DISCRIMINATOR MODIFICATION IN GAN FOR TEXT-TO-IMAGE GENERATION

Fei Fang[1]*, Ziqing Li*[1]*, Fei Luo*[1*]*, and Chunxia Xiao*[1*]

1. School of Computer Science, Wuhan University, Wuhan 430072, China.
fangfei369@163.com, thalialee@163.com, luofei_whu@126.com, cxxiao@whu.edu.cn

ABSTRACT

The existing Generative Adversarial Network-based text-to-image generation methods suffer from mode collapse and training instability. This paper relieves these problems by improving the discriminator ability from three aspects. First, we propose a diversity-sensitive conditional discriminator (D-SCD), which increases the diversity of the generated images by judging the combination of the generated image and mismatched text as false. Second, for the unconditional discriminator, we propose a contrastive searching gradient penalty (CSGP) strategy to measure the realism of the generated images and to penalize the gradients for stabilizing the training process. Finally, we introduce a multi-level images similarity (MLIS) loss for the discriminator feature extractor to further promote the high-level feature similarity between the real and generated images and objects. Extensive experimental results and ablation studies demonstrate that our modifications on the discriminators can effectively improve the quality of the generated images.

Index Terms— Text-to-image generation, discriminator, diversity, stability

1. INTRODUCTION & RELATED WORKS

Generating realistic images from text descriptions is an active research field in computer vision and multimedia communities. Our goal is to generate photo-realistic images that can exhibit as much semantic information of the text description as possible. Generative Adversarial Networks (GANs) have played an important role in text-to-image generation as the generator-discriminator structure is suitable for the cross-modality transformation task.

Many works [1–4] focus on the improvement of generators and fine-grained text-image consistency. In practice, the discriminators are important in providing the right guidance for the training of generators. The discriminator of GAN for the text-to-image generation task was firstly designed to judge whether the feature of the generated image (fake image) is consistent with the sentence feature vector [5]. Zhang et al. [6] proposed hierarchically-nested discriminators for multi-scale fake images to train the generator jointly. Zhang et al. [7] proposed joint conditional and unconditional (two-way) discriminators, which can determine whether a fake image matches the input description as [5] and distinguish realistic images from fake ones, respectively. The two-way discriminators are widely used in the subsequent works. The methods [3,4,8] proposed fine-grained discriminators accompanying with the two-way discriminators to estimate the local and global image quality of the fake images.

Text-to-image generation is a multi-modal task that more than one image can visualize the input text. However, the two-way discriminators ignore this feature and conduct overstrict punishment for the multi-modal task. The strict discriminators will give rise to mode collapse and training instability of GAN. To increase the diversity of the fake images and alleviate the instability of GAN's training, Lim et al. [9] proposed geometric GAN using the separating hyperplane from Support Vector Machine. Their proposed Hinge loss can stabilize the training of GAN but may not effectively measure the similarity between the real and fake images. Hu et al. [10] proposed DCGAN to increase the diversity of single-object fake images using multiple generators and one discriminator. Tao et al. [11] considered some fake images as real ones during the training process. However, this strategy applied a fixed number of false images as real ones at each training step, which reduced accuracy.

Improving the similarity between corresponding real and fake images can further increase the diversity among fake images. Recently, some works added fake-real images similarity loss to promote image similarity in high-level image features. For example, Zhang et al. [12] proposed contrastive losses between (1) image and text, (2) region and word, and (3) fake and real images to train the generator. They used discriminator network as image feature extractor. However, they did not train the discriminator specially as a qualified feature extractor to distinguish the high-level features of the real and fake images. For complex scene images, we need to improve the similarity between real and fake images, objects and object positions. Dong et al. [13] captured features at both image and object levels to generate image captions. Some methods [14, 15] built scene graphs or established rules to constraint the objects and their positions, but they are less efficient than

This work is partially supported by the Key Technological Innovation Projects of Hubei Province (2018AAA062) and NSFC (No. 61972298). ∗ Chunxia Xiao and Fei Luo are corresponding authors.

