2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00149

DiST-GAN: Distillation-based Semantic Transfer for Text-Guided Face Generation

Guoxing Yang[1] Feifei Fu[1] Nanyi Fei[1] Haoran Wu[2] Ruitao Ma[2] Zhiwu Lu[1]
[1] *Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China*
[2] *China Unicom Research Institute, Beijing, China*
luzhiwu@ruc.edu.cn

***Abstract*—Recently, large-scale pre-training has achieved great success in multi-modal tasks and shown powerful generalization ability due to superior semantic comprehension. In the field of text-to-image synthesis, recent works induce large-scale pre-training with VQ-VAE as a discrete visual tokenizer, which can synthesize realistic images from arbitrary text inputs. However, the quality of images generated by these methods is still inferior to that of images generated by GAN-based methods, especially in some specific domains. To leverage both the superior semantic comprehension of large-scale pre-training models and the powerful ability of GAN-based models in photorealistic image generation, we propose a novel knowledge distillation framework termed DiST-GAN to transfer the semantic knowledge of large-scale visual-language pre-training models (e.g., CLIP) to GAN-based generator for text-guided face image generation. Our DiST-GAN consists of two key components: (1) A new CLIP-based adaptive contrastive loss is devised to ensure the generated images are consistent with the input texts. (2) A language-to-vision (L2V) transformation module is learned to transform token embeddings of each text into an intermediate embedding that is aligned with the image embedding extracted by CLIP. With these two novel components, the semantic knowledge contained in CLIP can thus be transferred to GAN-based generator which preserves the superior ability of photorealistic image generation in the mean time. Extensive results on the Multi-Modal CelebA-HQ dataset show that our DiST-GAN achieves significant improvements over the state-of-the-arts.**

***Index Terms*—Text-to-image generation, Knowledge distillation, Large-scale pre-training**

I. Introduction

Generating images according to the given descriptive sentences is a fundamental and meaningful task, which has significant potential in many real-world applications. Due to these practical applications, text-to-image generation has become an active research area in recent years [1]–[3]. Particularly, with the tremendous success of Generative Adversarial Networks (GANs) [4] in image synthesis, GAN-based methods have dominated the text-to-image generation task [2], [5]–[8]. Although these methods can generate high-quality images, their performance is still limited by the dataset size as well as the model size, i.e., they tend to suffer from insufficient semantic learning and comprehension.

Recently, large-scale pre-training has achieved great success in multi-modal tasks (*e.g.*, image-text retrieval [9], [10]), and shown powerful generalization ability due to superior semantic comprehension. In the field of text-to-image synthesis, recent works [3], [11] induce large-scale pre-training by utilizing VQ-VAE [12] as a discrete visual tokenizer, which can synthesize high-quality images from arbitrary texts and compose unrelated objects in semantically plausible ways. Although these methods show superior performance in generating semantically reasonable images, the quality of generated images is still *inferior* to that of images generated by GAN-based methods, especially in some specific domains (*e.g.*, face). In addition, when generating images from text inputs, these VQ-VAE based methods rely on a large Transformer for predicting image tokens one by one, which leads to significantly higher computational cost than GAN-based methods.

In this paper, we propose a novel distillation-based framework termed DiST-GAN to induce large-scale pre-training into GAN-based generator for text-guided face image generation. Different from VQ-VAE based methods, we choose to exploit the semantic knowledge contained in the large-scale visual-language pre-training models (*e.g.*, CLIP [9]) instead of directly training our model on large-scale datasets. Note that although these pre-training models show superior performance on various downstream tasks without fine-tuning, it is still very challenging to generalize them to text-to-image generation because of the large gap between text-to-image generation and text-image retrieval (widely-used for pre-training). To address this challenge, we devise two key components for our DiST-GAN: (1) A new CLIP-based adaptive contrastive loss is devised to ensure that the generated images are consistent with the input texts. Concretely, an adaptive contrastive loss is defined over generated images and input texts, and the weight of each negative sample is adjusted automatically according to the similarity of the input texts. (2) A language-to-vision (L2V) transformation module is learned to transform token embeddings of each text into an intermediate embedding that is aligned with the image embedding extracted by CLIP. With these two novel components, we can transfer the semantic knowledge contained in CLIP to GAN-based generator through distillation (while the superior ability of GAN-based generator is preserved). Besides, without using large Transformer as the backbone, our DiST-GAN can generate images more efficiently than VQ-VAE based methods.

