TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 13/24 pp469–480
DOI: 10.26599/TST.2023.9010023
Volume 29, Number 2, April 2024

Feature-Grounded Single-Stage Text-to-Image Generation

Yuan Zhou*, Peng Wang, Lei Xiang, and Haofeng Zhang

Abstract: Recently, Generative Adversarial Networks (GANs) have become the mainstream text-to-image (T2I) framework. However, a standard normal distribution noise of inputs cannot provide sufficient information to synthesize an image that approaches the ground-truth image distribution. Moreover, the multistage generation strategy results in complex T2I applications. Therefore, this study proposes a novel feature-grounded single-stage T2I model, which considers the "real" distribution learned from training images as one input and introduces a worst-case-optimized similarity measure into the loss function to enhance the model's generation capacity. Experimental results on two benchmark datasets demonstrate the competitive performance of the proposed model in terms of the Frechet inception distance and inception score compared to those of some classical and state-of-the-art models, showing the improved similarities among the generated image, text, and ground truth.

Key words: text-to-image (T2I); feature-grounded; single-stage generation; Generative Adversarial Network (GAN)

1 Introduction

Text-to-image (T2I) synthesis, which considers text description as input and outputs an image with high semantic relevance corresponding to the description, connects natural language with computer vision, thereby promoting artificial intelligence in "looking" and "understanding". Automatic image generation from text descriptions has attracted considerable interest owing to its importance in many applications[1–6], such as generating portraits based on the appearance description[7], designing desired images with a given style label[8], synthesizing unseen features based on the class description in zero-shot learning[9]. Depending on whether Generative Adversarial Networks (GANs)[10] are used as the main framework, T2I methods can be roughly divided into two categories: non-GAN-based and GAN-based models.

In the early stage, images are generated using a set of simple phrases through spatial reasoning[11–15]. Zhu et al.[16] proposed a synthesis system to augment communication. This system first identifies and "picturizes" text units, and then searches for the most likely image parts conditioned on the text; the system finally optimizes the picture layout based on the text and image parts. Therefore, the non-GAN-based methods spatially arrange images through the correlations between source image parts and handcrafted keywords or key phrases. The major limitation of traditional learning based T2I synthesis approaches is that they cannot generate new image content; they can only change the characteristics of the training images[17]. Alternatively, the Variational AutoEncoder (VAE)[18] is used in T2I to create new visual content. alignDRAW[19] introduces recurrence to VAE to paint an image in multiple steps while checking relevant words in the description. Attribute-to-image[20] develops a layered generative model, which takes attribute descriptions as inputs and disentangles latent attribute variables using a VAE. However, obtaining realistic results using VAE remains a challenge.

• Yuan Zhou, Peng Wang, and Lei Xiang are with School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China. E-mail: zhouyuan@nuist.edu.cn; {wpeng633, xl294487391}@gmail.com.
• Haofeng Zhang is with School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. E-mail: zhanghf@njust.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2022-11-16; revised: 2023-03-22; accepted: 2023-03-26

By contrast, GANs and conditional GANs[21] are explicitly trained to generate the most plausible and realistic images based on text, which are difficult to distinguish from real data[22]. Therefore, since Reed et al.[23] proposed a simple and effective GAN architecture and training strategy in 2016, numerous GAN-based T2I models have sprung up and become mainstream[24–28].

GAN-INT-CLS[23], as a baseline, pursues generation of images semantically matched to the texts by conditioning the text embedding. Then, to obtain high-quality images that are semantically matched to the text, StackGAN[29] synthesizes coarse images in the first stage and then generates high-resolution images based on the coarse images in the second stage. Furthermore, AttnGAN[30] synthesizes fine-grained details at different subregions by focusing on relevant words in the natural language description. Moreover, Mirro-GAN[2] enhances the diversity and semantic consistency of the generated images in a text-image-text cycle manner. MA-GAN[28], XMC-GAN[31], and PCCM-GAN[5] introduce contrastive learning to T2I, which guarantees that the generated images are consistent with the text and have diversity.

Diffusion models have emerged as a promising generative framework, transforming T2I models from GAN-based to diffusion-based. GLIDE[32], DALL·E2[33], and Imagen[34] proposed the following similar ideas: a forward diffusion process generates image embeddings given a text caption, and a reverse diffusion process generates images conditioned on the image embedding and text caption. Although the diffusion-based model achieves high-fidelity image generation, the number of parameters is enormous. For example, GLIDE and DALL·E2 have 3.5 billion parameters, and Imagen has 4.6 billion. These models with enormous parameters are outside the scope of this study; we focus on GAN-based models and leave diffusion-based models for our future research.

The multistage or cycle training strategy of previous GAN-based works causes difficulty in applying T2I in the real world. The old saying goes, "A picture is worth a thousand words". These GAN-based models consider the text with a random sample from a standard normal distribution as input, and the text and random noise cannot cover adequate information compared with the ground-truth images. Therefore, using a sample from the standard normal distribution for generation is not the best choice[35]. Moreover, these models consider semantic matching between generated images and text but neglect consistency with the ground-truth images.

In this paper, we propose a feature-grounded single-stage model named FGSS-GAN. FGSS-GAN uses an image distribution encoder to generate a vector complementing a text, providing more information to the generator than a sample from a standard normal distribution. Furthermore, FGSS-GAN can generate plausible and reliable images by introducing a triple similarity measure into the objective function. Our contributions are summarized as follows:

- We propose a single-stage T2I framework that promotes generation of high-quality images while improving the consistency of images, text, and ground truth.
- We use a distribution encoder to approximate the distribution of the real image, which guarantees that the generated images obey the proper image distribution.
- We propose a similarity comparator that introduces a worst-case-optimized similarity loss to the objective function. This comparator reduces the difference between the generated images and the ground truth in semantic and visual spaces, thus ensuring that the generated images are in line with the textual description and the ground-truth images.
- We conducte extensive experiments and detailed analysis on two benchmark datasets, and experimental results showed the competitive performance of our single-stage model.

