Enhancing Text-to-Image Generation with Diversity Regularization and Fine-Grained Supervision

Yujuan Qi [a], Debao Guo [b]

College of Control Science and Engineering, China University of Petroleum, Qingdao, Shandong, China

[a] qiyj@upc.edu.cn, [b] guodb0956@outlook.com

Abstract. Generating high-quality and realistic objects in the field of generation poses a significant challenge in artificial intelligence. Text-to-image generation technology is one of the focal points of research in this cutting-edge area. Currently, Generative Adversarial Networks (GANs) for text-to-image generation face two major issues: the mode collapse problem in conditional GANs is increasingly severe, and some existing models rely solely on sentence-level features, resulting in a lack of detailed features in the generated images. To address these problems, we propose a concise and efficient generative adversarial network named Text-to-Image Generation with Fine-Grained Supervision (T2IG-DFGAN). Its innovations include: (1) Introducing a regularization loss that guides the generator to produce diverse images under similar noise inputs by calculating pixel-level image differences and input noise differences; (2) Incorporating word-level features and introducing a fine-grained text matching loss to enhance image details. Compared to the current state-of-the-art techniques, T2I-DFGAN performs better in synthesizing realistic images that match text descriptions and demonstrates superior performance on multiple commonly used datasets.

Keywords: Text-to-Image Generation; GANs; Diversity Regularization; Fine-Grained Supervision.

1. Introduction

Text-to-image generation, which precisely converts natural language descriptions into closely corresponding, lifelike, and detailed images, has become a frontier in scientific research due to its profound practical application potential. In recent years, this field has made significant progress, with GANs [1] demonstrating outstanding performance in various domains, as shown in the literature [2,3,4], especially in the field of text-to-image generation where GANs have truly shone. However, despite these achievements, text-to-image generation technology based on GANs still faces two core challenges.

The primary issue is the inherent mode collapse problem in GANs, which is particularly prominent in text-to-image generation models based on GANs. When processing high-dimensional conditional information (such as text descriptions), the generator often neglects the importance of low-dimensional noise vectors, leading to highly similar generated images that lack necessary diversity. This high degree of similarity not only weakens the discriminator's ability to distinguish between images but may also cause stagnation or trapping in local optima during the training process, thereby further exacerbating the phenomenon of mode collapse. Secondly, when constructing text encoders, many advanced GANs models overly rely on sentence-level features while neglecting word-level features. This over-reliance results in an inadequate and inaccurate understanding of text descriptions, affecting the semantic consistency between generated images and text. To overcome these challenges, researchers have continuously innovated and explored. They adopt stacked network architectures as the core framework for generating high-quality images, and to effectively fuse text and image features, these models introduce cross-modal attention mechanisms [5-15]. Furthermore, they utilize Deep Fusion Generative Adversarial Networks [41] to make the generation process more efficient. To further enhance the semantic consistency between generated images and text, researchers have also introduced advanced components or strategies, such as the Deep Attention Multimodal Similarity Model (DAMSM) network [19], cyclic consistency principles [11], or twin network structures [20].

Particularly noteworthy is that our proposal of an innovative method, T2IG-DFGAN, is based on the Deep Fusion Generative Adversarial Networks [41], as illustrated in Fig. 1. In Fig. 1, (a) shows the images generated by DFGAN, while (b) demonstrates how T2IG-DFGAN generates high-quality images by introducing diversity regularization loss and incorporating word-level features, along with text-image matching loss.

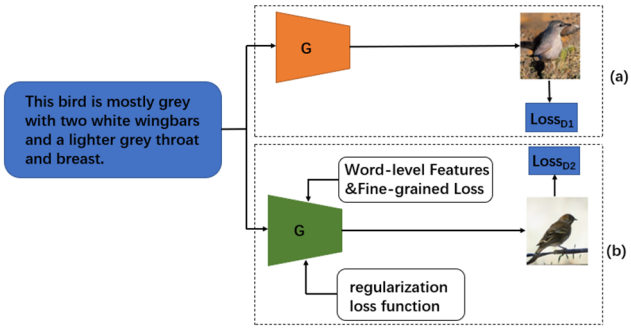

Fig 1. Existing text-to-image generation models versus our proposed T2IG-DFGAN for generating high-resolution images.