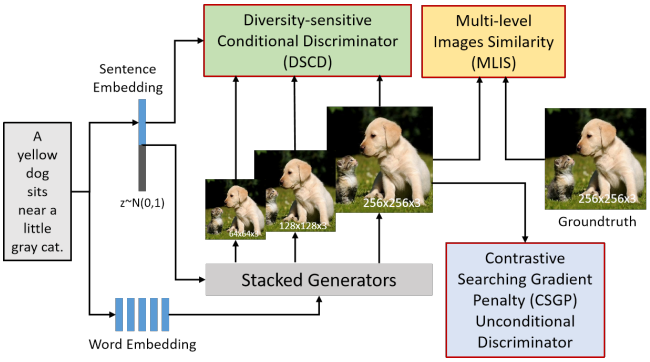

Fig. 1. The overall structure of our proposed method.

trained GANs.

To alleviate the above problems, we propose three solutions to improve conditional and unconditional discriminators. First, we change one of the judgments in the training of the conditional discriminator. Traditional methods judge the combinations of the generated images and *matched* texts as false, which ignore the diversity of the fake images. Instead, we judge the combinations of the generated images and *mismatched* texts as false. This modification can alleviate mode collapse and facilitate the generation of more diverse images. Second, we propose a contrastive searching gradient penalty strategy for the unconditional discriminator. We determine whether the fake image is real or fake by comparing its distances from both the constructed real and fake images. To stabilize the training of GAN, we also penalize the gradient explosion caused by undervaluing some contrastive real images. Finally, we propose multi-level images similarity loss to improve the similarity of the fake & real images and fake & real foreground objects. Before extracting image and object features using the discriminator network, we train the discriminator network to be a qualified feature extractor. The trained discriminator can distinguish the features from real and fake images by using triplet and center losses.

In conclusion, the contributions of our work are threefold:

(1) We propose a diversity-sensitive conditional discriminator (DSCD) to increase the diversity of the fake images and alleviate the mode collapse;

(2) We propose a contrastive searching gradient penalty (CSGP) strategy for the unconditional discriminator. It can better evaluate the realism of the fake images and penalize the gradient explosion for stabilizing the training of GAN;

(3) We propose a multi-level images similarity (MLIS) loss to improve the similarity measurement between the features of real and fake images and objects. Specially, we train the discriminator to effectively extract features, which are used to compute the MLIS loss.

2. METHOD

We adopt the widely used two-way (conditional and unconditional) discriminators to train the generators in GAN for the text-to-image generation task. The overall structure of our method is shown in Fig. 1.

2.1. Diversity-sensitive conditional discriminator (DSCD)

The conditional discriminator is trained to give correct judgments for three different combinations. It needs to judge the combined features of (1) the real images and *matched* texts as true, (2) the real images and *mismatched* texts as false and (3) the fake images and *matched* texts as false:

$$\begin{cases} D^{\mathrm{c}}(x_i, s_j) = True, i = j \\ D^{\mathrm{c}}(x_i, s_j) = False, i \neq j \\ D^{\mathrm{c}}(G(z_i, s_i), s_j) = False, i = j \end{cases} \tag{1}$$

where D^{c} is the conditional discriminator, G is the generator, $x_i \sim P_{\mathrm{data}}$ is the i-th real image, $z_i \sim P_{(z)}$ is the noise vector that normally sampled from standard normal distribution and s_j is the input text description.

First, it is unreasonable to judge the combination of the fake images and *matched* texts as false. On the one hand, a fake image with poor realism is unlikely to match the corresponding text, which means this judgment is barely used. In more cases, the fake images are neither realistic nor semantically match the input text, especially in the early stages of GAN's training. On the other hand, when the generated images have high quality in the late stages of GAN's training, it is more reasonable to remove this judgment. In practice, the qualified generated images will not look the same as the real images due to the multi-modal nature of the text-to-image generation task. When we have high-quality fake images that match the corresponding texts but different from the corresponding real images, the original D^{c} may become too strict to judge the combinations as true. Moreover, this overstrict D^{c} will further lead to mode collapse in the image generation [11].

