Enriching Attributes from Knowledge Graph for Fine-grained Text-to-Image Synthesis

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

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

Jieyu Huang
Shanghai Film Academy, Shanghai University, Shanghai, China
jieyu88@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

In this paper, we propose an Attribute-Rich Generative Adversarial Network (AttRiGAN) for text-to-image synthesis, which enriches the simple text description by associating knowledge graph and embedding it in the synthesis task in the form of an attribute matrix. Higher fine-grained images can be synthesized with AttRiGAN, and the synthesized sample are more similar to the objects that exist in the real world, since they are driven by attributes which are enriched from the knowledge graph. The experiments conducted on two widely-used fine-grained image datasets show that our AttRiGAN allows a significant improvement in fine-grained text-to-image synthesis.

CCS CONCEPTS

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

KEYWORDS

Text-to-image synthesis, Attribute embedding, Knowledge graph, Attention mechanism

ACM Reference Format:
Yonghua Zhu, Ning Ge, Jieyu Huang, Yunwen Zhu, Binghui Zheng, and Wenjun Zhang*. 2021. Enriching Attributes from Knowledge Graph for Fine-grained Text-to-Image Synthesis. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487075.3487155

CSAE 2021, October 19–21, 2021, Sanya, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487155

1 INTRODUCTION

As one of the research tasks that combine the two modalities of visual contents and natural languages, in recent years, text-to-image synthesis has focused on using Generative Adversarial Net-works (GANs) [1] to synthesize simple text descriptions into complex images. However, the text descriptions are limited in what they can provide, not only because of the redundancy of words when describing the synthesis requirements in sentence form, but also because the principle of text-to-image synthesis task is to guide the synthesis of complex images with simple text descriptions. The limitation of natural language information will directly result in the non-authenticity of synthesized images. Even if mature generative adversarial networks can synthesize high-resolution images which look "real", the synthesized contents don't objectively exist in the real world.

In visual question answering [2, 3] and image recognition [4], knowledge graph is often used for effective information integration, it brings a new idea for our research. The proposed method first builds a knowledge graph containing objects and attributes based on the dataset. Specifically, in the Caltech-UCSD Birds 200 dataset (CUB-200) [5], the relationship between bird species and various attributes are established to represent fine-grained details, then we attempt to use bird attributes to enhance the fine-grained image synthesis task. In the proposed AttRiGAN, the input text is represented as a graph structure, which also includes the bird species and attributes, to facilitate its match with the knowledge graph. Then the knowledge graph is used to enrich the attribute information for the graph structure to realize the finer-grained image synthesis. Because the knowledge graph is based on the real birds and their attributes, the enrichment of the text is knowledge supplement which is premised on authenticity and can better avoid the plight of synthetic birds that don't exist in the real world.

In the process of image synthesis, we learn from AttnGAN and embed the supplementary attributes into the AttRiGAN through the attention mechanism. The main contributions of the method we proposed are as follows:

1. We introduce the knowledge graph to enrich the simple text of the input and transform the text description into a graph structure to better match the two, the text description

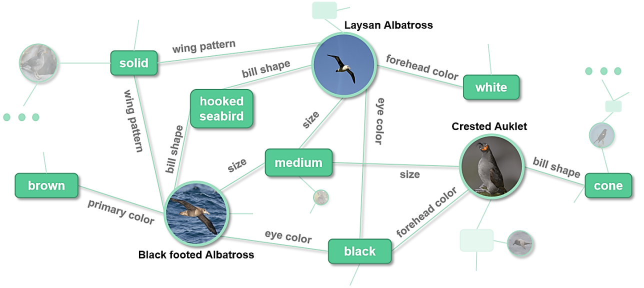

Figure 1: An Example Knowledge Graph for the CUB-200.

enriched by the knowledge graph is represented by objects and attributes.

2. We develop a text-to-image synthesis network AttRiGAN based on attribute attention, which can filter the important attributes to guide the image synthesis, so as to synthesize the real and high fine-grained images.
3. The experimental results demonstrate that our AttRiGAN outperforms the state-of-the-art approaches.

