Semantics Disentangling for Text-to-Image Generation

Guojun Yin[1,2], Bin Liu[1], Lu Sheng[2,4*], Nenghai Yu[1], Xiaogang Wang[2], Jing Shao[3]
[1]University of Science and Technology of China, Key Laboratory of Electromagnetic Space Information,
The Chinese Academy of Sciences, [2] CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[3]SenseTime Research, [4]College of Software, Beihang University
gjyin@mail.ustc.edu.cn, lsheng@buaa.edu.cn,
{flowice,ynh}@ustc.edu.cn, xgwang@ee.cuhk.edu.hk, shaojing@sensetime.com

Abstract

Synthesizing photo-realistic images from text descriptions is a challenging problem. Previous studies have shown remarkable progresses on visual quality of the generated images. In this paper, we consider semantics from the input text descriptions in helping render photo-realistic images. However, diverse linguistic expressions pose challenges in extracting consistent semantics even they depict the same thing. To this end, we propose a novel photo-realistic text-to-image generation model that implicitly disentangles semantics to both fulfill the high-level semantic consistency and low-level semantic diversity. To be specific, we design (1) a Siamese mechanism in the discriminator to learn consistent high-level semantics, and (2) a visual-semantic embedding strategy by semantic-conditioned batch normalization to find diverse low-level semantics. Extensive experiments and ablation studies on CUB and MS-COCO datasets demonstrate the superiority of the proposed method in comparison to state-of-the-art methods.

1. Introduction

The rapid progress of the Generative Adversarial Networks (GAN) [11, 21, 1, 20] brings a remarkable evolution in natural image generation with diverse conditions. In contrast to conditions such as random noises, label maps or sketches, it is a more natural but challenging way to generate an image from a linguistic description (text) since (1) the linguistic description is a natural and convenient medium for a human being to describe an image, but (2) cross-modal text-to-image generation is still challenging.

Existing text-to-image generation works [40, 37, 42, 14, 29] mainly focus on increasing the visual quality and resolution of the generated images by either a stacked coarse-to-fine generator structure [40, 14] or an attention-guided generation procedure [37]. However, these methods neglect one important phenomenon that the human descriptions for a same image are highly subjective and diverse in their expressions, it means that naively using these texts as unique descriptions to generate images would often produce unstable appearance patterns that are far apart from the ground-truth images. For example, when given different descriptions (Fig 1(a-ii)) for the same ground-truth image in Fig. 1(a-i), the generated images in Fig. 1(a-iii) by [37] present various appearance patterns apart from the groundtruth, not even similar to the same kind of bird. It

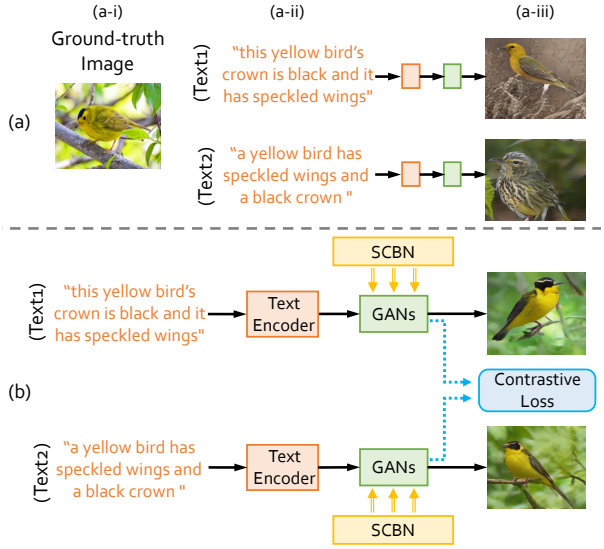

Figure 1. Given the language descriptions in (a-ii), their corresponding images are generated by existing GANs in (a-iii). Compared to the groundtruth image in (a-i), such holistic subjective text may lead generation deviation (a-iii) due to the lacking of common and distinct semantic meanings. The proposed SD-GAN in (b) distills the semantic commons by a Siamese structure and retains semantic diversities & details via a semantic-conditioned batch normalization.

