Lightweight Text-Driven Image Editing with Disentangled Content and Attributes

Bo Li, Xiao Lin, Bin Liu, Zhi-Fen He, Yu-Kun Lai

***Abstract*—Text-driven image editing aims to manipulate images with the guidance of natural language description. Text is much more natural and intuitive than many other interaction modes, and attracts more attention recently. However, compared with classical supervised learning tasks, there is no standard benchmark dataset for text-driven interactive image editing up to now. Therefore, it is hard to train an end-to-end model for pixel-aligned interactive image editing driven by text. Some methods follow the paradigm of text-to-image models by incorporating the target image into the process of text-to-image generation. However, these methods relying on cross-modal text-to-image generation involve complicated and expensive models, which can lead to inconsistent editing effects. In this paper, a novel text-driven image editing method is proposed. Our key observation is that this task can be more efficiently learned using image-to-image translation. To ensure effective learning for image editing, our framework takes paired text and the corresponding images for training, and disentangles each image into content and attributes, such that the content is maintained while the attributes are modified according to the text. Our network is a lightweight encoder-decoder architecture that accomplishes pixel-aligned end-to-end training via cycle-consistent supervision. Quantitative and qualitative experimental results show that the proposed method achieves state-of-the-art performance.**

***Index Terms*—interactive image editing, text-driven, disentanglement, cycle-consistency**

I. Introduction

Interactive image editing allows users to manipulate images according to their preferences to produce meaningful and creative results. It has been applied in numerous fields, such as digital advertisement, media production, etc. Users can interactively edit images in different ways, including scribble-guided methods [1]–[4], reference image based methods [5], [6] and text-driven methods [7]–[9]. However, professional knowledge and user interactions are required for scribble-guided methods. Although reference-based methods reduce the burdens of user interaction, a proper example image meeting the user's requirement can be difficult to find. Compared with scribbles and reference images, text is more natural and intuitive for users. Without any professional knowledge, a user can provide meaningful editing requests. Thus, text-driven image editing has been a popular research topic in the field of image processing.

Compared with classical supervised learning models, there is no standard benchmark dataset for text-driven interactive image editing up to now. For a given image with aligned text description, the pixelwise edited result with a given unaligned text is not available. Therefore, it is impossible to train a fully supervised text-driven interactive image editing model. An intuitive idea is to follow the paradigm of text-to-image models [10]–[13], and incorporate the target image information into the process of text-to-image generation. This kind of approaches [8], [9], [14] indeed generates some meaningful results. However, their approaches are designed to generate images from scratch, instead of modifying the image guided by text description. Furthermore, some recent large models based on VQGAN (Vector Quantized Generative Adversarial Network) [15] and DDPM (Denoising Diffusion Probabilistic Models) [16] like VQGAN-CLIP [17], DALL-E [18], [19], Imagen [20] and LDM [21] achieve great improvements on general text-to-image tasks. However, these models are extremely complex and expensive to train. In addition, these models are designed for general joint learning of text and images, and may fall short for some detailed editing tasks. For fair comparison, we adapt a big model LDEdit [22] based on LDM (Latent Diffusion Models) by training it on the same dataset as used in our paper.

Based on the above analysis, most existing works based on text-to-image generation suffer from the following two issues. On the one hand, these models are complex and difficult to train. On the other hand, these models suffer from inconsistent editing effects, since the text and image features are coupled in the text-to-image generation process. Some examples are shown in Fig. 1, and it can be easily found that numerous text-irrelevant attributes are falsely edited in the results of state-of-the-art methods SISGAN [7], TAGAN [14], SIMGAN [23], ManiGAN [8], L-ManiGAN [24] and LDEdit [22].

In this paper, a novel text-driven image editing method is proposed. Compared with the complicated text-to-image framework that most existing works adopted, the proposed framework is a lightweight encoder-decoder architecture. It accomplishes the pixel-aligned end-to-end training via a cycle-consistent supervision, taking two pairs of training samples with each pair involving an aligned image and corresponding text. First, a disentangling encoding module is designed to decompose the target image into two separate components: content features and attribute features. During the interactive editing process, the attribute features are modulated by the provided text description, while the content features which catch the text-irrelevant component will be kept as well as possible. Through the disentangling encoding, the controllability of interactive editing is improved effectively, as shown in Fig. 1, where the background and text-irrelevant objects are preserved well while text-relevant attributes are edited correctly. Second, based on our disentangled representation,

Bo Li, Xiao Lin, Bin Liu and Zhi-Fen He are with the School of Mathematics and Information Science, Nanchang Hangkong University, Nanchang, China.

Yu-Kun Lai is with the School of Computer Sciences and Informatics, Cardiff University, Cardiff, UK.

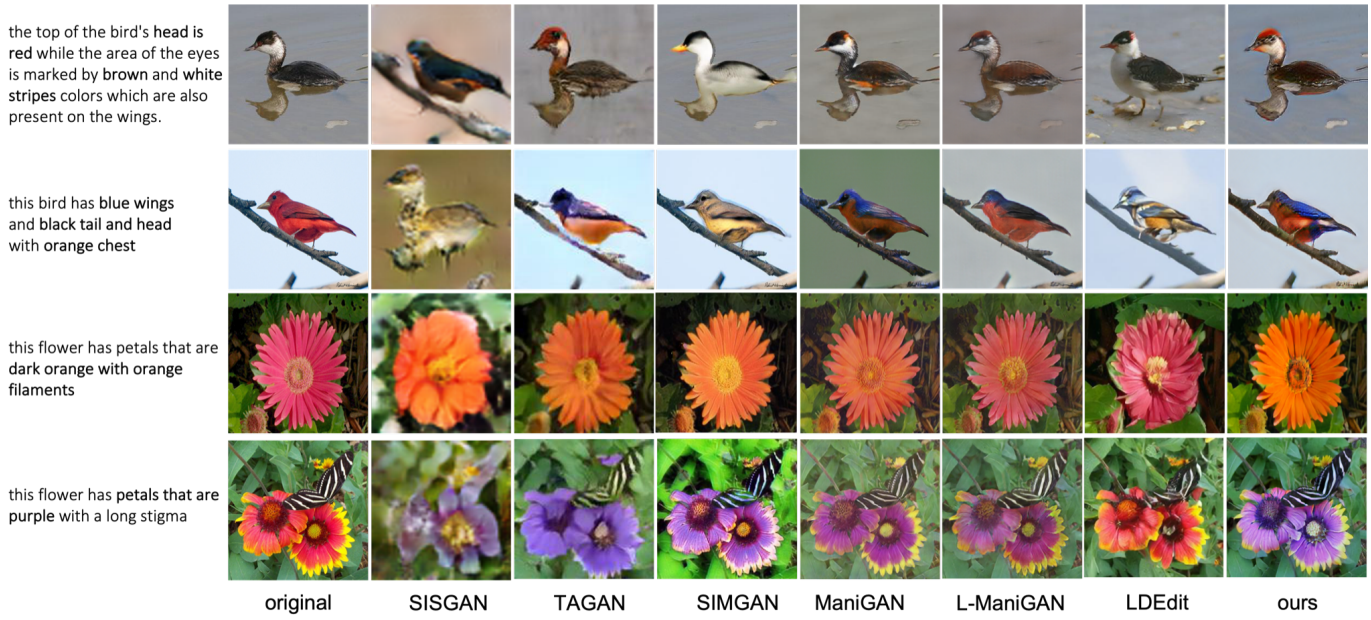

Fig. 1. Text-driven image editing results with different methods: SISGAN [7] (64×64), TAGAN [14] (128 128×128), SIMGAN [23], ManiGAN [8], L-ManiGAN (Lightweight ManiGAN) [24], LDEdit [22] and the proposed method.

a specifically designed cyclic training strategy is proposed to regularize and enable end-to-end training. Given an input image $\mathbf{I}_1$ (with aligned text $\mathbf{T}_1$) to be edited, unaligned text $\mathbf{T}_2$ (corresponding to another image $\mathbf{I}_2$) is utilized to edit its attributes, resulting in $\hat{\mathbf{I}}_1$. When further editing $\hat{\mathbf{I}}_1$ with the aligned text $\mathbf{T}_1$, the resulting image $\hat{\mathbf{I}}'_1$ should reconstruct the original attributes. The reconstruction loss will guide the model to align text space and image attribute space. Through the cycle-consistent training strategy, a lightweight encoder-decoder structure can be adopted, rather than the complicated text-to-image generative model. We evaluate the proposed model on widely used public datasets against six state-of-the-art methods. Quantitative and qualitative experimental results show that the proposed method outperforms previous state-of-the-art methods.

In summary, the main contributions of this paper include:

- A novel text-driven image editing model is proposed, in which a lightweight encoder-decoder model is designed rather than utilizing a complicated text-to-image generative network.
- A disentangling encoding module is designed to improve the consistency between text instructions and edited results.
- As there is no standard dataset for completely supervised text-driven editing, a cyclic training strategy is proposed to accomplish the pixel-aligned end-to-end training in the image domain.

