Each Attribute Matters: Contrastive Attention for Sentence-based Image Editing

Liuqing Zhao *[1]
liuqingzhao@post.usts.edu.cn
Fan Lyu *[2]
fanlyu@tju.edu.cn
Fuyuan Hu [1]
fuyuanhu@mail.usts.edu.cn
Kaizhu Huang [3]
kaizhu.huang@xjtlu.edu.cn
Fenglei Xu [1]
xufl@mail.usts.edu.cn
Linyan Li †[4]
lilinyan@szjm.edu.cn

[1] Suzhou University of
Science and Technology,
Suzhou, China
[2] College of Intelligence and Computing,
Tianjin University
Tianjin, China
[3] Xi'an Jiaotong-Liverpool University,
Suzhou, China
[4] Suzhou Institute of Trade and Commerce,
Suzhou, China

*L.Zhao and F.Lyu share equal contribution.
† L.Li is the corresponding author.

Abstract

Sentence-based Image Editing (SIE) aims to deploy natural language to edit an image. Offering potentials to reduce expensive manual editing, SIE has attracted much interest recently. However, existing methods can hardly produce accurate editing and even lead to failures in attribute editing when the query sentence is with multiple editable attributes. To cope with this problem, by focusing on enhancing the difference between attributes, this paper proposes a novel model called Contrastive Attention Generative Adversarial Network (CA-GAN), which is inspired from contrastive training. Specifically, we first design a novel contrastive attention module to enlarge the editing difference between random combinations of attributes which are formed during training. We then construct an attribute discriminator to ensure effective editing on each attribute. A series of experiments show that our method can generate very encouraging results in sentence-based image editing with multiple attributes on CUB and COCO dataset. Our code is available at https://github.com/Zlq2021/CA-GAN

1 Introduction

As billions of images are uploaded and shared every day [16, 32], image editing has become one of the most demanding tasks in social media. However, to edit an image as desired, one may have to master professional software such as Adobe PhotoShop. In contrast with manual editing, automatic image editing, has recently attracted much interest in computer vision. This paper studies the problem of Sentence-based Image Editing (SIE) [7, 19, 25] that intends to deploy natural language to assist image editing automatically. One main

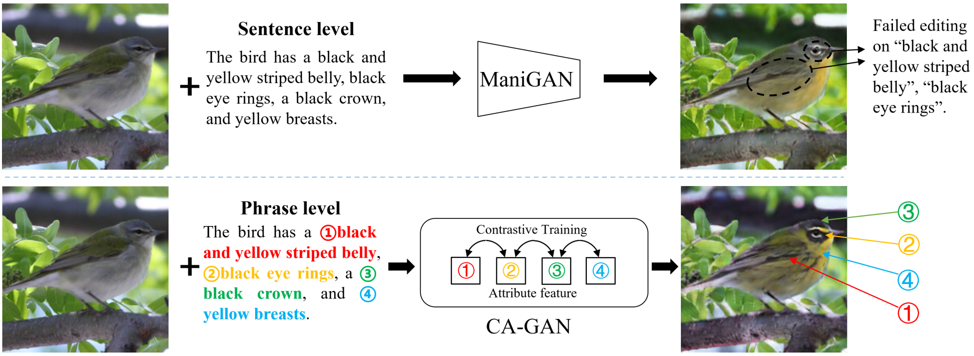
Sentence level
The bird has a black and yellow striped belly, black eye rings, a black crown, and yellow breasts.
+
ManiGAN
Failed editing on “black and yellow striped belly”, “black eye rings”.
Phrase level
The bird has a ①black and yellow striped belly, ②black eye rings, a ③ black crown, and ④ yellow breasts.
+
Contrastive Training
① ② ③ ④
Attribute feature
CA-GAN
③
②
④
①

Figure 1: Comparisons between the existing sentence-level editing and the proposed Contrastive Attention GAN. The state-of-the-art ManiGAN [19] cannot effectively parse different attributes from the given sentence, which yields the failed attribute editing on “black eye rings”. The proposed CA-GAN parses the sentence, learns to distinguish attributes from each other and edits successfully on all attributes.

challenge for SIE is to build the cross-modal mapping from the query sentence to the pixels in image. In the last decade, Deep Neural Networks [23, 24] enabling generative models to produce pixel-level manipulation from another image have become the main solution to SIE.

Based on Generative Adversarial Networks (GANs) [17, 30, 33, 39, 46, 48], recent works on SIE focus on combining sentence and image information. For example, AttnGAN [37] maps the query sentence and the image to be edited into a shared hidden space and minimizes the multi-modal similarity to improve the quality of text-to-image generation. TAGAN [25] provides word-level feedback to the generator through a fine-grained text discriminator. ManiGAN [19] combines language and image with a three-stage network structure, which progressively generates images from three different scales. These methods show impressive results with short query sentence or phrase.

Nevertheless, when the query sentence is long and contains multiple attributes to be edited, the existing methods can hardly produce effective editing for all attributes. As a typical example shown in Fig. 1, a query sentence “*The bird has a black and yellow striped belly, black eye rings, a black crown, and yellow breasts*” is used to guide the editing on a given bird image. Intuitively, the sentence has a few different editable attributes: “*black and yellow striped belly*”, “*black eye rings*”, “*black crown*” and “*yellow breasts*”. The current state-of-the-art ManiGAN [19] fails to edit the attribute “*black eye rings*”. The main reason for the failure is that the existing methods only focus on sentence-level editing rather than each attribute. Going further, we argue that there are three main obstacles for these GAN-based sentence-level editing methods: 1) they cannot parse sentences effectively and the attribute differences are indistinguishable; 2) they cannot build the attribute-pixel correspondence properly; 3) the sentence-level discriminator utilised in these methods is limited in detecting the failed attribute editing.