*Lu Sheng is the corresponding author.

shows that the rich variations of linguistic expressions pose challenges in extracting consistent semantic commons from different descriptions of the same image. Variations of descriptions may lead to deviated image generation even if they describe the same bird with very similar semantic expressions.

To address this issue, in this paper, we propose a novel photo-realistic text-to-image generation method that effectively exploit the semantics of the input text within the generation procedure, named as *Semantics Disentangling Generative Adversarial Network* (SD-GAN). The proposed SD-GAN distills the *semantic commons* from texts for image generation consistency and meanwhile retains the *semantic diversities & details* for fine-grained image generation.

Inspired by the advantages of Siamese structure used in different tasks [32, 33, 4, 10, 43] which can find the similarity between a pair of sequences, we treat our discriminator as an image comparator so as to preserve the semantic consistency among the generated images as long as their descriptions are comprehensive and refer to the same semantic contents. Specifically, the proposed SD-GAN uses a Siamese scheme with a pair of texts as input and trained with the contrastive loss shown in Fig. 1(b). Denote *intra-class* pair as the same groundtruth image with different descriptions while *inter-class* pair as the different groundtruth image with different descriptions. By the SD-GAN, the *intra-class* pairs with similar linguistic semantics should generate consistent images that have smaller distances in the feature space of the discriminator, while *inter-class* pairs have to bear much larger distances. Since we do not have text-to-semantic embedding structure before our image generator, this special training strategy also forces the text-to-image generator has an inherent distillation of semantic commons from diverse linguistic expressions.

To some extent, the Siamese structure indeed distills the *semantic commons* from texts but meanwhile ignores the *semantic diversities & details* of these descriptions even from the same image. To maintain the semantic diversities from the texts, the detailed linguistic cues are supposed to be embedded into visual generation. Previous works try to guide visual generation by taking the text features as the input to the generator [40, 41, 37]. From another perspective, we reformulate the batch normalization layer within the generator, denoted as *Semantic-Conditioned Batch Normalization* (SCBN) in Fig. 1(b). The proposed SCBN enables the detailed and fine-grained linguistic embedding to manipulate the visual feature maps in the generative networks.

Our contributions are summarized as follows:

1) *Distill Semantic Commons from Text-* The proposed SD-GAN distills semantic commons from the linguistic descriptions, based on which the generated images can keep generation consistency under expression variants. To our best knowledge, it is the first time to introduce the Siamese mechanism into the cross-modality generation.

2) *Retain Semantic Diversities & Details from Text-* To complement the Siamese mechanism that may lose unique semantic diversities, we design an enhanced visual-semantic embedding method by reformulating the batch normalization layer with the instance linguistic cues. The linguistic embedding can further guide the visual pattern synthesis for fine-grained image generation.

3) The proposed SD-GAN achieves the state-of-the-art performance on the CUB-200 bird dataset [34] and MS-COCO dataset [22] for text-to-image generation.

2. Related Works

Generative Adversarial Network (GAN) for Text-to-Image. Goodfellow *et al*. [11] first introduced the adversarial process to learn generative models. The Generative Adversarial Network (GAN) is generally composed of a generator and a discriminator, where the discriminator attempts to distinguish the generated images from real distribution and the generator learns to fool the discriminator. A set of constraints are proposed in previous works [28, 16, 26, 9, 36] to improve the training process of GANs, *e.g*., interpretable representations are learned by using additional latent code in [3]. GAN-based algorithms show excellent performance in image generation [21, 1, 20, 25, 35, 2, 23]. Reed *et al*. [30] first showed that the conditional GAN was capable of synthesizing plausible images from text descriptions. Zhang *et al*. [40, 41] stacked several GANs for text-to-image synthesis and used different GANs to generate images of different sizes. Their following works [42, 37] also demonstrated the effectiveness of stacked structures for image generation. Xu *et al*. [37] developed an attention mechanism that enables GANs to generate fine-grained images via word-level conditioning input. However, all of their GANs are conditioned on the language descriptions without disentangling the semantic information under the expression variants. In our work, we focus on disentangling the semantic-related concepts to maintain the generation consistency from complex and various natural language descriptions as well as the details for text-to-image generation.