The remainder of this paper is organized as follows. Section II briefly overviews previous work on text-to-image generation and text guided image manipulation. In section III, the proposed lightweight text-driven image editing model with disentangled content and attributes is presented in detail. The experimental results and analysis are reported in Section IV. Finally, we conclude this paper and present future research directions in Section V.

II. Related Work

Direct text-to-image generation (T2I). This task aims to synthesize an image using a single caption as input. One of the pioneering approaches was proposed in [25] as a natural extension of the cGAN (conditional Generative Adversarial Network) [26]. An additional auxiliary classification task is utilized to improve generation performance in [27]. Although these methods can generate realistic images, the generative quality and resolution are not satisfactory. To synthesize high-quality images, there are a series of works adopting stacked architectures [10]–[13], [28]. A critical problem of T2I synthesis is text-image alignment, and recently more and more researchers have incorporated attention techniques into their methods to improve the alignment between text space and image space [12], [13], [29]–[32]. Some other methods tried to incorporate multi-modal information such as scene graphs or other attribute labels for supervision [33]–[36] to achieve more controllable edited results. Recently, some large models [17]–[21] with billions of parameters and trained on huge datasets achieved great improvements. However, these models are too complex and expensive to train.

Text guided image manipulation. Different from T2I, text guided image manipulation aims to edit the target image to satisfy the semantic meanings of the text description while keeping text-irrelevant components, rather than generating a novel image from scratch. SISGAN [7] proposed an encoder-decoder architecture and built the model upon a conditional GAN framework conditioning on both images and text. However, since there is no standard pixel-wise aligned dataset for this task, this vanilla way has insufficient supervision information, leading to poor performance on aligning editing results with the corresponding text, as shown in Fig. 1. In

order to enhance the supervision, TAGAN [14] extended the original conditional discriminator into a number of word-level discriminators for a word-level constraint. Some other methods treat the text-guided image editing problem as a variant of text-to-image generation, such as [8], [9], [24]. Based on ControlGAN [12], an additional text-image affine combination module was proposed in ManiGAN [8] to refine the editing result by incorporating the target image into the T2I network. A lightweight version of ManiGAN was proposed in [24]. TEA-cGAN [9] implemented a multi-stage synthesis by fusing different attentions based on a similar network to ManiGAN. However, since this kind of methods generates the edited images from scratch, the models are complicated and expensive for both training and inference. To manipulate the accurate text-relevant regions, SIMGAN [23] proposed to first separate the instances and background, and narrow down the manipulation regions from external to internal to avoid inaccurate modifications. Despite the performance improvement, an extra segmentation mask is required in the training process. Wu et al. [37] introduced an editing proposal generator to generate edited images with and without semantic conditions, and also with a prediction of semantic mask, but the reorganization mindset of proposals is quite complicated. More recently, Yan et al. [38] introduced a differentiable architecture search approach named ZeroNAS for constructing GANs, and Zhang et al. [39] proposed a progressive meta-learning scheme. These are well-suited for zero-shot or few-shot learning, and showed powerful potentials to build a suitable GAN architecture optimized for image editing with textual descriptions. StyleCLIP [40] developed a text-based image manipulation method by leveraging the power of contrastive language-image pre-training model (CLIP) [41]. However, it is still a text-to-image strategy and cannot reveal the changes of text-irrelevant components. Recently, DDPMs have gained competitive performance on unconditional image generation, text-to-image generation and image editing, such as Blended-Diffusion [42], DiffusionCLIP [43] and LDEdit [22], etc. However, these models are generally complicated with billions of parameters, and do not account for the disentanglement of text-relevant and irrelevant information. Some other works studied image manipulation methods with relatively simple text input, such as short phrases [44], and auto-generated text with patterns and labels [45]. In this paper, a novel text-driven image editing method is proposed. Compared with the complicated text-to-image framework that most existing works adopted, the proposed framework is a lightweight encoder-decoder architecture, and accomplishes end-to-end training via the cycle-consistent supervision of two pairs of training data.

III. Approach

The goal of text-driven image manipulation is to edit an image based on the instruction of a user-provided text description. As shown in Fig. 1, the edited results should be consistent to the semantic guidance of the text instruction, while preserving the text-irrelevant regions as well as possible. In this section, a novel text-driven image editing method is proposed. On the one hand, a disentangling encoding strategy is proposed to guarantee the consistency between the text and edited results. On the other hand, a cycle-consistent reconstruction process is introduced to provide the image domain supervision, which is rarely used in existing text-driven image editing methods. Benefiting from the cyclic training process, the proposed method is lightweight and a real image editing model by fusing language instruction rather than a variant of a complex text-to-image generation model. The details of the proposed method are described as follows.

A. Architecture

The pipeline of the proposed algorithm is shown in Fig. 2. Let $(\mathbf{I}_1, \mathbf{T}_1)$, $(\mathbf{I}_2, \mathbf{T}_2)$ denote two pairs of training data, where $\mathbf{I}_i$ and $\mathbf{T}_i$ $(i = 1, 2)$ are respectively the input image and its aligned description text. Our training model is composed of two stages: editing stage and cycle-editing stage. In the editing stage, the target image $\mathbf{I}_1$ is first disentangled into two separate components: text-irrelevant content and text-relevant attributes. Then the unaligned text $\mathbf{T}_2$ is fused to the attribute features to guide the editing and produces the edited result $\hat{\mathbf{I}}_1$. Ideally, the edited image $\hat{\mathbf{I}}_1$ should share the content of image $\mathbf{I}_1$ and the attribute features of $\mathbf{T}_2$ corresponding to image $\mathbf{I}_2$. In order to improve the disentangling performance and editing consistency, a cycle-consistent reconstruction stage is proposed. The ground truth text $\mathbf{T}_1$ is utilized to modulate the attribute features of the edited image $\hat{\mathbf{I}}_1$ back to the original attribute of image $\mathbf{I}_1$, and the reconstruction loss in the image domain is used to guide the model to align text space and image attribute space. Since $\mathbf{I}_2$ is processed in the same way as $\mathbf{I}_1$, for simplicity we mainly describe our work with $\mathbf{I}_1$ as input.

B. Disentanglement of Content and Attributes

For the task of text-driven image manipulation, we assume that the editing should focus on text-relevant attributes of the image while keeping the content as well as possible. In order to reduce the text-irrelevant changing, a disentangling encoding strategy is proposed to first decompose the target image into two separate components: text-irrelevant content $\mathbf{v_c} \in \mathbb{R}^{C\times W\times H}$ and text-relevant attributes $\mathbf{v}_a \in \mathbb{R}^{(k\times k)\times d}$, where C, W and H are the number of channels, width and height for content features, $k\times k$ is the spatial size of attributes and d is the dimension of each attribute. Then the text features $\mathbf{w} \in \mathbb{R}^{L\times d}$ (where L is the length of text) is extracted to guide the semantic editing by modulating the attribute features $\mathbf{v}_a$ via a Fuser module as shown in Fig. 2(b), while keeping the text-irrelevant content features $\mathbf{v}_c$. Finally, the fused attribute features are combined with the content features to produce the edited result through a generator network $\mathbf{G}$. In detail, the content encoder $\mathbf{E}_C$ is composed of a shallow convolutional neural network (CNN) with residual connections, while a deeper network $\mathbf{E}_A$ is applied to encode attribute features. The text encoder $\mathbf{E}_T$ is an RNN (Recurrent Neural Network) based network which can output both word features and sentence features.

In order to improve the disentanglement performance of content-attribute space and enhance the alignment between text

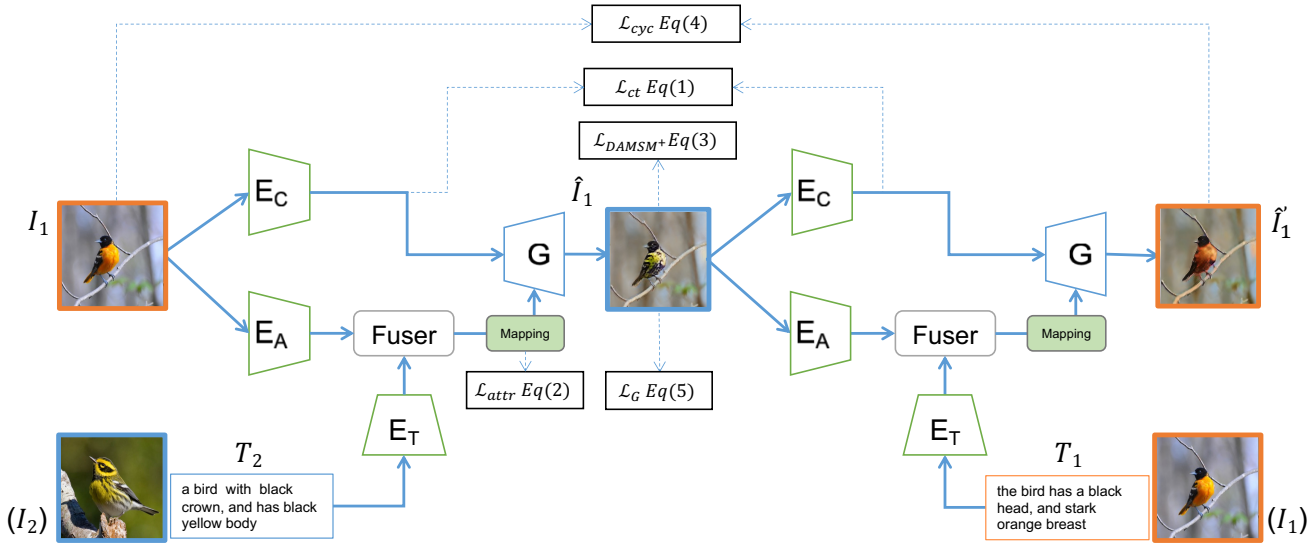

Fig. 2. The framework of the proposed method. (a) is the main framework, E_C and E_A are the disentangling encoders for content and attribute features respectively, E_T is the encoder for text and G is the decoder model. The target image $\mathbf{I}_1$ is first edited according to unaligned text $\mathbf{T}_2$ (its matched image is (I_2)), and then is cyclically reconstructed by aligning with the ground truth text $\mathbf{T}_1$ (matched (I_1)). (b) is detailed structure of the Fuser module. (c) is the sub-network for pre-training content-attribute disentanglement.

