Text Pared into Scene Graph for Diverse Image Generation

Yonghua Zhu
Shanghai Film Academy, Shanghai University, Shanghai, China
zyh@shu.edu.cn

Jieyu Huang
Shanghai Film Academy, Shanghai University, Shanghai, China
jieyu88@shu.edu.cn

Ning Ge
Shanghai Film Academy, Shanghai University, Shanghai, China
gening@shu.edu.cn

Yunwen Zhu
Shanghai Film Academy, Shanghai University, Shanghai, China
eilleen31@shu.edu.cn

Binghui Zheng
Shanghai Film Academy, Shanghai University, Shanghai, China
zhengbh@shu.edu.cn

Wenjun Zhang*
Shanghai Film Academy, Shanghai University; Information Technology Academy, Shanghai Jian Qiao University, Shanghai, China
wjzhang@shu.edu.cn

ABSTRACT

Although significant recent advances in condition generative model have shown remarkable improvements for controlled image generation, the image generation for multiple complex objects is still a challenge. To address the challenge, we propose a module of text description parsed into scene graph, which can generate reasonable scene layout to ensure the generated image and object realistic. Our proposed method enhances the interaction between objects and global semantics by concatenates each object embedding with text embedding To preserve the local image semantics, the Spatially-adaptive normalization(SPADE) layer is added into the generator of our model. We validate our method on Visual Genome and COCO-Stuff, where qualitative results and ablation study demonstrate the ability of our model in generating images with multiple objects and complex relationships.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Machine learning approaches**; • **Neural network**;

KEYWORDS

Text-to-image generation, Scene graph, Image-text retrieval

ACM Reference Format:
Yonghua Zhu, Jieyu Huang, Ning Ge, Yunwen Zhu, Binghui Zheng, and Wenjun Zhang*. 2021. Text Pared into Scene Graph for Diverse Image Generation. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487075.3487158

1 INTRODUCTION

Reed [1] pointed out text-to-image(text2img) generation has two tasks to solve: firstly learn a text feature representation to captures the important visual details and secondly use these feature to synthesize a compelling image that a human might mistake for real. However, there is another core challenging that the distribution of images conditioned on the text description is highly multi-modal, in the sense that there are many plausible configurations of pixels that correctly illustrate the description.

In other words, same sentence depicts image usually corresponds to objects with various poses and diverse appearances. The previous works of scene graph [2–4] and layout [5–7] can generate more specific image, but lack of creativity for compute aid design. And people [8] is more accustomed to employ natural language as form of input when search the web for images.

In this paper, we aim to generate images with multiple objects and complex relationships from text description. By extending the backbone of [4], we propose a new framework, which is Text Pared Into Scene Graph Generative Adversarial Network (TS-GAN). The sentence is a linear structure parsed into graph-structured scene graph, which could better represent inner-objects relationship. Recently, there are some image-text retrieval works [9–13] using scene graph to learn the comprehensive unified representations to express multi-modal data. Concretely, they apply scene graph on evaluating the similarity of the image-text pairs by dissecting the input image and text sentence into it. Enlightened by these works, we propose a Scene Graph Generation From Text (SGGFT) module to take scene graph as the intermediate feature representation of text description in our model. To best of our knowledge, it is the first time to propose the module of text parsed into scene graph in text2img task. The text encoder employ GRU as encoder network to make each embedding vector preserve global semantics. Each object embedding generated from semantic layout concatenate text embedding that the object in the scene usually is related to the environment.

The main contributions of our proposed method are as follows:

- We propose the SGGFT module for to generating scene graph from the text to more in line with the habits of users.
- Each object embedding concatenates text embedding by our method, which can make object interact with global image semantics.
- We also develop various extended experiments to demonstrate the capacity of our model to generate complex and realistic images.

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487158

2 RELATED WORK

The image synthesis method based on Generative Adversarial Networks GANs [14] gradually improving in ability of generating high resolution, visual quality and diversity of generated images. In this work, we address conditional GAN [15], which is an extension of GAN by providing labels as additional input to the generator and discriminator for creating images that matching the specific input. Based on variety of additional input, conditional image generation can be divided into different subtasks, including text2img, image generation from scene graph and layout-to-image(layout2img) and so on.