Our main contributions are three-fold: (1) We propose a simple yet effective distillation-based framework to transfer the semantic knowledge of large-scale visual-language pre-training models to GAN-based generator, namely DiST-GAN, for text-guided face image generation. (2) To better distill

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00149

the semantic knowledge contained in CLIP to GAN-based generator, we devise a novel CLIP-based adaptive contrastive loss and a language-to-vision (L2V) transformation module for our DiST-GAN. (3) Extensive results on the Multi-Modal CelebA-HQ show that our DiST-GAN can generate and high-quality images from texts and achieves new state-of-the-art.

II. Related Work

Text-to-Image Generation. Generating images from given descriptive texts has achieved great progress with the tremendous success of deep generative models [4], [5], [13]. Due to the powerful ability of GAN to generate photo realistic outputs, text-to-image generation has been dominated by GAN-based methods [1], [2], [6], [7], [14], [15]. However, GAN-based methods still suffer from object distortion and illogical object placement in some cases due to the limited dataset/model size. Recently, VQ-VAE [12] based methods [3], [11] have been proposed to introduce large-scale pre-training into text-to-image generation to improve the model performance. However, the quality of generated images is still inferior to that of images generated by GANs, especially in some specific domains (*e.g.*, face). In this work, our proposed DiST-GAN induces large-scale pre-training into GAN-based models to leverage both the superior semantics comprehension of large-scale pre-training models and the powerful ability of GANs in generating photorealistic images for text-guided face generation, a specific domain in text-to-image generation.

Large-Scale Pre-Training. Recently, many works introduce large-scale pre-training to multi-modal task and achieve great success [3], [9]–[11], [16], [17]. The models of these methods are typically trained on large-scale datasets that contain over one hundred million data, which show promising performance with superior semantic comprehension. According to the pre-training task, these methods can be roughly divided into two groups: (1) **Text-Image Retrieval Task**: Methods in this group [9], [10], [17] propose a two-stream framework with contrastive loss to learn the joint text-image embedding space and trains it on a dataset containing millions of image-text pairs, which achieves significant improvements over the state-of-the-arts on various down-stream tasks without fine-tuning. (2) **Text-to-Image Generation Task**: DALL-E [3] introduces large-scale pre-training into text-to-image generation with a variation of VQ-VAE [12] as the discrete visual tokenizer, which is trained on a dataset of 250 million text-image pairs. DALL-E can generate high-quality images from arbitrary texts and compose unrelated objects in semantically plausible ways. However, the quality of images generated by them is still inferior to that of images by traditional GAN-based methods, *especially in some specific domains*. In this work, our proposed DiST-GAN is a novel framework that introduces the first group of large-scale pre-training models into text-guided face image generation by transferring the semantics of large-scale pre-training to GAN-based generator through distillation.

Knowledge Distillation. Knowledge distillation is firstly proposed by [18], which aims to transfer knowledge from a teacher network (typically more complicated with better performance) to a student network (typically simple and concise with less parameters), resulting in the student network achieving great performance that similar to the teacher network. In addtion to classification, knowledge distillation has also been applied in other tasks in recent works (*e.g.*, detection [19]). In this paper, to the best of our knowledge, we firstly propose to transfer the semantic knowledge of CLIP to GAN-based generator through distillation for text-guided face image generation. This is *still challenging* because of the large gap between the pre-training task used in CLIP and target task.