To tackle these drawbacks, we aim to strengthen each editable attribute so as to attain an accurate SIE model. Concretely, inspired by the contrastive training, we propose a novel Contrastive Attention Generative Adversarial Network (CA-GAN) for SIE. Our proposed CA-GAN contains three main components: 1) **Sentence Parsing and Attribute Combination**. To facilitate the training for the attribute-level editing, we first parse the query sentence based on POS Tagging to ensure the attribute-object correspondence. Then we augment the query space by random attribute combinations, which prove to significantly highlight attribute-level information. 2) **Contrastive Training using attention**. Intuitively, based on the augmentation from random attribute combinations, different combinations yield different editing. Thus, we construct the Contrastive Attention for different combinations in the GAN architecture. Our model can then enlarge the editing difference between any two attribute combinations, whilst keeping the background invariant. 3) **Attribute-level Discriminator**. In the discriminator, we build an attribute-level discriminator for providing effective editing feedback on each attribute to the generator. With the proposed CA-GAN, the editable attributes in the sentence can be well distinguished via training. Thus an effective SIE model can be generated which emphasizes each attribute appropriately. To the best of our knowledge, this is the first work that proposes to separate attributes from long sentences to strengthen the attribute editing. We evaluate the proposed CA-GAN on two benchmark image editing datasets, *i.e.*, CUB and MS-COCO. The evaluation results show that our method can edit the attributes at the pixel level effectively and accurately.

2 Related Work

Sentence-based image editing. In recent years, based on Generative Adversarial Networks (GANs) [17, 30, 33, 39, 46, 48], researchers pay much attention to the image generation or transformation from text or image, such as Text-to-Image Generation [26, 31, 37, 42, 43] and Image-to-Image Translation [12, 29, 38, 40]. To make the transformation controllable, Text-based Image Editing will only edit the target area of the image through text description. Generally, the query text can be a word, a short phrase or a long sentence. Dong *et al.* proposed an encoder-decoder structure to edit images matched with a given text [7]. In order to keep the content irrelevant to the text in the original image, [34] proposed to construct foreground and background distribution with different recognizers. Nam *et al.* eliminated different visual attributes by introducing a text adaptive discriminator, which can provide more detailed training feedback to the generator [25]. Li *et al.* adopted the structure of the multi-level network, and could generate high-quality image content through the combination module of ACM and DCM [19]. However, the generators of these methods ignore the difference between long sentences with words and phrases that may contain multiple editable attributes as well. In this paper, we focus on the Sentence-based Image Editing, and propose to construct Contrastive Attention to enhance the attribute editing.

Contrastive training. For a given anchor point in the data, the purpose of contrast learning [4, 9, 41] is to bring the anchor point closer to the positive point and push the anchor point further away to the negative point in the representation space, thus enhancing the consistency of the feature representation. In previous vision tasks [5, 14, 18, 36], the idea of contrast learning is also applied by exploring the relationship between positive and negative samples. It was also demonstrated in [28] that contrast learning methods are effective in the task of image to image conversion. Some works also studied the contrastive training in natural language processing [9, 45]. CDL-GAN [47] add Consistent Contrastive Distance (CoCD) and Characteristic Contrastive Distance (ChCD) into a principled framework to improve GAN performance. CERT [8] uses back-translation for data augmentation. BERT-CT [3] uses two individual encoders for contrastive learning. In this work, we argue that different attributes should be edited discriminatively and this leads to the idea of contrast attention on SIE.

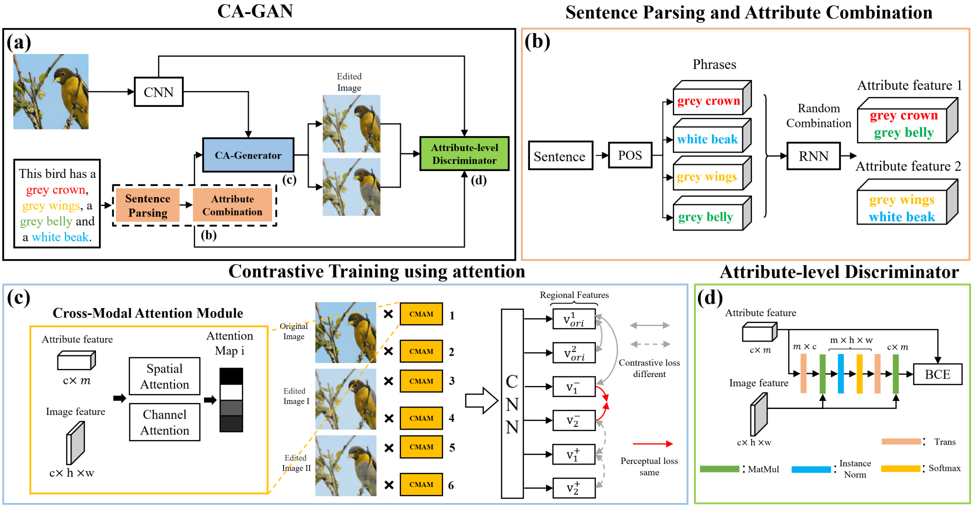

Figure 2: (a) Overview of the proposed CA-GAN. Sentence information is passed through (b), segmentation of editable attributes by sentence analysis, random combination for data augmentation. (c) Contrastive Training using attention module. The contrastive training between different attributes is constructed by editing images with different combinations of attributes. CMAM module outputs attention maps for different attributes. (d) Attribute-level Discriminator, which provides attribute-level feedback to the generator.

