Self-modulated feature fusion GAN for Text-to-image Synthesis

Wenhong Cai
College of Informatics
Huazhong Agricultural University
Hubei, China
wenhongc@126.com

Jinhai Xiang
College of Informatics
Huazhong Agricultural University
Hubei, China
jimmy_xiang@mail.hzau.edu.cn

Bin Hu
College of Informatics
Huazhong Agricultural University
Hubei, China
hubinwh@mail.hzau.edu.cn

2023 International Conference on Engineering and Emerging Technologies (ICEET) | 979-8-3503-1692-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEET60227.2023.10526149

***Abstract*—Text-to-image synthesis is one of the key tasks in the cross-modal generation, which aims to generate natural and realistic images under the condition of text description. The main challenge of this task is how to efficiently integrate text information into the process of image synthesis while satisfying a high degree of semantic consistency. Existing methods based on generative adversarial networks(GANs) use stacked network structures (2-3 groups of generators and discriminators) to generate high-resolution images in stages and add text information at different stages of generation. Most aforementioned methods may give rise to some issues such as generator interdependence and an unwarranted proliferation of network parameters. To address these limitations, we propose a generative adversarial network architecture with a single generator and discriminator. The Adaptive Semantic Image feature Fusion module in the generator can effectively compensate for the lack of fine-grained information caused by a single generator and discriminator. In addition, to stabilize the network training and enhance the semantic consistency between the text and the synthesized image, local spectral normalization is used in the discriminator, and contrastive loss is added to enhance the discriminator's ability to supervise the synthesis of the generator. Extensive experiments on CUB and COCO datasets demonstrate that the proposed model is superior to the existing models.**

***Index Terms*—Text-to-image synthesis, GANs, Cross-modal, Spectral normalized, Contrastive Learning, deep learning**

I. Introduction

Cross-modal research has advanced significantly in the past few years. Some impressive progress has emerged in text-to-image generation tasks. It aims to generate real images related to any given text description and shows a great potential application in image editing, video games, and computer-aided design. Generative adversarial networks [1] and variational autoencoders(VAE) [2] play an important role in the above fields. Recently, a number of GAN-based approaches have had advanced results [3] [4] [5] in text-to-image tasks.

BIGGAN [6] generates high-quality images by scaling up both the model size and batch size. SSAGAN [4] learns semantic adaptive transformation conditional on text to effectively fuse text features and image features. Although these network structures are very popular, the complex network structure and the abundance of parameters may engender the inadvertent loss of crucial feature information, and each part of the structure is isolated, which poses challenges in optimization.

The contributions of this paper are the following:

- We propose an adaptive semantic image feature fusion module that incorporates textual features into the process of image synthesis.
- To stabilize the network training process, we add local spectral normalization to the discriminator. The contrastive loss is proposed to make better alignment of semantics and images.
- The proposed model improves the visual quality and evaluation metrics on the CUB and COCO datasets.

II. Related Work

The emergence of Generative Adversarial Networks (GANs) has notably elevated the efficacy of image generation domain, including general image generation, image editing [7], and Text-image synthesis, etc.. Text-image synthesis represents a pivotal task within the domain of cross-modal image generation. Reed et al. [8] introduced the integration of textual descriptions as a supervised condition into the image generation process, and proposed GAN-INT-CLS, which compelled both the generator and discriminator not only to focus on authentic image characteristics but also to align the generated image with the provided textual input. The StackGAN [9] architecture, comprising two sequential stages, addresses the synthesis task by initially producing a rudimentary image consistent with the accompanying text and subsequently refining this image to achieve higher resolution. Subsequently, most of the methods are based on the structure of the generator and multiple discriminators to unfold. This structure will cause the entanglement between different scale generators, and degrade the performance of the network.

DAMSM (Deep Attentional Multimodal Similarity Model) is proposed in AttnGAN [10], which contains two encoder networks. Bi-LSTM network is used for text encoder to extract semantic features in the text description. DFGAN [11] proposed a fusion block that deeply fused text features and image features, and adopted a single generator discriminator structure, which reduced the network complexity and achieved good results. RAT-GAN [12] invented a recursive affine transformation (RAT) for generative adversarial networks, which connects all image fusion blocks with a recurrent neural network to model the long-term dependence between them. Since the advent of DALLE 2 [13], the diffusion model has

gradually replaced the status of GAN and has become the mainstream research method. Recently, the GigaGAN [14] has a parameter amount of 1 billion, and its image generation effect in the T2I task is close to that of the diffusion model, and its inference speed is faster.