Overall, our contributions can be summarized as follows:

• Introduce a regularization loss that calculates pixel-level differences between images generated from similar noise inputs, guiding the generator to produce distinct images, avoiding mode collapse.

•Incorporate word-level features and use a fine-grained text matching loss to enhance detailed features in generated images, ensuring high consistency with text descriptions and boosting imagery detail.

•Conduct experiments on two datasets with the current state-of-the-art methods. The experimental results show that the proposed network exhibits better performance.

2. Related Work

GANs [1] have demonstrated their impressive capability in simulating complex real-world data distributions through a minimax game between the generator and the discriminator [21-25]. The Conditional Generative Adversarial Networks (CGANs) [18,26] bridges text and image generation proposed by Reed et al. The StackGAN series [16,17] enhances image resolution by stacking generators and discriminators, combining text with noise to create realistic images. AttnGAN [19] introduces a cross-modal attention mechanism, guiding the generator to focus on key text information to produce detailed and text-consistent images. MirrorGAN [33] reinforces semantic consistency by regenerating text from images in a closed-loop feedback system [27]. SD-GAN [28] employs a Siamese network [29,30] to deeply mine semantic commonalities, ensuring high consistency during image generation. DM-GAN [15] further optimizes initial image generation with a Memory Network [31,32], storing and refining key information to yield clearer images. DFGAN [41] shines in the field of text-to-image generation with its single-stage non-entangled text-to-image generation backbone, an object-aware discriminator aimed at enhancing semantic consistency, and fusion blocks that deepen the integration of text and image.

Inspired by these advanced methods, we propose the T2IG-DFGAN model. T2IG-DFGAN significantly enhances image diversity by introducing regularization loss to minimize pixel differences arising from noise variations. Additionally, the model leverages word-level features and fine-grained matching loss to enrich image details, resulting in generated images that are not only more realistic but also more aligned with text descriptions, outperforming previous models. By integrating these innovative elements, T2IG-DFGAN opens up broader horizons for text-to-image generation.

3. The Proposed Method

3.1 Model Overview

The T2IG-DFGAN model comprises a generator, discriminator, and pre-trained text encoder (Fig. 2). The generator takes a sentence vector, a word vector, and a Gaussian noise vector. The noise vector is reshaped and fused with the sentence vector in a text fusion block using transformations and dynamic convolutions. Word features are further fused for detailed semantics. These features are converted to images via convolution. The discriminator downsamples images through DownBlocks, and computes adversarial loss with gradient penalty, distinguishing fakes from reals to enhance image quality and text-image alignment. The text encoder uses a pre-trained bidirectional LSTM from AttnGAN.

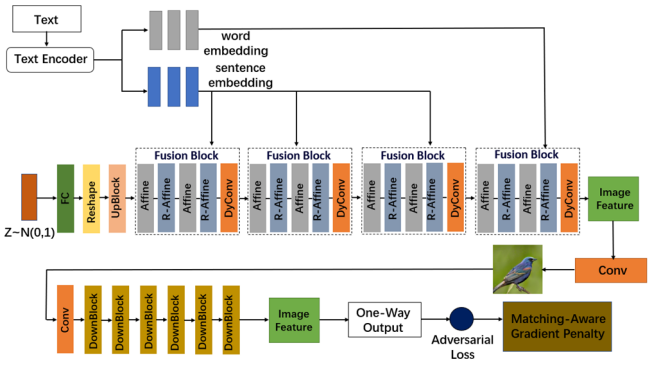

Fig 2. Structural Diagram of the T2IG-DFGAN Model.