3 Method

3.1 Overview

Given an image $\mathbf{I} \in \mathbb{R}^{c \times h \times w}$ and a query sentence $\mathcal{S}$, SIE aims to transform $\mathbf{I}$ guided by $\mathcal{S}$ to an edited image $\hat{\mathbf{I}}$. Our Contrastive Attention Generative Adversarial Network (CA-GAN) is based on the popular three-stage editing architecture [37]. To be more specific, there are usually three stages in the main module, and each stage contains a generator and a discriminator. Three stages are trained at the same time, and progressively generate images of three different scales, i.e., $64^2 \rightarrow 128^2 \rightarrow 256^2$. As shown in Fig. 2, CA-GAN contains three main components: (1) **Sentence Parsing and Attribute Combination** (Fig. 2(b)). The query sentence is parsed to multiple editable attributes based on a Lexical rules based on POS tagging [1]. Then, the attributes are randomly combined into two groups for augmentation. (2) **Contrastive Training using attention** (Fig. 2(c)). This module uses attention distinguishes different combinations, and each attribute can be learned. (3) **Attribute-level Discriminator** (Fig. 2(d)). This module is designed to provide the feedback if an attribute is edited well. We will elaborate these steps in the following.

3.2 Sentence Parsing and Attribute Combination

Some off-the-shelf methods can parse a sentence to multiple phrases, such as Topicrank [2] and Sentence Transformers [27] and BERT [6]. However, these methods suffer from two main problem: 1) task-specific design and not for image editing; 2) large-scare network with massive parameters. In contrast, we propose a parsing rule to effectively parse attributes from a sentence and perform in lightweight scale. Specifically, given a sentence $\mathcal{S}$, as shown

Table 1: Comparisons of attribute extraction between the present methods and ours.

Query Sentence	a grey bird with webbed feet, a short and blunt orange bill, grey head and wings and has white eyes, a white stripe behind its eyes and white belly and breast	the bird is black with a white belly and an orange bill
Method	**Attribute Extraction**	**Attribute Extraction**
Transformers: [27]	[‘grey bird’, ‘feet’, ‘short’, ‘blunt orange bill’, ‘grey head’, ‘wings’, ‘white eyes’, ‘white stripe’, ‘eyes’, ‘white belly’, ‘breast’]	[‘bird’, ‘black’, ‘white belly’, ‘orange bill’]
Topicrank: [2]	[‘grey bird with webbed feet’, ‘blunt orange bill’, ‘grey head’, ‘wings’, ‘white eyes’, ‘white stripe’, ‘eyes’, ‘breast’]	[‘bird’, ‘orange bill’]
Spacy: [11]	[‘orange bill’, ‘grey head wings’, ‘white belly’, ‘white eyes’, ‘white stripe’]	[‘white belly’, ‘orange bill’]
Ours	[‘grey bird’, ‘webbed feet’, ‘short blunt orange bill’, ‘grey head wings’, ‘white eyes’, ‘white stripe eyes’, ‘white belly breast’]	[‘bird black’, ‘white belly’, ‘orange bill’]

in Fig. 2(c), we first use POS tagging [1] (NLTK [22]) to label each word in a sentence with a lexical property (*e.g.* noun, adjective). With the lexical property, we need to further separate the attributes from the sentences in the form of “adjective-noun”. But it is difficult to determine the adjectives belongingness of two neighboring attributes in the sentence. This can be illustrated in the examples “*bird with a black wing*”→“[*bird black*], [*wing*]”, “*yellow belly and wings*”→“[*yellow belly*], [*wings*]”, where the adjective word is wrongly categorized. To cope with these problems, we leverage two state values $f_1, f_2 \in \{0,1\}$ to assist the attribute separation from sentence based on the “noun-adjective” rule. We have $f_1 = 1$ for noun word and $f_2 = 1$ for adjective word. If the next word of a noun is “*has*” and “ *with*”, then $f_1 = 0$. If the word is a conjunction and the previous word and the next word are both adjectives, then $f_2 = 0$. If and only if $f_1 * f_2 = 1$, all nouns and adjectives that have been traversed are classified as one attribute. By this simple rule, we can effectively divide $\mathcal{S}$ into M attributes, *i.e.*, $\hat{\mathcal{S}} = \{\mathcal{A}_1, \cdots, \mathcal{A}_M\}$. We compare several sentence parsing methods with the proposed rules with several multi-attribute sentence. As shown in Tab. 1, the parsing results of our work can effectively extract editable attributes against other related methods.

In the real world, the attribute distribution are highly imbalanced. For example, in CUB dataset [35], the attribute about “belly” appears 34,899 times, but the attribute about “eye-ring” only appears 3,125 times. This phenomenon leads to the poor editing on some kinds of attributes with fewer number. To this end, we propose to combine attributes randomly to augment data. Specifically, in the training phase, we randomly combine attributes from $\mathcal{S}$ to build $\hat{\mathcal{S}}_1 = \{\mathcal{A}_i\}_{i \in \mathcal{C}_1}$ and $\hat{\mathcal{S}}_2 = \{\mathcal{A}_i\}_{i \in \mathcal{C}_2}$, where $\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, \cdots, M\}$. By randomly combining attributes, we obtain more editing alternatives and each attribute will be learned without the limit of distribution imbalance. To further study the attribute-specific editing, we design a *contrastive training* strategy to make each kind of attribute trained effectively.