2 Related Work

The GAN-INT-CLS can synthesize 64 pixel × 64 pixel × 3-channel image conditioning on the text, which is the first employment of GAN in the T2I task and a milestone for the GAN-based T2I model. Similar to the naive GAN, GAN-INT-CLS has two components: the generator and the discriminator. The model first encodes the text using a text encoder and samples a noise from a standard normal distribution. Then, it concatenates text embedding and the noise and feeds these to the generator. Next, the discriminator takes image and text embedding as inputs to judge whether the image-text pair is genuine. The generator attempts to generate an image conditioning on the text to fool the discriminator. Furthermore, the discriminator attempts to distinguish whether the image and the text are a pair. Therefore, the model is optimized in an adversarial manner. To enhance the consistency between the synthesized image and the correct text, the discriminator in GAN-INT-CLS must recognize the fake image with the correct text and

the true image with the mismatched text. Meanwhile, Scott et al.[36] proposed the Generative Adversarial What-Where Network (GAWWN), which synthesizes 128 pixel × 128 pixel × 3-channel images conditioned on informal text descriptions and object location, which is similar to the current attention mechanism to a certain extent.

Aiming to generate high-quality images, Zhang et al.[29] proposed the multistage model StackGAN. StackGAN yields 64 pixel × 64 pixel × 3-channel images based on the given text description in Stage I and then inputs the low-resolution images and text descriptions into the GAN in Stage II to generate 256 pixel × 256 pixel × 3-channel images with photo-realistic details. The StackGAN++[25] still uses a multistage training strategy, which consists of multiple generators and discriminators arranged in a tree-like structure; images at multiple scales corresponding to the same scene are generated using different tree branches. Similar to the StackGAN family, HDGAN[37] proposes a hierarchical GAN framework accompanying hierarchical-nested adversarial objectives, applies adversarial training on different scales, and assists generator training in capturing complex image statistics. Based on multistage or multiscale refinement generation methods, Xu et al.[30] proposed the AttnGAN, which can synthesize fine-grained details at different image subregions by focusing on relevant words in the natural language description. AttnGAN synthesizes images in a multistage way. It has three generators and discriminators on different scales, and each stage has adversarial training. The image features of each stage combined with the corresponding word-context features are used to generate images at the next stage. Furthermore, a Deep Attentional Multimodal Similarity Model (DAMSM) was proposed to enforce fine-grained image-text matching. DF-GAN[38] is a single-stage AttnGAN. Moreover, to maintain the image-text semantic consistency and refine the synthesis quality, Qiao et al.[2] proposed MirroGAN, which has a cascaded attention-GAN architecture for generating target images from coarse to fine scales in a cyclic manner. MirroGAN regenerates text description from the generated image, which semantically aligns with the given text description.

Different from methods using single text descriptions, RiFeGAN[39], MA-GAN[28], and DAE-GAN[40] use auxiliary information to generate images from text and are also multistage models. RiFeGAN utilizes caption matching to enrich descriptions, exploits self-attentional embedding mixtures to extract features from multicaptions in an attentional framework, and finally synthesizes images using those features. Instead of using only one sentence in every pair, MA-GAN uses multiple sentences in the single-sentence generation and multisentence discrimination module, which contains three paired generators and discriminators corresponding to the generation of images with different resolutions. Ruan et al.[40] proposed the DAE-GAN, in which the "aspect" information in the text is often ignored but highly helpful for synthesis of image details. DAE-GAN represents text information comprehensively from multiple granularities, including sentence, word, and aspect levels. Then, at the two-stage generation, it generates a low-resolution image with sentence-level embedding at the initial stage. Next, at the refinement stage, viewing aspect-level features as central vision and word-level features as peripheral vision, it utilizes word-level embedding to enhance the previously generated images globally. Then, DAE-GAN dynamically utilizes aspect-level embedding to refine image details from a local perspective.

However, most of the methods mentioned above only consider text-image consistency and ignor the consistency of the image with the grouad truth. Furthermore, although the multistage framework obtains plausible results, it makes the training complex and cumbersome for application. Thus, we aim to find a simple T2I model to synthesize plausible images corresponding to the semantic description and ground-truth visual content. To this end, we propose a novel feature-grounded single-stage T2I synthesis method based on the attention mechanism, which considers the consistency among text, image, and ground-truth image.