Most of the subsequent methods are based on the structure of multiple generators and multiple discriminators. Different from previous research, our model adopts the one-stage structure to reduce the number of parameters and simplify the training process. The sentence vector and word vector are encoded by BERT and added to an adaptive image-text fusion module. At the same time, the contrastive loss is applied to improve the consistency of text and image, and high-quality images are generated.

III. Method

A. A generator for self-modulated feature fusion

To address the instability of GAN model training, previous models adopt stacked structure (generally three layers) as the backbone, DAMSM (AttnGAN [10]), circular structure (MirrorGAN [15]), twin structure (SD-GAN [16]) and other structures to maintain the semantic consistency of text and image. But the stacked structure causes entanglement between the different generators, resulting in the final optimized images look like a simple combination of fuzzy shapes and a few image subjects. we posit the deployment of a single-stage generator to generate images directly. This approach boasts a more straightforward architecture and a reduced parameter count in comparison to the multi-stage generator.

Through the text to generate images, is a mapping between different modal data, so how to give the generator efficient text encoding is one of the key points of research. Most of the existing works use bidirectional recurrent memory networks to extract text features. We consider a more advanced BERT [17] model to encode each text description as a sentence vector e^{sent} and a word vector e^{word}.

We propose an Adaptive Semantic Image feature Fusion module(ASIF) to enhance the consistency between fusion modules of different layers (see Fig. 1). In the previous work, there are many options to utilize ResBlock by adding Conditional Batch Normalization(CBN) [18]. For example, the "SPADE" module [19] integrates additional feature information into the network via the normalization method. Since BN converts feature maps into a normal distribution. It can be regarded as the unconditional reverse operation of affine transformation, which reduces the distance between each feature map in a batch. The principle of conditional generation is to distance the feature map for generating different samples, so it is unfavorable to conditional generation.

ASIF extracts the affine transformation method from CBN and uses an affine transformation to operate visual feature mapping based on the semantic description.

$$\gamma = MLP_1(e^{sent}), \beta = MLP_2(e^{word}) \tag{1}$$

where e^{sent} is the sentence vector, and γ, β are parameters predicted by two one-hidden-layer MLPs conditioned on e^{sent}. For the feature vector $h^{B\times C\times H\times W}$ input in the previous layer, the feature channel scale is changed first, and then the movement parameter is used for the offset operation.

$$AS(h_i, e) = \gamma_i \cdot h_i + \beta \tag{2}$$

As shown in Fig. 1, to increase the network depth and nonlinearity of the model, two ASIF modules are used in a ResBlock, which are stacked together with the convolution layer and the upsampling layer to form the UpBlock block. Finally, as shown in Fig. 2(a), we use a single-level generator consisting of six upper sample blocks to synthesize the fake image. The noise vectors $z \sim N(0, 1)$ are sampled from the standard Gaussian distribution and fed to the generator at the beginning. Each upsampling block is followed by an ASIF block to control the content of the image. The final 256×256 feature map is transformed by a hyperbolic tangent function into a false image of size 256×256×3.

B. Discriminator based on local spectral normalization

In the process of GAN training, there is always a problem that the better the discriminator training, the more serious the generator gradient disappears. The excellent Wasserstein distance in WGAN [20] replaces the J-S divergence in native GAN. Then the problem of Wasserstein distance is transformed into the problem of optimal Lipschitz continuous function based on the KR duality principle. To make the discriminator satisfy the Lipschitz continuity, the authors of WGAN directly clipped the oversized parameters below a threshold using "gradient clipping". The Lipschitz continuity constraint is introduced to make the neural network more insensitive to input disturbances, thus making the training process more stable and easier to converge. Spectral normalization for generative adversarial network [21] (herein after referred to as Spectral Norm) allowed the discriminator to satisfy the Lipschitz continuity in a more elegant way, limiting the severity of function changes, which makes the model more stable.

Therefore, we perform local spectral normalization on the parameter matrix of each layer convolution of discriminator D, as shown in Figure 2(b). The structure of the discriminator is dual to the generator and consists of six DownBlock layers. After the fifth DownBlock layer, the output feature h is coupled with the sentence feature vector e^{sent} for image truth and false discrimination and loss calculation.