Conditional Batch Normalization (CBN). Batch normalization (BN) is a widely used technique to improve neural network training by normalizing activations throughout the network with respect to each mini-batch. BN has been shown to accelerate training and improve generalization by reducing covariate shift throughout the network [17]. Dumoulin *et al*. [6] proposed a conditional instance normalization layer that learns the modulation parameters with the conditional cues. These parameters are used to control the behavior of the main network for the tasks such as image stylization [15], visual reasoning [27], video segmentation [38], question answering [5] and *etc*. In our work, conditional batch normalization is firstly adopted for visual

feature generation and the semantic-conditioned batch normalization layers enhance the visual-semantic embedding and the proposed layers are implemented in the generators of GANs for the purpose of the efficient visual generation based on the linguistic conditions.

3. Semantics Disentangling Generative Adversarial Network (SD-GAN)

In this paper, we propose a new cross-modal generation network named as Semantics Disentangling Generative Adversarial Network (SD-GAN) for text-to-image generation, as shown in Fig. 2. It aims at distilling the *semantic commons* from texts for image generation consistency and meanwhile retaining the *semantic diversities & details* for fine-grained image generation: (1) Taking the advantages of Siamese structure, the generated images are not only based on the input description at the current branch, but also influenced by the description at the other branch. In other words, the Siamese structure distills the common semantics from texts to handle the generation deviation under the expression variants. (2) To generate fine-grained visual patterns, the model also needs to retain the detailed and diverse semantics of the input texts. We modulate neural activations with linguistic cues by the proposed *Semantic-Conditioned Batch Normalization* (SCBN), which will be introduced in Sec. 3.2.

3.1. Siamese Structure with Contrastive Losses

Although existing methods [40, 37] achieved excellent performances on high-resolution image generation, the generation deviations from language expression variants still pose great challenges for the text-semantic image generation. To address the issues, the proposed SD-GAN adopts a Siamese structure for distilling textual semantic information for the cross-domain generation. The contrastive loss is adopted for minimizing the distance of the fake images generated from two descriptions of the same groundtruth image while maximizing those of different groundtruth images. During the training stage, the generated image is influenced by the texts from both two branches.

For constructing the backbone architecture for each Siamese branch, we adopt the sequential stacked generator-discriminator modules used in most previous works [40, 37, 14]. As shown in Fig. 2, it consists of 1) a text encoder $\mathtt{E}$ (in orange) for text feature extracting from descriptions, and 2) hierarchical generative adversarial subnets (in green) for image generation which contains a bunch of generators, *i.e.*, $\mathtt{G}_0, \mathtt{G}_1, \mathtt{G}_2$, and the corresponding adversarial discriminators, *i.e.*, $\mathtt{D}_0, \mathtt{D}_1, \mathtt{D}_2$.

Text Encoder. The input of each branch is a sentence of natural language description. The text encoder $\mathtt{E}$ aims at learning the feature representations from the natural language descriptions and following [40, 41, 37], we adopt a bi-directional Long Short-Term Memory (LSTM) [13] that extracts semantic vectors from the text description. Generally, in the bi-directional LSTM, the hidden states are utilized to represent the semantic meaning of a word in the sentence while the last hidden states are adopted as the global sentence vector, *i.e.*, w_t denotes the feature vector for the t^{th} word and $\bar{s}$ denotes the sentence feature vector.