features and attribute features, the training strategy is split into two stages. First, given two images $\mathbf{I}_1, \mathbf{I}_2$ with distinct attribute features, the content encoder $\mathbf{E}_C$ and attribute encoder $\mathbf{E}_A$ are optimized by minimizing the distance $\mathcal{L}_{ct}$ (Eq. 1) between content features and maximizing the semantic similarity of attribute features $\mathcal{L}_{attr}$ (Eq. 2) when exchanging the attribute features of two images as shown in Fig. 2(c).

$$\mathcal{L}_{ct} = \frac{1}{CWH}\|\mathbf{v}_{c_1} - \hat{\mathbf{v}}_{c_1}\|_2^2, \quad (\mathbf{v}_c = \mathbf{E}_C(\mathbf{I})) \tag{1}$$

$$\mathcal{L}_{attr} = \text{cos-sim}(\hat{\mathbf{v}}_{a_1}, \mathbf{v}_{a_2}), \quad (\mathbf{v}_a = \mathbf{E}_A(\mathbf{I})) \tag{2}$$

where $\mathbf{v}_{c_1}$ and $\hat{\mathbf{v}}_{c_1}$ are respectively the content features of image $\mathbf{I}_1$ and edited result $\hat{\mathbf{I}}_1$, $\hat{\mathbf{v}}_{a_1}$ and $\mathbf{v}_{a_2}$ are respectively the attribute features of $\hat{\mathbf{I}}_1$ and the reference image $\mathbf{I}_2$ as shown in Fig. 2(c), and cos-sim$(\cdot)$ is the discrete cosine distance. The first training stage strives to separate the semantic attribute features from the basic content features. Then, in order to accomplish meaningful text-guided attribute editing, the text encoder $\mathbf{E}_T$ and attribute encoder $\mathbf{E}_A$ will be optimized in the second stage to enhance the semantic alignment between the latent linguistic space and the visual attribute space. In detail, an unaligned text $\mathbf{T}_2$ is encoded by $\mathbf{E}_T$ to the linguistic space $\mathbf{w}_2$, and then a Fuser module (Fig. 2(b)) is designed to fuse the text features $\mathbf{w}_2$ with attribute features $\mathbf{v}_a$ to accomplish semantic attribute editing in the latent space. Finally, the text-guided edited attribute features will be combined with basic content to generate the edited image $\hat{\mathbf{I}}_1$. In order to guarantee the semantic consistency between the given text $\mathbf{T}_2$ and the corresponding editing result $\hat{\mathbf{I}}_1$, an improved DAMSM [13] loss function is proposed to minimize the gap between the image attribute distribution and the text distribution globally. DAMSM ranks the text-image matching by computing the following two posterior probabilities,

$$\mathcal{L}_{DAMSM}(I, T) = \mathcal{P}(\mathbf{I}|\mathbf{T}) + \mathcal{P}(\mathbf{T}|\mathbf{I})$$

where $\mathcal{P}(\mathbf{I}|\mathbf{T})$ denotes the probability that sentence $\mathbf{T}$ is matched with its corresponding image $\mathbf{I}$, and vice versa. Although DAMSM can reflect the matching between image and text globally, the absolute DAMSM loss is still not proper for our task. For two pairs of image-text data, the absolute DAMSM score ranks the similarity of a single image-text pair, but it ignores the relationship between two pairs of data. For example, $\mathcal{P}(\hat{\mathbf{I}}_1|\mathbf{T}_2)$ will benefit from the guidance of $\mathcal{P}(\mathbf{I}_2|\mathbf{T}_2)$ by the visual attribute correlation. Therefore, an improved DAMSM loss function is proposed to enhance the semantic consistency between images and text.

$$\mathcal{L}_{DAMSM^+}(\hat{\mathbf{I}}_1, \mathbf{T}_2) = \alpha\mathcal{P}(\hat{\mathbf{I}}_1|\mathbf{T}_2) + \beta\mathcal{P}(\mathbf{T}_2|\hat{\mathbf{I}}_1) \tag{3}$$

where $\alpha = \exp(-\mathcal{P}(\mathbf{I}_2|\mathbf{T}_2))$ and $\beta = \exp(-\mathcal{P}(\mathbf{T}_2|\mathbf{I}_2))$. The improved DAMSM$^+$ loss will guide the text to pay attention to the similar visual attributes on both images $\mathbf{I}_1$ and $\mathbf{I}_2$.

An experiment is conducted to validate the effectiveness of the proposed disentangling strategy. To demonstrate the disentanglement effects and the alignment performance between text and visual attributes, two text descriptions with distinct semantic attributes are provided, and the editing results are shown in the 3rd column of Fig. 3. It can be found that the proposed method can approximately align the text semantic instruction with the corresponding visual attribute features. In addition, according to the editing results with two distinct text descriptions, we can find that the text-irrelevant content features are preserved well while attribute features are edited respectively by corresponding text. The results help validate the effectiveness of the proposed content-attribute disentangling strategy. More experiments will be conducted to evaluate the proposed disentangling strategy in section IV-B3.

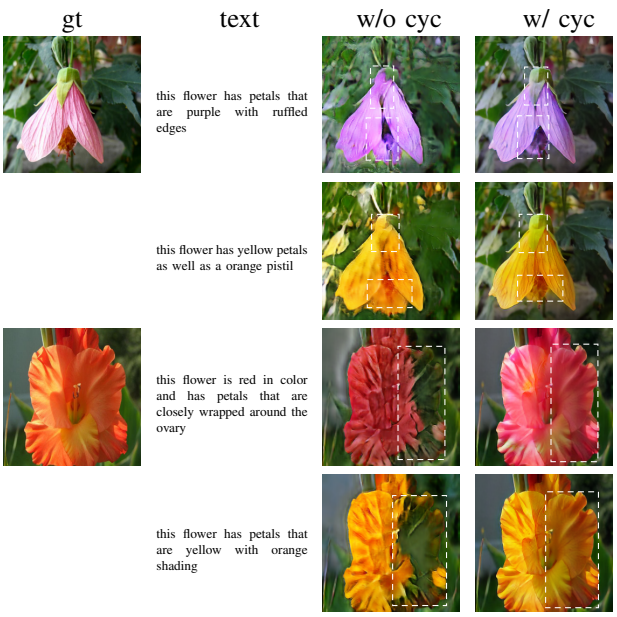

Fig. 3. Illustration of editing results w/o and w/ cycle-consistency training strategy.

C. Cycle-Consistency Training Strategy

Although the disentangling strategy and the improved DAMSM loss can help improve the performance of text-guided attribute-consistency interactive editing, the editing results may still suffer from the absence of pixel-aligned supervision, resulting in obvious artifacts in edited images. For example, the green calyx of the flower in the first row of Fig. 3 is wrongly edited to the color of the flower.

In this section, a cycle-consistency training strategy is proposed to fulfill the pixel-level supervision and enhance the disentangling performance. The detailed pipeline is shown in Fig. 2. In the first stage of the editing process, image $\mathbf{I}_1$ is disentangled into content features $\mathbf{v}_{c_1}$ and attribute features $\mathbf{v}_{a_1} \in \mathbb{R}^{(k\times k)\times d}$ separately, and then the text features $\mathbf{w}_2 \in \mathbb{R}^{L\times d}$ extracted from $\mathbf{T}_2$ is utilized to modify the visual attribute $\mathbf{v}_{a_1}$ via a Fuser module to produce the edited attribute features $\tilde{\mathbf{v}}_{a_1}$ complying with the semantic instruction of the text $\mathbf{T}_2$. Finally, the edited image $\hat{\mathbf{I}}_1$ will be generated by combining the edited attribute features $\tilde{\mathbf{v}}_{a_1}$ and the basic content features $\mathbf{v}_{c_1}$ through the generator G.