2 RELATED WORK

GANs show good performance in both the reality and resolution of synthesized images [6–9].

Knowledge is integrated into the image synthesis task in various contents and forms. Lifelong GAN [10] and KT-GAN [11] with knowledge distillation employed homologous or heterogeneous information distillation to learn the knowledge to synthesize realistic images. RiFeGAN [12] serves as an important example of our state-of-the-art approaches. It retrieved and refined compatible candidate captions from an image caption knowledge base based on the text description, using self-attentional embed-ding mixture algorithm to extract the features of multiple captions for image synthesis.

In the research of fine-grained image synthesis, AttnGAN [13] is a GAN model based attention mechanism. It embedded the text description into the image synthesis stage in the forms of sentence and word level features, then synthesized fine-grained details at different sub-regions of the image. However, the weight of words in AttnGAN was fixed and did not highlight the role of important words. SEGAN [14] attempted to employ the attention regularization term to highlight the important words, but the calculation of adaptive threshold had become a new difficulty.

Compared with the simple word vector in the sentence, in our work, the attribute information supplemented by the knowledge graph is the important information that has been screened, and directly embedded through the attention mechanism. Thus, we no longer need to consider the problem of representing important words through word weights.

3 METHODOLOGY

In this section, we will introduce our proposed Attribute-Rich Generative Adversarial Network, *i.e.* AttRiGAN.

3.1 Enriching Attributes from Knowledge Graph

Because the simple text description contains limited information and cannot provide rich constrained conditions for image synthesis, we introduce knowledge graph to enrich the text description.

We first build a fine-grained knowledge graph of bird attributes based on the attribute annotation of the dataset. The nodes represent the category and attribute values of birds, and the edges represent the attributes associated with them. The visual example for the CUB-200 is shown in Figure 1

Attributes are the intermediate semantic representations of objects, it is key to distinguish the subordination between them. For a given dataset, we build a knowledge graph KB, it contains nodes for object classes C^0, attribute values A^0, and edges R^0 .The correlation $r_j^0 \in R^0$ between an object class $c_i^0 \in C^0$ and an attribute value $a_j^0 \in A^0$ represents a type of attribute. Naturally, we can store the knowledge graph as RDF triples, *i.e.* $KB = (C^0, R^0, A^0)$.

As shown in Figure 2, for the input text description, we should establish its match with the knowledge graph. According to the literature [15] that text description is modeled by a graph-based semantic representation, we map the text into a graph structure $TG^1 = (C, R^1, A^1)$ through the dependency parse trees. Nodes $c \in C$ and $a_j^1 \in A^1$ represent an object class and an attribute value extracted from the text description, respectively, the edge $r_j^1 \in R^1$ indicates the attributes associated between them. It should be noted, this task is a fine-grained image synthesis of a single object, so the object that can be included in the text description is unique, *i.e.* there is only one element c in the aggregate C.

Enriching the text description with knowledge graph is to complete the attribute information of the text graph structure, specifically. Compared with the text graph structure, the knowledge graph has a larger volume, in order to facilitate the matching calculation, we take the object class c_i^0 of the knowledge graph as a unit to calculate the similarity of the attributes, which is between c_i^0 and the object class c from the text graph structure. The attribute similarity score $s_i \in S$ is calculated as

$$s_i = \sum\nolimits_j sim_{Jaccard}(a_j^0, a_j^1) \tag{1}$$

$$sim_{\text{Jaccard}}\left(a_j^0, a_j^1\right) = \frac{\left|a_j^0 \cap a_j^1\right|}{\left|a_j^0 \cup a_j^1\right|} \tag{2}$$

where $sim_{\text{Jaccard}}\,(a_j^0, a_j^1)$ represents the similarity score of the attribute values between c_i^0 and c under the same attribute(*i.e.* $r_j^0 \in r_j^1$), *Jaccard* coefficient is used to measure the similarity of the two aggregates, here we think of the string as an aggregate.

