11-28-2021

Improving Text-To-Image Synthesis Using Contrastive Learning

Hui Ye
Georgia State University, Atlanta, GA, United States

Xiulong Yang
Georgia State University, Atlanta, GA, United States

Martin Takáč
Mohamed bin Zayed University of Artificial Intelligence

Raj Sunderraman
Georgia State University, Atlanta, GA, United States

Shihao Ji
Georgia State University, Atlanta, GA, United States

Improving Text-to-Image Synthesis Using Contrastive Learning

Hui Ye[1], Xiulong Yang[1], Martin Takáč[2], Rajshekhar Sunderraman[1], Shihao Ji[1]
[1]Department of Computer Science, Georgia State University, Atlanta, GA, USA
[2]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Masdar City, Abu Dhabi, UAE
[1]{hye2, xyang22}@student.gsu.edu, [1]{rsunderraman, sji}@gsu.edu
[2]takac.MT@gmail.com

Abstract

The goal of text-to-image synthesis is to generate a visually realistic image that matches a given text description. In practice, the captions annotated by humans for the same image have large variance in terms of contents and the choice of words. The linguistic discrepancy between the captions of the identical image leads to the synthetic images deviating from the ground truth. To address this issue, we propose a contrastive learning approach to improve the quality and enhance the semantic consistency of synthetic images. In the pretraining stage, we utilize the contrastive learning approach to learn the consistent textual representations for the captions corresponding to the same image. Furthermore, in the following stage of GAN training, we employ the contrastive learning method to enhance the consistency between the generated images from the captions related to the same image. We evaluate our approach over two popular text-to-image synthesis models, AttnGAN and DM-GAN, on datasets CUB and COCO, respectively. Experimental results have shown that our approach can effectively improve the quality of synthetic images in terms of three metrics: IS, FID and R-precision. Especially, on the challenging COCO dataset, our approach boosts the FID significantly by 29.60% over AttnGAN and by 21.96% over DM-GAN.

1. Introduction

The objective of the text-to-image synthesis problem is to generate high-quality images from the specific text descriptions. It is a fundamental problem with a wide range of practical applications, including art generation, image editing, and computer-aided design. Most recently proposed text-to-image synthesis methods [45, 46, 42, 13, 47, 23, 48, 43, 18, 22, 17, 1, 12, 6, 19, 33, 4, 32, 14, 36, 44] are based on Generative Adversarial Networks (GANs) [7]. Conditioned on the text descriptions, the GAN-based models can generate realistic images with consistent semantic meaning. In practice, one image is associated to multiple captions in the datasets. These text descriptions annotated by humans for the same image are highly subjective and diverse in terms of contents and choice of words. Additionally, some text descriptions do not even provide sufficient semantic information to guide the image generation. The linguistic variance and inadequacy between the captions of the identical image leads to the synthetic images conditioned on them deviating from the ground truth.

To address this issue, we propose a novel contrastive learning approach to improve the quality and enhance the semantic consistency of synthetic images. In the image-text matching task, we pretrain an image encoder and a text encoder to learn the semantically consistent visual and textual representations of the image-text pair. Meanwhile, we learn the consistent textual representations by pushing together the captions of the same image and pushing way the captions of different images via the contrastive loss. The pretrained image encoder and text encoder are leveraged to extract consistent visual and textual features in the following stage of GAN training. Then we also utilize the contrastive loss to minimize the distance of the fake images generated from text descriptions related to the same ground truth image while maximizing those related to different ground truth images. We generalize the existing text-to-image models to a unified framework so that our approach can be integrated into them to improve their performance. We evaluate our approach over two popular base models, AttnGAN [42] and DM-GAN [48] on datasets CUB [37] and COCO [20]. The experimental results have shown that our approach can effectively improve the quality of the synthetic images in terms of Inception Score (IS) [27], Fréchet Inception Distance (FID) [11], and R-precisions [42].

The contributions of our work can be summarized as follows: 1) We propose a novel contrastive learning approach to learn the semantically consistent visual and textual representations in the image-text matching task. 2) We propose a novel contrastive learning approach to enhance the semantic consistency of the synthetic images in the stage of GAN training. 3) Our approach can be incorporated into the exist-

ing text-to-image models to improve their performance. Extensive experimental results demonstrate the effectiveness of our approach. Our source code is publicly available at https://github.com/huiyegit/T2I_CL.

2. Related Work

2.1. Text-to-Image Generation

Recently, a great number of studies [25, 45, 46, 42, 13, 47, 23, 48, 43, 18, 22, 17, 1, 12, 6, 19, 33, 4, 24, 32, 14, 36, 44] present promising results on the text-to-image synthesis task, most of which make use of GANs as the backbone model. We briefly summarizes some of them that are most related to our approach. Zhang et al. [45] propose the stacked GAN architecture which produces images from low-resolution to high-resolution. AttnGAN [42] presents an attention mechanism, where the Deep Attentional Multimodal Similarity Model (DAMSM) is able to compute the similarity between the generated image and the caption using both the global sentence level information and the fine-grained word level information. DM-GAN [48] introduces a dynamic memory generative adversarial network to generate high-quality images. It utilizes a dynamic memory module to refine the initial generated image, a memory writing gate to highlight the relevant text information and a response gate to update image representations. SD-GAN [43] employs a Siamese structure with a pair of texts as input and trains the model with the contrastive loss. The conditional batch normalization is adopted for fine-grained image generation. Compared with the Siamese structure in SD-GAN, our approach is derived from the recent development of the contrastive learning paradigm, therefore, it has the advantage of better performance and less computational cost. Furthermore, we generalize our approach so that it can be applied to the existing GAN-based models for text-to-image synthesis. XMC-GAN [44] has also applied the contrastive learning approach in the text-to-image generation. However, the objectives of contrastive loss in our approach are different from those of XMC-GAN. We compute the contrastive losses of the caption-caption pair and the fake-fake pair, which are complementary to the contrastive losses in XMC-GAN. In this work, we choose AttnGAN and DM-GAN as the base models to evaluate our approach.