3.2 A Special Regularization Loss Function

CGANs enhance model flexibility and controllability by adding conditional input. This allows generated data to match user-specified conditions but also intensifies mode collapse. Mode collapse occurs when the generator produces limited or identical samples due to focusing on the conditional context, which contains complex semantic information, rather than maintaining diversity through the noise vector like in traditional GANs.

To tackle this challenge, this paper introduces a unique regularization loss l_z, as shown in (1), to balance conditional control and sample diversity. Where l_z measures the ratio of pixel-level differences between generated images to differences in their input noise, indicating the generator's efficiency in using noise for diversity. High l_z penalizes over-reliance on conditional context, urging the generator to maintain conditional accuracy while leveraging noise for diversity.

$$l_z = max\left(\frac{d_I\left(G(z_1,c),\mathrm{G}(z_2,c)\right)}{d_z(z_1,z_2)}\right) \tag{1}$$

Where z is a noise vector following a Gaussian distribution, c is a condition vector, d_I denotes the distance between the generated images, and d_z represents the distance between the noise vectors.

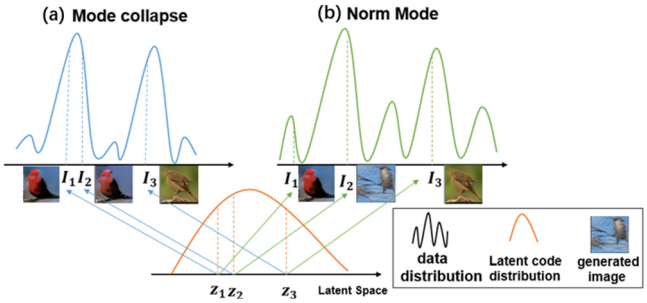

Fig 3. Comparison between Mode Collapse and Normal Mode.

In the latent space, similar noise elements (such as z_1 and z_2 in Fig. 3) may lead the generator to produce highly similar images due to the shared, similar information they contain, resulting in mode collapse, as shown in Fig. 3(a). By introducing and maximizing the diversity loss, specifically maximizing l_z in (1), the generator is encouraged to produce diverse images, thereby mitigating the mode collapse problem, as illustrated in Fig. 3(b).

During loss function regularization, as shown in (2), we use a small constant *eps* to ensure numerical stability, preventing division by zero when l_z is zero.

$$L_{ms}=\frac{1}{l_z+eps} \tag{2}$$

Where l_z is the original regularization term, and *eps* is a very small positive number used to ensure that the denominator does not become zero.

In addition, the original loss function of the model, as shown in (3), consists of two parts: one part is the logarithmic loss for real data, and the other part is the logarithmic loss for generated data.

$$L_{ori}=E_{c,y}[log(c,y)]+E_{c,z}\left[log\left(1-D\left(c,G(c,z)\right)\right)\right] \tag{3}$$

Where c is the conditional context, z is Gaussian noise, D is the discriminator, and G is the generator. We add the modified regularization L_{ms} to L_{ori}, forming L_{new1}. L_{new1}retains original info and enhances stability with *eps*, guiding model training better.

3.3 Word-level Features and Fine-grained Loss

When the text encoder is confined solely to utilizing sentence-level features while completely neglecting word-level and finer-grained features, the generated images inevitably fall into the trap of lacking details. While these images can roughly outline the scene contours or emotional atmosphere described by the sentence, they appear hollow and lack depth due to the absence of specific and vivid detail descriptions. Consequently, they fail to meet users' basic demands for image content richness and realism. Furthermore, due to the severe deficiency in detail representation, these images often lack complex textures, lighting changes, and subtle interactions between objects common in natural images. These shortcomings not only make the images appear unlifelike but, more importantly, serve as obvious indicators distinguishing "real" from "fake." When subjected to human eyes or advanced image recognition technologies, their unnaturalness becomes apparent, making them easily identifiable as "fake" images generated by some algorithm.

(a) (b)

Fig 4. Comparison Lacks in Fine-Grained Image & Fused Word-Vector Generation.