III. Methodology

A. Framework Overview

As illustrated in Fig. 1, the main components of our DiST-GAN model are text encoder *TE*, image encoder *IE*, Language-to-Vision Transformation module (L2V module) *T*, generator *G*, and discriminator *D*. Given an input text *t*, we firstly embed them into a sequence of token embeddings $v_w = (v_w^1, v_w^2, \ldots, v_w^n)$ by text encoder. Meanwhile, we use the image encoder to embed image *x* corresponding to *t* into image embedding v_x. In this work, our text and image encoders can be formed with the large-scale vision-language pre-training model CLIP, both of which are frozen during training. To better meet the two requirements of text-to-image generation (*i.e.*, photo realism as well as alignment with text), we propose to transfer the semantic knowledge contained in CLIP to our novel L2V module, which transforms v_w into an intermediate embedding v_m to align with the image embedding v_x. The generator *G* generates the output image $\tilde{x}$ according to the intermediate embedding. Our model for text-guided face image generation is defined as:

$$\tilde{x} = G(v_m) = G(T(TE(t))). \tag{1}$$

Note that the paired image x is only used for training. Our model does generate image without using it in the test phase.

B. CLIP-Based Adaptive Contrastive Learning

To transfer the semantic knowledge contained in CLIP to the generator *G* so that the generated images can be more consistent with input texts in semantics, we apply an InfoNCE loss [20] over the generated images and input texts. Let the set of generated images and input texts be respectively denoted as $\tilde{\mathcal{X}} = \{\tilde{x}_i | i = 1, ..., N\}$ and $\mathcal{T} = \{t_i | i = 1, ..., N\}$, where $N = |\tilde{\mathcal{X}}| = |\mathcal{T}|$. The i-th generated image corresponds to the i-th input text. The InfoNCE loss is defined as:

$$\mathcal{L}_{NCE} = -\mathbb{E}\big[\log \frac{\exp(S(\tilde{x}_i, t_i))}{\sum_{j=1}^{N} \exp(S(\tilde{x}_i, t_j))}\big], \tag{2}$$

where $S(\cdot, \cdot)$ denotes the score function to measure the similarity between two vectors. Specifically, we define the score function as the cosine similarity between text and image vectors extracted by CLIP [9].

Note that, for each generated image $\tilde{x}_i \in \tilde{\mathcal{X}}$, contrastive learning with the InfoNCE loss considers all texts in $\mathcal{T}$ as the negative samples except the corresponding input text t_i, and