2.2. Image-Text Matching

The text-to-image synthesis involves the subtask image-text matching, which refers to learning the joint image-text representation to maximize the semantic similarity for an image-sentence pair. Liwei et al. [38] and Michael et al. [40] have leveraged the triplet loss to learn the joint image-text embedding for the image-text retrieval and the video-text representation for the video-text action retrieval task, respectively. Tao et al. [42] propose the Deep Attentional Multimodal Similarity Model (DAMSM) to learn the fine-grained image-text representation for text-to-image synthesis. DAMSM (Figure 1a) trains an image encoder and a text encoder jointly to encode sub-regions of the image and words of the sentence to a common semantic space, and computes a fine-grained image-text matching loss for image generation. However, the variations exist in the text representations corresponding to the same image, which leads to the generated images deviating from the ground truth image. To address this issue, we utilize the contrastive learning approach to push together the text representations related to the identical image and push away the text representations related to different images.

2.3. Contrastive Learning

Contrastive learning has recently attracted great interest due to its empirical success in self-supervised representation learning in computer vision. In the last two years, various contrastive methods [28, 2, 9, 3, 16, 35, 26, 5, 8, 10, 15, 21, 34, 39] for visual representations have been proposed. SimCLR [2] presents three major findings to learn better representations, including composition of data augmentations, a learnable nonlinear transformation between the representation and the contrastive loss, and large batch size and training step. Similar to SimCLR, we adopt the simple contrastive learning framework. In order to integrate the contrative learning approach into the GAN-based models with simple implementation and small computational cost, our approach does not have the learnable nonlinear transformation or large batch size. We set the same training batch sizes as our baselines.

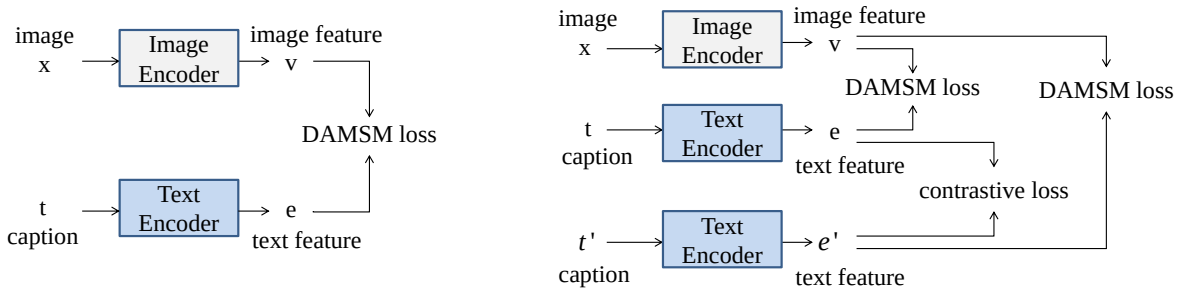
3. Method

3.1. Contrastive Learning for Pre-training

In the text-to-image synthesis task, the purpose of image-text matching is to learn the text representations which are semantically consistent with the corresponding images. Since the text representations will be leveraged as the conditions to guide the image generation, it is beneficial to develop a more effective pre-training approach to improve the quality of synthetic images.

Inspired by recent contrastive learning algorithms, we propose a novel approach for the pre-training of image-text matching. We learn the textual representations to match the visual representations via the DAMSM loss. Moreover, we train the textual representations by pushing together the captions corresponding to the same image and pushing away the captions corresponding to different images via the contrastive loss. As illustrated in Figure 1b, our framework consists of the following three major components.

Data sampling. At each training step, we sample a



(a) DAMSM (original)

(b) DAMSM + contrastive learning (ours)

Figure 1. Architectures of original DAMSM and our approach for image-text matching.

minibatch of images $\boldsymbol{x}$, captions $\boldsymbol{t}$ and captions $\boldsymbol{t}'$, where both captions $\boldsymbol{t}$ and $\boldsymbol{t}'$ are corresponding to images $\boldsymbol{x}$. For image-text matching, we consider two positive image-caption pairs (x_i, t_i) and (x_i, t_i') for each image x_i to calculate the DAMSM loss. Furthermore, we consider the caption-caption pair (t_i, t_i') as the positive pair to calculate the contrastive loss.

Image encoder f and text encoder g. We adopt an image encoder f to extract the visual vector representations and sub-region features from the image samples. Furthermore, we utilize a text encoder g to extract the textual vector representations and word features from the text samples. The text encoder g is shared in the framework. Our architecture has the flexibility of allowing various choices of deep neural network models. To have a fair comparison with the baselines, we adopt the same Inception-v3 [31] and Bi-directional Long Short-Term Memory (Bi-LSTM) [29] to instantiate the image encoder f and text encoder g.