3 Proposed Method

As illustrated in Fig. 1, our proposed FGSS-GAN consists of three main modules: (1) Encoders: the text encoder E_T extracts sentence- and word-level features from the text. The image encoder E_I extracts image features that align with the text features. Then, the distribution encoder E_z encodes the image features to a Gaussian vector z, augmenting the visual features that are the main focus of this study. (2) Generator and discriminator: generator G takes the sentence feature and the augmented vector as input, and generates images using word features to refine the generation; and discriminator D distinguishes whether the images are true or false. (3) Similarity comparator C aligns the

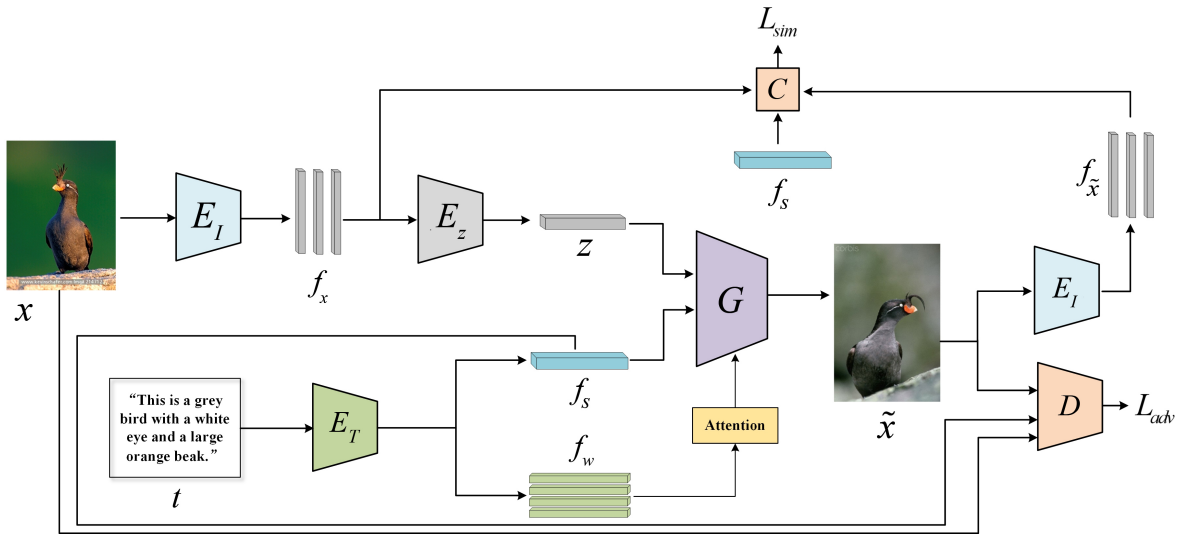

Fig. 1 Framework of the proposed feature-grounded single-stage T2I generation method. E_I is the image encoder, E_z is the real image distribution encoder, and C is the similarity comparator. The generator G considers the augmented vector z and the sentence feature f_s as inputs. The discriminator D judges the synthesized image $\tilde{x}$ and the real image x conditioned on f_s.

visual and the semantic features.

3.1 Encoders

(1) Text and image encoders

FGSS-GAN uses a pretrained text encoder E_T in CLIP[41] to encode the text description t into a sentence-level feature $f_s \in \mathbf{R}^{512}$ and a word-level feature $f_w \in \mathbf{R}^{512\times seq\_len}$, where seq_len is the length of a sentence. Then, we define

$$(f_s, f_w) = E_T(t) \tag{1}$$

FGSS-GAN uses a pretrained ViT-B/32 as the image encoder E_I, which maps the ground-truth image x into an image feature $f_x \in \mathbf{R}^{512}$,

$$f_x = E_I(x) \tag{2}$$

The text and image encoders are fixed when training.

(2) Distribution encoder

Different images contain different visual information providing more information than a standard normal distribution noise. Therefore, considering visual embedding as the input rather than a standard normal noise is reasonable. The FGSS-GAN first extracts the training image features f_x using the image encoder E_I, and then uses the distribution encoder E_z to map the visual features f_x to a continuous manifold represented by an independent Gaussian distribution $\mathcal{N}(\mu(f_x), \sigma(f_x))$ in the same manner as the VAE. Furthermore, regularization is added to E_z to ensure the smoothness of this manifold,

$$D_{KL}(\mathcal{N}(\mu(f_x), \sigma(f_x))||\mathcal{N}(0, I)) \tag{3}$$

A sample from this manifold provides visual information for generation, and we call it the augmented vector z,

$$z = \mu(f_x) + \epsilon \times \sigma(f_x) \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, 1)$.

3.2 Generator and discriminator

Our FGSS-GAN uses BigGAN-Deep[42] as the backbone based on a previous study[27]; generator G and discriminator D have the same architecture. Unlike prior GAN-based models, the generator of FGSS-GAN uses sentence embedding with the augmented vector z as the input. Furthermore, the model computes the attention based on the word embedding f_w and the feature map from the penultimate layer of the generator, and applies attention to the feature map. Then, the feature map is sent to the last layer of the generator to refine the generated image.

The generator aims to confuse the discriminator to consider the fake image as a real one. However, the discriminator attempts to distinguish the ground-truth image and the text as a real pair and the pair of the generated image with the given text as fake. Therefore, the adversarial loss $\mathcal{L}_{adv}$ is as follows:

$$\begin{aligned}\mathcal{L}_{adv} =& E_{(x,t)\sim p_{data}} \log[D(x, f_s)] + \\ & E_{(x,t)\sim p_{data}} \log[1 - D(\tilde{x}, f_s)]\end{aligned} \tag{5}$$

where $D()$ denotes the output of the discriminator.

The generator and discriminator are optimized by alternatively maximizing $\mathcal{L}_{adv}$ and minimizing $\mathcal{L}_{adv}$.