3.3 Contrastive Training using attention

In a sentence $\mathcal{S}$, using different attribute combinations to edit image will yield differnt results. Contrastive training [9] aims to learn a representation to pull “positive" pairs in certain metric space and push apart the representation between “negative" pairs. That means, based on this observation, we can impose contrastive training on the network to control the editing difference between two attribute combinations. We implement our design based on the famous AttenGAN [37], which calculates the spatial attention of the image w.r.t each word for text-to-image generation. However, the proposed CA-GAN is quite different from AttenGAN because of the difference on how to construct attention. In specific, we construct Contrastive Attention for different attribute combinations in the proposed CA-GAN, and with the contrastive training, each attribute editing can be enhanced.

The generator of CA-GAN (See Fig. 2) has two inputs, image feature $\mathbf{v}$ by CNN and attribute combination features $\mathbf{s}_1$ and $\mathbf{s}_2$ by RNN. We construct cross-modal attention matrix $\mathbf{C} \in \mathbb{R}^{c \times hw}$ using Cross-Modal Attention Module (CMAM) a as shown in Fig. 2(e), where $\mathbf{C}_{i,j}$ is calculated as follows:

$$\mathbf{C}_{i,j} = \frac{\exp(\mathbf{s}_j^\top \mathbf{v}_i)}{\sum_k \exp(\mathbf{s}_k^\top \mathbf{v}_i)}, \tag{1}$$

where $i \in \{1, \cdots, c\}$ is the channel index, $j \in \{1,2\}$ is the attribute combination index. Then, we can easy to obtain two attention maps w.r.t the attribute combination $\mathcal{S}_1$ and $\mathcal{S}_2$ by

$$\hat{\mathbf{C}}_j = \mathbf{s}^\top \mathbf{C}, \forall j \in \{1,2\}. \tag{2}$$

The Contrastive Attention contains both spatial and channel attention map. Because the attention map represents the area of the image to be edited by attributes, with the attention matrix, we can get the attended feature from different attribute combination attention maps by the means of Hadamard Product. Thus, given $\mathbf{s}_1$ and $\mathbf{s}_2$, we can easily get six kinds of attention-image pairs using Eq. (1) with the original image $\mathbf{I}$, the edited image $\hat{\mathbf{I}}_1$ (from combination $\hat{\mathcal{S}}_1$) and $\hat{\mathbf{I}}_2$ (from combination $\hat{\mathcal{S}}_2$). This can be seen in Fig. 2(b). The six pairs are denoted as

(1) $\mathbf{I}_1^+ = \hat{\mathbf{I}}_1 \times \hat{\mathbf{C}}_1$: **positive** sample for the **first** editing attribute combination;
(2) $\mathbf{I}_1^- = \hat{\mathbf{I}}_1 \times \hat{\mathbf{C}}_2$: **negative** sample for the **first** editing attribute combination;
(3) $\mathbf{I}_2^+ = \hat{\mathbf{I}}_2 \times \hat{\mathbf{C}}_2$: **positive** sample for the **second** editing attribute combination;
(4) $\mathbf{I}_2^- = \hat{\mathbf{I}}_2 \times \hat{\mathbf{C}}_1$: **negative** sample for the **second** editing attribute combination;
(5) $\mathbf{I}_{\text{ori}}^1 = \mathbf{I} \times \hat{\mathbf{C}}_1$: editing areas for the **first** attribute combination on the original image;
(6) $\mathbf{I}_{\text{ori}}^2 = \mathbf{I} \times \hat{\mathbf{C}}_2$: editing areas for the **second** attribute combination on the original image.

The six attended images are fed to a pretrained vgg-16 and get features $\mathbf{v}_1^+$, $\mathbf{v}_1^-$, $\mathbf{v}_2^+$, $\mathbf{v}_2^-$, $\mathbf{v}_{\text{ori}}^1$ and $\mathbf{v}_{\text{ori}}^2$. Then, we construct the contrastive loss between different features in the image:

$$\mathcal{L}_{\text{diff}} = -\log \frac{\exp(\cos(\mathbf{v}_1^-, \mathbf{v}_{\text{ori}}^1))}{\sum_{p=1}^N \exp(\cos(\mathbf{v}_2^+, \mathbf{v}_{\text{ori}}^1)} - \log \frac{\exp(\cos(\mathbf{v}_2^-, \mathbf{v}_{\text{ori}}^1))}{\sum_{p=1}^N \exp(\cos(\mathbf{v}_1^+, \mathbf{v}_{\text{ori}}^1)}. \tag{3}$$

Through contrastive training, the generator can learn the distribution of each attribute, and establish an accurate association between attribute and image. In addition, to preserve text-independent background regions, we build the perceptual loss [13] to reduce the randomness in the generation process as

$$\mathcal{L}_{\text{per}} = \frac{1}{c \times h \times w} ||\mathbf{v}_1^- - \mathbf{v}_2^-||_2^2. \tag{4}$$

3.4 Attribute-level Discriminator

To encourage generators to edit multiple attributes based on sentences, the discriminator should provide attribute-level training feedback to the generator. Previous work attempted to use sentence-level discriminator [7] or word-level discriminator [19, 25], but they cannot establish an exact connection between the image area and each attribute. For instance, in the sentence of "*the bird has black wings, a black head and a red belly*", when the "*black*" attribute is passed through the discriminator, sentence-level discriminators do not provide the exact area of the feature in the image, and word-level discriminators localize to both “*head*" and “*wing*" regions. In order for the discriminator to provide feedback related to each attribute, we propose to develop the attribute-level discriminator.