To provide pixel-level constraints for interactive editing, a cycle editing process is conducted in a similar way. Image $\hat{\mathbf{I}}_1$ is first disentangled into text-irrelevant content and text-relevant attributes, and then the aligned text $\mathbf{T}_1$ corresponding to image $\mathbf{I}_1$ is utilized to recover the original attributes of $\mathbf{I}_1$ cyclically. Finally, the reconstructed image $\hat{\mathbf{I}}'_1$ will be generated by combining the content features through the generator G, which should be close to the image $\mathbf{I}_1$. The detailed pipeline can be seen in Fig. 2.

The cycle-consistency training loss is defined as follows,

$$\mathcal{L}_{cyc}(\hat{\mathbf{I}}'_1, \mathbf{I}_1) = \|\hat{\mathbf{I}}'_1 - \mathbf{I}_1\|_1 + \mathcal{L}_{ct}(\hat{\mathbf{I}}'_1, \mathbf{I}_1) \tag{4}$$

The cycle-consistency training strategy helps improve the editing performance in the following two respects. First, cycle reconstruction loss $\|\hat{\mathbf{I}}'_1 - \mathbf{I}_1\|_1$ provides pixel-level constraints to enhance the semantic alignment of text features and visual attribute features in the latent space. Second, cycle reconstruction helps improve the disentangling performance of basic content features and text-relevant attribute features by minimizing the perturbation of content features $\mathcal{L}_{ct}(\hat{\mathbf{I}}'_1, \mathbf{I}_1)$ during the first editing stage and the cycle reconstruction stage.

From the last column of Fig. 3, it is easy to find that the editing results are improved obviously when equipped with the cycle-consistency training strategy. A detailed ablation study is later shown in section IV-B4. We evaluate the performance with and without cycle-consistency training strategy. Experimental results shown in Fig. 9 validate the positive effects in both the robust training and editing performance.

D. Objective Function

In addition to the above loss functions, adversarial losses w.r.t. the discriminator and generator training are introduced, as follows:

$$\begin{aligned}\mathcal{L}_D =& -\frac{1}{2}(\mathbb{E}_{\mathbf{I}_1\sim P_{data}}[\log D(\mathbf{I}_1)] + \mathbb{E}_{\hat{\mathbf{I}}_1\sim P_G}[\log(1 - D(\hat{\mathbf{I}}_1))]\\ \mathcal{L}_G =& -\mathbb{E}_{\hat{\mathbf{I}}_1\sim P_G}[\log D(\hat{\mathbf{I}}_1)]\end{aligned} \tag{5}$$

where $D(\cdot)$ denotes the output score of the discriminator.

Finally, the whole objective function of the proposed model for the generator training can be formulated as follows:

$$\mathcal{L}_{train} = \mathcal{L}_G + \lambda_1\mathcal{L}_{cyc} + \lambda_2\mathcal{L}_{DAMSM} + \lambda_3(1 - \mathcal{L}_{attr}) \tag{6}$$

IV. Experiments

In this section, we evaluate the performance of the proposed method on public CUB [46] and Oxford-102 [47] datasets against six state-of-the-art methods with open source: SISGAN [7][1], TAGAN [14][2], SIMGAN [23][3], ManiGAN [8][4], L-ManiGAN [24][5] and LDEdit [22][6]. For methods without official pre-trained checkpoints, we train them on the corresponding dataset with their official configurations. For LDEdit, we first reduce the scale of the LDM model to a similar level to other methods, and then train LDEdit on CUB and Oxford-102 datasets respectively.

The hyperparameters of the proposed model are set as follows: λ_1 controlling cyclic reconstruction $\mathcal{L}_{cyc}$ is set to 10, the weight λ_2 controlling $\mathcal{L}_{DAMSM}$ is set to 5, and λ_3 for attribute similarity is set to 1. All experiments are implemented using PyTorch and trained on a single Nvidia GeForce RTX 2080ti GPU.

[1]https://github.com/woozzu/dong_iccv_2017
[2]https://github.com/woozzu/tagan
[3]https://github.com/meluffy/SIMGAN
[4]https://github.com/mrlibw/ManiGAN
[5]https://github.com/mrlibw/Lightweight-Manipulation/
[6]https://github.com/CompVis/latent-diffusion

TABLE I
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON THE CUB DATASET: INCEPTION SCORE (IS), FINE-TUNED INCEPTION SCORE (IS*), TEXT-IMAGE SIMILARITY (SIM), L1 PIXEL DIFFERENCE (DIFF), MANIPULATIVE PRECISION (MP), NUMBER OF PARAMETERS (#PARAMS), INITIALIZATION TIME (INIT-TIME) AND INFERENCE TIME (RUNTIME).

	CUB								
Method	IS↑	IS*↑	CLIPScore↑	sim↑	diff↓	MP↑	#params↓	init-time(s)↓	runtime(s/img)↓
SISGAN [7]	5.63	3.98	.72	.133	.498	.067	31.57M	45.99	1.66
TAGAN [14]	5.60	14.23	.73	.307	.152	.260	**28.74M**	46.1	1.88
SIMGAN [23]	4.29	19.24	.70	.211	**.086**	.193	55.87M	5.27	1.10
ManiGAN [8]	5.95	24.33	.71	.298	.289	.212	160.28M	3.82	3.00
L-ManiGAN [24]	5.69	19.15	.73	.342	.192	.277	55.11M	3.49	2.42
LDEdit [22]	5.24	**28.69**	.72	.208	.130	.181	453.54M	21.07	18.96
ours	**6.85**	25.40	**.74**	**.350**	.150	**.298**	50.51M	**2.03**	**0.92**

TABLE II
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON OXFORD-102 DATASET.

	Oxford-102					
Method	IS↑	IS*↑	CLIPScore↑	sim↑	diff↓	MP↑
SISGAN	3.61	7.51	.74	.318	.500	.159
TAGAN	3.05	8.76	.74	.295	.176	.243
SIMGAN	3.58	28.74	.72	.228	.191	.185
ManiGAN	3.77	23.62	.71	.435	.188	.353
L-ManiGAN	**3.83**	22.74	.70	.214	.198	.172
LDEdit	3.03	**30.11**	.69	.151	**.156**	.130
ours	3.77	29.13	**.75**	**.559**	.223	**.435**

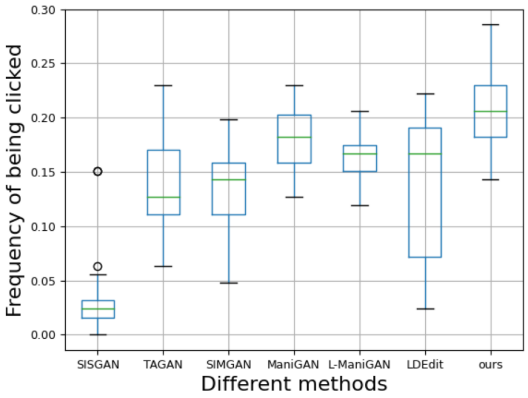

Fig. 4. Subjective user study results on 20-group edited results on CUB and Oxford-102 datasets.

A. Experimental Setup

Datasets: CUB is composed of 8,855 training images and 2,933 test images, and there are 10 text descriptions corresponding to each image.

To analyze the descriptions, we find that among the top 100 most frequent words, more than 80% descriptions are shape-irrelevant. As a result, learned models tend to bias towards editing color and texture attributes rather than shapes, which is essentially a limitation of the dataset.

Oxford-102 is composed of 8,189 images. We manually split them into 6,551 images for training, and 1,638 as test images. There are 10 text descriptions corresponding to each image.

B. Comparison with state-of-the-art approaches

*1) **Quantitative comparison**:* Five quantitative evaluation metrics are adopted for comparing the performance of different methods, including inception score (IS) [48], CLIPScore [49], text-image similarity (*sim*), L_1 pixel difference (*diff*) and manipulative precision (MP) [8]. IS is used to evaluate the perceptual quality of the edited image. We configure IS evaluation directly based on the official implementation provided in [48] and also fine-tuned the inception-v3 model additionally on CUB and Oxford-102 to get fine-tuned IS scores (denoted as IS*). CLIPScore [49] is a metric for image-text compatibility, and we get CLIPScore based on the official ViT-B-32 pre-trained model [50]. MP metric is introduced in [8], and it is defined and computed by MP=(1-*diff*) $\times$ *sim*, where *diff* refers to the pixel-level L_1-distance between the original image and edited image, and *sim* is the similarity between input text and edited image. Specifically, *sim* is computed by cosine distance between image features and text features, which are encoded by our aligned image and text encoders.