3.3 Similarity comparator

FGSS-GAN has a similarity comparator to measure

the pariwise similarity of the ground-truth image x, generated image $\tilde{x}$, and text t. Feature vectors in the feature space can depict the i-th image's normalized semantic features f_{s_i} and visual features (f_{x_i} and $f_{\tilde{x}_i}$). The three feature vectors should coincide and have the same orientation if the generated image is consistent with the text and ground-truth image. To this end, their three distances should be small. We define the distance between a and b as $dist(a,b)$, which equals to $1 - cos\_dist(a,b)$, where $cos\_dist(a,b)$ is the cosine distance between a and b. A large $dist(a,b)$ indicates a small similarity between a and b. Therefore, $dist(\tilde{x}_i, t_i)$ and $dist(x_i, \tilde{x}_i)$ are the distances from the i-th generated image $\tilde{x}_i$ to the corresponding text t_i and ground truth x_i , respectively. We use $dist(\tilde{x}_i, t_i)$ and $dist(x_i, \tilde{x}_i)$ to denote the mean distances of a batch,

$$d_{x\tilde{x}} = \frac{1}{n}\sum_{i=1}^{n} dist(x_i, \tilde{x}_i),$$
$$d_{\tilde{x}t} = \frac{1}{n}\sum_{i=1}^{n} dist(\tilde{x}_i, t_i) \quad (6)$$

where n is the batch size.

Encoders E_T and E_I are fixed while training, and the ground-truth image and text pair guarantees that their feature vectors are close to each other. Therefore, we only consider the worst case of the two distances in Eq. (6), which means the largest distance. With the supervision of the ground-truth image and the text, minimizing the worst case pushes the generated image to approach the text and ground-truth image in feature space. The similarity comparison loss is expressed as follows:

$$\mathcal{L}_{sim} = \max\{d_{x\tilde{x}}, d_{\tilde{x}t}\} \quad (7)$$

3.4 Objective function

The overall objective function is defined to generate a photo-realistic image aligned with the text description and ground-truth image,

$$\min_{G,E_z} \max_{D} \lambda_{adv}\mathcal{L}_{adv} + \lambda_{sim}\mathcal{L}_{sim} + D_{KL} \quad (8)$$

where λ_{adv} and λ_{sim} are the hyper-parameters to control the impact of loss terms. Here, we set them at 0.5.

The distribution encoder and generator are optimized by minimizing Formula (8), and the discriminator is optimized by maximizing Formula (8).

4 Experiment

To validate the proposed FGSS-GAN, we conduct extensive experiments on two benchmark datasets: Caltech-UCSD Birds (CUB-200)[43] and MS COCO (COCO)[44]. We compare our methods with classical and state-of-the-art T2I methods, and perform quantitative and qualitative analyses.

4.1 Datasets and evaluation metrics

(1) Datasets

The CUB-200 dataset contains 11 788 images of 200 subcategories belonging to birds, and each image has ten sentence descriptions. A total of 5994 images are used for training and 5794 for testing in CUB-200. The COCO dataset is a large-scale dataset with 82 783 natural images for training and 40 504 images for validation, and each image corresponds to no less than five caption descriptions. Moreover, each COCO image contains multiple objects and various backgrounds, making it more challenging for T2I generation.

(2) Evaluation metrics

Following previous works, we use Inception Score (IS)[45] and Frechet Inception Distance (FID)[46] to evaluate the generation performance, and cosine similarity to evaluate the text-image consistency. The lower the FID, the better the evaluation determining whether the generated and the ground-truth images follow the same distribution. In addition, IS is used to evaluate whether the generator can synthesize diverse and meaningful images. A higher IS value means excellent performance. Moreover, a large cosine similarity means high text and image alignment.

4.2 Implementation details

During training, we used Adam optimization with $\beta_1 = 0$ and $\beta_2 = 0.99$. For CUB-200, we set the batch size to 32, and the learning rates of G and D are 1×10^{-4} and 4×10^{-4}, respectively. Moreover, the batch size for COCO is 16, and the learning rates of G and D are 5×10^{-5} and 2×10^{-4}, respectively. The input and output image size is 256 pixel × 256 pixel × 3-channel.

Empirically, we set the network widths of G and D as 96. Aiming to refine the generation performance, the generator G of FGSS-GAN pays attention to the word-level semantic feature. We then determine the layer of the generator that should consider the attention mechanism. We conduct experiments on CUB-200, with a sole focus on the generator of BigGAN. Figure 2 shows that introducing attention to the 13-th layer of G obtains the best FID and IS for the best performance. FGSS-GAN is trained on NVIDIA RTX 3090 GPU for 700 epochs on the CUB-200 dataset and 500 epochs on the COCO dataset.

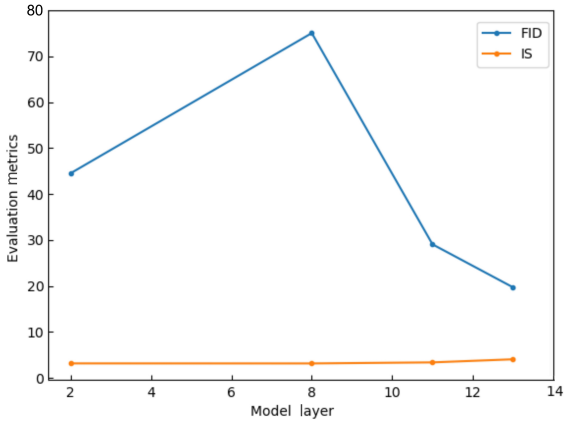

Fig. 2 Comparative results of introducing attention to different layers of the generator. When the attention mechanism worked on the 13-th layer, FGSS-GAN obtains the best performance under the FID and IS metrics.

4.3 Comparison analysis

We compare our FGSS-GAN method with nine other classical and state-of-the-art methods. Table 1 shows the network structures and training strategies. Multistage models have multiple generators and discriminators adopted in different stages. By contrast, single-stage models, such as the proposed FGSS-GAN, only has one generator and one discriminator, making training more convenient.