Meanwhile, we find it is beneficial to judge the combination of the fake images and *mismatched* texts as false:

$$D^{\mathrm{c}}(G(z_i, s_i), s_j) = False, i \neq j. \tag{2}$$

The modified D^{c} will penalize the fake images that mismatch the corresponding input texts. When the mode collapse appears, many generated fake images are visually similar and tend to be semantically inconsistent with the corresponding input texts. The modified D^{c} will penalize this phenomenon and generate more reasonable adversarial losses to relieve the mode collapse and help generate more diverse images.

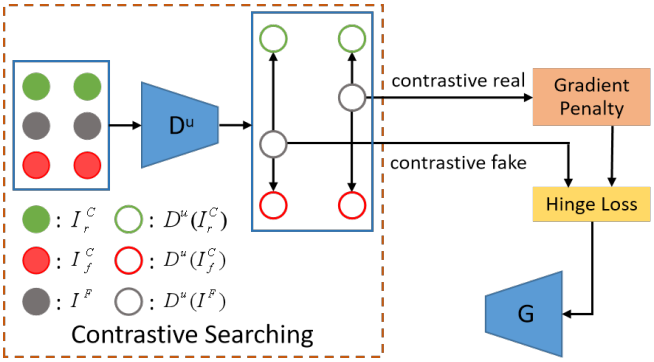

Fig. 2. The main approach of the contrastive searching gradient penalty (CSGP) strategy for unconditional discriminator.

2.2. Contrastive searching gradient penalty (CSGP) for unconditional discriminator

In this work, we propose a contrastive searching gradient penalty (CSGP) strategy for unconditional discriminator, which is shown in Fig. 2. We propose a contrastive searching approach to judge the realism of the generated images. Then we penalize the gradients produced by the undervalued fake images to stabilize the training of GAN.

2.2.1. Contrastive searching for unconditional discriminator

We propose a more reasonable approach for the unconditional discriminator to estimate the realism of the fake images. Specifically, we first construct real and fake image sets, and then compute the distances from the fake image to both the constructed image sets. We estimate a fake image as a contrastive real image if it is close to the constructed real image. Otherwise, we estimate it as a contrastive fake image.

Formally, we construct a real image set $I_{\mathrm{r}}^{\mathrm{C}}$ for a batch of fake images by:

$$I_{\mathrm{r}}^{\mathrm{C}} = \lambda_1 I^{\mathrm{R}} + (1-\lambda_1) I^{\mathrm{F}}, \tag{3}$$

where I^{R} is the ground truth real images and I^{F} is the generated fake images. In our implementation, we randomly set $\lambda_1 \in [0.85, 0.99]$. Likewise, we construct the fake image set $I_{\mathrm{f}}^{\mathrm{C}}$ by:

$$I_{\mathrm{f}}^{\mathrm{C}} = \lambda_2 I^{\mathrm{N}} + (1-\lambda_2) I^{\mathrm{F}}, \tag{4}$$

where I^{N} is the random noise images with the same sizes as the fake images. Their pixel values are randomly sampled from $Uniform(-1, 1)$. The value of λ_2 is randomly sampled from $[0.4, 0.6]$.

For a generated fake image I_i^{F}, we compare its distances from both the corresponding constructed real and fake images $I_{\mathrm{r}i}^{\mathrm{C}}$ and $I_{\mathrm{f}i}^{\mathrm{C}}$ to decide whether I_i^{F} is a contrastive real image. Formally, we compute both the distances $d_{\mathrm{r}i}$ and $d_{\mathrm{f}i}$ utilizing the output values from the unconditional discriminator D^{u}:

$$d_{\mathrm{r}i} = |D^{\mathrm{u}}(I_i^{\mathrm{F}}) - D^{\mathrm{u}}(I_{\mathrm{r}i}^{\mathrm{C}})| \tag{5}$$

$$d_{\mathrm{f}i} = |D^{\mathrm{u}}(I_i^{\mathrm{F}}) - D^{\mathrm{u}}(I_{\mathrm{f}i}^{\mathrm{C}})| \tag{6}$$

where $|\cdot|$ means the absolute value of the differences. Finally, we judge the generated image I_i^{F} as a contrastive real image if $d_{\mathrm{r}i} < d_{\mathrm{f}i}$ and $D^{\mathrm{u}}(I_i^{\mathrm{F}}) \geq \alpha$. The parameter α is used to ensure the realism of I_i^{F}, and we normally set it to be 0.