In this experiment, we evaluate the performance of SISGAN [7] (64 $\times$ 64 resolution), TAGAN [14] (128 $\times$ 128 resolution), SIMGAN [23], ManiGAN [8], L-ManiGAN [24], LDEdit [22] and the proposed model on the CUB and Oxford-102 test sets with these six metrics. For each test image, a text description is randomly selected from the whole text collections, and 50k-pair tests are conducted for each method. The average quantitative metric scores are shown in Table I and Table II. It is obvious that the proposed method achieves almost the best performance on all of the metrics on the CUB dataset and gains competitive performance on most metrics on Oxford-102 data. The high IS (or IS*) value demonstrates that our model can produce more realistic edited results. CLIPScore evaluates the compatibility between image and

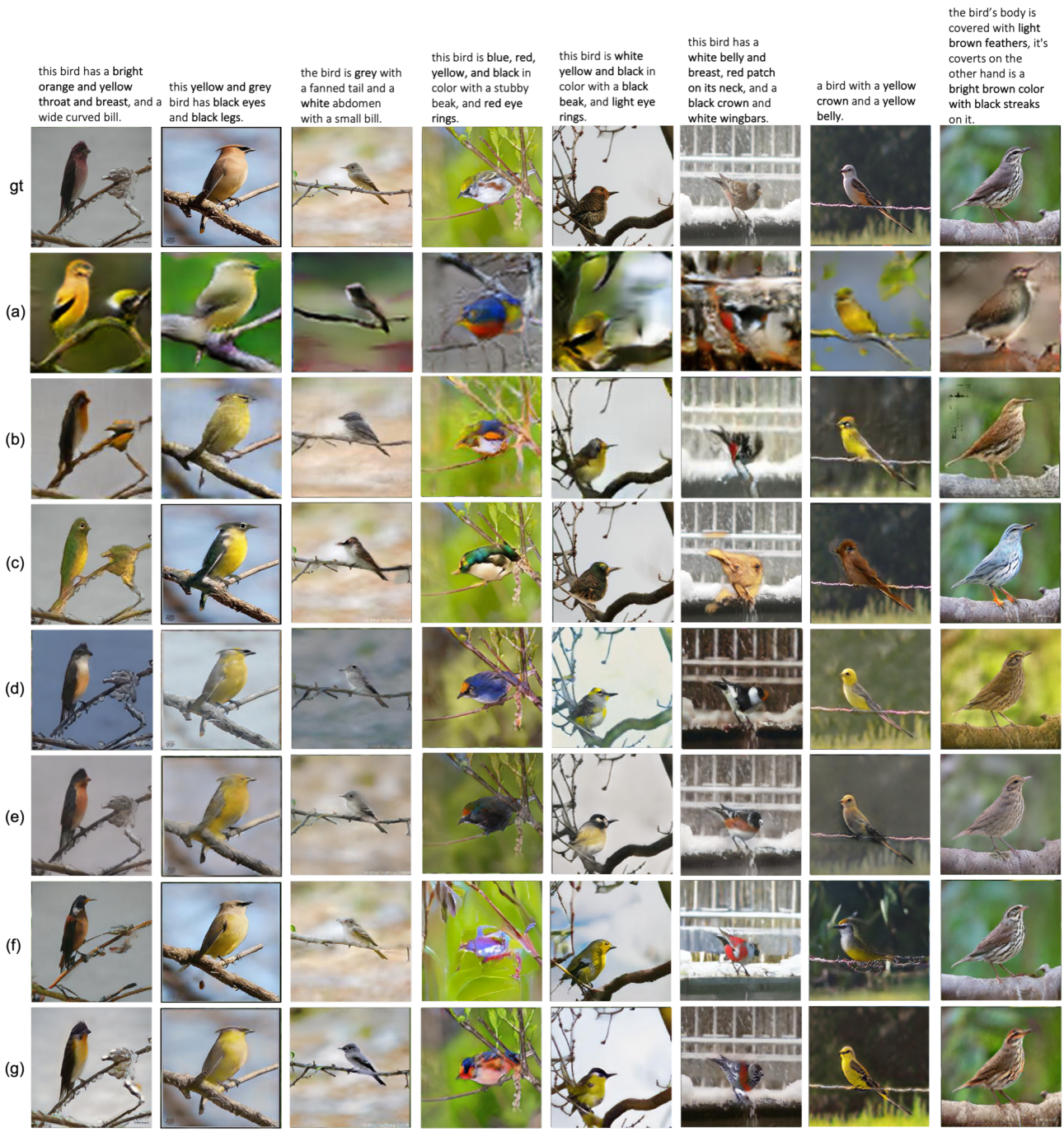

Fig. 5. Qualitative comparison of (a) SISGAN [7](64 × 64), (b) TAGAN [14] (128 × 128), (c) SIMGAN [23], (d) ManiGAN [8], (e) L-ManiGAN [24], (f) LDEdit [22] and (g) the proposed method on CUB dataset.

text, and the highest CLIPScore indicates that the editing results by the proposed method meet the semantic instruction of the given text best. MP metric measures the semantic consistency between the input text and the edited image, and also characterizes the disturbance of text-irrelevant regions. The highest MP value indicates that the proposed model can achieve realistic editing results aligned with the semantic demands of the given text, while preserving the text-irrelevant regions well. The quantitative results validate the effectiveness of the proposed text-driven image editing framework.

In addition, the model complexity and running time are also compared (Table I, #params, init-time and runtime). SISGAN and TAGAN are the lightest models due to its static word embedding approach (which is over 8GB). However, their model loading time-cost (init-time) are both longer than any other methods we compared. SISGAN's performance is far from other methods. TAGAN gets better performance than SISGAN, nevertheless it is unstable due to the generative framework and the network complexity for a higher resolution is restricted by its huge static language model. Compared with

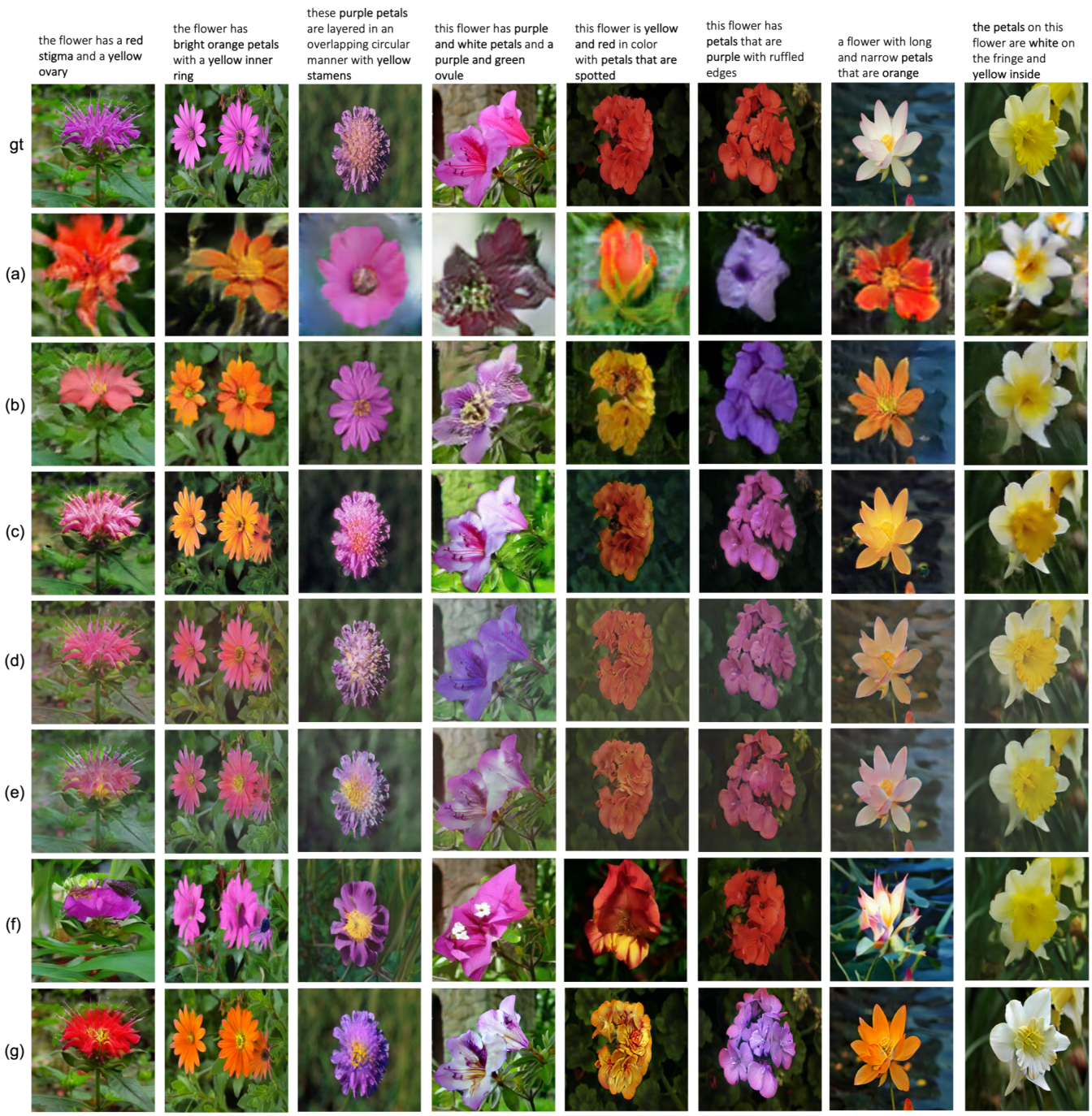

Fig. 6. Qualitative comparison of (a) SISGAN [7](64 × 64), (b) TAGAN [14] (128 × 128), (c) SIMGAN [23], (d) ManiGAN [8], (e) L-ManiGAN [24], (f) LDEdit [22] and (g) the proposed method on Oxford-102 dataset.