Hierarchical Generative Adversarial Networks. Inspired by [40, 37, 14, 41], we adopt hierarchical stages from low-resolution to high-resolution for the photo-realistic image generation. Given the sentence feature $\bar{s}$ from the text encoder $\mathtt{E}$ and a noise vector z sampled from a standard normal distribution, the low resolution (64×64) image is generated at the initial stage, as shown in Fig. 3 (a). (The SCBN layer in Fig. 3 will be introduced in Sec. 3.2.) The following stage uses the output of the former stage as well as the sentence feature $\bar{s}$ to generate the image with higher-resolution, as shown in Fig. 3 (b). At each stage, the generator is followed by a discriminator that distinguishes whether the image is real or fake. These discriminators $\mathtt{D}_0, \mathtt{D}_1, \mathtt{D}_2$ are independent for extracting the visual features and will not share parameters.

Contrastive Loss. In our work, the purpose of the proposed Siamese structure is to enhance the generation consistency regardless of the expression variants of the input descriptions during the training procedure. We input two different text descriptions to the two branches of the Siamese structure respectively. If the visual features generated from two branches are textual semantic-aware, the two generated images should be similar (*i.e.* with a small distance). Otherwise, the two generated images should be different (*i.e.* with a large distance). To this end, we adopt the contrastive loss to distill the semantic information from the input pair of descriptions.

The contrastive loss is firstly introduced in [12] and the loss function is formulated as

$$L_c = \frac{1}{2N} \sum_{n=1}^{N} y \cdot d^2 + (1-y) \max(\varepsilon - d, 0)^2, \qquad (1)$$

where $d = \|v_1 - v_2\|_2$ is the distance between the visual feature vectors v_1 and v_2 from the two Siamese branches respectively, and y is a flag to mark whether the input descriptions are from the same image or not, *i.e.*, 1 for the same and 0 for different. The hyper-parameter N is the length of the feature vector and its value is set as 256 empirically in the experiments. The hyper-parameter ε is used to balance the distance value when $y = 0$ and its value is set as 1.0 in the experiments.

With the contrastive loss, the Siamese structure is optimized by minimizing the distance between the generated images from the descriptions of the same image and maximizing the distance of those generated from the descriptions of different images. Note that due to the input noises, even

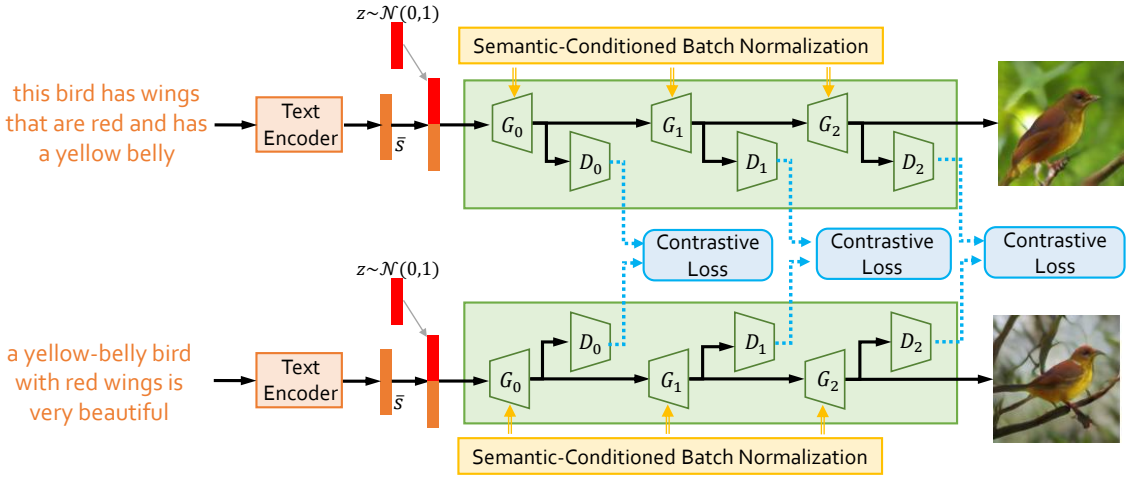

Figure 2. The architecture of SD-GAN. The robust semantic-related text-to-image generation is optimized by contrastive losses based on a Siamese structure. The Semantic-Conditioned Batch Normalization (SCBN) is introduced to further retain the unique semantic diversities from text and embed the visual features modulated to the textual cues.