2.2.2. Gradient penalty of the contrastive real images

We improve Hinge loss [9] to train the D^{u} and G more effectively and stabilize the training process. At every training step, we update D^{u} twice and then update G once. Specifically, at step t, we first search for the contrastive real images M_{cr} and contrastive fake images M_{cf} from a batch of fake images generated by G_t. Then we update the current D_t^{u} to $D_t^{'\mathrm{u}}$ by normal Hinge Loss:

$$\begin{aligned}\mathcal{L}_{D_t^{'\mathrm{u}}} = &\mathbb{E}_{x\sim P_{\mathrm{data}}}[\max(0, 1 - D_t^{\mathrm{u}}(x))] \\ &+\mathbb{E}_{\hat{x}\in M_{\mathrm{cf}}}[\max(0, 1 + D_t^{\mathrm{u}}(\hat{x}))],\end{aligned} \tag{7}$$

where $\hat{x}$ is a fake image. In practice, the normally updated $D_t^{'\mathrm{u}}$ may undervalue a subset of the contrastive real images $N_{\mathrm{cr}} \subseteq M_{\mathrm{cr}}$ to be less than α. The undervaluation of the contrastive real images can cause gradient increase or even gradient explosion, thus causing the training instability of GAN. Therefore, we punish the undervaluation of N_{cr} and update the $D_t^{'\mathrm{u}}$ to be D_{t+1}^{u} with the following loss:

$$\mathcal{L}_{D_{t+1}^{\mathrm{u}}} = \mathbb{E}_{\hat{x}\in N_{\mathrm{cr}}}[(\alpha - D_t^{'\mathrm{u}}(\hat{x}))\sigma(\beta D_t^{\mathrm{u}}(\hat{x}))], \tag{8}$$

where the σ is the sigmoid function, and the value of $\sigma(\cdot)$ is the weight of this punishment. The larger value of $D_t^{\mathrm{u}}(\hat{x})$ means the higher realism of the fake image $\hat{x}$. The hyper-parameter β is used to differentiate the weight values.

Moreover, we improve the adversarial Hinge loss to update the generator with the updated D_{t+1}^{u}:

$$\begin{aligned}\mathcal{L}_{G_{t+1}^{\mathrm{u}}} = &\mathbb{E}_{z\sim P_{(z)}}[\max(0, \gamma - D_{t+1}^{\mathrm{u}}(G_t(z)))] \\ &+\mathbb{E}_{z\sim P_{(z)}}[\max(0, D_{t+1}^{\mathrm{u}}(G_t(z)) - 1)],\end{aligned} \tag{9}$$

where γ is an adaptive parameter and we clip it to be $0.5 \leq \gamma \leq 1.0$. For a batch of k real images $\mathbf{x} \sim P_{\mathrm{data}}$, we set γ to be the smallest value of $D^{\mathrm{u}}(\mathbf{x})$. In the above adversarial loss, we promote the value of $D^{\mathrm{u}}(G(z))$ to be between γ and 1. This approach can reduce the distances between $D^{\mathrm{u}}(G(z))$ and $D^{\mathrm{u}}(x)$, thus reducing the differences between the fake and real images.

2.3. Multi-level images similarity (MLIS)

We compute multi-level images similarity loss to promote the real and fake similarity at both image and object levels. Before computing the similarity, we train the discriminator network to extract features that can reflect the differences between real and fake images. The computation approach of

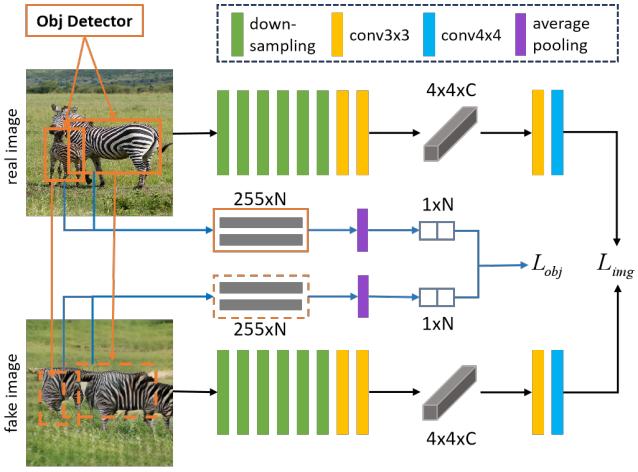

Fig. 3. The computation approach of MLIS and the discriminator networks.