2.1 Text2img

Reed [1] first introduced generative adversarial model to solve the task of text2img, but only achieve 64×64 resolution in image generation. Their follow-up work GAWWN [16] solved the problem how to synthesis images given text describing what to draw in which location and got better performance. Zhang [17, 18] extended generative model as multi-stages model StackGAN and StackGAN++ focus on generating photographic images. AttnGAN [19] leveraging the attention mechanism, each word in an input description had a different level of information depicting the image content, refines the images to high-resolution. Based on the multi-stage generative network, HD-GAN [20] used a hierarchical nesting model to encourage more effective use of text and image information without multi-stage generator and discriminator.

2.2 Scene Graph

Image generation [2–4] from scene graph used scene layout as an intermediate layer behind text input. They can generate multi-object and complex relationship images and ensure both image and object are realistic. The method based on scene graph generated less diversity images which means the pixel configuration dimension of images conforming to text semantics is low. Same sentence usually corresponds to objects with various poses and appearances, so the same scene also could be depicted in various ways.

2.3 Layout2img

Instead of learning a direct mapping from textual description or scene graph to an image, layout2img method generates high-quality and multiple-object images directly. Zhao [5] proposed image generation for layout, which comprising bounding boxes and object categories. LostGAN [6] and ObjGAN [7] improved the ability of generating image from layout. Although layou2img solves the problems of high-resolution and credibility, the input of natural language is more in line with the users' habits of searching and computer-aid design.

3 METHOD

The overall network architecture of our proposed model TS-GAN illustrated in Figure 1. Given the text input, our model first represents the text in two different embedding vectors. Text encoder represent text description as embedding φ_t capturing global semantic context feature. Then F_{ca} module takes φ_t as input to generate conditioning latent variable c_0 that enrich discontinuity in the latent data manifold, which is beneficial for learning the generator. The another representation of text description is semantic scene graph S though our proposed SGGFT. Semantic layout embedding was generated by the graph convolution network [2], M and B are the network respectively generate object's mask and corresponding coordinates for each object. Next, it is spatially concatenate to the global context vector φ_t. Generator is the encoder-decoder network architecture that output the results images based on input vector. Our generative model ensures that the generated images and objects are realistic and conform to the text semantics by adversarial training a pair of discriminators networks D_{img} and D_{object} [21]. Each of these components is described in more detail below.

3.1 Text Encoder

Let φ_t be the text embedding of the given description, we encode the text description by text encoder consists of a single layer bidirectional recurrent network with Gated Recurrent Units (GRUs). Where e_i is a global sentence vector and e is the matrix of word embedding, h_i^E is a hidden vector encoding the current word and its context. F_{ca}represents the Conditioning Augmentation [17] that converts the hidden vector h_i^E to the conditioning latent vector c_0. Further, the global context embedding φ_t conforming to specifications:

$$h_i^E = BiGRU\left(e_i, h_{i-1}^E, h_{i+1}^E\right) \tag{1}$$

$$\varphi_t = z \oplus F_{ca}\left(h_i^E\right)) \tag{2}$$

z is a noise vector sampled from a standard normal distribution and $\oplus$ is a spatial concatenation.

3.2 Scene Graph Generation from Text

A text description includes many objects and complex relationships. Hence, the information conveyed by a sentence can often be more explicitly represented as a scene graph of objects and their relationships. Firstly, we obtain the whole pre-trained network model through joint learning on pairs of text-image data, and secondly introduce the part of the text description parsed into scene graph into our network, SGGFT. The overall training pipeline of SGGFT is illustrated in Figure 2.

The scene graph generation from image SGGFI is an off-the-shelf scene graph generation method [11]. The semantic scene graph which are built from semantic triplets parsed by dependency parse trees [13, 22].

3.3 Generator

Our image generation method builds on and improves the method proposed in [4], as such it shares some of the basic architecture design principles outlined in Oron Ashual et al [4], A scene graph is represented as $G = \{O, E\}$, where $O = \{o_1, ..., o_n\}$ is the node-set of objects with each $o_i \in C$ and C is category of the object, and $E \subseteq O \times R \times O$ is the directed edge-set of form (o_i, r_i, o_j) where $o_i, o_j \in O$ and $r_i \in R$, R is a set of relationship categories.