Loss function. Similar to the baselines, we adopt the DAMSM loss as image-text matching loss. Moreover, we define the contrastive loss on pairs of two branches of input captions. We compute the contrastive loss to minimize the distance of textual representations related to the same image while maximizing those related to different images. We utilize the Normalized Temperature-scaled Cross Entropy Loss (NT-Xent) [30, 41, 2] as the contrastive loss. Given a pair, let $sim(a, b) = a^T b/\|a\|\|b\|$ denote the dot product between l_2 normalized a and b. Then the loss function for the ith sample is defined as

$$L(i) = -\log \frac{\exp(sim(u_i, u_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(sim(u_i, u_k)/\tau)}, \qquad (1)$$

where the ith and jth sample make the positive pair, $\mathbb{1}_{k \neq i}$ is an indicator function whose value is 1 iff $k \neq i$, τ denotes a temperature parameter and N is the batch size (e.g., N images and $2N$ captions). The overall contrastive loss is computed across all positive pairs in a minibatch, which can be defined as

$$L_c = \frac{1}{2N} \sum_{i=1}^{2N} L(i) \qquad (2)$$

Algorithm 1 summarizes the proposed method.

Algorithm 1 Contrastive learning for image-text matching

Input: Batch size N, temperature τ, image encoder f, text encoder g

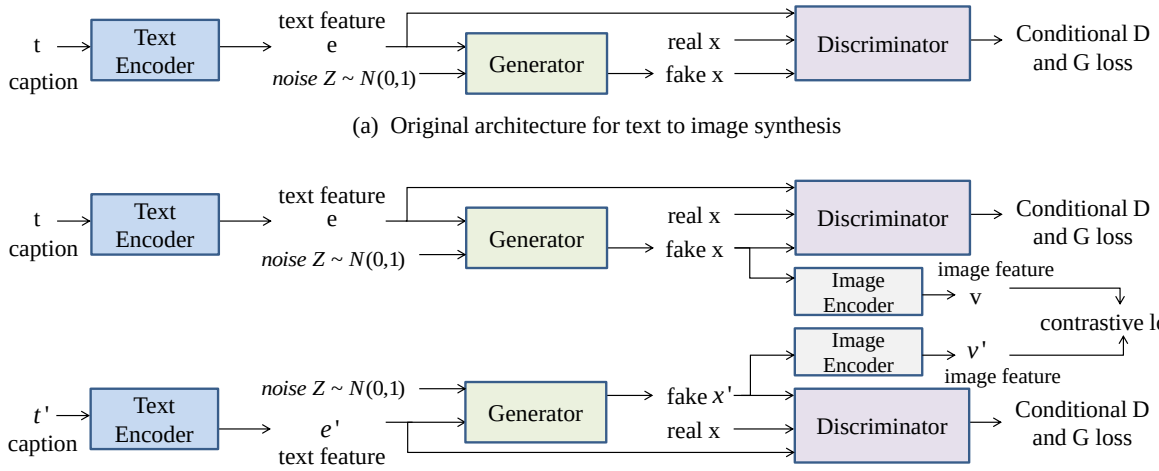
Output: Optimized image encoder f and text encoder g

for $\{1, \cdots,$ # of training iterations$\}$ do
  Sample a minibatch of images $\boldsymbol{x}$
  Sample a minibatch of captions $\boldsymbol{t}$ associated with $\boldsymbol{x}$
  Sample another minibatch of captions $\boldsymbol{t}'$ associated with $\boldsymbol{x}$
  $\boldsymbol{v} = f(\boldsymbol{x})$ // image representation
  $\boldsymbol{e} = g(\boldsymbol{t})$ // text representation
  $\boldsymbol{e}' = g(\boldsymbol{t}')$ // text representation
  $\mathcal{L}_1 = \text{DAMSM}(\boldsymbol{v}, \boldsymbol{e})$ // image-text matching loss
  $\mathcal{L}_2 = \text{DAMSM}(\boldsymbol{v}, \boldsymbol{e}')$ // image-text matching loss
  $\mathcal{L}_c = \text{NT-Xent}(\boldsymbol{e}, \boldsymbol{e}')$ // Equation 2
  $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_c$
  Update networks f and g to minimize $\mathcal{L}$
end for

3.2. Contrastive learning for GAN training

In practice, the captions annotated by humans for the same image have large variance in terms of contents and the choice of words, especially when the scenes are complex. The linguistic discrepancy between the captions of the identical image leads to the synthetic images conditioned on them deviating from the ground truth. Inspired by recent contrastive learning approaches, we apply the contrastive learning method to enhance the consistency between the generated images from the captions related to the same image and motivate them to be closer to the ground truth. Our approach consists of the following three major components.