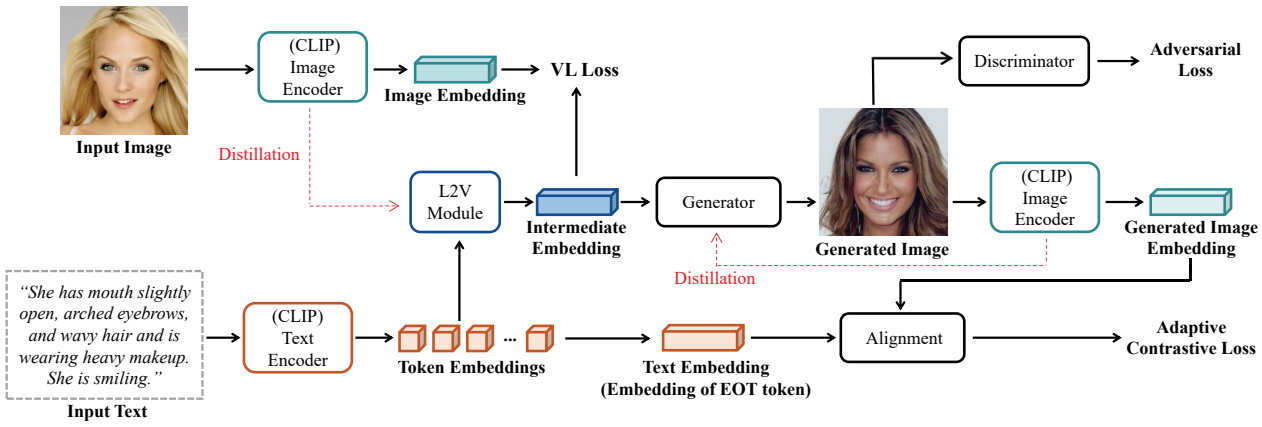

Fig. 1. A schematic illustration of our DiST-GAN model. Our novel CLIP-based adaptive contrastive loss and L2V module are the key components for knowledge distillation in text-guided face image generation. The red dotted lines denote the direction of knowledge distillation. Note that the input image is *only used* in the training phase (but not in the test phase).

simply minimizes the scores of the obtained negative image-text pairs with the same weight. However, when it is applied to the face datasets (*e.g.*., Multi-Modal CelebA-HQ [21]) which contain many similar texts, the texts similar to the input text t_i should not be seen as negative samples (for $\tilde{x}_i$) as the other texts. In this case, simply minimizing the scores of all these pairs with the same weight in the InfoNCE loss is problematic. To alleviate this issue, we thus propose a novel adaptive contrastive loss:

$$\mathcal{L}_{con} = -\mathbb{E}\Big[\log \frac{\exp(S(\tilde{x}_i, t_i))}{\sum_{j=1}^{N} \sigma_{i,j} \exp(S(\tilde{x}_i, t_j))}\Big], \tag{3}$$

where $\sigma_{i,j}$ denotes the weight for the image-text pair $(\tilde{x}_i, t_j)$. In this work, we choose to adjust $\sigma_{i,j}$ according to the similarity between t_i and t_j. Concretely, we first compute the pairwise cosine similarity matrix $M = [m_{i,j}]_{N\times N}$ between all texts in $\mathcal{T}$, where $m_{i,j}$ is the similarity between the i-th and j-th texts. We then define the weight $\sigma_{i,j}$ as:

$$\sigma_{i,j} = \frac{1}{m_{i,j}}. \tag{4}$$

That is, the negative pair $(\tilde{x}_i, t_j)$ becomes less important when the corresponding two texts are more similar.

C. Language-to-Vision Transformation

Besides transferring the semantic knowledge contained in CLIP to the generator G with our CLIP-based adaptive contrastive loss, we also propose to transfer that from CLIP to our novel L2V module T based on distillation. To make L2V module learn the semantic knowledge from CLIP better, we adopt the feature-based and response-based distillation. Concretely, the L2V module T takes the token embeddings v_w as the input and transforms it into the intermediate embedding $v_m = T(v_w)$. For feature-based distillation, the intermediate embedding v_m is subject to the L_1 constraint w.r.t. the corresponding image embedding v_x (paired with v_w) following previous works. For response-based distillation, we consider the similarity between the corresponding image and the texts in a batch as the key information for semantic comprehension. Therefore, we compute the cosine similarity matrix M_m and M_i respectively for intermediate embedding and corresponding image embedding, and then minimize the L1 distance between these two matrices. The VL loss can be defined as:

$$\mathcal{L}_{vl} = \mathbb{E}\|T(TE(t)) - IE(x)\|_1 + \mathbb{E}\|M_m - M_i\|_1. \tag{5}$$

Our overall loss can then be summarised as:

$$\min_{G,TE,IE,T} \max_{D} \quad \mathcal{L}_{adv} + \lambda_{con}\mathcal{L}_{con} + \lambda_{vl}\mathcal{L}_{vl}, \tag{6}$$

where λ_{con} and λ_{vl} denote the weight hyperparameters and $\mathcal{L}_{adv}$ is the adversarial loss:

$$\mathcal{L}_{adv} = \mathbb{E}\big[\log D(x) + \log(1 - D(G(T(TE(t)))))\big]. \tag{7}$$