Our attribute-level discriminator has two inputs, attribute combination features $\mathbf{s}$ and image feature $\mathbf{v}$. We use $\mathbf{\Delta}_j \in \mathbb{R}^{c\times 2}$ represents the correlation between the j^{th} ($j \in \{1,2\}$) attribute combination and the whole image

$$\mathbf{\Delta}_j = \frac{\exp((\mathbf{s}_j^\top \mathbf{v}_i)^\top \mathbf{v}_i)}{\sum_{k=1}^{L} \exp((\mathbf{s}_j^\top \mathbf{v}_i)^\top \mathbf{v}_i)}. \tag{5}$$

Next, we sum the attribute-weighted image feature $\mathbf{\Delta}_j$ at the C dimension to get $\hat{\mathbf{\Delta}}_j$.

Finally, the attribute-level feedback between $\mathbf{s}$ and $\hat{\mathbf{\Delta}_j}$ is calculated by Binary Cross-Entropy (BCE) loss as

$$\mathcal{L}_{\text{attr}} = \sum_j \text{BCE}(\mathbf{s}, \hat{\mathbf{\Delta}}_{\mathbf{j}}). \tag{6}$$

By calculating BCE loss, the discriminator is able to provide attribute-level training feedback to the generator, thus benefiting the alignment between the different attribute features and visual features in the sentence.

3.5 Objective Function

The generator and discriminator are trained alternatively by minimizing both the generator loss $\mathcal{L}_{\text{G}}$ and discriminator loss $\mathcal{L}_{\text{D}}$. The generator of the whole network contains unconditional adversarial loss and conditional adversarial loss, contrastive loss $\mathcal{L}_{\text{diff}}$, perceptual loss $\mathcal{L}_{\text{per}}$ and text-image matching loss $\mathcal{L}_{\text{DAMSM}}$.

$$\mathcal{L}_{\text{G}} = -\frac{1}{2} E_{\hat{\mathbf{I}}\sim P_{\text{G}}}[\log(\text{D}(\hat{\mathbf{I}}))] - \frac{1}{2}\mathbb{E}_{\hat{\mathbf{I}}\sim P_G}[\log(\text{D}(\hat{\mathbf{I}}, \mathcal{S}))] + \lambda_1 \mathcal{L}_{\text{diff}} + \lambda_2 \mathcal{L}_{\text{per}} + \lambda_3 \mathcal{L}_{\text{DAMSM}}, \tag{7}$$

where $\mathbf{I}$ is the real image sampled from the original image distribution, and $\hat{\mathbf{I}}$ is the generated image sampled from the training model distribution, λ_1, λ_2 and λ_3 are hyperparameters controlling different losses. $\mathcal{L}_{\text{DAMSM}} = \frac{\exp(\gamma \mathbf{R}(\mathbf{s},\mathbf{v}))}{\sum_{k=1}^{M} \exp(\gamma \mathbf{R}(\mathbf{s},\mathbf{v}))}$ is used to calculate the matching score between image and text, where $\mathbf{R} = (c_i^T e_i)/(||c_i||||e_i||)$, γ is the smoothing factor, c denotes the picture feature corresponding to the word, and e denotes the feature of the whole sentence. The complete discriminator objective is defined as:

$$\begin{aligned}\mathcal{L}_{\text{D}} = &-\frac{1}{2}\mathbb{E}_{\mathbf{I}\sim \text{P}_{data}}[\log(\text{D}(\mathbf{I}))] - \frac{1}{2}\mathbb{E}_{\hat{\mathbf{I}}\sim P_G}[\log(1-\text{D}(\hat{\mathbf{I}}))] \\ &-\frac{1}{2}\mathbb{E}_{\hat{\mathbf{I}}\sim \text{P}_{\text{G}}}[\log(1-D(\hat{\mathbf{I}}, \mathcal{S}))] - \frac{1}{2}\mathbb{E}_{\hat{\mathbf{I}}\sim \text{P}_{\text{G}}}[\log(\text{D}(\mathbf{I}, \mathcal{S}))] + \lambda_4 \mathcal{L}_{attr},\end{aligned} \tag{8}$$

where λ_4 is the hyperparameters controlling $\mathcal{L}_{attr}$.

4 Experiments

4.1 Dataset and implementation detail

Dataset: Our model is evaluated on the Caltech-UCSD Birds (CUB) [35] and MS COCO [20] datasets, where the query sentences of CUB and COCO are the bird description and

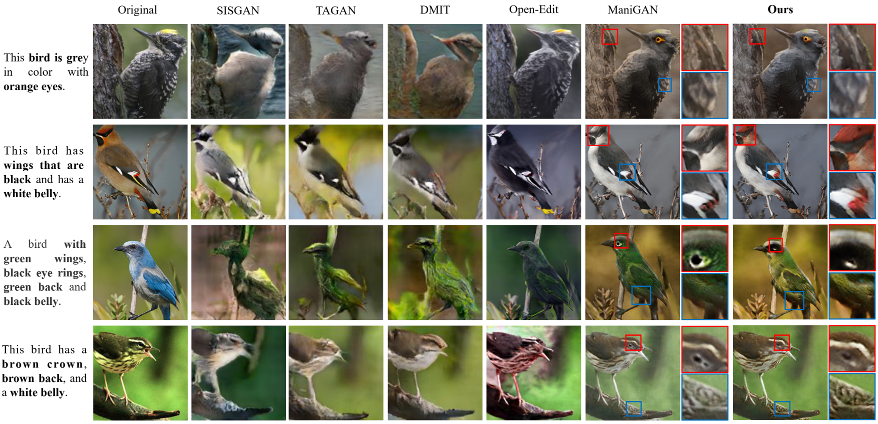
Original
SISGAN
TAGAN
DMIT
Open-Edit
ManiGAN
Ours
This bird is grey in color with orange eyes.
This bird has wings that are black and has a white belly.
A bird with green wings, black eye rings, green back and black belly.
This bird has a brown crown, brown back, and a white belly.