Data sampling. The data sampling approach is similar to the one in the pre-training stage. At each training step, we sample a minibatch of images $\boldsymbol{x}$, captions $\boldsymbol{t}$ and captions $\boldsymbol{t}'$ in the same way as in Section 3.1. The deep



(a) Original architecture for text to image synthesis

(b) Our architecture with contrastive learning for text to image synthesis

Figure 2. Architectures of original approach and our approach for text to image synthesis.

generative model outputs the fake images $\boldsymbol{x}$ and $\boldsymbol{x}'$ conditioned on captions $\boldsymbol{t}$ and $\boldsymbol{t}'$, respectively. Then we consider the image-image pair (x_i, x_i') as the positive pair in the contrastive learning.

Model architecture. In this section, we derive our framework from the vanilla GANs step by step. GANs are a family of powerful generative models that estimate the data distribution through an adversarial learning process, in which a generator network G produces synthetic data given the input noise z and a discriminator network D distinguishes the true data from the generated data. The generator G is optimized to output realistic samples to fool the discriminator D. Formally, the generator G and discriminator D are following the minimax objective:

$$\min_G \max_D \mathop{\mathbb{E}}_{x\sim P_r}[\log(D(x))] + \mathop{\mathbb{E}}_{z\sim P_z}[\log(1-D(G(z)))], \quad (3)$$

where x is a real sample from the data distribution P_r, and the input z is sampled from some prior distribution P_z, such as a uniform or Gaussian distribution.

Most of the recent works [45, 46, 42, 13, 23, 48, 43, 18, 22, 17, 1, 12, 6, 19, 33, 4] on the text-to-image synthesis problem are based on GANs. We generalize these approaches to a unified framework, as shown in Figure 2a. The framework is extended from the GANs with the auxiliary information e, which is encoded from the input caption t by a pre-trained text encoder g. Then both the generator G and discriminator D are conditioned on this textual condition e. The training process is similar to the standard GANs with the following objective:

$$\min_G \max_D \mathop{\mathbb{E}}_{x\sim P_r}[\log(D(x,e))] + \mathop{\mathbb{E}}_{z\sim P_z}[\log(1-D(G(z,e),e))] \quad (4)$$

Algorithm 2 Contrastive learning for GAN training

Input: Batch size N, temperature τ, coefficient λ_c, generator G, discriminator D, pre-trained image encoder f, pre-trained text encoder g.

Output: Optimized G and D.

for $\{1, \cdots,$ # of training iterations$\}$ do
  Sample a minibatch of images $\boldsymbol{x} \sim P_r$
  Sample a minibatch of latent variable $\boldsymbol{z} \sim P_z$
  Sample a minibatch of captions $\boldsymbol{t}$ associated with $\boldsymbol{x}$
  Sample another minibatch of captions $\boldsymbol{t}'$ associated with $\boldsymbol{x}$
  $\boldsymbol{e} = g(\boldsymbol{t})$
  $\boldsymbol{e}' = g(\boldsymbol{t}')$
  $\mathcal{L}_{D1} = \frac{1}{N}\sum_{i=1}^{N}[\log D(x_i, e_i) + \log(1 - D(G(z_i, e_i), e_i))]$
  $\mathcal{L}_{D2} = \frac{1}{N}\sum_{i=1}^{N}[\log D(x_i, e_i') + \log(1 - D(G(z_i, e_i'), e_i'))]$
  $\mathcal{L}_D = \mathcal{L}_{D1} + \mathcal{L}_{D2}$
  Update D to minimize $\mathcal{L}_D$
  Sample noise $\boldsymbol{z}$, captions $\boldsymbol{t}$ and $\boldsymbol{t}'$ as step 3, 4 and 5
  Compute $\boldsymbol{e}$, $\boldsymbol{e}'$ as step 6 and 7
  $\mathcal{L}_{G1} = \frac{1}{N}\sum_{i=1}^{N}\log(1 - D(G(z_i, e_i), e_i))$
  $\mathcal{L}_{G2} = \frac{1}{N}\sum_{i=1}^{N}\log(1 - D(G(z_i, e_i'), e_i'))$
  $\boldsymbol{v} = f(G(\boldsymbol{z}, \boldsymbol{e}))$ // image representation
  $\boldsymbol{v}' = f(G(\boldsymbol{z}, \boldsymbol{e}'))$ // image representation
  $\mathcal{L}_c$ = NT-Xent $(\boldsymbol{v}, \boldsymbol{v}')$ // Equation 2
  $\mathcal{L}_G = \mathcal{L}_{G1} + \mathcal{L}_{G2} + \lambda_c \mathcal{L}_c$
  Update G to minimize $\mathcal{L}_G$
end for

We further extend the generalized text-to-image framework to a Siamese structure and integrate the contrastive learning approach into it. As shown in Figure 2b, the image encoder f takes the fake images x and x' as input, and extracts the visual representations v and v' to compute the contrastive loss. The two branches of the architecture share

the identical generator G, discriminator D, image encoder f and text encoder g. The image encoder f and text encoder g are pre-trained in the image-text matching task and work in the evaluation mode in the phase of GAN training.

Loss function. In addition to the adversarial losses from Equation 4, we define the contrastive loss on pairs of fake images generated from two branches of input captions. We utilize the contrastive loss to minimize the distance of the fake images generated from two text descriptions related to the same image while maximizing those related to different images. We apply the same NT-Xent loss in Section 3.1. Algorithm 2 summarizes the proposed method.