Experiments

D. Evaluation Metrics

We adopt the Frechét Inception Distance (FID) [23] to evaluate the image quality, which calculates the Frechét distance between two multivariate Gaussians fit to Inception [24] features of generated and real images. Moreover, as in [1], [2], [8], we adopt R-precision to assess how the generated images are aligned with the input texts in semantics. Concretely, for each generated image, we use it to retrieve the paired input text from a subset of 100 candidate texts from the test set, including the paired input text and another 99 texts. We then examine if the input text falls in the top-5 ranked retrieval results and set the accuracy as 1 or 0. R-precision is the average accuracy of all generated images. Note that R-precision is originally computed with the image and text encoders trained on a small dataset in [1], [14], [15]. For better evaluation, we follow GODIVA [25] to adopt CLIP to calculate the similarity for retrieval instead. Both ViT [26] and ResNet-50×4 (not used for training) from CLIP are used to compute R-precision,

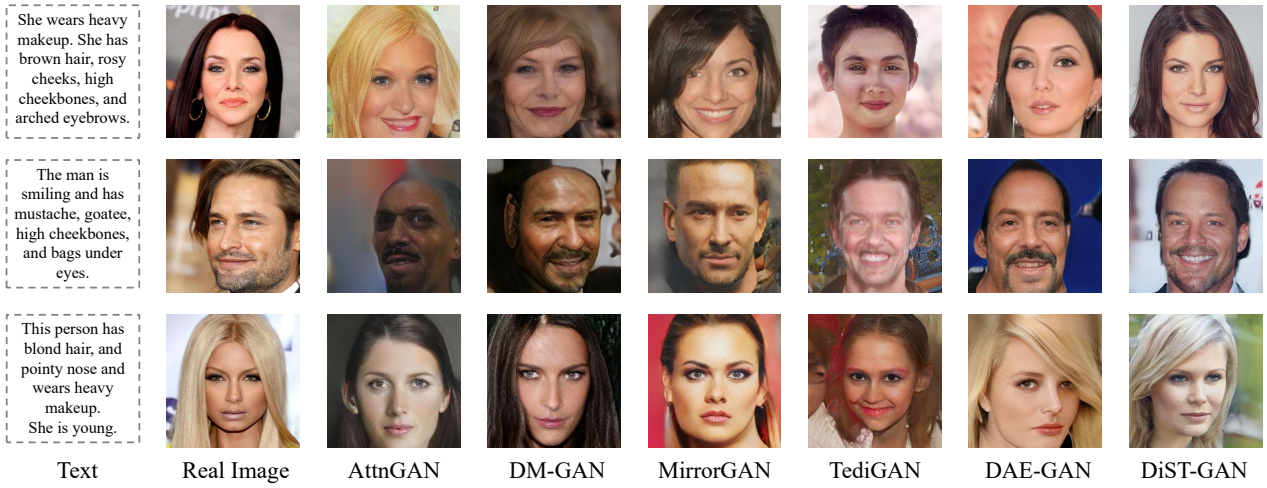

Fig. 2. Qualitative results for text-guided face image generation on Multi-Modal CelebA-HQ [21]. The first two columns show the input texts and corresponding images, respectively. The other columns are the generation results of all competitors.

TABLE I
QUANTITATIVE RESULTS FOR TEXT-GUIDED FACE IMAGE GENERATION ON MULTI-MODAL CELEBA-HQ [21]. R-PREC (VIT) AND R-PREC (RN50×4) DENOTE THE R-PRECISION COMPUTED WITH VIT AND RESNET-50×4 FROM CLIP [9] AS IMAGE ENCODER, RESPECTIVELY. THE SUM OF REALISM OR ACCURACY IS NOT 100% DUE TO THE CHOICE OF "NONE OF THESE METHODS PERFORM WELL". * DENOTES THAT CLIP (WITH VIT AS IMAGE ENCODER) IS USED TO DIRECTLY OPTIMIZE THE FINAL GENERATED IMAGES DURING THE TEST PHASE.