MLIS and the discriminator networks are shown in Fig. 3. We use Siamese-structured networks as the image feature extractors. The networks share the layers and parameters for obtaining the $4 \times 4 \times C$ feature maps with the conditional and unconditional discriminators.

2.3.1. *Training of discriminator*

We construct positive and negative samples and use triplet and center losses to train the discriminator networks in Fig. 3. Our goal is to reduce the distances between real and positive samples and increase the distances between real and negative samples.

For a real image I_i^{R}, the positive samples include m images with high realism and consistent semantic with I_i^{R}, and the negative samples include m images with low realism and inconsistent semantic with I_i^{R}. We input I_i^{R} and all the positive and negative samples into the discriminator network in Fig. 3 and generate output values. The output value of I_i^{R} is $V_{\mathrm{r}i}$, and the output values of the positive and negative samples are Ω_{p} and Ω_{n}, respectively. We compute the absolute differences between $V_{\mathrm{r}i}$ and Ω_{p} and represent the *largest* difference value as V_{p}. Similarly, we compute the absolute differences between $V_{\mathrm{r}i}$ and Ω_{n} and represent the *smallest* difference value as V_{n}. (See more details in the supplementary material.)

To train the discriminator, we compute the triplet loss to make the maximum distance between the I_i^{R} and the positive samples to be still smaller than the minimum distance between the I_i^{R} and the negative samples:

$$\mathcal{L}_{\mathrm{trip}} = \max(0, V_{\mathrm{p}} - V_{\mathrm{n}} + \theta), \tag{10}$$

where θ is the margin of the triplet loss and we set $\theta = 0.1$. In our implementation, computing the center losses can improve the differentiation ability and robustness of the discriminator network:

$$\mathcal{L}_{\mathrm{c1}} = \frac{1}{m}\sum_{j=1}^{m} \| \Omega_{\mathrm{p}j} - V_{\mathrm{r}i} \|_2^2, \tag{11}$$

$$\mathcal{L}_{\mathrm{c2}} = \frac{1}{m}\sum_{j=1}^{m} \| \Omega_{\mathrm{n}j} - V_{\mathrm{r}i} \| - c_{\mathrm{n}} \|_2^2, c_{\mathrm{n}} = \frac{1}{m}\sum_{k=1}^{m} | \Omega_{\mathrm{n}k} - V_{\mathrm{r}i} | \tag{12}$$

where c_{n} is the mean distances between $V_{\mathrm{r}i}$ and Ω_{n}.

2.3.2. *Multi-level images similarity*

As shown in Fig. 3, for real and fake image similarity, we first feed the $256 \times 256 \times 3$ real and fake images into the discriminator network and extract $4 \times 4 \times C$ feature maps by several down-sampling layers and convolution layers, where C is the number of channels. The $4 \times 4 \times C$ feature maps are then processed by other two convolution layers with kernels of 3×3 and 4×4, respectively. Finally, we compute the absolute differences between the real and fake output values to get the images similarity loss $\mathcal{L}_{\mathrm{img}}$.

For the similarity of real and fake foreground objects, we first use YOLOv3 [16] as an object detector to detect N foreground objects in the real images, where N is a variable and $N \geq 1$. We also use the Non-max suppression technique to merge the redundant detected positions. We then extract the object region features from the fake images using the same positions as detected in the real images (dashed orange boxes in Fig. 3). For each real and fake object region, we extract a feature of 255 dimensions by YOLOv3 and use an average pooling layer to reduce the feature dimension to a single value. Finally, we compute the mean absolute differences between the real and false feature values to get the real-fake object similarity loss $\mathcal{L}_{\mathrm{obj}}$.