To address the aforementioned issues, this paper innovatively introduces word-level features and designs a fine-grained text-matching loss mechanism, aimed at enhancing the detailed features of the generated images, with specific effects illustrated in Fig. 4. In Fig. 4 (a), the drawbacks of lacking fine-grained processing can be clearly observed. Of particular note is the image in the first column of

the second row in (a), where the bird's color is nearly indistinguishable from the tree trunk, blending into one another. Similarly, in the image in the third column of the second row in (a), the overall tone appears dull and lifeless, and an unexpected occlusion area appears at the bottom of the image, which not only significantly reduces the image's clarity but also greatly impairs the viewing experience. In contrast, Fig. 4 (b) incorporates the processing of word-level vectors during the generation process, achieving significant improvements in fine-grained details. This ensures that the generated images not only align with the overall text description but also exhibit a high degree of consistency in even the most subtle details, thereby greatly enhancing the level of detail and realism of the images.

During image generation, we introduce fine-grained text-to-image matching loss (4 & 5) to measure matching degrees between images and text, enhancing image quality.

$$L_1^w = -\sum_{i=1}^{M} logP(D_i|Q_i) \tag{4}$$

$$L_2^w = -\sum_{i=1}^{M} logP(Q_i|D_i) \tag{5}$$

Where M is the word count in the text description. D_i represents an image region, with Q_i its corresponding text. L_1^w. represents the logarithm of the probability of predicting the corresponding image region D_i given the text Q_i , L_2^w does the same for predicting Q_i given D_i.

4. Experiment

In this section, we will introduce the datasets, training details, and evaluation metrics used in our experiments, followed by assessments on T2IG-DFGAN and its variants.

Datasets: The proposed model is evaluated on two challenging datasets: CUB bird [36] and COCO [37]. CUB contains 11,788 images of 200 bird species with ten descriptions per image. COCO comprises 80,000 training and 40,000 testing images, each with five descriptions.

Training Details: Our model use the Adam optimizer [38] with β1=0.0 and β2=0.9. Following TTUR [12], the generator's learning rate is 0.0001, and the discriminator's is 0.0004. Our program runs on Python 3.8, PyTorch 1.9, and CUDA 11.1.

Evaluation Details: Our model use the Inception Score (IS) [39] and Frechet Inception Distance (FID) [40] for evaluation. IS assesses image quality by calculating KL divergence, with higher scores indicating better quality. FID measures the distance between synthetic and real images in the feature space of Inception v3, with lower scores indicating more realistic images.

4.1 Quantitative Evaluation

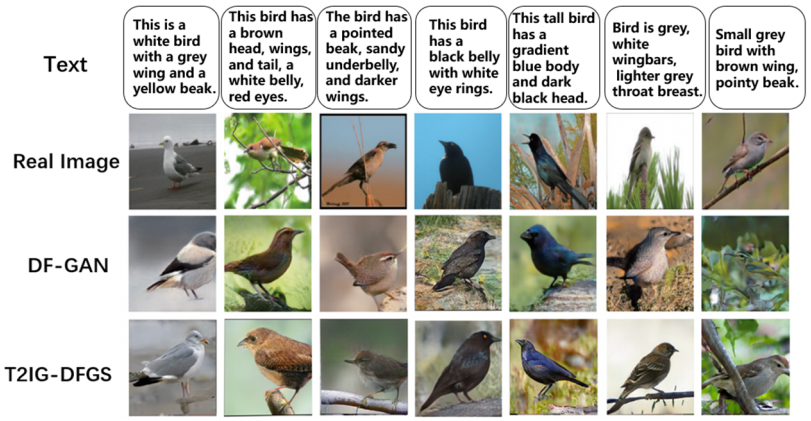

Fig 5. Comparison of text-generated image examples using DF-GAN and T2IG-DFGAN methods, along with real images, on the CUB dataset

In this section, we conduct an in-depth comparison of the proposed text-to-image synthesis method with several state-of-the-art techniques that all employ stacked structures, specifically StackGAN [16], StackGAN++[17], AttnGAN[19], MirrorGAN[12], SD-GAN[28], DM-GAN[15], and DF-GAN[41]. The performance and characteristics of these methods are detailed and compared in Table 1, providing a clear evaluation framework.