Figure 3: Editing comparisons on CUB datasets.

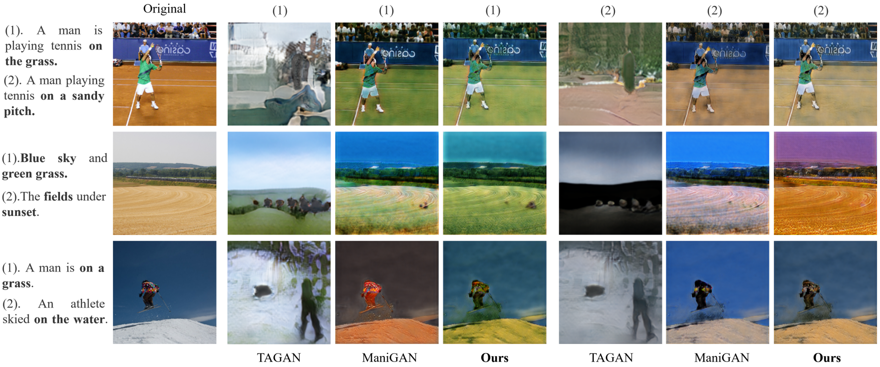
Original
(1)
(1)
(1)
(2)
(2)
(2)
(1). A man is playing tennis on the grass.
(2). A man playing tennis on a sandy pitch.
(1).Blue sky and green grass.
(2).The fields under sunset.
(1). A man is on a grass.
(2). An athlete skied on the water.
TAGAN
ManiGAN
Ours
TAGAN
ManiGAN
Ours

Figure 4: Editing comparisons on COCO datasets.

image captions provided by themself. CUB contains 200 bird species with 11,788 images where each has 10 sentence deceptions. We pre-encode the sentence by a pretrained text encoder following AttnGAN [37]. The COCO [20] dataset contains 82,783 training and 40,504 validation images, each of which has 5 corresponding text descriptions including word, phrase and sentence. Both the datasets have images to be edited by query sentences with multiple attributes.

Implemenation detail. CA-GAN is optimized by Adam [15] and the learning rate is empirically set to 0.0002. The model trains 600 and 120 epochs for CUB and COCO dataset, respectively. The hyperparameters λ_1, λ_2, λ_3 and λ_4 are set to 0.7, 0.6, 1 and 0.9 empirically.

Comparing method. The comparing state-of-the-art approaches include SISGAN [7], TAGAN [25], DMIT [41], Open-edit [21], and ManiGAN [19]. Note that these comparing methods never consider the attribute editing but focus on the sentence editing.

Evaluation Metric. We use the Fréchet Inception Distance (FID) [10] and the Learned Perceptual Image Patch Similarity (LPIPS) [44] as the evaluation metrics. FID calculates the distance between two multidimensional variable distributions, representing the image generation quality. LPIPS represents the diversity of the generated images by calculating the L1 distance of the features extracted from AlexNet pre-trained in ImageNet. We also report the performance by human ranking on each dataset. We test edited accuracy (Acc.) and realism (Real) by randomly sampling 100 images with the same conditions and collect more than 20 surveys from different population. Specially, for the COCO dataset, each pair of image and sentence is from the same category.

Table 2: Quantitative comparison: Fréchet inception distance (FID), Learned Perceptual Image Patch Similarity (LPIPS), Acc. and Real of various methods on CUB and COCO.

	CUB				COCO			
Method	FID ↓	LPIPS ↓	Acc.(%) ↑	Real(%) ↑	FID ↓	LPIPS ↓	Acc.(%) ↑	Real(%) ↑
SISGAN [7]	36.69	0.7169	5.25	5.3	-	-	-	-
TAGAN [25]	28.92	0.7160	9.3	9.7	31.85	0.7802	8	6.2
DMIT [41]	28.01	0.7102	14.05	16.65	-	-	-	-
Open-edit [21]	23.57	0.7131	16.8	13.85	-	-	-	-
ManiGAN [19]	21.74	0.7059	24.05	23.5	25.96	0.7637	42.6	44.2
Ours	**20.08**	**0.6893**	**30.4**	**30.85**	**24.03**	**0.7438**	**49.4**	**49.6**

Table 3: Ablation study quantitative comparison: Fréchet inception distance (FID), Learned Perceptual Image Patch Similarity (LPIPS) on CUB.

