a−) Generated fake image that corresponds to the caption of (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.

synthetic contents will make people spot the fake. For example, since bridges are always 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 GAN (StrucGAN) to generate remote sensing images in a structured way given the text descriptions. The proposed StrucGAN uses AttnGAN as the backbone to achieve multistage 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 data set [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.

1) We shed light on the possibility of improving the structural rationality of contents to synthesize realistic remote sensing images.
2) The StrucGAN 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 (RPM) 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 bidirectional 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 $w \in \mathbb{R}^{M\times N_w}$, where $M$ is the dimension of the word feature vector and $N_w$ is the number of words. The sentence feature vector, $s \in \mathbb{R}^M$, is the concatenated last hidden states of the bidirectional LSTM. The conditioning augmentation module [4] converts the sentence vector $s$ to the conditioning vector $\bar{s}$, which is a latent variable randomly sampled from an independent Gaussian distribution $\mathcal{N}(\mu(s)), \Sigma(s))$.

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 StackGANs 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 upsampling module $F_i$, which is defined as follows:

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

where $z$ is a vector sampled from a standard normal distribution. The upsampling 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. $F_i^{\text{att}}$ 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 normalizes the product through the softmax function as attention weight, which is then used to weight word features to generate the word-context vector. Finally, the previous hidden state and the corresponding word-context features are added together to input the upsampling 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 the semantic vector $v \in \mathbb{R}^{M\times N_p}$ and the global semantic vector $\bar{v} \in \mathbb{R}^M$, 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:

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

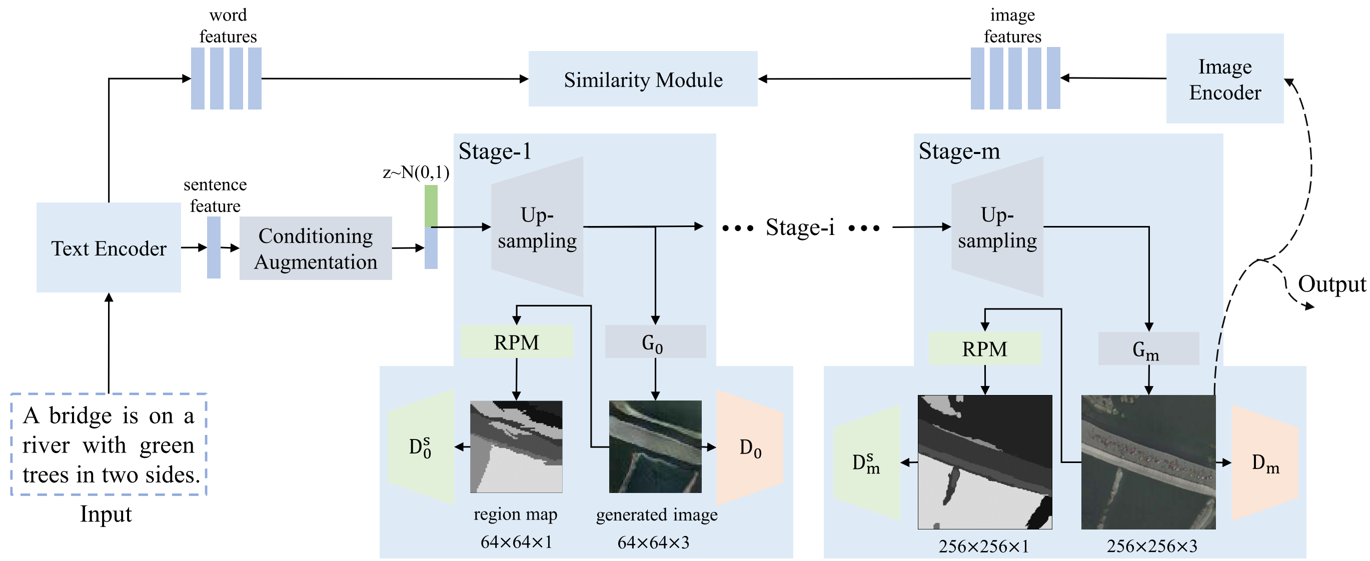

Fig. 2. Overview of the proposed StrucGAN for text-to-remote-sensing-image generation.