The highest score is s_{max} when $i = k$, at this point, we select c_k^0 to enrich the text graph structure, concretely, it completes the attributes $r \in R$ and the attribute values $a \in A$ for $c \in C$, and then we can gain an enriched graph $TG = (C, R, A)$.

3.2 Fine-grained Text-to-Image Synthesis

AttnGAN and related methods derived from it [14, 16] are the most typical methods in fine-grained text-to-image synthesis. But AttnGAN has an obvious problem with word redundancy and the

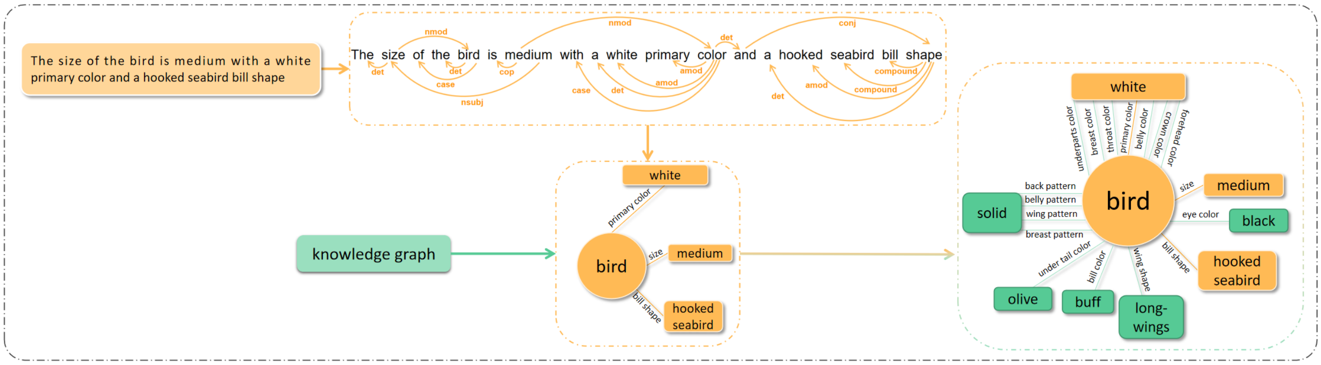

Figure 2: Text Graph Structure: The Text Description Gains Rich Attributes from the Knowledge Graph.

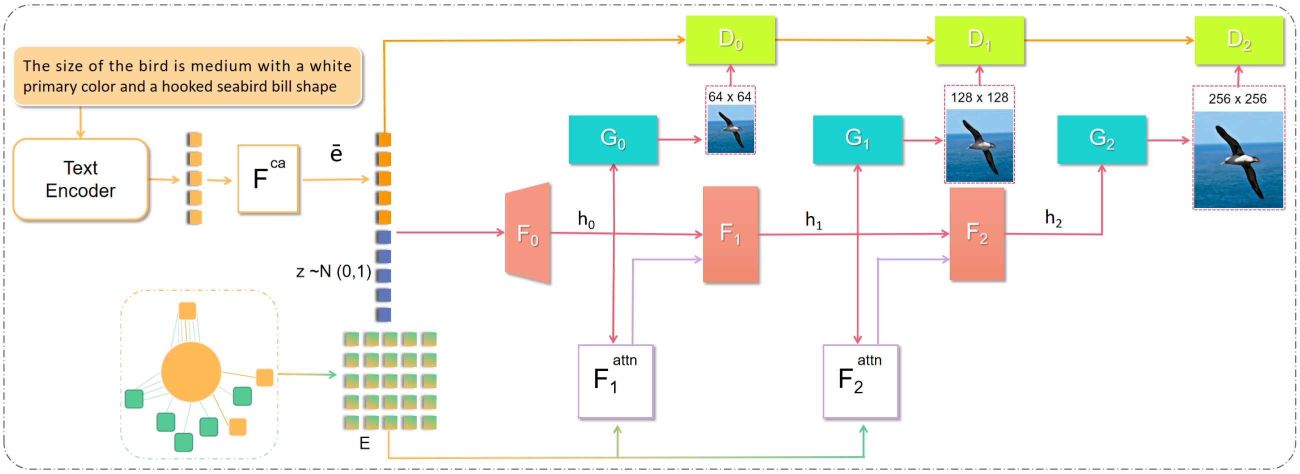

Figure 3: AttRiGAN: Attribute-Rich Generative Adversarial Network.