SISGAN, ManiGAN can accomplish more meaningful editing and performs much better in quantitative studies. However, ManiGAN suffers from the complex and heavy model caused by the framework that performs text-to-image generation from scratch. Although the lightweight improvement is proposed in L-ManiGAN [24], it is still with ten percent more parameters than the proposed method with almost equivalent init-time to its original version, and cannot perform as well as the proposed method on most of the main metrics. Benefiting from the predicted mask of target region, SIMGAN can preserve background better and get lower *diff* values. However, the low *sim* and CLIPScore demonstrate that SIMGAN suffers from the semantic consistent editing to the given text. Even though the scale of LDM model is reduced a lot, it is still the largest one among these methods, and as a result it is powerful enough at generation to reach highest IS* and reconstruct well with low *diff*. However, it is hard to capture and align details between image and text, resulting in lower MP and CLIPScore than other methods. The runtime of LDEdit costs about 20 times of the proposed method. Among the baselines,

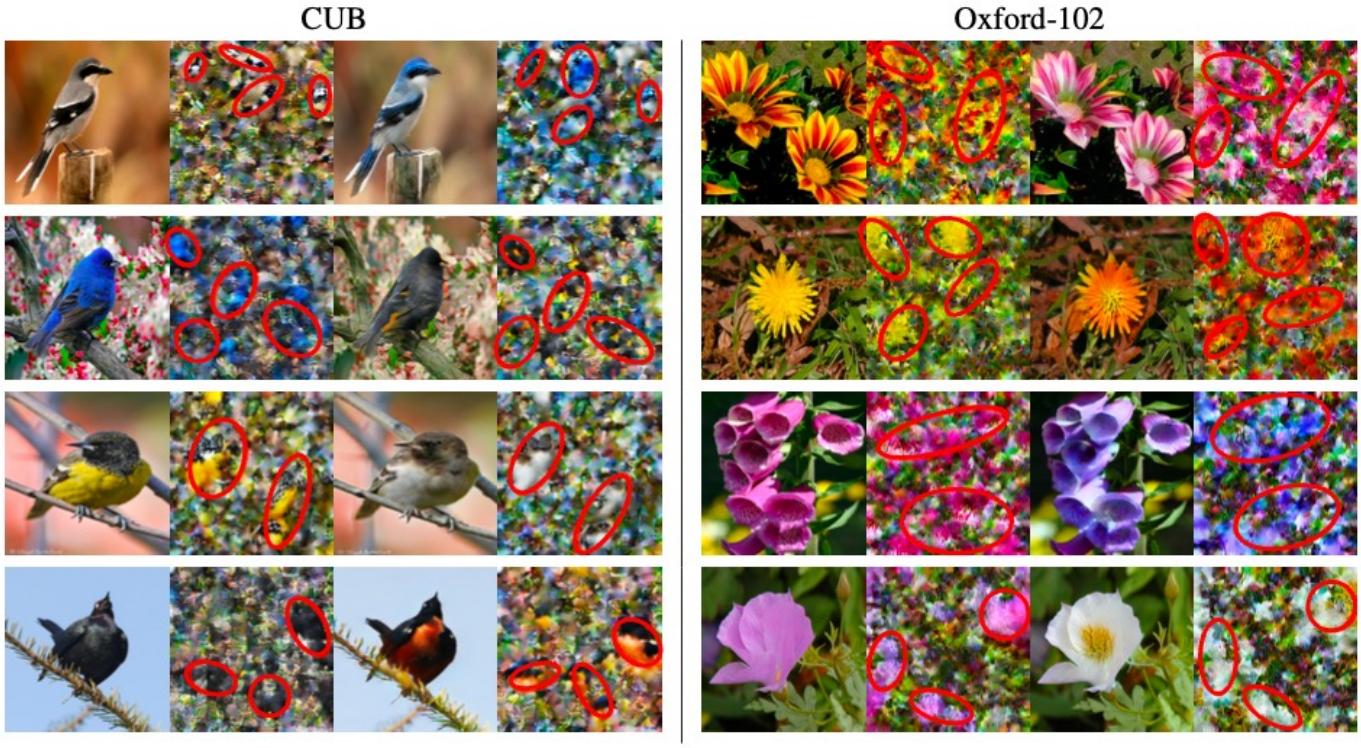

Fig. 7. Attribute disentanglement. The first two columns are the ground truth images and the corresponding attribute features, and the last two columns are the edited results and the corresponding attribute images. The obvious attribute features are marked in red ovals.

the proposed method achieves the fastest init-time and runtime, and gains competitive performance under most metrics with a lightweight model.

Finally, an additional subjective user study is designed to evaluate the performance of different methods. The user study is designed using the 2AFC (Two-Alternative Forced Choice) paradigm. 20 test images randomly selected from CUB and Oxford-102 test sets respectively are involved in the experiment. For each test image, a pair of editing results are randomly shown to the participants. As there are 7 different editing results for each image, 420 clicks are required for each participant. 60 users with age between 15 and 50 were invited to participate in the user study. The distribution of user preference is shown in Fig. 4. We can see that more users prefer the edited results of the proposed method.

*2) **Qualitative comparison**:* In this section, the interactive editing results are evaluated by visual inspection. Fig. 5 and Fig. 6 show the editing results of different methods. As SISGAN just adopted the adversarial learning to automatically learn implicit disentanglement of image and text, numerous artifacts can be found in the editing results as shown in the third and fourth columns of both Fig. 5 and Fig. 6. With the similar framework, TAGAN achieves better performance against SISGAN. However, the generated results still suffer from numerous artifacts especially on the CUB dataset (Fig. 5(b)). Although ManiGAN can generate more meaningful results, it performs poorly at disentangling text-irrelevant regions, and the semantic text-image alignment is not satisfactory. In most of the examples shown in Fig. 5 and Fig. 6, the text-irrelevant background of ManiGAN editing results are incorrectly tangled with the text-relevant features. Although L-ManiGAN improves the performance of ManiGAN to some extent, numerous artifacts and unaligned text-image editing can be found, as shown in the first three columns of Fig. 6. Benefiting from the separation of target instances and background, SIMGAN can keep the text-irrelevant background well in most cases. However, as the performance of SIMGAN highly depends on the accuracy of mask prediction, the method will produce numerous artifacts when the mask is not accurately predicted, such as shown in the first, six and seven columns in Fig. 5. LDEdit is essentially a conditional generation model based on LDM by learning a mapping between two distributions with the guidance of text. Despite the high-quality image generation capability, it falls short for attribute disentanglement and accurate editing of local attributes while keeping text-irrelevant components. For example, the text-irrelevant object shape and background are wrongly edited in most cases in Fig. 5 and Fig. 6. Compared with the state-of-the-art methods, the proposed model can disentangle text-irrelevant features well and achieve better performance at text-image consistency as shown in the last row.

TABLE III
QUANTITATIVE COMPARISON OF CONTENT PRESERVATION.

	L1-distance between content features		
Dataset	$\mathcal{L}(a,b)$	$\mathcal{L}(b1,b2)$	$\mathcal{L}(a,ref)$
CUB	.06	.05	.21
Oxford-102	.05	.05	.14

Fig. 8. Content disentanglement. (a) ground truth images, (b1) and (b2) are the corresponding editing results by exchanging the attribute features of (a) with reference images (c1) and (c2).

*3) **Disentanglement visualization**:* To evaluate the performance of disentangling strategy, we visualize the learned attribute features before and after the editing process. To eliminate the effect of content features, a random noise is sampled to substitute the content, and the visualization of the attribute is generated by combining the noise vector and the learned attribute features through the generator network G as shown in Fig. 2(a). The visualization results of the attribute features before and after attribute editing are respectively shown in the second and fourth columns in Fig. 7. The key regions of the attribute are marked by red ovals, and they can be clearly observed corresponding to the image parts. By comparing the attribute images before and after editing by text, the key regions are almost consistent on both the size and shape, while only the texture and color are changed. This phenomenon implies that the text features learned by our strategy are aligned well with the visual attribute features. In other words, the attribute encoder is correctly trained to learn the text-relevant features. From the final edited images shown in the fourth column, we can also find that the text-irrelevant features are preserved well in most cases. This phenomenon demonstrates the effectiveness of the disentangling strategy between text-relevant attributes and basic content features.