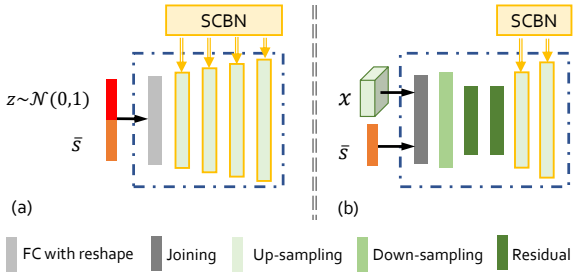

Figure 3. Illustration of the generators in the proposed SD-GAN: (a) G_0, the generator at the initial stage from the linguistic to vision, (b) G_1/G_2, the generator at the second/third stage for generating higher-resolution images based on generated visual features at the former stage. The SCBNs operate at the end of each up-sampling layer.

though the input descriptions are exactly the same, the generated images might be different more or less in appearance, *e.g.*, pose, background and *etc*. To avoid collapsed nonsensical mode in the visualization (*i.e.*, the generated images are too close in appearance), the distance of their feature vectors are not required to be "zero". Therefore, we modify the Eq. 1 as

$$L_c = \frac{1}{2N}\sum_{n=1}^{N} y\max(d,\alpha)^2 + (1-y)\max(\varepsilon - d, 0)^2, \quad (2)$$

where α is a hyper-parameter to avoid the fake images generated too closely even though the input two descriptions are from the same image. We set $\alpha = 0.1$ in the experiments.

3.2. Semantic-Conditioned Batch Normalization (SCBN)

In this work, we consider the linguistic concepts as the kernels of visual representations for cross-domain generation from linguistic to vision. Inspired by the instance normalization in the existing works [15, 5, 38], we modulate the conditional batch normalization with the linguistic cues from the natural language descriptions, defined as *Semantic-Conditioned Batch Normalization* (SCBN). The purpose of SCBN is to reinforce the visual-semantic embedding in the feature maps of the generative networks. It enables the linguistic embedding to manipulate the visual feature maps by scaling them up or down, negating them, or shutting them off, *etc*. It complements to the Siamese structure introduced in Sec. 3.1 which only focuses on distilling semantic commons but ignore the unique semantic diversities in the text.

Batch Norm - Given an input batch $x \in \mathbb{R}^{N\times C\times H\times W}$, BN normalizes the mean and standard deviation for each individual feature channel as

$$\mathrm{BN}(x) = \gamma \cdot \frac{x-\mu(x)}{\sigma(x)} + \beta, \quad (3)$$

where $\gamma, \beta \in \mathbb{R}^C$ are affine parameters learned from data, and $\mu(x), \sigma(x) \in \mathbb{R}^C$ are the mean and standard deviation which are computed across the dimension of batch and spatial independently for each feature channel.

Conditional Batch Norm - Apart from learning a single set of affine parameters γ and β, Dumoulin *et al*. [6] proposed the Conditional Batch Normalization (CBN) that learns the modulation parameters γ_c and β_c with the conditional cues c. The CBN module is a special case of the more general scale-and-shift operation on feature maps. The modified normalization function is formatted as

$$\mathrm{BN}(x|c) = (\gamma+\gamma_c)\cdot\frac{x-\mu(x)}{\sigma(x)} + (\beta+\beta_c). \quad (4)$$

Semantic-Conditioned Batch Normalization - To reinforce the visual-semantic embedding for the visual genera-

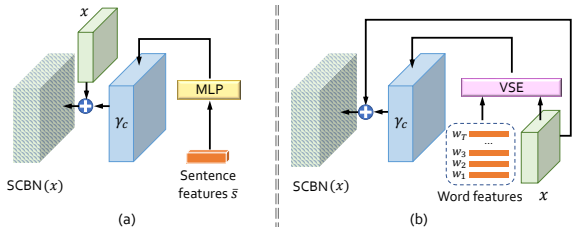

Figure 4. Semantic-conditioned batch normalization (SCBN) with (a) sentence-level cues that consists of a one-hidden-layer MLP to extract modulation parameters from the sentence feature vector; and (b) word-level cues that uses VSE module to fuse the visual features and word features. Note that the illustration only takes γ_c as the example and the implementation for β_c is alike.