fixed weights of the words. In order to solve these problems, we use the graph structure TG to replace the word features in AttnGAN for attention embedding.

We transform the attributes and attribute values of the graph structure TG into an embedding matrix $E \in \mathbb{R}^{M \times N}$. The row of the embedding matrix is M dimension, which denotes attribute categories, such as color, size and so on. Analogously, the column is N dimension, which denotes attribute objects, such as wing, bill and so on. Whereupon the corresponding positions of elements in the embedding matrix represent attribute values.

Figure 3 shows the framework of our AttRiGAN. The text is first encoded as an embedding vector $\bar{e}$ by a bi-directional Long Short-Term Memory Network (LSTM) text encoder [17], $\bar{e}$ does Conditioning Augmentation [7] via F^{ca}. z is a noise vector sampled from the standard normal distribution. $\bar{e}$ is concatenated with z and they are transformed into the initial image feature $h_0 \in \mathbb{R}^{\hat{M} \times P}$ by a network F_0 composed of a series of upsampling blocks. Then a multi-scale method is used to gradually synthesize the fine-grained image $\hat{x}_i$ through the image feature $h_i \in \mathbb{R}^{\hat{M} \times P}$, specifically,

$$h_i = F_i\left(h_{i-1}, F_i^{attn}\left(h_i, E\right)\right) \in \mathbb{R}^{\hat{M} \times P} \text{ for } i = 0, 1, 2 \tag{3}$$

$$\hat{x}_i = G_i\left(h_i\right) \tag{4}$$

F_1 and F_2 are upsampling modules which consist of several residual networks and an upsampling module. G_i converts h_i into an image by using a 3×3 convolutional layer and a $\tan h$ activation function. D_i is a discriminator constructed by several convolutional layers, batch normalization layers, and leaky rectified linear units.

F_i^{attn} is the attention module, which takes the attribute embedding matrix E and the hidden state $h_i \in \mathbb{R}^{\hat{M} \times P}$ as input, and calculates as follows,

$$F_i^{attn}\left(h_i, E\right) = \left(e_0, e_1, \ldots e_{P-1}\right) \in \mathbb{R}^{\hat{M} \times P} \tag{5}$$

$$e_j = \sum\nolimits_{i=0}^{N-1} \beta_{j,i} E'_i \tag{6}$$

$$E' = UE \tag{7}$$

where E' is the attribute matrix corresponding to E in the image feature space through a perception layer, $E_i^{'}$ is the i^{th} column vector of E', $U \in \mathbb{R}^{\hat{M} \times M}$, $\beta_{j,i}$ is the attention weight, which is defined as

$$\beta_{j,i} = \frac{exp(S_{j,i})}{\sum_{k=0}^{N-1} exp(S_{j,k})} \tag{8}$$

$$s_{j,i} = h_j^T E_i^{'} \tag{9}$$

The final objective function of the AttRiGAN is defined as

$$L = L_G + \lambda L_{DAMSM}, \text{ where } L_G = \sum_{i=0}^{2} L_{G_i} \tag{10}$$

where λ is a hyper parameter to balance the two terms of Eq. 10). The first term is the GAN loss, which includes the unconditional loss that determines whether the generated image is true, and the conditional loss that determines whether the image matches the text description. In the i^{th} stage of the AttRiGAN, the generator G_i corresponds to the discriminator D_i. The adversarial loss of G_i is defined as