To further validate the effectiveness of our method, experiments are conducted on the CUB [36] dataset and the visualization results generated by DF-GAN are compared with those by our method, using real images as a reference, as shown in Fig. 5.

Table 1. Comparison of IS, FID, and results on the test sets of CUB and COCO with state-of-the-art methods.

Model	CUB		COCO
	***FID*↓**	***IS*↑**	***FID*↓**
StackGAN [16]	-	3.70	-
StackGAN++[17]	-	3.84	-
AttnGAN [19]	23.98	4.36	35.49
MirrorGAN [12]	18.34	4.56	34.71
SD-GAN [28]	-	4.67	-
DM-GAN [15]	16.09	4.75	32.64
CPGAN [43]	-	-	55.80
DAE-GAN [44]	15.19	4.42	28.12
TIME [42]	14.30	4.91	31.14
DF-GAN [41]	14.81	5.10	19.32
T2IG-DFGAN	**12.25**	**5.22**	**17.12**

As shown in Table I, T2IG-DFGAN demonstrates competitive performance compared to other mainstream models. Compared to AttnGAN[19], T2IG-DFGAN improves the IS from 4.36 to 5.22 and reduces the FID [40] from 23.98 to 12.25 on the CUB [36] dataset. On the COCO [37] dataset, our model (assuming it refers to a related variant) reduces the FID from 35.49 to 17.12. Compared to MirrorGAN[12] and SD-GAN[28], our model enhances the IS on the CUB dataset from 4.56 and 4.67 to 5.22, respectively. Compared to DM-GAN[15], our model improves the IS from 4.75 to 5.22 and decreases the FID [40] from 16.09 to 12.25 on the CUB [36] dataset, while also reducing the FID [40] from 32.64 to 17.12 on the COCO dataset [37].

4.2 Ablation Study

In this section, a thorough ablation study on the test sets of the CUB [36] and COCO [37] datasets to verify the effectiveness of each component in the proposed T2IG-DFGAN model. These key components include the diversity regularization loss, word-level loss, and SE attention mechanism. DF-GAN [41] is selected as the baseline model and the output image size is adjusted to 64.

On the CUB dataset, the baseline model is first compared the baseline model with the introduction of a regularization loss term, and meticulously recorded the best Inception Score (IS) [39] achieved during 500 training epochs. Subsequently, diversity regularization loss and word-level loss are further incorporated, and the results are compared with the baseline model once again. The best Fréchet Inception Distance (FID) [40] achieved during 300 training epochs is meticulously noted. The specific data are presented in Table II.

Table 2. Performance of Different Components of Our Model on the CUB Test Set.

Architecture	IS↑	FID↓
Baseline	3.93±0.36	144.22
+LDR Loss	**4.21±0.50**	149.81
+word Loss	-	**143.73**

On the COCO dataset, the research delved further. Not only are the regularization loss, word-level loss, and SE attention mechanism introduced, but the combined effect of the regularization term and SE attention mechanism is also explored. By comparing with the baseline model, the best FID results for various metrics during 165 training epochs are meticulously recorded. The specific data are presented in Table III.

Table 3. Performance of Different Components of Our Model on the COCO Test Set.

Architecture	FID↓
Baseline	118.33
+LDR Loss	117.27
+word Loss	118.70
+SE Attention	117.89
+ LDR Loss& SE Attention	**106.91**

Through the results of the above ablation studies, it can be clearly seen that the various components of the T2IG-DFGAN model have indeed brought improvements and enhancements in the text-to-image generation task.

5. Conclusion

In this paper, a method named T2IG-DFGAN is proposed for the task of text-to-image generation. We incorporate a diversity regularization mechanism within a single-stage text-to-image generation backbone network to encourage the generator to create diverse images. Additionally, word-level features are integrated to make the generated images not only more refined but also more aligned with the text descriptions. This process is achieved through a carefully designed text-image loss function. Extensive experimental results demonstrate that our proposed T2IG-DFGAN method significantly outperforms the current state-of-the-art models on both the CUB dataset and the more challenging COCO dataset.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672–2680, 2014.