	Ours	Ours w/o SPAC	Ours w/o CA	Ours w/o AD
FID	**20.08**	30.17	22.36	24.32
LPRPS	**0.6893**	0.7153	0.7085	0.7147

4.2 Comparisons with the state-of-the-arts

Qualitative comparison. Fig. 3 and Fig. 4 show the edit comparisons on the CUB and COCO datasets. It can be seen that, on CUB, the three methods, SISGAN, TAGAN, and DMIT, are unable to well generate bird contour information and lead to large blurred areas. These can be particularly observed in Fig. 4 where the tree trunk color is changed, black eye orbits are lost, and the orange head is modified. Although ManiGAN shows good image editing performance to some extent, it cannot edit multiple attribute regions better and the background regions change. We believe that this is due to the fact that ManiGAN is not trained to split and compare attribute features, and thus lacks attribute-level discriminators to offer training feedback related to each feature in the sentence.

Quantitative comparison. As shown in Table 2, compared with several existing state-of-the-art methods, our method leads to the best FID and LPIPS values on both CUB and COCO, implying that our method is able to achieve high-quality editing images. Moreover, our method also has the highest values for Acc. and Real, indicating that our model generates more favorable images for people. Our method produces higher quality images. This means people feel that our editing is better and the images are more realistic.

4.3 Ablation Studies

In this section, we evaluate the main modules in CA-GAN and analyze their impact. The result can be seen in Fig. 5 (b) and Table 3. First, without using the Sentence Parsing and Attribute Combination (w/o SPAC) module, we alternatively combine every two neighboring words in the sentence to construct an editable attribute. As shown in Fig. 5 (b), in the absence of rules, randomly dividing sentences may perform random editing with no relevant to the query sentence and achieve the worst FID and LPIPS. Second, we evaluate the effect of Contrastive Attention (w/o CA) module. We leverage a whole sentences as input, removing the contrastive attention module and contrastive loss. This confirms our hypothesis that by generating different editable information through sentence parsing and attribute combination, the differences between attributes are amplified by different contrastive attention, which helps the model to focus on the corresponding attributes in a given sentence. Finally, we evaluate the effectiveness of the proposed Attribute-level Discriminator (w/o AD), and the model cannot effectively operate on the image content based on the sentence information. For example in Fig. 5 (b), the bird's torso shows blurring and artifacts, and the color of

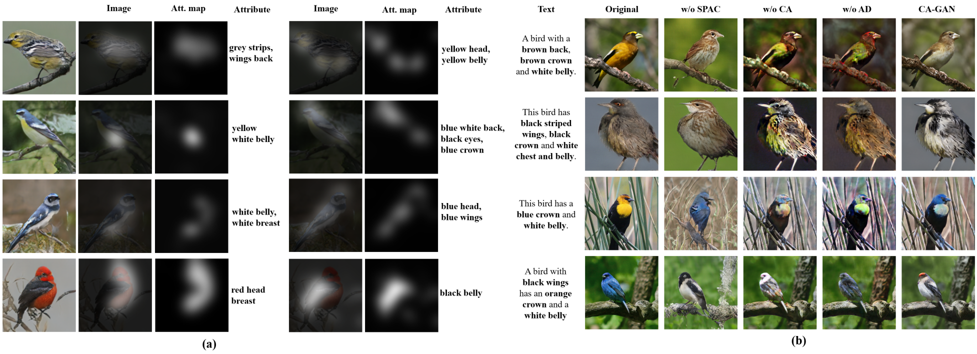
Image
Att. map
Attribute
grey strips, wings back
yellow head, yellow belly
yellow white belly
blue white back, black eyes, blue crown
white belly, white breast
blue head, blue wings
red head breast
black belly
(a)
Text
Original
w/o SPAC
w/o CA
w/o AD
CA-GAN
A bird with a brown back, brown crown and white belly.
This bird has black striped wings, black crown and white chest and belly.
This bird has a blue crown and white belly.
A bird with black wings has an orange crown and a white belly
(b)

Figure 5: (a) Visualisation of attention maps. (b) Ablation and comparison studies by removing proposed Sentence Parsing and Attribute Combination (**w/o SPAC**), removing contrastive attention (**w/o CA**), removing proposed attribute-level discriminator (**w/o AD**).

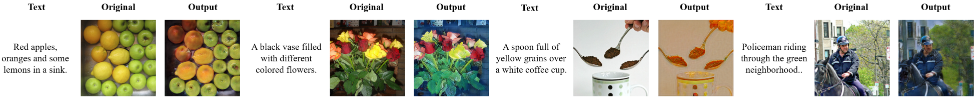
Text
Original
Output
Red apples, oranges and some lemons in a sink.
A black vase filled with different colored flowers.
A spoon full of yellow grains over a white coffee cup.
Policeman riding through the green neighborhood..

Figure 6: Some failure cases on COCO.

the feathers changes considerably. This indicates that the generator failed to decompose the different visual attributes due to the lack of attribute-level training feedback, and thus cannot effectively establish attribute-region connections to edit the images.

4.4 Visualization of contrastive attention and failure cases

In this section, we visualize the generated results of the attention maps corresponding to the different attributes in the CMAM from the third stage in Fig. 5 (a). We can observe that the model can better generate the accurate attention maps based on the attribute information after the sentence parsing and attribute combination, and with better accurate position, finer shape, and better semantic consistency between the attributes and the editable content. We show some failure cases in Fig. 6 on COCO. We find that the description semantics maybe fuzzy when there exist multiple categories, and the model may fail to edit the image.