tion, we implement the proposed SCBN layers in the generators, as shown in Fig. 3. Firstly, we recap the text encoder (*i.e.*, bi-directional LSTM) to obtain the linguistic features from the input description. Denote the linguistic features of the t^{th} word as w_t. The last hidden states are adopted as the global sentence vector $\bar{s}$. Therefore, the linguistic cues for SCBN can be obtained from two aspects, *i.e.*, sentence-level and word-level.

(1) *Sentence-level Cues.* In order to embed the sentence feature, we adopt a one-hidden-layer multi-layer perceptron (MLP) to extract modulation parameters γ_c and β_c respectively from the sentence feature vector $\bar{s}$ of the input description, as shown in Fig. 4 (a).

$$\gamma_c = f_\gamma(\bar{s}), \beta_c = f_\beta(\bar{s}), \tag{5}$$

where $f_\gamma(\cdot)$ and $f_\beta(\cdot)$ denote the one-hidden-layer MLPs for γ_c and β_c respectively. Then we extend the dimension of $f_\gamma(\bar{s})$ and $f_\beta(\bar{s})$ to the same size as x for embedding the linguistic cues and visual features with Eq. 4. Then the instance sentence features modulate the neural activations of the generated visual features by channel-wise.

(2) *Word-level Cues.* Denote $\mathcal{W} = \{w_t\}_{t=1}^{T} \in \mathbb{R}^{D\times T}$ as the set of word features, where w_t is the feature of the t-th word, and $\mathcal{X} \in \mathbb{R}^{C\times L}$ as the visual features where C is the channel size and $L = W \times H$. Inspired by [39, 8, 7, 37], the visual-semantic embedding (VSE) module is adopted for mutual fusion of word features and visual features, as shown in Fig. 4 (b). We first use a perception layer (*i.e.*, $f(w_t)$) to match the dimension of textual features and visual features. Then the VSE vector $\texttt{vse}_j$ is computed for each sub-region j of the image based on its embedded features v_j which is a dynamic representation of word vectors $\{w_t\}_{t=1}^{T}$ relevant to its visual feature v_j.

$$\texttt{vse}_j = \sum_{t=0}^{T-1} \sigma(v_j^\top \cdot f(w_t)) f(w_t), \tag{6}$$

where $\sigma(v_j^\top \cdot f(w_t))$ indicates the visual-semantic embedding weight of t^{th} word vector w_t for the j^{th} sub-region v_j of visual feature maps, similar as the dot-product similarity of cross correlation. $\sigma(\cdot)$ is the $\texttt{softmax}$ function in the experiments. We then adopt two $\texttt{conv\_1} \times 1$ layers for computing the word-level modulation parameters γ_c and β_c respectively from the VSE matrix.

4. Experiments

4.1. Experiment Settings

Datasets. Following previous text-to-image methods [37, 40, 41], our method is evaluated on CUB [34] and MS-COCO [22] datasets. The CUB dataset contains 200 bird species, it includes 11788 images with 10 language descriptions for each image. Following the settings in [37, 40, 41], we split the CUB dataset into class-disjoint training and test sets, *i.e.*, 8855 images for training and 2933 for test. All images in CUB dataset are preprocessed and cropped to ensure that bounding boxes of birds have greater-than-0.75 object-image size ratios. The MS-COCO dataset is more challenging for text-to-image generation. It has a training set with $80k$ images and a validation set with $40k$ images. It has 5 language descriptions for each image.

Training Details. Apart from the contrastive losses introduced in Sec. 3.1, the generator and the discriminator losses of the proposed SD-GAN follow those in [37] due to its excellent performance. The text encoder and inception model for visual features used in visual-semantic embedding are pretrained by [37] and fixed during the end-to-end training. [1] The network parameters of the generator and discriminator are initialized randomly.