[2] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. IEEE transactions on pattern analysis and machine intelligence, 43(10):3365–3387, 2020.

[3] Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati, and Jiaying Liu. Fashion meets computer vision: A survey. ACM Computing Surveys (CSUR),54(4):1–41, 2021.

[4] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative adversarial networks for image and video synthesis: Algorithms and applications. Proceedings of the IEEE, 109(5):839–862, 2021.

[5] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. Rifegan: Rich feature generation for text-toimage synthesis from prior knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10911–10920, 2020.

[6] Yuchuan Gou, Qiancheng Wu, Minghao Li, Bo Gong, and Mei Han. Segattngan: Text to image generation with segmentation attention. arXiv preprint arXiv:2005.12444, 2020.

[7] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical textto-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7986–7994, 2018.

[8] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In Advances in Neural Information Processing Systems, pages 2065–2075,2019.

[9] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12174–12182, 2019.

[10] Ruifan Li, Ning Wang, Fangxiang Feng, Guangwei Zhang, and Xiaojie Wang. Exploring global and local linguistic representation for text-to-image synthesis. IEEE Transactions on Multimedia, 2020.

[11] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Learn, imagine and create: Text-to-image generation from prior knowledge. In Advances in Neural Information Processing Systems, pages 887–897, 2019.

[12] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1505–1514, 2019.

[13] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092, 2021.

[14] Mingkuan Yuan and Yuxin Peng. Ckd: Cross-task knowledge distillation for text-to-image synthesis. IEEE Transactions on Multimedia, 2019.

[15] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for textto-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5802– 5810, 2019.

[16] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 5907– 5915, 2017.

[17] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE TPAMI, 41(8):1947–1962, 2018.

[18] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Proceedings of the International Conference on Machine Learning, pages 1060– 1069, 2016.

[19] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Finegrained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1316– 1324, 2018.

[20] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-toimage generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2327– 2336, 2019.

[21] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12174–12182, 2019.

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4401–4410, 2019.

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8110–8119, 2020.

[24] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In ECCV, 2020.

[25] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In International conference on machine learning, pages 7354– 7363. PMLR, 2019.

[26] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In Advances in neural information processing systems, pages 217–225, 2016.

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycleconsistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223– 2232, 2017.

[28] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-toimage generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2327– 2336, 2019.

[29] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In European conference on computer vision, pages 135–153. Springer, 2016.

[30] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In European conference on computer vision, pages 135–153. Springer, 2016.

[31] Caglar Gulcehre, Sarath Chandar, Kyunghyun Cho, and Yoshua Bengio. Dynamic neural turing machine with continuous and discrete addressing schemes. Neural computation, 30(4):857–884, 2018.

[32] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In International Conference on Learning Representations, 2015.

[33] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. arXiv preprint arXiv:2105.13290, 2021.

[34] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. arXiv preprint arXiv:2103.00823, 2021.

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092, 2021.

[36] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence ´ Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

[38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015.

[39] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Advances in neural information processing systems, pages 2234–2242, 2016.

[40] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in neural information processing systems, pages 6626–6637, 2017.

[41] Tao M, Tang H, Wu F, et al. Df-gan: A simple and effective baseline for text-to-image synthesis [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16515-16525.

[42] Bingchen Liu, Kunpeng Song, Yizhe Zhu, Gerard de Melo, and Ahmed Elgammal. Time: text and image mutual-translation adversarial networks. arXiv preprint arXiv:2005.13192, 2020.

[43] Jiadong Liang, Wenjie Pei, and Feng Lu. Cpgan: Contentparsing generative adversarial networks for text-to-image synthesis. In European Conference on Computer Vision, pages 491–508. Springer, 2020.

[44] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspectaware gan for text-to-image synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13960–13969, 2021.