C. Objective Function

To simplify the network structure and improve the efficiency of network training, we choose the discriminator as the image encoder, and no additional image encoder is set. Two types of contrastive loss functions are used to update the parameters of the discriminator and generator. Similar to [22] [23], firstly, we define two sample pairs with corresponding relations :(image, text description), (generated image, real image), and use InfoNCE loss to maximize the mutual information between these sample pairs. For the first sample pair (image and text description), we set it as image x and its corresponding text

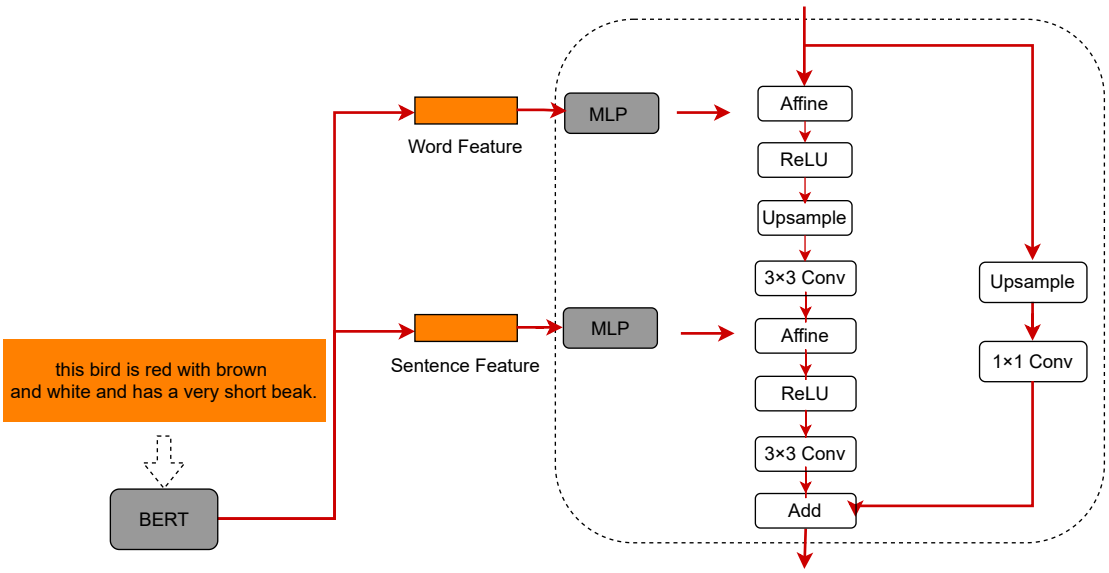

Fig. 1. Adaptive Semantic Image feature Fusion Module

description s respectively, and adopt cosine similarity as the scoring function:

$$S_{des}(x,s) = \cos(f_{img}(x), f_{sent}(s))/\tau \tag{3}$$

$$\cos(f_{img}(x), f_{sent}(s)) = \frac{f_{img}(x)^{\mathrm{T}} f_{sent}(s))}{\|f_{img}(x)\|\|f_{sent}(s)\|} \tag{4}$$

where, τ is a hyperparameter, $f_{img}(-)$ is an image encoder, and $f_{sent}(-)$ is a text encoder. Combined with the form of InfoNCE loss, the Contrastive loss of (image, text description) sample pair can be defined as:

$$L_{des}(x_i, s_i) = -log\frac{exp(\cos(f_{img}(x_i), f_{sent}(s_i))/\tau)}{\sum_{j=1}^{M} exp(\cos(f_{img}(x_i), f_{sent}(s_j))/\tau)} \tag{5}$$

This form of contrastive loss definition is also known as normalized temperature scale cross-entropy loss (NT-XENT). Similarly, the scoring function of the second sample pair is defined as follows:

$$S_{img}(x, x^{'}) = \cos(f_{img}(x), f_{img}(x^{'}))/\tau \tag{6}$$

In the sample pair, the real image is defined as x, the generated image is defined as $G(z,s)$, z is the random noise that follows the standard normal distribution, and s is the corresponding text description. The comparison loss of the sample pair is:

$$L_{img}(x_i, G(z,s^i)) = -log\frac{exp(S_{img}(x_i, G(z,s_i)))}{\sum_{j=1}^{M} exp(S_{img}(x_i, G(z,s_j)))} \tag{7}$$