Table 1 Comparisons of model structures and training strategies. *#G* denotes the number of generators, and *#D* denotes the number of discriminators.

Method	*#G*	*#D*	Multi-stage
GAN-INT-CLS[23]	1	1	No
GAWWN[36]	1	1	No
StackGAN[29]	2	2	Yes
StackGAN++[25]	3	3	Yes
HDGAN[37]	3	3	Yes
AttnGAN[30]	3	3	Yes
DF-GAN[38]	1	1	No
MA-GAN[28]	3	3	Yes
DAE-GAN[40]	3	3	Yes
FGSS-GAN (our proposed)	1	1	No

4.3.1 Quantitive results

We compared our FGSS-GAN with the nine T2I methods using the CUB-200 and COCO datasets, and Table 2 shows the comparison results in terms of FID and IS. FGSS-GAN outperformed the other three single-stage methods with higher FID and IS. Therefore, we only compare the similarities between the proposed FGSS-GAN and the three state-of-the-art methods (Table 3).

Table 2 shows that for CUB-200, FGSS-GAN achieves an FID of 19.08 with the third performance, and the top two FIDs are obtained by multistage models DAE-GAN and StackGAN++. Furthermore, FGSS-GAN surpasses all other methods with the highest IS of 4.79, which is higher than that achieved by the state-of-the-art DAE-GAN by 8.37%. Notably, the value of IS of FGSS-GAN is only 0.03, which means the FGSS-GAN shows high robustness with respect to CUB-200. Meanwhile, for COCO, FGSS-GAN decreases the best-reported FID of DAE-GAN from 28.12 to 27.89, attaining a 21.41% improvement compared to AttnGAN. FGSS-GAN obtains a lower IS on COCO, and the value is poorer than those of DAE-GAN and AttnGAN. However, FGSS-GAN still performs better than StackGAN and StackGAN++. Table 3 shows that the generated image using FGSS-GAN has the highest similarity with the ground truth and slightly lower similarity with the text compared with the state-of-the-art DAE-GAN.

To our knowledge, CUB is a fine-grained dataset

Table 2 Comparison results of different methods with respect to CUB and COCO datasets. "↓" means the smaller value, the better, and "↑" means the larger value, the better. A method name with "⁎" indicates a single-stage model; otherwise, it is a multistage model.

Method	FID ↓		IS ↑	
	CUB-200	MS-COCO	CUB-200	MS-COCO
GAN-INT-CLS[23]⁎	–	–	2.88±0.04	7.88±0.07
GAWWN[36]⁎	–	–	3.60±0.07	–
StackGAN[29]	55.28	74.05	3.70±0.04	8.45±0.03
StackGAN++[25]	15.30	81.59	4.04±0.06	8.30±0.10
HDGAN[37]	–	–	4.15±0.05	11.86±0.18
AttnGAN[30]	23.98	35.49	4.36±0.03	23.87±0.42
DF-GAN[38]⁎	19.24	28.92	4.56±0.04.	–
MA-GAN[28]	21.66	–	4.76±0.09	–
DAE-GAN[40]	**15.19**	28.12	4.42±0.04	**34.97±0.84**
FGSS-GAN (our proposed) ⁎	19.08	**27.89**	**4.79±0.03**	14.75±0.32

Table 3 Similarity comparison results of different methods. $(\tilde{x}, x)$ and $(\tilde{x}, t)$ represent the similarity between the generated image $\tilde{x}$ and ground-truth image x, and between the generated images $\tilde{x}$ and the text t, respectively.

Method	CUB-200		MS-COCO	
	$(\tilde{x}, x)$	$(\tilde{x}, t)$	$(\tilde{x}, x)$	$(\tilde{x}, t)$
AttnGAN[30]	0.041	0.027	0.003	0.037
DF-GAN[38]	0.102	0.087	0.106	0.006
DAE-GAN[40]	0.267	**0.234**	0.141	**0.219**
FGSS-GAN	**0.531**	0.181	**0.476**	0.201

with detailed descriptions, and COCO is a dataset with multiple objects and complex layouts. Therefore, the noise encoder in FGSS-GAN encodes more informative noise using CUB training images than standard normal distribution noise. Furthermore, FGSS-GAN can synthesize diverse CUB-200 images based on augmented noise, which leads to achieving the highest IS value with respect to CUB-200. Moreover, FGSS-GAN considers the similarity between the generated image, text, and ground-truth image in the loss function. Thus, when synthesizing an image, FGSS-GAN attempts to find the "best" position where the distance between the synthesized image, ground-truth image, and text is optimal. FGSS-GAN makes the synthesized COCO image consistent with the real image in terms of FID. However, DAE-GAN and MA-GAN achieve better FID on CUB-200 and IS on COCO than FGSS-GAN because these methods use auxiliary information, such as aspect-level information or multiple sentences complementing the image details, to generate images.

Therefore, our method FGSS-GAN realizes a compromise between text and the ground-truth image, achieves comparable performance with multistage T2I models, and outperforms classical single-stage models.

4.3.2 Qualitative results

We evaluate the visual quality of generated images. Figures 3 and 4 show the synthesized images using our method FGSS-GAN and the other T2I methods. Because FGSS-GAN has a noise encoder and a similarity comparator, it synthesizes vivid images consistent with the text descriptions, especially performing well on details.