Given a scene graph $G = \{O, E\}$ generated by SGGFT as input, our model employed graph convolution network [2] outputs a semantic layout, which aggregates information across all objects and edges in the graph as coarse 2D layout. The layout embedding size is set to 64×64. M is the mask regression network generate the object's mask m_i of shape $H \times W$. The network B predicts corresponding

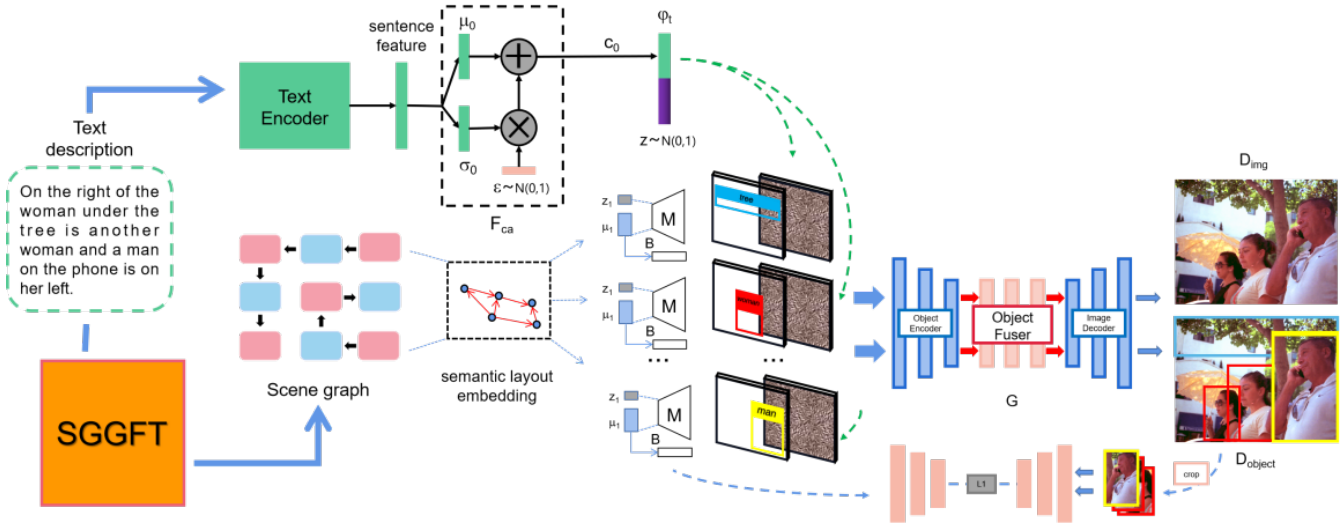

Figure 1: The Framework of TS-GAN. The Text Description Respectively through Text Encoder to Capture the Context Semantics and Sggft to Generate Scene Graph that Better Represent Inner-Object Relationship. The Text Embedding φ_t Concatenate the Object Embedding from Scene Graph to Generate The Results Image.

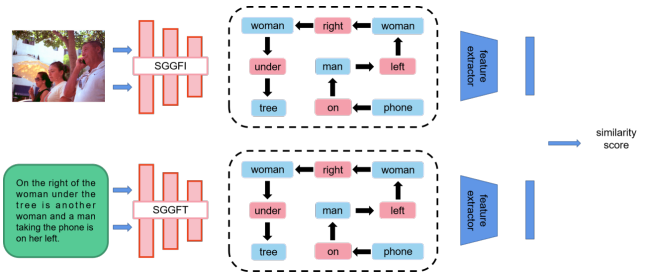

Figure 2: The Overall Training Pipeline of SGGFT as Illustrated. SGGFI and SGGFT Generate Scene Graph by Input Image and Text. Two Output Scene Graph Respectively Parsed into Feature Extractor to Encode. Feeds Them to a Multi-Layer Perceptron (MLP) to Computes Similarity Score.

bounding box of the object $b_i = \{x_0, y_0, x_1, y_1\} \in [0,1]^4$, encodes the coordinates as a ratio of image scale.
The text embedding φ_t spatially replicated to form $H \times W$ (e. g, of size 64×64) as image canvas and concatenate every object embedding vector. Not only generator needn't interpolation to bridge the image patch on blank canvas, but also enhance interaction between object and global semantics.

$$I = f\left(\varphi_t \oplus (m_i, b_i), \ldots, \varphi_t \oplus (m_n, b_n)\right) \tag{3}$$

The network Generator f is the encoder-decoder architecture. The dimension of the input vector H×W×D, which consist of all object embedding and φ_t, D is twice the number of object embedding. Specially, we add Spatially-adaptive normalization layer (SPADE) [23] in our image decoder, which could preserve local semantic information in image layout to make the results image realistic and the object recognizable.