Different from Reed et al. [8], when defining adversarial loss, we do not use the unmatched text as the negative sample, because it has been reflected in the comparison loss proposed earlier. Excessive loss function will cause high learning efficiency of the discriminator and result in serious disappearance of generator gradient. The overall loss function adopts Hinge loss [8] proposed by Reed et al. The exact training objective loss of the discriminator is defined as follows:

$$\begin{aligned} L_{adv}^{D} = &- E_{x\sim p_{real}}[min(0, -1 + D(x,s))] \\ &- E_{x\sim p_{fake}}[min(0, -1 + D(G(z),s))] \end{aligned} \tag{8}$$

where s is the given text description, $G(z)$ is the image generated by the generator, D is the discriminator, and x is the real image. Accordingly, the objective function loss of the generator is defined as follows:

$$L_{adv}^{G} = -E_{G(z)\sim p_{fake}}[D(G(z),s)] \tag{9}$$

IV. Experiments

Datasets: We trained and tested our model on the datasets CUB-200-2011 and COCO-2014. The CUB-200-2011 contains a total of 11,788 images of 200 bird species. The training set contains 5,994 images and the test set contains 5,794 images. Each image has 1 category label, 15 part positions, 312 binary attributes, and a bounding box. Also, Reed et al. [24] collected ten captions for each image. The COCO-2014 dataset is the first version of the COCO dataset, a large-scale dataset for object detection, segmentation, keypoint detection, and captioning. It covers 82,783 training images, 40504 validation images, and 40775 test images. Each image has 5 labels and 250,000 keypoints.

Experimental details: The experimental code was implemented using the Pytorch v1.9.0 framework and trained on an Ubuntu 16.04 LTS system using two NVIDIA TESLA V100s. For the text encoder Bert, we used the interface provided by hugging face to directly call the pre-trained Bert model. For the training strategy, we trained 600 epochs on the CUB-200-2011 dataset and 300 epochs on the COCO-2014 dataset, with a Batch Size of 48. The initial learning rate of the generator was set to 0.001 and was increased by 10% every 50 epochs