4. Experiments

Datasets. Following the previous works, we evaluate our approach on datasets CUB [37] and COCO [20]. The CUB dataset contains 200 bird species with 11,788 images, where 150 species with 8,855 images are used as the training data, and the remaining 50 species with 2,933 images as the test data. Each image has 10 related captions in the CUB dataset. The COCO dataset has 80k images for training and 40k images for evaluation. Each image has 5 related captions in the COCO dataset.

Evaluation Metric. We choose the Inception Score (IS) [27], Fréchet Inception Distance (FID) [11], and R-precisions [42] as the quantitive metrics to evaluate the performance. The IS calculates the KL-divergence between the conditional and marginal probability distributions. In general, a larger IS indicates the generative model can synthesize fake images with better diversity and quality. The FID computes the Fréchet distance between synthetic and real images in the feature space extracted from the pre-trained Inception v3 model. A smaller FID indicates the synthetic data is more realistic and similar to the true data. R-precision calculates the precision of the image-text retrieval task to evaluate to what extent the synthetic images match the input captions. The higher R-precision indicates the generated images have greater consistency with the text descriptions. After training, the model generates 30,000 images conditioned on the captions in the test set for evaluation. The source code to calculate the three metrics is from the public website[1].

Note that several previous works [18, 33] have pointed out that IS can not provide useful guidance to evaluate the quality of the synthetic images on dataset COCO. Nevertheless, we still report the IS on COCO as the auxiliary metric. We consider the FID as the primary metric among the three in terms of robustness and effectiveness.

Implementation details. We choose two popular text-to-image synthesis models, AttnGAN [42] and DM-GAN [48], to evaluate our approach. Note that both AttnGAN and DM-GAN are the stacked architecture with 3 generator-discriminator pairs. To reduce the computational cost, we only calculate the contrastive loss of the fake images of size 256x256 from the last generator.

[1] https://github.com/MinfengZhu/DM-GAN

Method	Dataset	IS ↑	FID ↓	R-Precision ↑
AttnGAN*	CUB	4.33 ± .07	20.85	67.09 ± .83
AttnGAN + CL	CUB	**4.42 ± .05**	**16.34**	**69.64 ± .63**
AttnGAN*	COCO	23.71 ± .38	33.99	83.97 ± .78
AttnGAN + CL	COCO	**25.70 ± .62**	**23.93**	**86.55 ± .51**

Table 1. Comparison of our approach and AttnGAN over the datasets CUB and COCO. ↑ denotes the higher value the better quality. ↓ denotes the lower value the better quality. * indicates the results are obtained from the pre-trained model released publicly by the authors. The bold font represents better performance. CL denotes the proposed constrastive learning approach in this work.

Method	Dataset	IS ↑	FID ↓	R-Precision ↑
DM-GAN*	CUB	4.66 ± .06	15.10	75.86 ± .83
DM-GAN + CL	CUB	**4.77 ± .05**	**14.38**	**78.99 ± .66**
DM-GAN*	COCO	32.37± .29	26.64	92.09 ± .50
DM-GAN + CL	COCO	**33.34 ± .51**	**20.79**	**93.40 ± .39**

Table 2. Comparison of our approach and DM-GAN over the datasets CUB and COCO. The notations ↑, ↓, *, bold font and CL have the same meanings as the ones in Table 1

We retain the setting of parameters in the original baselines except the λ value used in AttnGAN. We find that $\lambda = 5$ is reported in the paper and used in the source code. However, when we ran the source code of AttnGAN with this value, the R-precision is about 58.80, which has a large difference from 67.21 reported in the paper. When we changed λ to 10, we got 67.00 for R-precision, which is consistent with the one reported in the paper as well as the IS score and FID. We believe there is a typo of λ value in the AttnGAN paper. Therefore, we set $\lambda = 10$ and adopt it in all of our experiments.

Following the same setting of the configuration files of the baselines, we train our novel model based on AttnGAN with 600 epochs on CUB and 120 epochs on COCO, and the other model based on DM-GAN with 800 epochs on CUB and 200 epoch on COCO. We evaluate the IS, FID and R-precision of the checkpoint every 50 epochs on CUB and 10 epochs on COCO. We choose the checkpoint with the best FID and report the corresponding IS and R-precision.

4.1. Text-to-Image Quality

We apply our contrastive learning approach to two baselines AttnGAN [42] and DM-GAN [48], and compare the performances with them over datasets CUB and COCO. The experimental results are reported in Table 1 and 2.

As shown in Table 1, our approach improves the three metrics IS, FID and R-precision over the datasets CUB and COCO. The IS is improved from 4.33 to 4.42 on CUB and

Figure 3. Comparison of example images between our approach and baselines on the CUB dataset.