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

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

### 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 an RPM and a structured discriminator.

The RPM takes the image as input and produces the region map

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

Specifically, we first apply the guided image filter [16] to smooth the generated image while keeping its edges and structures. Then, we use a predefined method “Selective Search” [17] as the RPM 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 $\tau$ of the Gaussian filter, a parameter $m_{\text{size}}$ that controls the minimum bounding box size of the proposals, and a scale parameter $s_{\text{scale}}$ that controls the initial segmentation scales. These parameters are set as $\tau = 0.8$, $m_{\text{size}} = 100$, and $s_{\text{scale}} = 100$.

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

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

where $F_i^r$ is the downsampling 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 $1 \times 1$ convolutional layer followed by a fully connected layer. The square brackets denote channel dimensionwise 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

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

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

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

The structural loss function for each generator $G_i$ is defined as

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

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

$$\begin{aligned} L_{D_i^s} = &-\mathbb{E}_{r_i \sim p_{\text{data}_i}}[\log D_i^s(r_i)] - \mathbb{E}_{\hat{r}_i \sim p_{G_i}}[\log(1 - D_i^s(\hat{r}_i))] \\ &- \mathbb{E}_{r_i \sim p_{\text{data}_i}}\left[\log D_i^s(r_i, \bar{s})\right] - \mathbb{E}_{\hat{r}_i \sim p_{G_i}}[\log(1 - D_i^s(\hat{r}_i, \bar{s}))] \end{aligned} \tag{10}$$

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

$$\begin{aligned} L_m = &-\sum_i^K \log\left(\sigma\{R(\hat{x}_i, y_j)\}_{j=1}^K\right) - \sum_i^K \log\left(\sigma\{R(\hat{x}_j, y_i)\}_{j=1}^K\right) \\ &- \sum_i^K \log\left(\sigma\{R'(\hat{x}_i, y_j)\}_{j=1}^K\right) - \sum_i^K \log\left(\sigma\{R'(\hat{x}_j, y_i)\}_{j=1}^K\right) \end{aligned} \tag{11}$$

where $\sigma$ 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

$$\sigma\{R(\hat{x}_i, y_j)\}_{j=1}^K = \frac{\exp(R(\hat{x}_i, y_i))}{\sum_j^K \exp(R(\hat{x}_i, y_j))}. \tag{12}$$

The total loss function of the generator is defined as

$$L_G = \sum_i^m (L_{G_i} + L_{G_i^s}) + \lambda L_m \tag{13}$$

where $\lambda$ 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, and 256 × 256 pixels, respectively.

### A. Data Set and Metrics

Experiments are conducted on the remote sensing captioning data set named RSICD that 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, and the validation set and the test set contain 2187 images. Each image is labeled with five description sentences, and there are 3323 different label words in the label file altogether.

TABLE I

Evaluation Scores of Different Methods on the RSICD Data Set [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** |

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

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

where $x$ denotes the generated image, and $y$ is the class label predicted by the inception model. The 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.

Since the inception score cannot 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 $Y_i^k = \{y_1, y_2, \ldots, y_k\}$. The R-precision is defined as follows:

$$\text{R-precision}_k = \frac{1}{n} \sum_n I\left(t_i \in Y_i^k\right) \tag{15}$$

where $I$ is the indicator function and $k = 1, 3, 5, 10$ in our 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 that 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