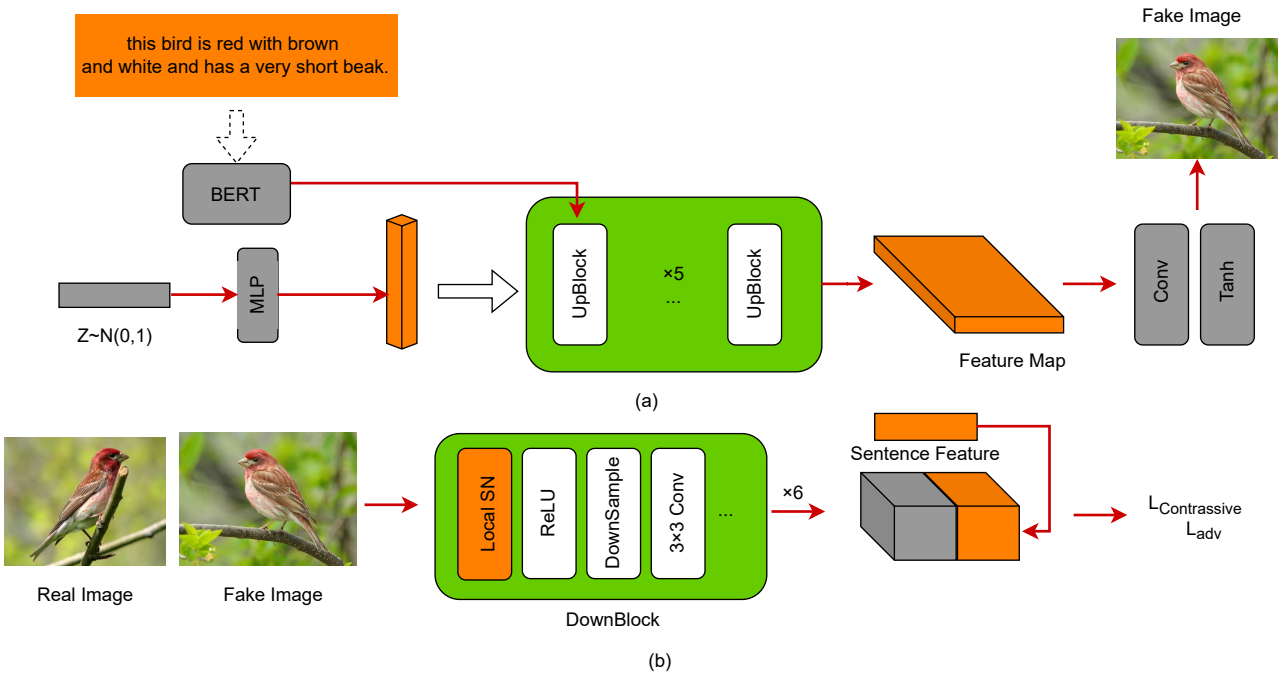

Fig. 2. (a): Generator network structure, the semantic vectors encoded by BERT will be combined with features in UpBlock to generate images (b): Discriminator network structure, with local spectral normalization added to each DownBlock

starting from the 250th epoch. The initial learning rate of the discriminator was set to 0.00097, decreasing by 10% for every 50 epochs in the first 250 epochs. Training took approximately 5 days on the CUB-2014 data and approximately 20 days on the COCO-2014 training set.

Evaluation metrics: We used the Inception Score (IS) [25] and Frechet Inception Distance (FID) [26] metrics to quantify our results. Inception Score (IS) was calculated by classifying the generated images using a pre-trained Inception-v3 network. The higher the IS score, the higher the image quality. The Frechet Inception Distance score (FID) is a measure to calculate the distance between the feature vectors of the generated image and the real image. First, the pre-trained Inception V3 network extracts the high-dimensional features of the images. Since the distribution is multi-dimensional, the covariance matrix is considered to measure the correlation between two distributions, while the mean matrix and covariance matrix are used to measure the distance between distributions. A smaller FID score indicates that the image is closer to the real image. To test our model, we will generate 30K images and calculate the IS as well as the FID based on the method provided by AttnGAN [10].

Compared methods: We will compare four novel and advanced methods with high visibility and authority in the field: AttnGAN [10], DM-GAN [27], DF-GAN [11], DAE-GAN [28].

A. Quantitative comparison

In this section, we quantitatively compare AS-GAN with several of the more advanced T2I methods. In Table I we show the IS scores and FID scores for each model on the dataset CUB-200-2011 for bird images and the MS COCO-2014 dataset for common object images.

For the COCO dataset, previous works [4] [29] [11] reported that the IS metric completely fails in evaluating the synthesized images. Therefore, we no longer calculate the IS score of the model on the COCO dataset. The bolded data indicates that the proposed method achieved the best score in the comparison. As can be seen, our model achieves the highest Inception score (5.26) compared to other state-of-the-art models in both the CUB and the COCO datasets. The majority of existing methodologies primarily concentrate on augmenting the alignment between image and textual content information. On the other hand, our work achieves significant improvements in the quality of the image (image sharpness, edge sharpness, etc.). At the same time, we also maintained a high score in the FID metric, which judges the authenticity of an image. To enhance the generalization ability as well as the robustness, contrastive learning is used during training to enhance the performance of the model. We fine-tuned the hyperparameters on the CUB dataset and migrated the model untuned to the COCO dataset and still achieved better results, showing the effectiveness of our approach.

B. Qualitative comparison

In this section, we show the qualitative comparison results of our method with state-of-the-art methods such as DF-GAN [11], AttnGan [10], and DM-GAN [27] on the bird dataset CUB-200-2011 and the generic object dataset COCO-2014, as shown in Fig. 3.