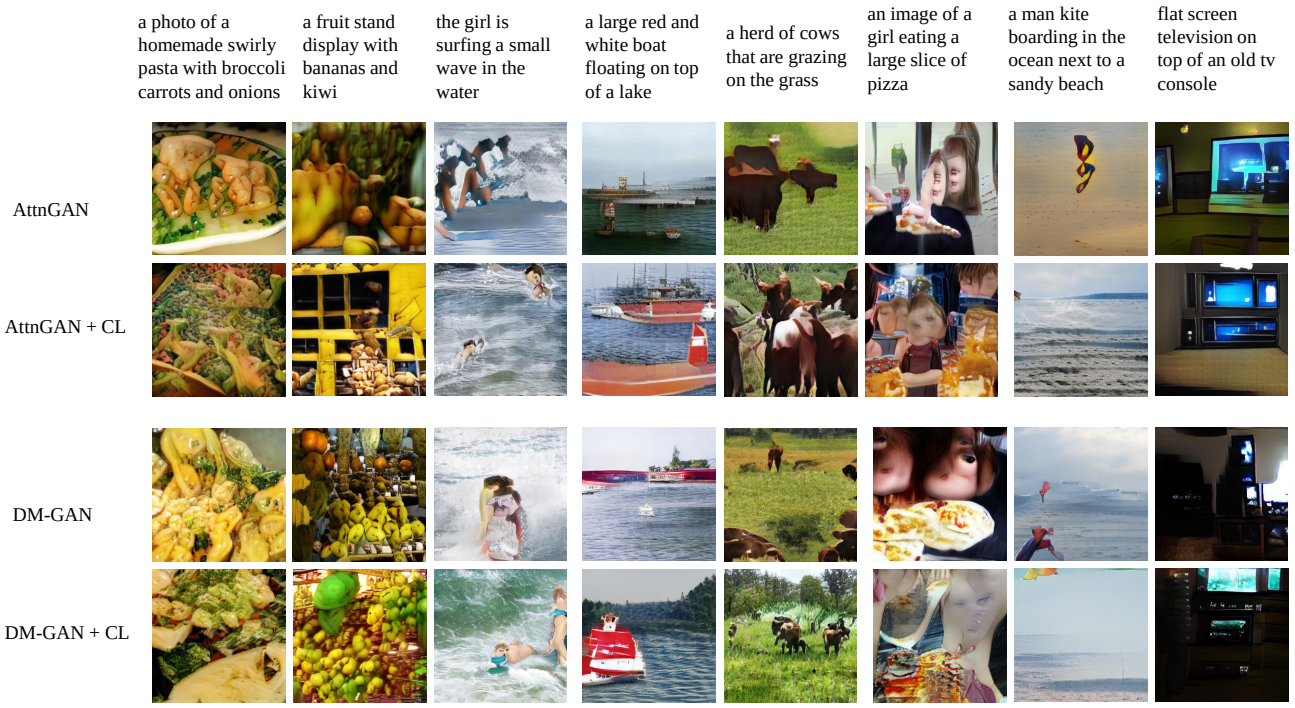


Figure 4. Comparison of example images between our approach and baselines on the COCO dataset.

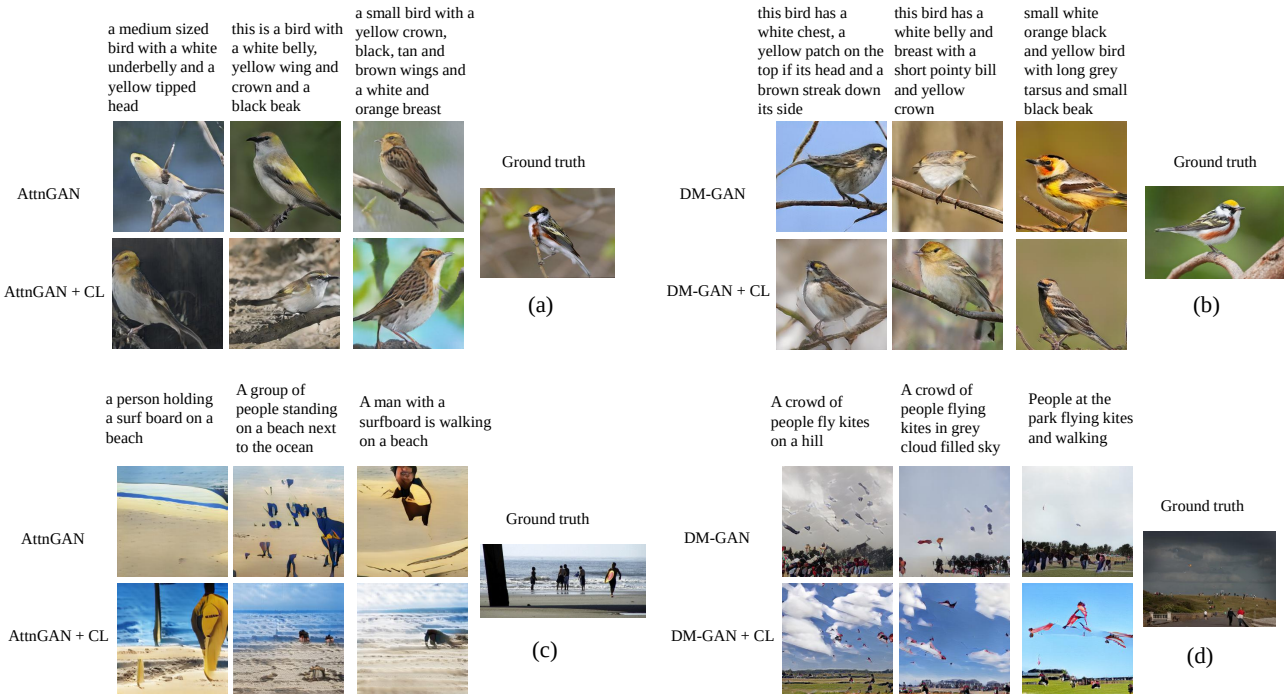


Figure 5. Comparison of example images between our approach and baselines.