In order to further evaluate the content-attribute disentangling performance, an experiment is designed to measure the preservation of the content features after editing by exchanging the attribute features with the reference image (Fig. 2(c)). The editing results are shown in Fig. 8. To quantitatively evaluate the content-preserving performance, the distance between content features is computed by $\mathcal{L}(x, y) = \|v_c(x) - v_c(y)\|_1$, where v_c is the output of content encoder E_c. For each image a in the test dataset, the edited result b is generated by exchanging the attribute features with a randomly sampled reference c, and the similarity is computed by $\mathcal{L}(a, b)$. Similarly, to evaluate the content similarity between edited results with two distinct reference images c_1, c_2, the distance $\mathcal{L}(b_1, b_2)$ between the corresponding edited results b_1, b_2 is computed. The average results computed on the test dataset are shown in Table III. Small $\mathcal{L}(a, b)$ indicates the content similarity between the ground truth image and the edited result is relatively high. Through the values of $\mathcal{L}(b_1, b_2)$, we can also find the content features are preserved well after editing with two distinct attributes, which partially validate the effectiveness of the proposed content-attribute disentangling strategy.

*4) **Ablation study on cycle-consistency training**:* An ablation study is designed to verify the effectiveness of cycle-consistency training strategy. In this experiment, the intermediate training results with and without the cyclic training stage are demonstrated in Fig. 9, and the corresponding quantitative metrics are also presented as shown in Table. IV and Fig. 10. In order to evaluate the disentanglement performance of these two methods, we also generate the cyclic reconstruction results for the method without the cycle-consistency strategy by passing editing results through a frozen cyclic process. From the editing results shown in Fig. 9, we can find that the method without the cycle-consistency training step can approximately map the semantic attributes of the text to the editing results with the guidance of the improved $\mathcal{L}_{DAMSM}$ function. However, as there is no pixel-aligned supervision, and the DAMSM function only measures the matching between images and text globally, the editing results cannot be as accurate as the method equipped with the cycle-consistency training strategy. From the cyclic reconstruction experimental results (Fig. 9), the method with the cycle-consistency training strategy can effectively preserve the text-irrelevant contents while reconstructing the original attributes, implying good disentangling performance. Moreover, from the intermediate visual and metric results (Fig. 10), we can find that cycle-consistency training strategy can also help speed up the convergence rate of the training dramatically.

TABLE IV
QUANTITATIVE COMPARISON ON THE PERFORMANCE OF TRAINING PROCESS W/ AND W/O CYCLE-CONSISTENCY TRAINING STRATEGY

	w/ cyc			w/o cyc		
epoch	5	50	150	5	50	150
IS*	10.41	17.79	25.53	5.14	9.85	10.59
MP	0.079	0.393	0.424	0.021	0.308	0.349
sim	0.098	0.493	0.543	0.034	0.457	0.540
diff	0.191	0.202	0.221	0.386	0.326	0.354

C. Ablation study for improved DAMSM

In this section, an experiment is conducted to validate the effectiveness of the improved DAMSM$^+$ loss against the original DAMSM function. The convergence curves (on training set) of DAMSM and DAMSM$^+$ are shown in Fig. 11(a). It is obvious that the proposed DAMSM$^+$ loss converges steadily during 100k iterations, whereas the original DAMSM loss stays at a relevant high level, and cannot converge to the loss as small as the proposed method. The performance of models trained with both losses can also be verified in the

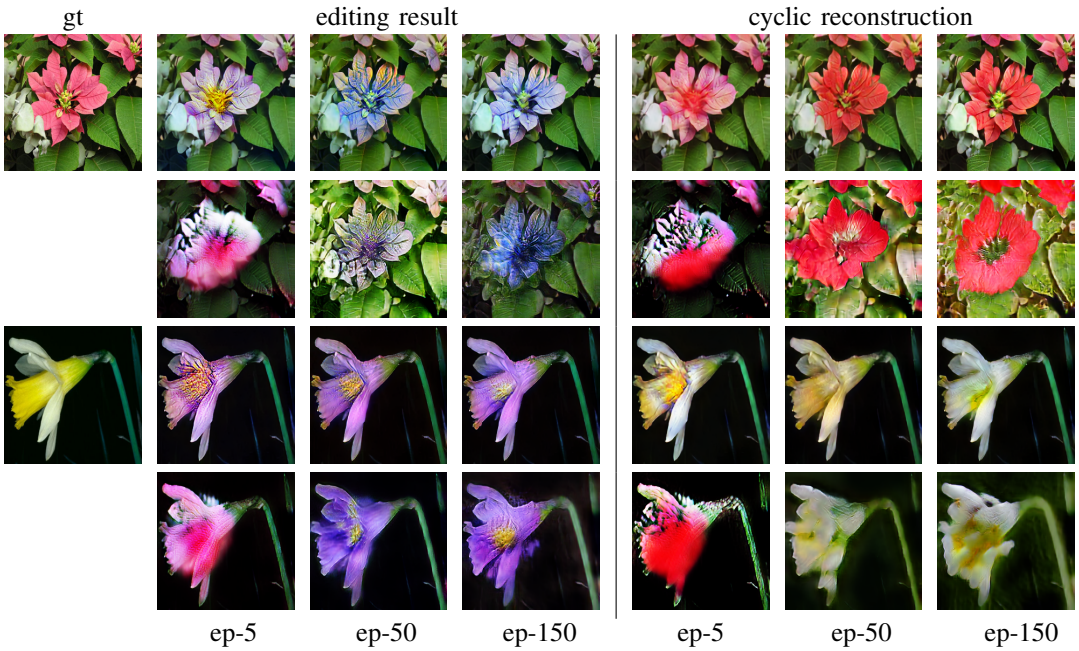

Fig. 9. Ablation study of the cycle-consistency training strategy. For each example, the first row shows the edited or reconstructed results by methods with cycle-consistency training strategy, while the second row are the corresponding results without cyclic training. ep-n represents the output after n training epochs.

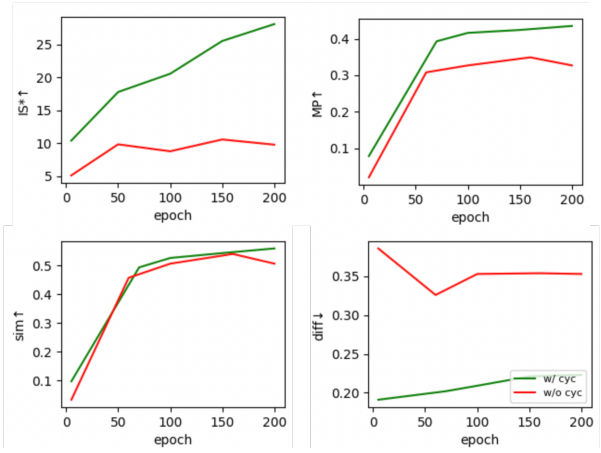

Fig. 10. Tendency comparison on the performance of training process w/ and w/o cycle-consistency training strategy.

visualized editing results in Fig. 11(b). The editing results with the original DAMSM are similar to reconstruction rather than text-related editing. As discussed in section III-B, the original DAMSM is good at estimating the matching-score on a pair of aligned data. However, it does not account for the relationship between two pairs of data. Instead, the proposed DAMSM$^+$ (Fig. 11(b)) can provide a better measurement for unpaired data, and produce more meaningful editing results.

D. Failure cases

There are mainly two types of failure cases: collapse in out-of-scope description and insensitiveness to shape editing. Some failure examples of the proposed method are shown in Fig. 12. The collapse situation (Fig. 12(a)) is caused by the limitation of dataset. The realistic text-image dataset is from the real world, specifically to CUB (bird) and flowers, and their appearances are limited, e.g., flowers in black color are rare. In these cases, the editing process will collapse and randomly choose a semantic feature. The insensitive shape editing (Fig. 12(b)) originates from the fact that more than 80% descriptions are shape-irrelevant in the dataset. In the shape-relevant descriptions, the shape is tangled with color words together, and as a result it will lead to the attribute encoder biased towards color.

V. Conclusion

A novel cycle-consistent text-driven image editing model is proposed in this paper. A disentangling encoding strategy is utilized to improve the controllability of interactive editing, i.e., modifying text-relevant attributes while keeping text-irrelevant image content. A cycle-consistent editing strategy is proposed to enhance the text-image alignment and provide pixel-level supervision. Compared with the complicated T2I framework that most existing works adopted, the proposed framework is a lightweight encoder-decoder architecture, and accomplishes the pixel-aligned end-to-end training. Quantitative and qualitative experimental results indicate that the proposed method outperforms the state-of-the-art methods. However, existing datasets tend to focus on color and texture attributes rather than shapes in text descriptions, and therefore there are still some limitations on editing the shape of a target.