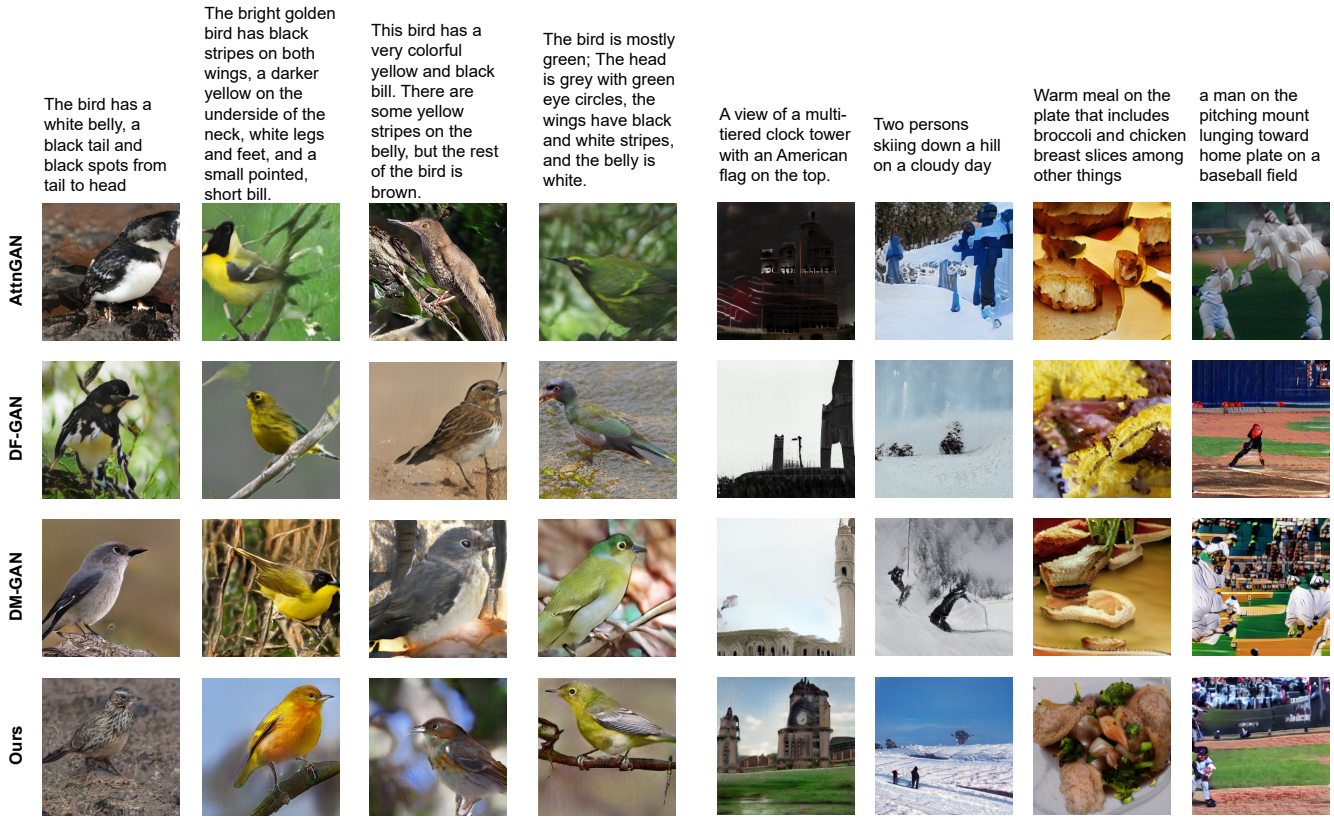

Fig. 3. Qualitative comparison on the CUB and COCO dataset. The input text descriptions are given in the first row and the corresponding generated images from different methods are shown in the same column. Best view in color.

TABLE I
PERFORMANCE OF IS AND FID OF ATTNGAN, DM-GAN, DF-GAN AND THE PROPOSED METHOD ON THE CUB AND COCO TEST SET. THE BEST RESULTS ARE IN BOLD.

	IS↑		FID↓	
Methods	CUB	COCO	CUB	COCO
AttnGAN	4.36	25.89 ± 0.07	23.98	35.49
DM-GAN	4.75	30.49 ± 0.5	16.09	32.64
DF-GAN	5.10	-	14.81	21.42
DAE-GAN	4.42	-	15.19	28.12
Ours	**5.26**	-	15.56	**20.12**

On the CUB dataset, the bird image that AS-GAN generated has a clearer picture, sharper contours, and more detailed feather textures compared to all other methods. For example in the fourth image, the bird feathers are smeared in the DM-GAN generated image. In addition, the background quality of our generated images is higher than that of DM-GAN and AttnGAN, and the number of artifacts is significantly lower than that of these two methods. It is worth noting that the proposed method generated significantly better bird shapes than DM-GAN and AttnGAN, due to the stacking structure of the networks in both methods, which resulted in distorted bird shapes, and the better background performance of DF-GAN, but in the third figure, the part of bird's tail is lacking. Our method also achieves a significant advantage in terms of the degree of matching between image and text content, as the text features contain more semantic information due to the encoder via Bert. In the first image, our bird image can be seen as a clear speckle, whereas of the other three methods, only DF-GAN generates an image with a small number of speckles, the other two methods just generate simple black and white colors.