from 23.71 to 25.70 on COCO. For the relatively more suitable metric FID, our approach boosts the baseline AttnGAN significantly by 21% on CUB and 29.60% on COCO, respectively. Meanwhile, our approach achieves higher R-precision with the gain of 2.55 on CUB and 2.58 on COCO. As shown in Table 2, in comparison to DM-GAN, our approach also improves the three metrics IS, FID and R-precision over CUB and COCO. The IS is improved from 4.66 to 4.77 on CUB and from 32.37 to 33.34 on COCO. Regarding the metric FID , our approach has the value of 14.38 with a small gain of 0.72 on CUB, while boosts the baseline DM-GAN significantly by 21.96% on COCO. Furthermore, our approach achieves better R-Precision with the improvement of 3.13 on CUB and 1.31 on COCO.

The dataset COCO is more challenging than CUB, as it has more complex scenes and the captions have greater variance to describe the identical image. However, it is noteworthy that our approach significantly improves the FID by 29.60% over the baseline AttnGAN and 21.96% over DM-GAN. In summary, the quantitative experimental results demonstrate that our contrastive learning approach can effectively improve the quality and enhance the consistency of the synthetic images generated from diverse captions.

4.2. Visual Quality

To further compare our proposed approach with the baselines, we visualize the synthetic images generated from the typical example captions. As shown in Figure 3, compared with the baseline AttnGAN, the images generated from our approach are more realistic and better match with the text descriptions in most cases. In the 8th column, the bird in the image from AttnGAN fails seriously with two heads, while the one from our approach has the reasonable appearance. In the 2nd column, we can see the vivid green crown in the bird from our approach, which matches the description “green crown” well, while the image from AttnGAN does not show this feature of the bird. As shown in Figure 3, the comparison between our approach and the baseline DM-GAN is similar to previous comparison. In the 3rd column, the image from our approach has the correct white belly to match the text description “while belly”, while the image from DM-GAN has the additional incorrect red color in the belly. Figure 4 shows the example images on COCO from our approach and the baselines AttnGAN and DM-GAN. It is challenging to generate photo-realistic images for the models showed in the figure. However, compared with the baselines, the images generated from our approach are more realistic and better match with the text descriptions in some cases. In the 4th column, the boat in the image from our approach has the red and while color, which aligns with the caption, while the image from AttnGAN does not show the red boat. In the 5th column, the image from our approach has better shape of cows than the one from AttnGAN. As shown in the 3rd column, the image from our approach has the basic shape of a girl, while it can not be observed in the image from DM-GAN at all.

Method	Dataset	IS ↑	FID ↓	R-Precision ↑
AttnGAN†	CUB	4.29 ± .05	19.16	68.02 ± .98
+ CL1	CUB	4.31 ± .02	17.97	69.11 ± .63
+ CL1 + CL2	CUB	**4.42 ± .05**	**16.34**	**69.64 ± .63**
AttnGAN†	COCO	25.05 ± .64	30.67	84.24 ± .58
+ CL1	COCO	**25.87 ± .41**	26.89	85.93 ± .63
+ CL1 + CL2	COCO	25.70 ± .62	**23.93**	**86.55 ± .51**

Table 3. Ablation study of our approach on AttnGAN over the datasets CUB and COCO. † indicates we retrain the model with the same setting of hyperparameters. CL1 and CL2 denote the constrastive learning approach in the pre-training and GAN training, respectively.

Method	Dataset	IS ↑	FID ↓	R-Precision ↑
DM-GAN†	CUB	4.67 ± .06	15.55	75.88 ± .89
+ CL1	CUB	4.71 ± .05	14.56	76.74 ± .88
+ CL1 + CL2	CUB	**4.77 ± .05**	**14.38**	**78.99 ± .66**
DM-GAN†	COCO	31.53 ± .39	27.04	91.82 ± .49
+ CL1	COCO	30.98 ± .69	25.29	92.10 ± .56
+ CL1 + CL2	COCO	**33.34 ± .51**	**20.79**	**93.40 ± .39**

Table 4. Ablation Study of our approach on DM-GAN over the CUB and COCO datasets. †, CL1 and CL2 have the same meanings as the ones in Table 3.

We also visualize the example images generated from multiple captions corresponding to the same ground truth image. Compared with the baselines AttnGAN and DM-GAN, the images generated from our approach are more realistic and closer to the ground truth images. As shown in the 1st column of Figure 5(a), the image from our approach has the black color in the wings, which is consistent with the ground truth image. Although the 1st caption does not have the text description of "black color" explicitly, our model is still able to train its semantic text embedding to contain this information from other captions related to the same image via our contrastive learning approach. Another similar example is shown in Figure 5(c). The 3rd caption does not provide the text description of "the ocean", the image from our approach can still have the ocean scene to be consistent with the ground truth, while the baseline DM-GAN is not able to achieve this.