In conclusion, the overall objective loss functions for the training of the two-way discriminators and the generators are:

$$\mathcal{L}_{\mathrm{D}} = \mathcal{L}_{\mathrm{D^c}} + \mathcal{L}_{\mathrm{D^u}} + \lambda_3 \mathcal{L}_{\mathrm{D'^u}} + \lambda_4 \mathcal{L}_{\mathrm{trip}} + \lambda_5 (\mathcal{L}_{\mathrm{c1}} + \mathcal{L}_{\mathrm{c2}}) \tag{13}$$

$$\mathcal{L}_{\mathrm{G}} = \mathcal{L}_{\mathrm{G^c}} + \mathcal{L}_{\mathrm{G^u}} + \lambda_6 \mathcal{L}_{\mathrm{img}} + \lambda_7 \mathcal{L}_{\mathrm{obj}} \tag{14}$$

where $\mathcal{L}_{\mathrm{D^c}}$ is the training loss for D^{c} using the proposed judgments in DSCD, and $\mathcal{L}_{\mathrm{G^c}}$ is the adversarial loss provided by D^{c} to train the generators. (See more details in the supplementary material.)

3. EXPERIMENTS

We use Microsoft COCO 2014 [17] (MSCOCO14) to evaluate the performance of our proposed method and other state-of-the-art methods. We use Inception score [18], *Fréchet inception distance* [19] to evaluate the performance of the experimental methods. Note that lower FID indicates higher image quality and diversity. Our experiments are conducted on the GPU of GeForce RTX 3090Ti with a memory capacity of 24GB.

Fig. 4. Comparison of the resulting images of AttnGAN [1], DMGAN [2], and our improved versions on the MSCOCO14 dataset.

Table 1. Performance of different text-to-image generation methods on two evaluation metrics.

Method	Venue	Year	IS↑	FID↓
AttnGAN [1]	CVPR	2018	25.89	33.10
AttnGAN$^+$ [1]	CVPR	2018	26.43	31.87
DMGAN [2]	CVPR	2019	30.49	32.64
ObjGAN [3]	CVPR	2019	27.37	25.85
AttnGAN+OP [8]	ICLR	2019	24.76	33.35
CPGAN [4]	ECCV	2020	**52.73**	48.87
AttnGAN+CL [20]	arXiv	2021	25.70	23.93
DMGAN+CL [20]	arXiv	2021	33.34	20.79
AttnGAN$^+$+Ours	ICME	2022	31.70	19.05
DMGAN+Ours	ICME	2022	37.23	**18.76**

Table 2. The ablation study results on the performance of different proposed modules.

Method	IS↑	FID↓
AttnGAN$^+$(Baseline)	26.43	31.87
Baseline+CSGP	28.62	26.35
Baseline+DSCD+CSGP	29.97	23.83
Baseline+DSCD+CSGP+MLIS	31.70	19.05

3.1. Quantitative and qualitative evaluations

We first improve the performance of AttnGAN [1] to AttnGAN$^+$ by using transformer as the text encoder. Then we apply all our proposed modifications of discriminators to the baseline methods of AttnGAN$^+$ and DMGAN [2]. Results in Table 1 demonstrate that our proposed modifications can significantly improve the performance of the baseline methods to relative high levels among the state-of-the-art methods. We improve the IS scores of the two baseline methods by 19.93% and 22.10%. The FID scores of the two baseline methods are reduced by 63.43% and 73.98%. According to the principles of IS and FID, IS does not penalize the intra-class mode collapse of the generated images, while FID is more sensitive to the diversity of the generated images. Therefore, we think FID can better reflect the diversity of the fake images and real-fake images similarity. Although our method does not reach the highest IS as CPGAN [4], we have much better performance on the more reliable FID score. Moreover, our improved version of DMGAN reaches the best FID score than all the comparative methods.

From the qualitative results in Fig. 4 we can see that our improved methods can generate more realistic and reasonable images than the baseline methods. Please refer to the supplementary material for the qualitative comparisons between some state-of-the-art and our improved methods.