Method	Automated Metrics			User Study	
	FID ↓	R-prec (ViT) ↑	R-prec (RN50×4) ↑	Realism (%) ↑	Accuracy (%) ↑
AttnGAN [1]	25.73	21.97	22.83	12.70	13.10
DM-GAN [14]	27.90	24.70	27.07	5.20	8.40
MirrorGAN [15]	27.07	18.70	20.60	10.40	8.40
TediGAN* [21]	54.51	**91.73**	29.37	6.40	2.50
DAE-GAN [22]	23.53	23.77	25.37	13.10	12.60
DiST-GAN (ours)	**19.32**	57.00	**47.87**	**40.90**	**38.90**

and the obtained results are denoted as R-prec (ViT) and R-prec (RN50×4), respectively.

Although the above automated metrics are useful to evaluate the effectiveness of generation models, the gap between them and human evaluation still exists. Therefore, we also conduct the user study to evaluate the generated images under human perception. The results of user study are reported as the Realism and Accuracy for image quality and semantic consistency evaluation, respectively.

E. Comparison to State-of-the-Arts

Qualitative Results. The qualitative results on Multi-Modal CelebA-HQ [21] are shown in Figure 2. We can obeserve that: (1) AttnGAN, DM-GAN, MirrorGAN and DAE-GAN can generate images in relatively high quality in most cases, but they always make mistakes in generating the color of hair according to the texts and tend to generate blurs and artifacts. (2) TediGAN easily generates low-quality images with overfitted semantics of texts due to its instance-level optimization. In addition, the quality of generated results of TediGAN highly depend on the initialization of latent codes. (3) Our DiST-GAN transfers the semantic comprehension ability of CLIP to GAN-based model, and thus generates high-quality images with sharper details w.r.t. given texts precisely. More qualitative results can be found in the supp. material.

Quantitative Results. The quantitative results are shown in Table I. It can be seen that: (1) Our DiST-GAN outperforms the other methods with large margins on FID and R-prec (RN50×4), indicating that it can generate images with the highest quality that are more aligned with the input texts. (2) When R-prec (ViT) is concerned, TediGAN achieves the highest score, but at the cost of image quality degradation (see its FID $= 54.51$ and its qualitative results in Figure 2). This is mainly due to the fact that TediGAN adopts instance-level optimization during synthesizing images and directly maximizes the cosine similarity between generated images and input texts using CLIP (with ViT as image encoder). Therefore, TediGAN tends to overfit on R-precision computed by the same model (see R-prec (ViT) vs. R-prec (RN50×4)), where the much lower R-prec (RN50×4) means that TediGAN does not perform semantic alignment well. Ignoring TediGAN, our DiST-GAN outperforms the second best competitor with over 30% improvement on R-prec (ViT).

We also conduct human evaluation in addition to automated metrics on Multi-Modal CelebA-HQ. Table I shows the user study results. It can be clearly seen that our DiST-GAN

TABLE II
ABLATION STUDY RESULTS FOR OUR FULL DISTGAN ON MULTI-MODAL CELEBA-HQ [21]. PRE-TRAINED DENOTES EMPLOYING PRE-TRAINED CLIP AND STYLEGAN V2 [27] WITHOUT FURTHER TRAINING. BASE DENOTES THE MODEL WITH ONLY STYLEGAN V2. CL DENOTES THE INFONCE LOSS. L2V AND ADA DENOTE OUR PROPOSED L2V MODULE AND ADAPTIVE CONTRASTIVE LOSS, RESPECTIVELY.

Method	FID ↓	R-prec (ViT) ↑	R-prec (RN50x4) ↑
Pre-trained	54.30	4.97	4.87
Base+CL	20.85	39.17	34.50
Base+Ada	19.92	41.23	35.60
Base+L2V+Ada (Full)	**19.32**	**57.00**	**47.87**

significantly outperforms all competitors in both Realism and Accuracy. Although TediGAN achieves the highest score in semantic consistency on R-prec (ViT), it leads to the lowest accuracy (only 2.5%) in user study, which further shows that TediGAN cheats CLIP on R-prec (ViT) but can not generate images aligned with input texts well. More details of our experiments can be found in the supp. material.