For example, in Fig. 3, all images show good quality at first glance. However, not every method can generate details corresponding to the text description. For example, in the first two columns of the left part in Fig. 3, "stubby beak" and "webbed feet" do not appear in every image from the four models, but they are presented in all the synthesized images of FGSS-GAN. In the first column of the right part in Fig. 3, the text description is "This black bird has a large, lighter beak that is curved...". Only FGSS-GAN can synthesize "a large, lighter beak that is curved". Furthermore, in the right part of Fig. 3, "the grey eyebrow" in the third column and "black strips

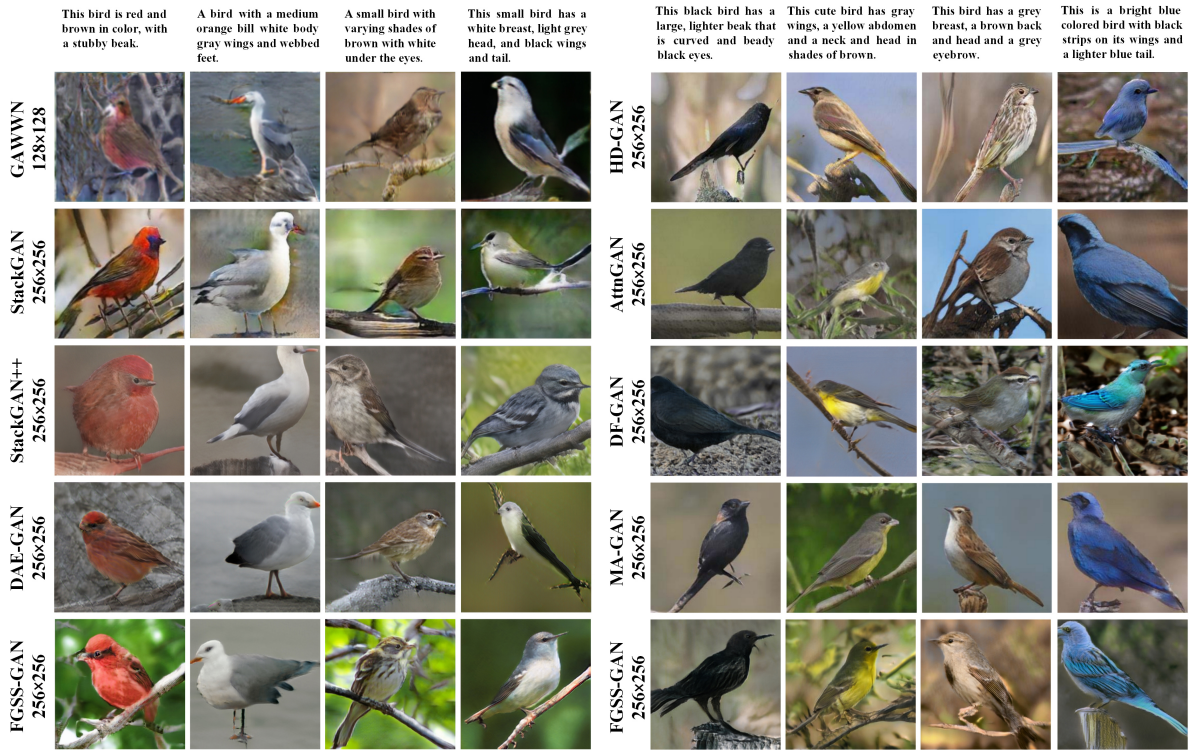

Fig. 3 Synthesized CUB-200 images using different T2I models.

Fig. 4 Synthesized COCO images from different T2I models.

on the wings and a lighter blue tail" in the fourth column only appeare in the FGSS-GAN synthesized images. Moreover, the image background of FGSS-GAN is not only blurred but also contains more information, such as branches, leaves, etc., owing to the noise augmentation feature in FGSS-GAN.

We can still observe the visual details consistent with the semantics in the challenging COCO dataset. In Fig. 4, the details are presented clearly in the images using FGSS-GAN. e.g., only FGSS-GAN presents "a group of people on skis" in the second column and "the middle of a day" in the fourth column. Meanwhile, "the middle of a day" is the background for an image, and the noise augmentation mechanism of FGSS-GAN guarantees that the synthesized image contains various backgrounds. FGSS-GAN also presents "two people in a speed boat" in the fifth column and "a family standing in front of a sign wearing skis and holding ski poles" in the sixth column. Particularly, the seventh column shows an image of fruits and vegetables. FGSS-GAN can synthesize colorful fruits and vegetables, unlike other models that only generate green and yellow objects.

4.4 Ablation study

To verify the effect of the distribution encoder and the similarity comparator, we conduct ablation studies on CUB-200 and COCO, and the quantitative results are shown in Table 4.

The distribution encoder and similarity comparator can improve the alignment between generated images, ground-truth images, and texts. Moreover, using CUB-200 can improve the model performance, that is, increasing FID by 2.74% and IS by 5.65% over the baseline. The similarity comparator can improve the FID by 0.54% and IS by 11.39%. Furthermore, for COCO, the distribution encoder can improve the performance by 7.97% and 2.89% in FID and IS over the baseline, respectively. Meanwhile, the similarity comparator can improve the performance by 13.46% and 1.03% in FID and IS, respectively. Therefore, the similarity comparator substantially influences diverse and meaningful generations for fine-grained datasets, such as CUB-200. The noise augmentation mechanism plays a vital role in synthesizing realistic images for datasets with multiple objects and complex layouts, such as COCO. This conclusion is consistent with the comparison results in Table 2. This explains the comparison results from another perspective, proving that the generated images of FGSS-GAN are compromises between texts and real images.