Acknowledgments

The work was funded by Natural Science Foundation of China (NSFC) under Grant 62172198, 61976040, 61762064

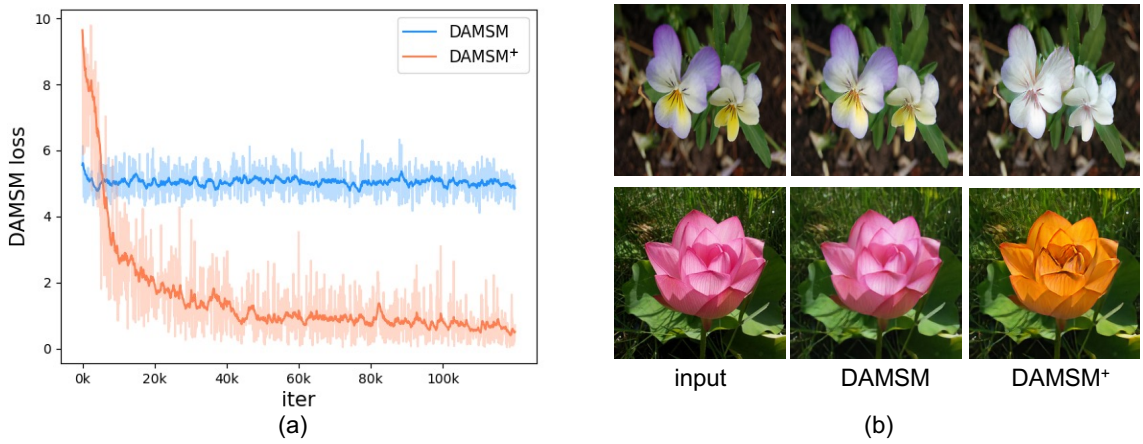

Fig. 11. Ablation study of DAMSM module. (a) loss convergence curves of the original DAMSM (blue) and the proposed DAMSM$^+$ (orange). (b) editing results of the models trained with the two different losses. Text descriptions are 1st-row: "flower with white petals", 2nd-row: "this flower with orange petals".

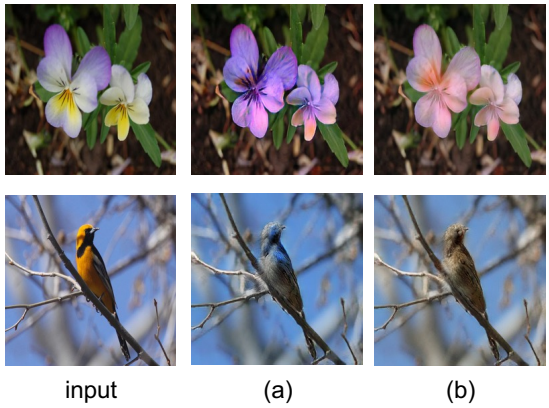

Fig. 12. Failure cases. (a) out of scope collapsing, (b) insensitive to shape editing. Input descriptions: (a) flower: this flower's petals are black. bird: a bird in purple feather. (b) flower: the flower has round petals. bird: the bird body is round.

and 62041604, Key Project of Jiangxi Natural Science Foundation 20224ACB202008, and the Opening Project of Nanchang Innovation Institute, Peking University.

REFERENCES

[1] Y. Jo and J. Park, "Sc-fegan: face editing generative adversarial network with user's sketch and color," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1745–1753.

[2] S. Yang, Z. Wang, J. Liu, and Z. Guo, "Deep plastic surgery: Robust and controllable image editing with human-drawn sketches," in European Conference on Computer Vision. Springer, 2020, pp. 601–617.

[3] K. Olszewski, D. Ceylan, J. Xing, J. Echevarria, Z. Chen, W. Chen, and H. Li, "Intuitive, interactive beard and hair synthesis with generative models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7446–7456.

[4] C. Deng, X. Xu, H. Wang, M. Yang, and D. Tao, "Progressive cross-modal semantic network for zero-shot sketch-based image retrieval," IEEE Transactions on Image Processing, vol. 29, pp. 8892–8902, 2020.

[5] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, "Drit++: Diverse image-to-image translation via disentangled representations," International Journal of Computer Vision, vol. 128, no. 10, pp. 2402–2417, 2020.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

[7] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic image synthesis via adversarial learning," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5706–5714.

[8] B. Li, X. Qi, T. Lukasiewicz, and P. H. Torr, "Manigan: Text-guided image manipulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7880–7889.

[9] D. Zhu, A. Mogadala, and D. Klakow, "Image manipulation with natural language using two-sided attentive conditional generative adversarial network," Neural Networks, vol. 136, pp. 207–217, 2021.

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

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

[12] B. Li, X. Qi, T. Lukasiewicz, and P. H. Torr, "Controllable text-to-image generation," in Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 2065–2075.

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

[14] S. Nam, Y. Kim, and S. J. Kim, "Text-adaptive generative adversarial networks: manipulating images with natural language," Advances in neural information processing systems, vol. 31, 2018.

[15] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12 873–12 883.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851, 2020.

[17] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "Vqgan-clip: Open domain image generation and editing with natural language guidance," in European Conference on Computer Vision. Springer, 2022, pp. 88–105.

[18] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in International Conference on Machine Learning. PMLR, 2021, pp. 8821–8831.

[19] 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.

[20] 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., "Photoreal-

istic text-to-image diffusion models with deep language understanding," arXiv preprint arXiv:2205.11487, 2022.
[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 684–10 695.
[22] P. Chandramouli and K. V. Gandikota, "Ldedit: Towards generalized text guided image manipulation via latent diffusion models," arXiv preprint arXiv:2210.02249, 2022.
[23] L. Gao, Q. Zhao, J. Zhu, S. Su, L. Cheng, and L. Zhao, "From external to internal: Structuring image for text-to-image attributes manipulation," IEEE Transactions on Multimedia, 2022.
[24] B. Li, X. Qi, P. Torr, and T. Lukasiewicz, "Lightweight generative adversarial networks for text-guided image manipulation," Advances in Neural Information Processing Systems, vol. 33, 2020.
[25] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in International Conference on Machine Learning. PMLR, 2016, pp. 1060–1069.
[26] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014.
[27] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, "Tac-gan-text conditioned auxiliary classifier generative adversarial network," arXiv preprint arXiv:1703.06412, 2017.
[28] X. Huang, M. Wang, and M. Gong, "Hierarchically-fused generative adversarial network for text to realistic image synthesis," in 2019 16th Conference on Computer and Robot Vision (CRV). IEEE, 2019, pp. 73–80.
[29] W. Huang, R. Y. Da Xu, and I. Oppermann, "Realistic image generation using region-phrase attention," in Asian Conference on Machine Learning. PMLR, 2019, pp. 284–299.
[30] H. Tan, X. Liu, X. Li, Y. Zhang, and B. Yin, "Semantics-enhanced adversarial nets for text-to-image synthesis," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10 501–10 510.
[31] Y. Yang, L. Wang, D. Xie, C. Deng, and D. Tao, "Multi-sentence auxiliary adversarial networks for fine-grained text-to-image synthesis," IEEE Transactions on Image Processing, vol. 30, pp. 2798–2809, 2021.
[32] H. Tan, X. Liu, M. Liu, B. Yin, and X. Li, "Kt-gan: knowledge-transfer generative adversarial network for text-to-image synthesis," IEEE Transactions on Image Processing, vol. 30, pp. 1275–1290, 2020.
[33] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 1–26, 2021.
[34] H. Dhamo, A. Farshad, I. Laina, N. Navab, G. D. Hager, F. Tombari, and C. Rupprecht, "Semantic image manipulation using scene graphs," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5213–5222.
[35] S. Su, L. Gao, J. Zhu, J. Shao, and J. Song, "Fully functional image manipulation using scene graphs in a bounding-box free way," in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1784–1792.
[36] J. Song, J. Zhang, L. Gao, Z. Zhao, and H. T. Shen, "Agegan++: Face aging and rejuvenation with dual conditional gans," IEEE Transactions on Multimedia, vol. 24, pp. 791–804, 2021.
[37] F. Wu, L. Liu, F. Hao, F. He, and J. Cheng, "Language-based image manipulation built on language-guided ranking," IEEE Transactions on Multimedia, 2022.
[38] C. Yan, X. Chang, Z. Li, W. Guan, Z. Ge, L. Zhu, and Q. Zheng, "ZeroNAS: Differentiable generative adversarial networks search for zero-shot learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 9733–9740, 2021.
[39] J. Zhang, J. Song, L. Gao, Y. Liu, and H. T. Shen, "Progressive meta-learning with curriculum," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 9, pp. 5916–5930, 2022.
[40] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," arXiv preprint arXiv:2103.17249, 2021.
[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," arXiv preprint arXiv:2103.00020, 2021.
[42] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18 208–18 218.
[43] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2426–2435.
[44] X. Liu, Z. Lin, J. Zhang, H. Zhao, Q. Tran, X. Wang, and H. Li, "Open-edit: Open-domain image manipulation with open-vocabulary instructions," in European Conference on Computer Vision. Springer, 2020, pp. 89–106.
[45] Y. Liu, M. De Nadai, D. Cai, H. Li, X. Alameda-Pineda, N. Sebe, and B. Lepri, "Describe what to change: A text-guided unsupervised image-to-image translation approach," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1357–1365.
[46] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.
[47] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. IEEE, 2008, pp. 722–729.
[48] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," Advances in neural information processing systems, vol. 29, pp. 2234–2242, 2016.
[49] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," arXiv preprint arXiv:2104.08718, 2021.
[50] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.