F. Ablation Study

To demonstrate the contributions of the proposed components, we conduct ablation study for our full DiST-GAN. Firstly, follow StyleCLIP [28], we use the frozen pre-trained StyleGAN v2 [27] and CLIP to directly generate images from given texts, whose results are denoted as Pre-trained. Note that StyleCLIP focuses on attribute manipulation but can not directly generate images from given texts. Therefore, we random initialize a latent code to generate a initial image with tyleGAN v2 and then follow StyleCLIP to edit it with the given text. Secondly, on the top of StyleGAN v2, we add various components and train the models *from scratch*. We denote the model with only StyleGAN v2 as Base. CL denotes the InfoNCE loss defined with CLIP. L2V and Ada denote our L2V module and CLIP-based adaptive contrastive loss, respectively. Our full DiST-GAN is actually Base+L2V+Ada.

The results of ablation study are shown in Table II. We can see that: (1) Although StyleCLIP achieves great success in text-guided image manipulation, it produces unsatisfied results in text-guided face image generation due to the large gap between text-guided face image generation and generic text-image retrieval (used for pre-training CLIP). In addition, when we generalize StyleCLIP to text-guided face image generation, we need to carefully select initial images so that it can generate plausible images. (2) Compared to the simple contrastive loss, our proposed CLIP-based adaptive contrastive loss is more beneficial to both the semantic consistency and the visual quality for text-guided face image generation. (3) The combination of L2V module and CLIP-based adaptive contrastive loss further improves R-precision with large margins. More ablation studies are given in the supp. material.

G. Further Evaluations

Results for Similar Texts. Although we only apply the constraint of text-vision alignment on sentence and image level, our DiST-GAN can capture the fine-grained semantic

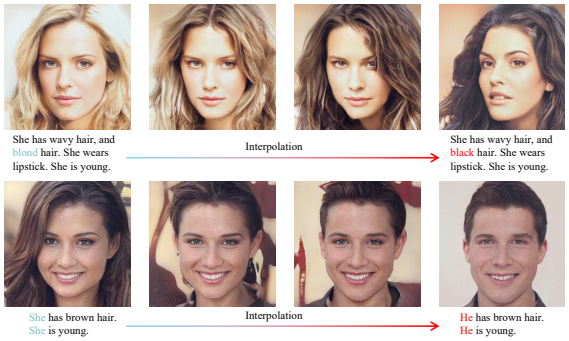

Fig. 3. Results for similar texts of our DiST-GAN. The first and last columns show the generated images with the texts below them. The other columns show the interpolation results.

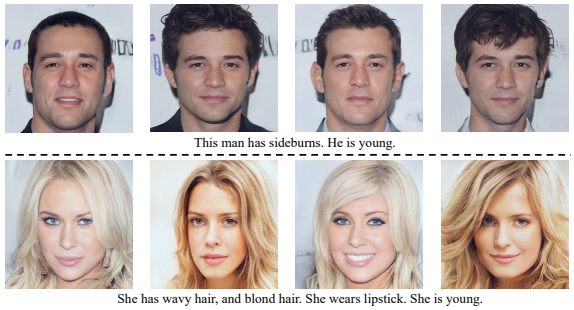

Fig. 4. Diverse results obtained by our DiST-GAN. Each row shows the images generated by the text below them.

TABLE III
RESULTS OF TEXT-TO-IMAGE GENERATION ON CUB.