3.2. Ablation study

We conduct the ablation studies on the validation set of MSCOCO14 to show the effectiveness of each proposed modification. Our baseline method is our improved version of AttnGAN [1] which uses three stacked GANs to generate fake images with the final size of $256 \times 256 \times 3$. Starting with the baseline method AttnGAN$^+$, we add one of the three proposed modifications of CSGP, DSCD, and MLIS to the previous method at each time. The quantitative results are shown in Table 2.

Compared with the baseline method, the proposed CSGP-based unconditional discriminator gains the maximum improvements on IS and FID scores. The CSGP strategy improves the IS by 2.19 points and reduces the FID by 5.52 points. The enhancement on IS verifies that the contrastive searching method can better evaluate the realism of the fake images. Meanwhile, the enhancement on FID proves that the proposed gradient penalty can help relieve the mode collapse caused by gradient instability. Adding the DSCD to the conditional discriminator further improves the IS by 1.35 points and reduces the FID by 2.52 points. The results prove that penalizing the combination of fake images and the mismatched

texts can improve the diversity of the generated images. Finally, adding the MLIS to the experimental model reduces the FID by 4.78 points. This result demonstrates that the proposed MLIS can effectively promote the similarity between real and fake images and objects, and the trained discriminator can correctly reflect the differences between real and fake images.

4. CONCLUSIONS

In this paper, we focus on improving the discriminators in GAN for text-to-image generation task. First, we propose diversity-sensitive conditional discriminator to alleviate the mode collapse of the fake images. Then we propose a contrastive searching gradient penalty strategy for the unconditional discriminator to judge the realism of fake images and stabilize GAN's training. Finally, we propose a multi-level images similarity loss to promote the real and fake similarity at both image and object levels. The experimental results demonstrate the proposed three modifications can promote GAN to generate better scene images.

5. REFERENCES

[1] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2018, pp. 1316–1324.

[2] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang, "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2019, pp. 5802–5810.

[3] 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/CVF Conference on CVPR*, 2019, pp. 12174–12182.

[4] Jiadong Liang, Wenjie Pei, and Feng Lu, "Cp-gan: Content-parsing generative adversarial networks for text-to-image synthesis," in *ECCV*. Springer, 2020, pp. 491–508.

[5] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee, "Generative adversarial text to image synthesis," in *ICML*. PMLR, 2016, pp. 1060–1069.

[6] Zizhao Zhang, Yuanpu Xie, and Lin Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2018, pp. 6199–6208.

[7] 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 transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.

[8] T. Hinz, S. Heinrich, and S. Wermter, "Generating multiple objects at spatially distinct locations," in *ICLR*, 2019.

[9] Jae Hyun Lim and Jong Chul Ye, "Geometric gan," *arXiv preprint arXiv:1705.02894*, 2017.

[10] Tao Hu, Chengjiang Long, and Chunxia Xiao, "A novel visual representation on text using diverse conditional gan for visual recognition," *IEEE TIP*, vol. 30, pp. 3499–3512, 2021.

[11] Song Tao and Jia Wang, "Alleviation of gradient exploding in gans: Fake can be real," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2020, pp. 1191–1200.

[12] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang, "Cross-modal contrastive learning for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2021, pp. 833–842.

[13] Xinzhi Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao, "Dual graph convolutional networks with transformer and curriculum learning for image captioning," in *Proceedings of the 29th ACM MM*, 2021, pp. 2615–2624.

[14] Fei Fang, Miao Yi, Hui Feng, Shenghong Hu, and Chunxia Xiao, "Narrative collage of image collections by scene graph recombination," *IEEE TVCG*, vol. 24, no. 9, pp. 2559–2572, 2017.

[15] Fei Fang, Fei Luo, Hong-Pan Zhang, Hua-Jian Zhou, Alix LH Chow, and Chun-Xia Xiao, "A comprehensive pipeline for complex text-to-image synthesis," *JCST*, vol. 35, no. 3, pp. 522–537, 2020.

[16] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[18] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," *arXiv preprint arXiv:1606.03498*, 2016.

[19] 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," *arXiv preprint arXiv:1706.08500*, 2017.

[20] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji, "Improving text-to-image synthesis using contrastive learning," *arXiv preprint arXiv:2107.02423*, 2021.