4.3. Ablation Study

In this work, the contrastive learning approach is applied in two stages: image-text matching and the training of GANs. We combine our novel approach for the image-text matching with the baselines AttnGAN and DM-GAN, and conduct experiments to evaluate the effectiveness of it. The experimental results are shown in Table 3 and 4. Compared with the two baselines AttnGAN and DM-GAN, our contrastive learning approach for the image-text matching task can help improve the performance in terms of the IS, FID and R-precision on datasets CUB and COCO, with one exception that the IS of our approach is about 0.55 smaller than DM-GAN on COCO. When we add the contrastive learning approach for GAN training into our previous approach, our complete approach shows further improvements in terms of the IS, FID and R-precision on datasets CUB and COCO, with the exception that the IS of our complete approach based on AttnGAN is about 0.11 smaller than our previous one on COCO. The experimental results demonstrate that our contrastive learning approach in the image-text matching task and GAN training can help improve the performance of text-to-image synthesis, respectively.

Method	λ_c	IS ↑	FID ↓	R-Precision ↑
AttnGAN + CL	0.1	4.28 ± .05	17.60	70.06 ± .84
	0.2	4.42 ± .05	**16.34**	69.64 ± .63
	0.5	4.35 ± .05	17.76	69.08 ± .63
	1.0	4.31 ± .07	16.72	69.28 ± .77
	2.0	4.41 ± .05	16.90	68.79 ± .38
	5.0	4.49 ± .08	17.19	67.64 ± .85
	10.0	4.43 ± .07	17.53	70.25 ± .55

Table 5. Ablation study on different choices of weight λ_c for contrastive loss.

Method	τ	IS ↑	FID ↓	R-Precision ↑
AttnGAN + CL	0.1	4.44 ± .06	17.12	69.62 ± .78
	0.2	4.44 ± .05	17.55	68.50 ± .85
	0.5	4.42 ± .05	**16.34**	69.64 ± .63
	1.0	4.43 ± .05	17.68	69.86 ± .92

Table 6. Ablation study on different choices of temperature τ for contrastive loss.

We adjust the hyperparameters to investigate the impact to the performance of our approach. We tune the weight λ_c in $\{0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0\}$ and the temperature τ in $\{0.1, 0.2, 0.5, 1.0\}$ for the contrastive loss. The experiments are conducted on the CUB dataset. In each case, we evaluate the checkpoint every 50 epochs and choose the one with the best FID. Table 5 shows the results of weight λ_c. We find that FID is not very sensitive to this hyperparameter as well as IS and R-precision. When λ_c is 0.2, the model has the best FID score 16.34, which is improved by 1.42, compared with the worst value 17.76. Table 6 shows the results of temperature τ. Similar to the weight λ_c, it can be observed that τ has a small impact to the performance of the model. The difference between the largest FID and smallest one is 1.34.

5. Conclusion

In this paper, we have shown how to incorporate the contrastive learning method into prior text-to-image models to improve their performance. Firstly, we train the image-text matching task to push together the textual representations corresponding to the same image through the contrastive loss. Furthermore, we employ the contrastive learning method to enhance the consistency between generated images from the captions related to the same image. We

propose a generalized framework for the existing text-to-image models, and evaluate our approach on two baselines AttnGAN and DM-GAN. Extensive experiments demonstrate that our approach outperforms the two strong baselines in terms of three metrics. Especially, on the challenging COCO dataset, our approach boosts the FID significantly by 29.60% over AttnGAN and by 21.96% over DM-GAN. Since the image-text representation learning is a fundamental task, we believe our approach has potential applicability in a wide range of cross domain tasks, such as visual question answering, image-text retrieval as well as text-to-image synthesis. We leave the extension to these tasks as a future work.

Acknowledgements

We would like to thank the anonymous reviewers for their comments and suggestions, which helped improve the quality of this paper. We would also gratefully acknowledge the support of VMware Inc. for its university research fund to this research.

References

[1] Miriam Cha, Youngjune L Gwon, and HT Kung. Adversarial learning of semantic relevance in text to image synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3272–3279, 2019.

[2] 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.

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[4] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10911–10920, 2020.

[5] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *arXiv preprint arXiv:2007.00224*, 2020.

[6] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10304–10312, 2019.

[7] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[8] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pages 4116–4126. PMLR, 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] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6629–6640, 2017.

[12] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Generating multiple objects at spatially distinct locations. *arXiv preprint arXiv:1901.00686*, 2019.

[13] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7986–7994, 2018.

[14] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

[15] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *arXiv preprint arXiv:2010.01028*, 2020.

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[17] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[18] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019.

[19] Jiadong Liang, Wenjie Pei, and Feng Lu. Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In *European Conference on Computer Vision*, pages 491–508. Springer, 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] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[22] 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.

[23] 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.

[24] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[25] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

[26] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[27] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2234–2242, 2016.

[28] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019.

[29] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[30] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems*, 29:1857–1865, 2016.

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[32] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6710–6719, 2019.

[33] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Fei Wu, and Xiao-Yuan Jing. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020.

[34] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[35] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

[36] Hung-Yu Tseng, Hsin-Ying Lee, Lu Jiang, Ming-Hsuan Yang, and Weilong Yang. Retrievegan: Image synthesis via differentiable patch retrieval. In *European Conference on Computer Vision*, pages 242–257. Springer, 2020.

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

[38] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[39] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[40] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 450–459, 2019.

[41] 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.

[42] 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.

[43] 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.

[44] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 833–842, 2021.

[45] 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.

[46] 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.

[47] 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.

[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.