5 Conclusion

In this paper, we studied the task of Sentence-based Image Editing. In contrast to existing methods that cannot produce accurate editing in case of a query sentence with multiple editable attributes, we proposed to enhance the difference between attributes and attained much better performance consequently. Particularly, we developed a novel model called CA-GAN by designing a contrastive attention mechanism on Generative Adversarial Network. We first parsed attributes from the sentence with POS Tagging and generated different attribute combinations. Then, a contrastive attention module was built to enlarge the editing difference between the combinations. Last, we constructed an attribute discriminator to ensure the effective editing on each attribute. Extensive experiments show that our model can lead to effective editing for sentences with multiple attributes on CUB and COCO datasets.

Acknowledgments

This work was supported by the Natural Science Foundation of China (No. 61876121), Primary Research and Development Plan of Jiangsu Province (No. BE2017663), Natural Science Foundation of Jiangsu Province (19KJB520054), Graduate Research Innovation Program of Jiangsu Province (SJCX20-1119), Scientific Research Project of School of Suzhou Institute of Trade and Commerce (KY-ZRA1805), National Natural Science Foundation of China(No. 61876155), Jiangsu Science and Technology Programme (Natural Science Foundation of Jiangsu Province) (No. BE2020006-4), Key Program Special Fund in XJTLU (No. KSF-T-06).

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[2] Adrien Bougouin, Florian Boudin, and Béatrice Daille. Topicrank: Graph-based topic ranking for keyphrase extraction. In *International joint conference on natural language processing (IJCNLP)*, pages 543–551, 2013.

[3] Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*, 2020.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017.

[8] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*, 2020.

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

[11] Te-En Huang, Tao-Hsing Chang, ADAT Co, Jon-Fan Hu, et al. A tool to analyze verb phrase and noun phrase relationship in sentences. In *41st Annual Meeting of the Cognitive Science Society*, 2019.

[12] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[14] Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. 2020.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Smith Kit. 126 amazing social media statistics and facts. [EB/OL]. https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/.

[17] Qicheng Lao, Mohammad Havaei, Ahmad Pesaranghader, Francis Dutil, Lisa Di Jorio, and Thomas Fevens. Dual adversarial inference for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7567–7576, 2019.

[18] Kwot Sin Lee, Ngoc-Trung Tran, and Ngai-Man Cheung. Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3942–3952, 2021.

[19] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020.

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

[21] Xihui Liu, Zhe Lin, Jianming Zhang, Handong Zhao, Quan Tran, Xiaogang Wang, and Hongsheng Li. Open-edit: Open-domain image manipulation with open-vocabulary instructions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 89–106. Springer, 2020.

[22] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[23] Fan Lyu, Qi Wu, Fuyuan Hu, Qingyao Wu, and Mingkui Tan. Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks. *IEEE Transactions on Multimedia*, 21(8):1971–1981, 2019.

[24] Fan Lyu, Shuai Wang, Wei Feng, Zihan Ye, Fuyuan Hu, and Song Wang. Multi-domain multi-task rehearsal for lifelong learning. In *Proceedings of The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[25] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *arXiv preprint arXiv:1810.11919*, 2018.

[26] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4467–4477, 2017.

[27] Narjes Nikzad-Khasmakhi, Mohammad-Reza Feizi-Derakhshi, Meysam Asgari-Chenaghlu, Mohammad-Ali Balafar, Ali-Reza Feizi-Derakhshi, Taymaz Rahkar-Farshi, Majid Ramezani, Zoleikha Jahanbakhsh-Nagadeh, Elnaz Zafarani-Moattar, and Mehrdad Ranjbar-Khadivi. Phraseformer: Multimodal key-phrase extraction using transformer and graph embedding. *arXiv preprint arXiv:2106.04939*, 2021.

[28] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

[29] Ori Press, Tomer Galanti, Sagie Benaim, and Lior Wolf. Emerging disentanglement in auto-encoder based unsupervised image content transfer. *arXiv preprint arXiv:2001.05017*, 2020.

[30] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Learn, imagine and create: Text-to-image generation from prior knowledge. *Advances in Neural Information Processing Systems*, 32:887–897, 2019.

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

[32] Christian Rupprecht, Iro Laina, Nassir Navab, Gregory D Hager, and Federico Tombari. Guide me: Interacting with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8551–8561, 2018.

[33] Hongchen Tan, Xiuping Liu, Xin Li, Yi Zhang, and Baocai Yin. Semantics-enhanced adversarial nets for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10501–10510, 2019.

[34] Duc Minh Vo and Akihiro Sugimoto. Paired-d gan for semantic image synthesis. In *Asian Conference on Computer Vision*, pages 468–484. Springer, 2018.

[35] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[36] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

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

[38] Zihan Ye, Fan Lyu, Linyan Li, Yu Sun, Qiming Fu, and Fuyuan Hu. Unsupervised object transfiguration with attention. *Cognitive Computation*, 11(6):869–878, 2019.

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

[40] Xiaoming Yu, Xing Cai, Zhenqiang Ying, Thomas Li, and Ge Li. Singlegan: Image-to-image translation by a single-generator network using multiple generative adversarial learning. In *Asian Conference on Computer Vision*, pages 341–356. Springer, 2018.

[41] Xiaoming Yu, Yuanqi Chen, Thomas Li, Shan Liu, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. *arXiv preprint arXiv:1909.07877*, 2019.

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

[43] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[45] Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. An unsupervised sentence embedding method bymutual information maximization. *arXiv preprint arXiv:2009.12061*, 2020.

[46] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6199–6208, 2018.

[47] Yingbo Zhou, Pengcheng Zhao, Weiqin Tong, and Yongxin Zhu. Cdl-gan: Contrastive distance learning generative adversarial network for image generation. *Applied Sciences*, 11(4):1380, 2021.

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