3.4 Discriminator and Loss Function

We trained the generator network f adversarially against a pair of discriminator D_{img} and D_{object}. The discriminator is trained to classify an input x as real or fake by maximizing the objective [14].

$$L_{GAN} = \underset{x\sim p_{real}}{E} logD\,(x) + \underset{x\sim p_{real}}{E} \log(1 - D\,(x)) \tag{4}$$

The image discriminator D_{img} ensure the whole image appearance look realistic and is implemented as a fully convolutional network used in [25]. The loss of D_{img} is given as:

$$L_{D_{img}} = \lambda_1 L_{rec} + \lambda_4 L_{perceptual} \tag{5}$$

The reconstruction loss L_{rec} penalizes the $L1$ difference between the ground-truth image $\hat{\mathrm{I}}$ and the generated image I. The perceptual loss $L_{perceptual}$ compares the generated image with the ground truth image using the activation $\mathrm{F_u}$ of the VGG network [26] at layer u in a set of predefined layers U.
The object discriminator D_{object} encourage each object appear realistic and generate in desired region. The loss of D_{object} is a compound loss:

$$L_{D_{obj}} = \lambda_3 L_{mask} + \lambda_4 L_{box} + \lambda_5 L_{AC}^{obj} \tag{6}$$

Mask loss L_{mask} penalizing differences between ground-truth and predicted masks with pixel-wise cross-entropy; not used for models trained on Visual Genome. Box loss $L_{box} = bi - \hat{b}_1$ penalizing the L1 difference between ground-truth and predicted bounding boxes. Loss L_{AC}^{obj} Auxiliary Classifier [21] from D_{object} ensure each generated object is recognizable as the their corresponding category. Therefore, the final loss function L is defined as:

$$L = \lambda_1 L_{rec} + \lambda_2 L_{perceptual} + \lambda_3 L_{mask} + \lambda_4 L_{box} + \lambda_5 L_{AC}^{obj} \tag{7}$$

Where, each λ_i is the hyperparameter that we set $\lambda_1 = \lambda_2 = \lambda_3 = 10$, $\lambda_4 = 0.1$ and $\lambda_5 = 1$.

Table 1: Performance on COCO and VG in Inception, Frechet Inception Distance (FID) and Diversity Score

Method	Inception ScoreCOCO VG		FID ScoreCOCO VG		Diversity ScoreCOCO VG	
Real Images	16. 3 ± 0. 4	13. 9 ± 0. 5	-	-	-	-
sg2im	7. 3 ± 0. 1	6. 3 ± 0. 2	67. 96	74. 61	0. 02 ± 0. 01	0. 15 ± 0. 12
layout2im	**9. 1 ± 0. 1**	8. 1 ± 0. 1	38. 14	31. 25	0. 15 ± 0. 06	0. 17 ± 0. 09
Our method	8. 7± 0. 2	**8. 7 ± 0. 1**	**65. 3**	**48. 7**	**0. 43 ± 0. 07**	**0. 37 ± 0. 01**

4 EXPERIMENT

We evaluate our model on COCO-Stuff [27] and Visual Genome [28] datasets. We preprocess and split Visual Genome dataset following the settings of [2, 5]. In total, we have 62,565 training, 5,506 validation and 5,088 testing images. And we divide the COCO-Stuff 2017 val set into our own val and test sets, leaving us with 24972 train, 1024 val, and 2048 test images. Each experiment perform on the 2017 COCO-Stuff datasets which using images with 3 to 8 objects from 91 categories. We using images contain 3 to 30 objects from 178 categories on Visual Genome version 1. 4. We compare our model with two previous state of the art method: sg2im[2] and layout2im[5] .