$$L_{G_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}} \left[\log\left(D_i\left(\hat{x}_i\right)\right]\right.}_{unconditional\, loss} \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}} \left[\log\left(D_i\left(\hat{x}_i, \bar{e}\right)\right]\right.}_{conditional\, loss} \tag{11}$$

where $\hat{x}_i$ is from the model distribution p_{G_i} at the i^{th} scale. Meanwhile, the discriminator D_i is trained, and its loss is defined as

$$L_{D_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{x_i \sim p_{data_i}} \left[\log D_i\left(x_i\right)\right] - \frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}} \left[\log\left(1 - D_i\left(\hat{x}_i\right)\right]\right.}_{unconditional\, loss}$$
$$\underbrace{-\frac{1}{2}\mathbb{E}_{x_i \sim p_{data_i}} \left[\log D_i\left(x_i, \bar{e}\right)\right] - \frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}} \left[\log\left(1 - D_i\left(\hat{x}_i, \bar{e}\right)\right]\right.}_{conditional\, loss} \tag{12}$$

where $\hat{x}_i$ is from the true image distribution p_{data_i} at the i^{th}scale. L_{DAMSM} is an attribute fine-grained image-text matching loss computed by the DAMSM and defined as

$$L_{DAMSM} = L_1^E + L_2^E + L_1^{\bar{e}} + L_2^{\bar{e}} \tag{13}$$

where L_1^E, L_2^E, $L_1^{\bar{e}}$ and $L_2^{\bar{e}}$are the loss functions of the supplement attributes and the text description, which are described by the matching probabilities compared to the image feature between the attribute matrix and the text embedding vector. The image feature is extracted locally and globally by an image encoder built on the Inception-v3 model [18], then it performs through a 1 × 1 convolutional layer and a multi-layer perceptron, respectively.

4 EXPERIMENTS

To validate our AttRiGAN, we conduct extensive quantitative and qualitative evaluations. We analyze different components of AttRiGAN on CUB-200 dataset with our baseline models. And then we compare our AttRiGAN with state-of-the-art GAN models for text-to-image synthesis.

4.1 Datasets and Evaluation Metrics

By referring to several datasets currently used in text-to-image synthesis and combining with the characteristics of our task, we select two widely used fine-grained image datasets, CUB-200 and Oxford-Flower-102 (Oxford-102) [19]. The CUB-200 contains 11,788 images of 200 species of birds, 27 kinds of attributes (attribute object - attribute category), and a total of 312 species of attribute relations (attribute - attribute value), we use the same method as [7] to preproccess CUB-200. Oxford-102 contains 8189 images of 102 kinds of flowers. Each image in both datasets has a description of 10 captions provided by [18] to describe fine-grained visual details. We divide CUB-200 and Oxford-102 into class-disjoint training set and test set.

Table 1: The Inception Score and the R-precision Rate of Each AttRiGAN Model on CUB-200 Test Set

Method	Inception score	R-precision(%)
AttRiGAN(full model)	5.25±0.05	**79.02±3.55**
Ours(No F^{attn})	3.98±0.04	10.37±5.88
Ours(No F_1^{attn})	4.34±0.05	65.89±4.57
Ours(Add F_3^{attn})	**5.27±0.01**	77.23±0.61

To assess the quality and variety of the images synthesized by AttRiGAN as accurately as possible, Inception score [20] is utilized, which can show the quality and the diversity of the image, a higher score means a better synthesized result. We fine-tune the inception model for different datasets. In addition, we use R-precision to evaluate how well the synthesized images conform to the text description. Specifically, we employ the synthesized image to query its corresponding text description. For one ground-true caption description and 99 randomly mismatched captions, we rank these candidate captions in descending order of the cosine similarities, there are r captions associated with the query image in the first R, so we set R-precision as r/R.

4.2 Ablation Study

According to the proposed AttRiGAN, we design several ablated versions of it, as shown in Table 1, the comparison of image quality synthesized by different models proves the necessity of each component of our model. In view of the characteristics of AttRiGAN, we conduct ablation experiments on the attribute matrix.