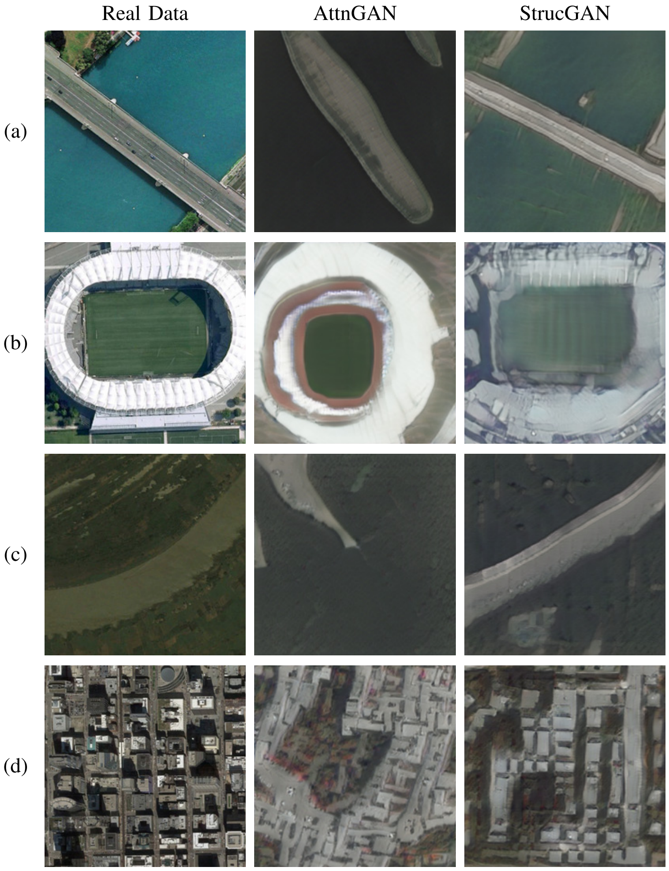

Fig. 3. Example results of generated images by AttnGAN (second column) and generated images by the proposed StrucGAN (third column) based on (a)–(d) input text description. These results all have a spatial size of 256 × 256 pixels. The real images are in the first column. (a) Bridge is built over the river. (b) Green football field is in the gym. (c) River flows through the wide land. (d) Many buildings are in commercial areas.

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 five different input text descriptions. The results in the second column would be easily identified as fake images, while the third column images look more realistic. In the first row, for example, the generated bridge by StrucGAN is long and straight, which is much more realistic than that one generated by AttnGAN. In the remaining examples, the playground generated by StrucGAN is elliptical, the generated river is winding, and the generated building is square; all of which are more realistic than those produced by AttnGAN.

The results in Fig. 3 show that the proposed StrucGAN could generate high-resolution images that are more structural reasonable than those generated by AttnGAN. These results also show that StrucGAN can generate semantically consistent images based on the input text descriptions.

## IV. Conclusion

A StrucGAN is proposed to synthesize high-resolution remote sensing images given the text description. The proposed StrucGAN extracts structured information of images with the selective search method to enable the discriminators to distinguish the image vraisemblance in a structured way. In this way, the generators are forced to synthesize more structural reasonable contents. The structural rationality of ground objects is critical to judge whether remote sensing images are real or not. Since StrucGAN utilizes structured information effectively, it achieves synthesizing high realistic remote sensing images. The quantitative and qualitative experiments' results show that StrucGAN can synthesize high resolution, realistic, structural reasonable remote sensing images, and these images are semantically consistent with the given text descriptions. The comparison experiment with AttnGAN shows that StrucGAN achieves better performance.

## References

[1] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016, *arXiv:1605.05396*. [Online]. Available: http://arxiv.org/abs/1605.05396

[3] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 217–225.

[4] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[5] H. Zhang *et al.*, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019.

[6] T. Xu *et al.*, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[7] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1505–1514.

[8] M. B. Bejiga, F. Melgani, and A. Vascotto, "Retro-remote sensing: Generating images from ancient texts," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 3, pp. 950–960, Mar. 2019.

[9] M. B. Bejiga, G. Hoxha, and F. Melgani, "Retro-remote sensing with Doc2 Vec encoding," in *Proc. Medit. Middle-East Geosci. Remote Sens. Symp. (M2GARSS)*, Mar. 2020, pp. 89–92.

[10] M. B. Bejiga, G. Hoxha, and F. Melgani, "Improving text encoding for retro-remote sensing," *IEEE Geosci. Remote Sens. Lett.*, early access, Apr. 14, 2020, doi: 10.1109/LGRS.2020.2983851.

[11] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[12] Z. Zheng, J. Chen, X. Zheng, and X. Lu, "Remote sensing image generation from audio," *IEEE Geosci. Remote Sens. Lett.*, early access, May 15, 2020, doi: 10.1109/LGRS.2020.2992324.

[13] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017.

[14] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[15] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[16] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[17] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[18] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2234–2242.