4.1 Quantitative Results

Table 1 summarizes comparison results of the Inception [29, 30], Frechet Inception Distance (FID) [24] and Diversity [31] Score. Inception Score is adopted to measure the quality and diversity of generated images. FID uses 2nd order information of the final layer of the inception model and calculates the similarity of generated images to real ones. Diversity Score computes the perceptual similarity between two images in deep feature space. The higher score of Inception and Diversity Score is better, FID Score is totally reverse. For 64×64 images, our proposed model outperforms other models in terms of FID score. Although, the TS-GAN performed a little bit worse in IS, in Diversity score our model notably showed the highest score. It demonstrate our model's capability to generate complex and diverse images with multiple objects.

4.2 Qualitative Results

We compare our model with baselines using the same input, the ground-truth layout. As we can see in Figure 3, it shows examples of generated images from our mode trained on Visual Genome datasets, as well as baselines. From these examples, it is clear that our model can generate complex images with multiple objects. Figure 3(a) shows two elephants, (b) contains a person, (c) contains food and (d) contains two buses. These examples also show that our method generates images which respect the relationships of the input layout. As we can see in Figure 3(d), layout2im fail to generate a meaningful image, due to the extreme difficulty of directly mapping layout to a real image without detailed instance segmentation. Given that the same layout may have many different possible real image, the ability to generate diverse and realistic images is a key advantage of our model.

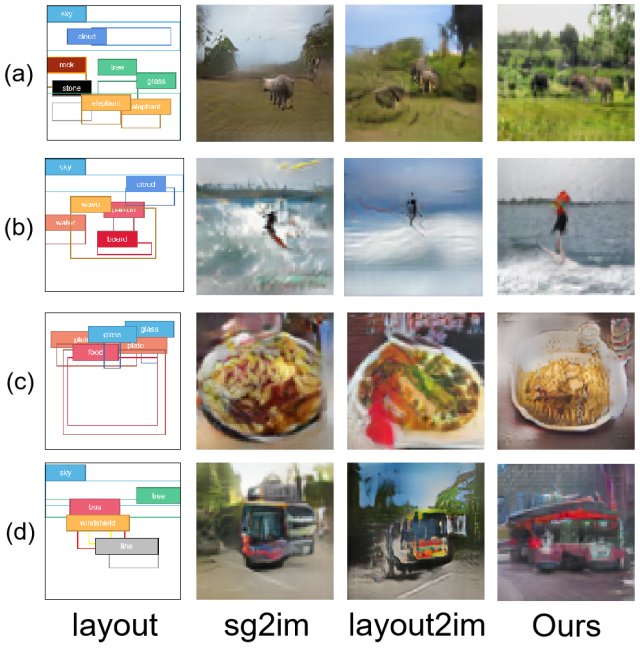

Figure 3: Examples of 64 × 64 Generated Images from Complex Layouts on Visual Genome Datasets by Our Proposed Method and Baselines.

Table 2: Ablation Study of Our Method

Method	Inception score	Diversity score
w/o φ_t	6. 9 ± 0. 2	0. 23 ± 0. 02
w/o SPADE	7. 1 ± 0. 2	**0. 39 ± 0. 01**
w/o SGGFT	6. 7 ± 0. 1	0. 31 ± 0. 05
full model	**7. 4 ± 0. 1**	0. 37 ± 0. 01

4.3 Ablation Study

Table 2 demonstrates the necessity of our key components by comparing scores of several ablated models trained on Visual Genome datasets. Although removing the SPADE that Diversity Score performs higher, reduce the generated image quality. Not surprisingly, the concatenation of text embedding and object embedding and SGGFT are detrimental to the model's performance. Without SGGFT module, it decreases the overall performance of our model. It is clear that out full model achieves a good balance across all these metrics.

5 CONCLUSION

In this paper, we develop an end-to-end method for generating diverse images from text description. Compared to the previous works of generating images from unstructured text, our model allowsus to generate realistic images and recognizable objects in the reasonable location. The qualitative results show that our model improves the generation quality compare to the baseline models. From the ablation study, the concatenation of text embedding and each object embedding help the object interact with the whole image semantic. The drawback of our model's ability is that can't generate enough high resolution images. It will be our future work.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Plan of China (No. 2017YFD0400101).