For Ours(No F^{attn}) which is without attribute embedding matrix, its conditioning variables is only the text embedding vector, it is a far cry from AttRiGAN both in Inception score and R-precision. Ours(No F_1^{attn}) has only one attention embedding, which can supplement effective information for image synthesis, however, it seems that the less embedding times are unable to make the network receive the supplementary attribute information completely.

In addition to the ablation study, in order to explore the appropriate attention embedding mechanism, we conduct more times of attention embedding. Ours(Add F_3^{attn}) has attribute attention embedding thrice, it appears to be able to synthesize finer-grained images, with a 0.02 higher than AttRiGAN on Inception score. But actually, from the comparison of R-precision, we can see that too much emphasis on attribute supplement makes the match between the synthesized image and text description not perfect. Above all, adding more layers of attention will consume more experimental resources and significantly reduce the computational efficiency of the model.

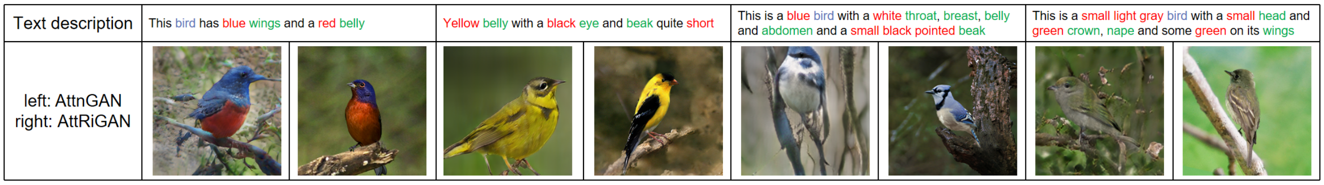
Text description
This bird has blue wings and a red belly
Yellow belly with a black eye and beak quite short
This is a blue bird with a white throat, breast, belly and abdomen and a small black pointed beak
This is a small light gray bird with a small head and green crown, nape and some green on its wings
left: AttnGAN
right: AttRiGAN

Figure 4: Several Results of AttGAN and AttRiGAN Performed on the CUB-200. The Text Descriptions Are Colored with Blue for the Object, Green for the Attributes, and Red for the Attribute Values.

Table 2: Inception Scores on CUB-200 and Oxford-102 Datasets

Dataset	StackGAN	AttnGAN	DM-GAN	RiFeGAN	Our AttRiGAN
CUB-200	3.70±0.04	4.36±0.03	4.75±0.07	5.23±0.09	**5.30±0.05**
Oxford-102	3.20±0.01	3.91±0.05	4.03±0.05	**4.53±0.05**	4.35±0.03

Table 3: R-precision Rates on CUB-200 Datasets

Method	R-precision(%)
AttnGAN	67.82±4.43
DM-GAN	72.31±0.91
Our AttRiGAN	**79.02±3.55**

4.3 Comparison with Previous Methods

We compare our AttRiGAN with state-of-the-art models on two datasets, CUB-200 and Oxford-102, and the comparison results are shown in Table 2 and Table 3

Figure 4 shows some results of AttnGAN and our AttRiGAN performed on the CUB-200, it is obvious that our model shows better results from these images. And in Table 2, the Inception score of our AttRiGAN is 5.30 which is far superior to a series of methods such as AttnGAN. Even for RiFeGAN based on external knowledge, which has an Inception score of 5.23, AttRiGAN still outperforms better than it. We suspect that the reason may be that RiFeGAN used the given text description to supplement the captions, and the supplementary captions are close to the text description, so cannot provide the network with more content containing other attributes. In other words, the extended attribute information is not as comprehensive as AttRiGAN's to synthesize finer-grained images. For the Oxford-102 dataset, our AttRiGAN is also more capable of synthesizing high-quality images than most methods.