Figure 5 visualizes the generated CUB-200 images under different ablation settings, showing the positive

Table 4 Ablation results of FID and IS. E_z is the distribution encoder, and C is the similarity comparator. $(\tilde{x}, x)$ and $(\tilde{x}, t)$ represent the similarity between the generated image $\tilde{x}$ and ground-truth image x and between the generated images $\tilde{x}$ and text t, respectively.

Method	CUB-200				MS-COCO			
	FID↓	IS↑	$(\tilde{x}, x)$	$(\tilde{x}, t)$	FID↓	IS↑	$(\tilde{x}, x)$	$(\tilde{x}, t)$
FGSS-GAN (without E_Z and C)	19.728	4.07± 0.11	0.026	0.015	35.26	14.19± 0.19	0.026	0.082
FGSS-GAN (without C)	19.186	4.30± 0.16	0.085	0.052	32.45	14.60± 0.25	0.057	0.093
FGSS-GAN	**19.082**	**4.79± 0.03**	**0.531**	**0.181**	**28.08**	**14.75± 0.32**	**0.476**	**0.201**

Fig. 5 Effects of the distribution encoder E_z and similarity comparator C on the CUB dataset. The baseline is FGSS-GAN without E_z and C.

effect of the distribution encoder and similarity comparator. The baseline images exhibit neither background nor blurring. After the adding the distribution encoder, the backgrounds details, such as the appearance of branches, leaves, and stones, are considerably improved. Then, the similarity comparator guarantees the generation of high-quality and sharp images. We further investigate the effect of distribution encoders on image generation, and Fig. 6 shows the images generated with the same text and different real images. The noises from various real images encoded by the noise encoder contain more information, such as the background and the relation between the object and the surrounding environment. For example, a bird standing on a branch in a natural image makes a bird do the same action in the synthesized image.

5 Conclusion

This study argues that real images contain more information than a standard normal distribution for the T2I task. We employ a real image distribution encoder that extracts the informative noises from the training images as the model input. Moreover, we assume that the synthesized images are consistent with the text description and ground-truth image. Therefore, we use a similarity comparator to introduce a worst-case-optimized similarity to the objective function, which guarantees the alignment of visual and semantic features. With these two components, we propose a novel feature-grounded single-stage T2I model FGSS-GAN, which achieves the balance between visual and semantic perspectives. Extensive experiments demonstrated that the proposed model has a competitive performance relative to multistage models and is substantially better than the performance of single-stage models.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61872187).

References

[1] X. Wu, K. Xu, and P. Hall, A survey of image synthesis and editing with generative adversarial networks, *Tsinghua Science and Technology*, vol. 22, no. 6, pp. 660–674, 2017.

[2] T. Qiao, J. Zhang, D. Xu, and D. Tao, MirrorGAN: Learning text-to-image generation by redescription, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 1505–1514.

[3] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, Semantics disentangling for text-to-image generation, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 2322–2331.

[4] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, et al., CogView: Mastering text-to-image generation via transformers, arXiv preprint arXiv: 2105.13290, 2021.

[5] Z. Qi, J. Sun, J. Qian, J. Xu, and S. Zhan, PCCM-GAN: Photographic text-to-image generation with pyramid contrastive consistency model, *Neurocomputing*, vol. 449, pp. 330–341, 2021.

[6] Z. Zhang and L. Schomaker, DiverGAN: An efficient and effective single-stage framework for diverse text-to-image generation, *Neurocomputing*, vol. 473, pp. 182–198, 2022.

[7] W. Xia, Y. Yang, J. H. Xue, and B. Wu, TediGAN: Text-guided diverse face image generation and manipulation, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 2256–2265.

[8] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone, CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms, arXiv preprint arXiv: 1706.07068, 2017.

[9] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, A generative adversarial approach for zero-shot learning from noisy texts, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 1004–1013.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Proc. 27^{th} Int. Conf. Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 2672–2680.

[11] G. Adorni, M. Di Manzo, and F. Giunchiglia, Natural language driven image generation, in *Proc. 10^{th} Int. Conf. Computational Linguistics* and 22^{nd} Annu. Meeting of the Association for Computational Linguistics, Stanford, CA, USA, 1984, pp. 495–500.

Fig. 6 Generated images using FGSS-GAN based on the same text description and different real images.

[12] A. Yamada, T. Yamamoto, H. Ikeda, T. Nishida, and S. Doshita, Reconstructing spatial image from natural language texts, in *Proc. 14th Conf. Computational Linguistics*, Nantes, France, 1992, pp. 1279–1283.

[13] S. R. Clay and J. Wilhelms, Put: Language-based interactive manipulation of objects, *IEEE Comput. Grap. Appl.*, vol. 16, no. 2, pp. 31–39, 1996.

[14] B. Coyne and R. Sproat, WordsEye: An automatic text-to-scene conversion system, in *Proc. 28th Annu. Conf. Computer Graphics and Interactive Techniques*, Los Angeles, CA, USA, 2001, pp. 487–496.

[15] R. Johansson, A. Berglund, M. Danielsson, and P. Nugues, Automatic text-to-scene conversion in the traffic accident domain, in *Proc. 19th Int. Joint Conf. Artificial Intelligence*, Edinburgh, UK, 2005, pp. 1073–1078.

[16] X. Zhu, A. B. Goldberg, M. Eldawy, C. R. Dyer, and B. Strock, A text-to-picture synthesis system for augmenting communication, in *Proc. 22nd National Conf. Artificial Intelligence*, Vancouver, Canada, 2007, pp. 1590–1595.