Method	FID ↓	R-prec (ViT) ↑	R-prec (RN50x4) ↑
AttnGAN	23.45	13.10	13.17
DM-GAN	24.05	13.86	13.86
MirrorGAN	24.16	10.14	11.09
DAE-GAN	30.30	6.03	6.00
DiST-GAN (ours)	**22.67**	**42.55**	**28.79**

information in words. To demonstrate this advantage, we synthesize images with two similar input texts. Specifically, we edit only one word from each given text and synthesize images according to the resultant two texts, respectively. In addition, we also generate the interpolation results of the two images generated with similar texts. As shown in Figure 3, the generated images with two similar texts are both well aligned with the input texts, providing evidence that our DiST-GAN can capture the fine-grained semantic information in words. Moreover, the interpolation results show that the generated images can gradually change from one to another as the text transformation, which means the latent space is continuous and semantically meaningful.

Diversity of Generated Images. Given an input text, our DiST-GAN can synthesize diverse images (while coherent with the text) by simply injecting different noises into the intermediate embedding and different intermediate feature

maps during generation process. Note that our model can achieve this without refining the generated images in multi-stages. The diverse generated results are shown in Figure 4.

Results on the CUB Dataset. To further show the effectiveness of our DiST-GAN, we also make evaluation on the widely-used CUB [29] dataset. We follow the standard split of CUB. The quantitative results on the CUB [29] dataset are shown in Table III. Note that TediGAN [21] are not considered as a competitor on CUB because it focuses on face image generation. We can observe that our DiST-GAN outperforms the other methods on both FID and R-precision, indicating that our model can generate images with best photo-realism and semantic consistency on this non-face dataset. The qualitative results on CUB can be found in the supp. material.

CONCLUSION

In this work, we have proposed a novel distillation-based framework to transfer the knowledge of semantics in large-scale visual-language pre-training models to GAN-based generator, called DiST-GAN. It is a simple but effective framework for face image synthesis according to textual descriptions. Due to the proposed CLIP-based adaptive contrastive loss and language-to-vision transformation module, our DiST-GAN can effectively learn the semantic comprehension from CLIP, which thus generates higher-quality images that better match the corresponding input texts. Extensive results show that our DiST-GAN can generate diverse and high-quality images and significantly outperforms the state-of-the-arts in both automated metrics and human evaluation.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (61976220), and China Unicom Innovation Ecological Cooperation Plan. Zhiwu Lu is the corresponding author.

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 *CVPR*, 2018, pp. 1316–1324.

[2] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang, "Cross-modal contrastive learning for text-to-image generation," in *CVPR*, 2021, pp. 833–842.

[3] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, "Zero-shot text-to-image generation," in *ICML*, Marina Meila and Tong Zhang, Eds., 2021, vol. 139, pp. 8821–8831.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680.

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

[6] Han Zhang, Tao Xu, and Hongsheng Li, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017, pp. 5908–5916.

[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 Trans. Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, 2019.

[8] 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 *CVPR*, 2019, pp. 12174–12182.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021, pp. 4904–4916.

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

[12] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural discrete representation learning," in *NeurIPS*, 2017, pp. 6306–6315.

[13] Diederik P Kingma and Max Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[14] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang, "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," in *CVPR*, 2019, pp. 5802–5810.

[15] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *CVPR*, 2019, pp. 1505–1514.

[16] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[17] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al., "WenLan: Bridging vision and language by large-scale multi-modal pre-training," *arXiv preprint arXiv:2103.06561*, 2021.

[18] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[19] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou, "General instance distillation for object detection," in *CVPR*, 2021.

[20] Aäron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[21] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu, "TediGAN: Text-guided diverse face image generation and manipulation," in *CVPR*, 2021, pp. 2256–2265.

[22] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen, "DAE-GAN: Dynamic aspect-aware GAN for text-to-image synthesis," *arXiv preprint arXiv:2108.12141*, 2021.

[23] 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 *NeurIPS*, 2017, pp. 6626–6637.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[25] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan, "GODIVA: generating open-domain videos from natural descriptions," *arXiv preprint arXiv:2104.14806*, 2021.

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in *CVPR*, 2020, pp. 8107–8116.

[28] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski, "StyleCLIP: Text-driven manipulation of StyleGAN imagery," *arXiv preprint arXiv:2103.17249*, 2021.

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