As shown in Table 3, compared with DM-GAN, AttRiGAN improves the R-precision from 72.31 to 79.02 on the CUB-200 dataset (6.71% improvement). The significant improvement in the R-precision indicates that the image synthesized by our AttRiGAN can better adapt to the given text description, and further verifies the effectiveness of attribute supplement using attention embedding. To sum up, our AttRiGAN model performs favorably against the state-of-the-art approaches.

5 CONCLUSION

In this paper, in order to synthesize high fine-grained images with limited simple text descriptions, we propose a text-to-image synthesis model called AttRiGAN. The AttRiGAN uses know-ledge graph to supplement attributes and combines rich attribute information by embedding into the network through the attention mechanism. Experiments on widely used datasets have shown that our AttRiGAN is effective not only in improving its Inception score and R-precision, but also in synthesizing fine-grained images that more closely resemble those that exist in the real world.

ACKNOWLEDGMENTS

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

REFERENCES

[1] I Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, and Y Bengio (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

[2] P Wang, Q Wu, C Shen, A Hengel, and A Dick (2015). Explicit knowledge-based reasoning for visual question answering. arXiv preprint arXiv:1511.02570.

[3] P Wang, Q Wu, C Shen, A Dick, and A Hengel (2017). Fvqa: Fact-based visual question answering. IEEE transactions on pattern analysis and machine intelligence, 40(10), 2413-2427.

[4] T Chen, L Lin, R Chen, Y Wu, and X Luo (2018). Knowledge-embedded representation learning for fine-grained image recognition. arXiv preprint arXiv:1807.00505.

[5] C Wah, S Branson, P Welinder, P Perona, and S Belongie (2011). The caltech-ucsd birds-200-2011 dataset.

[6] M Mirza, S Osindero (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

[7] H Zhang, T Xu, H Li, S Zhang, X Wang, X Huang, and D Metaxas, (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 5907-5915).

[8] Z Zhang, Y Xie, and L Yang (2018). Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6199-6208).

[9] H Zhang, T Xu, H Li, S Zhang, X Wang, X Huang, and D Metaxas, (2018). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE transactions on pattern analysis and machine intelligence, 41(8), 1947-1962.

[10] M Zhai, L Chen, F Tung, J He, M Nawhal, and G Mori (2019). Lifelong gan: Continual learning for conditional image generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2759-2768).

[11] H Tan, X Liu, M Liu, B Yin, and X Li (2020). KT-GAN: knowledge-transfer generative adversarial network for text-to-image synthesis. IEEE Transactions on Image Processing, 30, 1275-1290.

[12] J Cheng, F Wu, Y Tian, L Wang, and D Tao (2020). RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10911-10920).

[13] T Xu, P Zhang, Q Huang, H Zhang, Z Gan, X Huang, and X He (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1316-1324).

[14] H Tan, X Liu, X Li, Y Zhang, and B Yin (2019). Semantics-enhanced adversarial nets for text-to-image synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10501-10510).

[15] P Anderson, B Fernando, M Johnson, and S Gould (2016, October). Spice: Semantic propositional image caption evaluation. In European conference on computer vision (pp. 382-398). Springer, Cham.

[16] M Zhu, P Pan, W Chen, and Y Yang (2019). Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5802-5810).

[17] S Reed, Z Akata, H Lee, and B Schiele (2016). Learning deep representations of fine-grained visual descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 49-58).

[18] L Karacan, Z Akata, A Erdem, and E Erdem (2019). Manipulating attributes of natural scenes via hallucination. ACM Transactions on Graphics (TOG), 39(1), 1-17.

[19] M Nilsback, and A Zisserman (2008, December). Automated flower classification over a large number of classes. In 2008 Sixth Indian Conference on Computer Vision, Graphics and Image Processing (pp. 722-729). IEEE.

[20] T Salimans, I Goodfellow, W Zaremba, V Cheung, A Radford, and X Chen (2016). Improved techniques for training gans. Advances in neural information processing systems, 29, 2234-2242.