We also migrated our methods to the more complex and sharply larger COCO dataset for training and comparison. As can be seen, the duplication of frames due to stacked stitching is more apparent in the images generated by DM-GAN and AttnGAN, again due to their network structure. In terms of fine-grained matching of text content, DF-GAN completely ignores the "Two person" in the third image. AS-

GAN achieves excellent results in terms of image quality, confidence, and matching to text content.

We first conducted ablation experiments on the various innovations proposed to demonstrate the effectiveness of each part, as shown in table II.

TABLE II
ABLATION STUDY. THE IS AND FID SCORES ON THE CUB DATASET AFTER ADDING BERT, CONTRASTIVE LOSS, BN, SN, AND ASIF MODULES TO THE MODEL

Architecture	IS↑	FID↓
Original	4.45	34.51
Original+B	4.54	25.34
Original+B+C	-	-
Original+B+C+BN	3.91	47.99
Original+B+C+SN	5.06	19.23
Original+B+C+SN+AS	5.26	17.87

V. CONCLUSIONS

The main challenge of text-to-image generation lies in the proficient integration of textual information into the image synthesis. Conventional models usually employe stacked modules to gradually fuse text information to generate images with different resolutions in stages. In contrast, the proposed AS-GAN adopts a single generator and discriminator to enhance the fusion degree of semantic features and image features via the ASIF module and successfully enhances image quality. The architectural not only ensures the preservation of semantic information integrity but also mitigates the difficulty of training. In addition, to improve semantic consistency, we incorporate contrastive loss and employ local spectral normalization to stabilize the training of the network. Experiments on different datasets show that the proposed model significantly improves image quality.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," *Advances in neural information processing systems*, vol. 29, 2016.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[4] W. Liao, K. Hu, M. Y. Yang, and B. Rosenhahn, "Text to image generation with semantic-spatial aware gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 187–18 196.

[5] M. L. Olson, S. Liu, R. Anirudh, J. J. Thiagarajan, P.-T. Bremer, and W.-K. Wong, "Cross-gan auditing: Unsupervised identification of attribute level similarities and differences between pretrained generative models," *arXiv preprint arXiv:2303.10774*, 2023.

[6] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[7] Y. Liu, H. Fan, F. Ni, and J. Xiang, "Clsgan: Selective attribute editing model based on classification adversarial network," *Neural Networks*, vol. 133, pp. 220–228, 2021.

[8] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*. PMLR, 2016, pp. 1060–1069.

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

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

[11] M. Tao, H. Tang, S. Wu, N. Sebe, X.-Y. Jing, F. Wu, and B. Bao, "Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis," *arXiv preprint arXiv:2008.05865*, 2020.

[12] S. Ye, F. Liu, and M. Tan, "Recurrent affine transformation for text-to-image synthesis," *arXiv preprint arXiv:2204.10482*, 2022.

[13] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[14] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, "Scaling up gans for text-to-image synthesis," *arXiv preprint arXiv:2303.05511*, 2023.

[15] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1505–1514.

[16] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, "Semantics disentangling for text-to-image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2327–2336.

[17] J. D. M. C. K. Lee and K. Toutanova, "Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[19] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Gaugan: semantic image synthesis with spatially adaptive normalization," in *ACM SIGGRAPH 2019 Real-Time Live!*, 2019, pp. 1–1.

[20] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[21] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[22] H. Ye, X. Yang, M. Takac, R. Sunderraman, and S. Ji, "Improving text-to-image synthesis using contrastive learning," *arXiv preprint arXiv:2107.02423*, 2021.

[23] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 833–842.

[24] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 49–58.

[25] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[26] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[27] M. Zhu, P. Pan, W. Chen, and Y. Yang, "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*, 2019, pp. 5802–5810.

[28] S. Ruan, Y. Zhang, K. Zhang, Y. Fan, F. Tang, Q. Liu, and E. Chen, "Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 960–13 969.

[29] Z. Zhang and L. Schomaker, "Dtgan: Dual attention generative adversarial networks for text-to-image generation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.