[17] J. Agnese, J. Herrera, H. Tao, and X. Zhu, A survey and taxonomy of adversarial neural networks for text-to-image synthesis, *WIRs*: *D*ata Mining and Knowledge Discovery, vol. 10, no. 4, p. e1345, 2020.

[18] D. P. Kingma and M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv: 1312.6114, 2022.

[19] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, Generating images from captions with attention, arXiv preprint arXiv: 1511.02793, 2016.

[20] X. Yan, J. Yang, K. Sohn, and H. Lee, Attribute2Image: Conditional image generation from visual attributes, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, the Netherlands, 2016, pp. 776–791.

[21] M. Mirza and S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv: 1411.1784, 2014.

[22] G. Antipov, M. Baccouche, and J. L. Dugelay, Face aging with conditional generative adversarial networks, in *Proc. 2017 IEEE Int. Conf. Image Processing* (*ICIP*), Beijing, China, 2017, pp. 2089–2093.

[23] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, Generative adversarial text to image synthesis, in *Proc. 33rd Int. Conf. Machine Learning*, New York, NY, USA, 2016, pp. 1060–1069.

[24] A. Odena, C. Olah, and J. Shlens, Conditional image synthesis with auxiliary classifier GANs, in *Proc. 34th Int. Conf. Machine Learning*, Sydney, Australia, 2017, pp. 2642–2651.

[25] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, StackGAN++: Realistic image synthesis with stacked generative adversarial networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, 2019.

[26] M. Cha, Y. L. Gwon, and H. T. Kung, Adversarial learning of semantic relevance in text to image synthesis, in *Proc. 33rd AAAI Conf. Artificial Intelligence and 31st Innovative Applications of Artificial Intelligence Conf. and 9th AAAI Symp. Educational Advances in Artificial Intelligence*, Honolulu, HI, USA, 2019, pp. 3272–3279.

[27] D. M. Souza, J. Wehrmann, and D. D. Ruiz, Efficient neural architecture for text-to-image synthesis, in *Proc. 2020 Int. Joint Conf. Neural Networks* (*IJCNN*), Glasgow, UK, 2020, pp. 1–8.

[28] Y. Yang, L. Wang, D. Xie, C. Deng, and D. Tao, Multi-sentence auxiliary adversarial networks for fine-grained text-to-image synthesis, *IEEE Trans. Image Process.*, vol. 30, pp. 2798–2809, 2021.

[29] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks, in *Proc. 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 5908–5916.

[30] 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 *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 1316–1324.

[31] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, Cross-modal contrastive learning for text-to-image generation, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 833–842.

[32] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models, arXiv preprint arXiv: 2112.10741, 2022.

[33] 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.

[34] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al., Photorealistic text-to-image diffusion models with deep language understanding, arXiv preprint arXiv: 2205.11487, 2022.

[35] J. Liu, H. Bai, H. Zhang, and L. Liu, Near-real feature generative network for generalized zero-shot learning, in *Proc. 2021 IEEE Int. Conf. Multimedia and Expo* (*ICME*), Shenzhen, China, 2021, pp. 1–6.

[36] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, Learning what and where to draw, in *Proc. 30th Int. Conf. Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 217–225.

[37] Z. Zhang, Y. Xie, and L. Yang, Photographic text-to-image synthesis with a hierarchically-nested adversarial network, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6199–6208.

[38] M. Tao, H. Tang, S. Wu, N. Sebe, F. Wu, and X. Y. Jing, DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis, arXiv preprint arXiv: 2008.05865, 2022.

[39] J. Cheng, F. Wu, Y. Tian, L. Wang, and D. Tao, RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10908–10917.

[40] 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 *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision*, Montreal, Canada, 2021, pp. 13940–13949.

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in *Proc. 38th Int. Conf. Machine Learning*, Virtual Event, 2021, pp. 8748–8763.

[42] A. Brock, J. Donahue, and K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, arXiv preprint arXiv: 1809.11096, 2019.

[43] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, The CALTECH-UCSD birds-200-2011 dataset, http://www.vision.caltech.edu/datasets/cub_200_2011, 2011.

[44] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, Microsoft COCO: Common objects in context, in *Proc. 13th European Conf. Computer Vision*, Zurich, Switzerland, 2014, pp. 740–755.

[45] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, Improved techniques for training GANs, in *Proc. 30th Int. Conf. Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 2234–2242.

[46] 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, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6629–6640.

Yuan Zhou received the PhD degree from Nanjing University of Aeronautics and Astronautics, China in 2016. From 2017 to 2018, she was an academic visitor at the University of East Anglia, UK. She is currently an associate professor at School of Artificial Intelligence, Nanjing University of Information Science and Technology, China. Her research interests include machine learning and computer vision. She is a member of Chinese Association for Artificial Intelligence (CAAI).

Peng Wang received the BEng degree of electronic science and technology from Ningxia Normal University, China in 2021. He is currently a master student at School of Artificial Intelligence, Nanjing University of Information Science and Technology, China. His research interests include computer vision and image generation.

Lei Xiang received the BEng degree of computer science and technology from Hubei Polytechnic University, China in 2021. He is currently a master student at School of Artificial Intelligence, Nanjing University of Information Science and Technology, China. His research interests include zero-shot learning and domain adaptation.

Haofeng Zhang received the BEng and PhD degrees from Nanjing University of Science and Technology, China in 2003 and 2007, respectively. He is currently a professor at School of Computer Science and Engineering, Nanjing University of Science and Technology, China. From Dec. 2016 to Dec. 2017, he was an academic visitor at University of East Anglia, UK. His research interests include computer vision and mobile robot.