REFERENCES

[1] Reed S, Akata Z, Yan X, *et al.* (2016). Generative adversarial text to image synthesis[C]//International Conference on Machine Learning. PMLR, 1060-1069.
[2] Johnson J, Gupta A, Fei-Fei L (2018). Image generation from scene graphs[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 1219-1228..
[3] Li Y, Ma T, Bai Y, *et al.* (2019). Pastegan: A semi-parametric method to generate image from scene graph[J]. Advances in Neural Information Processing Systems, 32: 3948-3958..
[4] Ashual O, Wolf L (2019). Specifying object attributes and relations in interactive scene generation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 4561-4569.
[5] Zhao B, Meng L, Yin W, *et al.* (2019). Image generation from layout[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8584-8593.
[6] Sun W, Wu T (2019). Image synthesis from reconfigurable layout and style[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 10531-10540.
[7] Sylvain T, Zhang P, Bengio Y, *et al.* (2020). Object-centric image generation from layouts[J]. arXiv preprint arXiv:2003.07449, 1(2): 4..
[8] Tan F, Feng S, Ordonez V (2019). Text2scene: Generating compositional scenes from textual descriptions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6710-6719
[9] Mikolov T, Sutskever I, Chen K, *et al.* (2013). Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems, 3111-3119.
[10] Lee K H, Palangi H, Chen X, *et al.* (2019). Learning visual relation priors for image-text matching and image captioning with neural scene graph generators[J]. arXiv preprint arXiv:1909. 09953.
[11] Li Y, Ouyang W, Zhou B, *et al.* (2017). Scene graph generation from objects, phrases and region captions[C]//Proceedings of the IEEE international conference on computer vision, 1261-1270.
[12] Cha M, Gwon Y L, Kung H T (2019). Adversarial learning of semantic relevance in text to image synthesis[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 33(01): 3272-3279.
[13] Schuster S, Krishna R, Chang A, *et al.* (2015). Generating semantically precise scene graphs from textual descriptions for improved image retrieval[C]//Proceedings of the fourth workshop on vision and language, 70-80.
[14] Goodfellow I, Pouget-Abadie J, Mirza M, *et al.* (2014). Generative adversarial nets[J]. Advances in neural information processing systems, 27.
[15] Mirza M, Osindero S (2014). Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411. 1784.
[16] Reed S E, Akata Z, Mohan S, *et al.* (2016). Learning what and where to draw[J]. Advances in neural information processing systems, 29: 217-225..
[17] Zhang H, Xu T, Li H, *et al.* (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of the IEEE international conference on computer vision, 5907-5915.
[18] Zhang H, Xu T, Li H, *et al.* (2018). Stackgan++: Realistic image synthesis with stacked generative adversarial networks[J]. IEEE transactions on pattern analysis and machine intelligence, 41(8): 1947-1962.
[19] Xu T, Zhang P, Huang Q, *et al.* (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 1316-1324.
[20] Zhang Z, Xie Y, Yang L (2018). Photographic text-to-image synthesis with a hierarchically-nested adversarial network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6199-6208.
[21] Odena A, Olah C, Shlens J (2017). Conditional image synthesis with auxiliary classifier gans[C]//International conference on machine learning. PMLR, 2642-2651.
[22] Anderson P, Fernando B, Johnson M, *et al.* (2016). Spice: Semantic propositional image caption evaluation[C]//European conference on computer vision. Springer, Cham, 382-398.
[23] Park T, Liu M Y, Wang T C, *et al.* (2019). Semantic image synthesis with spatially-adaptive normalization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2337-2346.
[24] Heusel M, Ramsauer H, Unterthiner T, *et al.* (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 30.
[25] Isola P, Zhu J Y, Zhou T, *et al.* (2017). Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 1125-1134.
[26] Johnson J, Alahi A, Fei-Fei L (2016). Perceptual losses for real-time style transfer and super-resolution[C]//European conference on computer vision. Springer, Cham, 694-711.
[27] Lin T Y, Maire M, Belongie S, *et al.* (2014). Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 740-755.
[28] Krishna R, Zhu Y, Groth O, *et al.* (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International journal of computer vision, 123(1): 32-73.
[29] Salimans T, Goodfellow I, Zaremba W, *et al.* (2016). Improved techniques for training gans[J]. Advances in neural information processing systems, 29: 2234-2242
[30] Simonyan K, Zisserman A. (2014). Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556.
[31] Zhang R, Isola P, Efros A A, *et al.* (2018). The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 586-595.