Evaluation Details. It is not easy to evaluate the performance of the generative models. Following prior arts on text-to-image generation [37, 40, 41, 14, 42, 18], we apply the numerical assessment approach "inception score" [31] for quantitative evaluation. In our experiments, we directly use the pre-trained Inception model provided in [40] to evaluate the performance on CUB and MS-COCO datasets.

Although the inception score has shown well correlated with human perception on visual quality [31], it cannot tell whether the generated images are well conditioned on the text descriptions. Therefore, as a complementary, we also design a subject test to evaluate the generation performance. We randomly select 50 text descriptions for each class in the CUB test set and 5000 text descriptions in the MS-COCO test set. Given the same descriptions, 50 users (not including any author) are asked to rank the results by different methods. The average ratio ranked as the best by human users are calculated to evaluate the compared methods.

[1] We also finetuned these models with the whole network, however the performance was not improved.

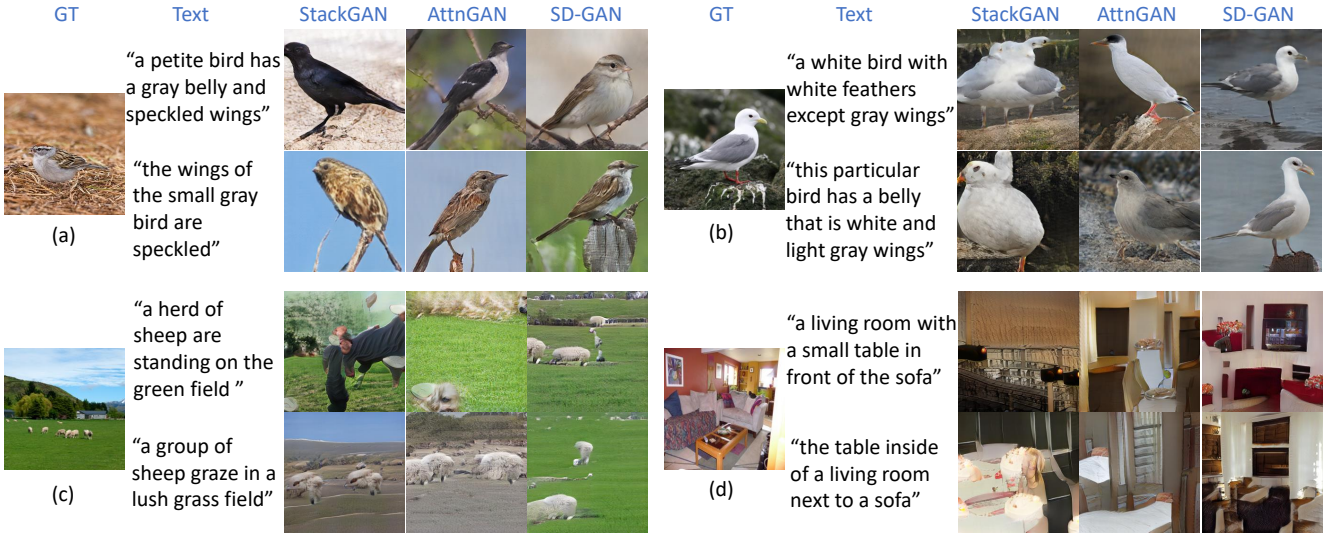

Figure 5. Qualitative examples of the proposed SD-GAN comparing with StackGAN [40] and AttnGAN [37] on CUB (top) and MS-COCO (bottom) test sets. For each example, the images are generated by the methods based on two randomly-selected descriptions (Text) from the same ground-truth image (GT).

4.2. Comparing with the state-of-the-arts

We compare our results with the state-of-the-art text-to-image methods on CUB and MS-COCO dataset. The inception scores for our proposed SD-GAN and other compared methods are listed in Tab. 1. On the CUB dataset, our SD-GAN achieves the inception score $4.67 \pm .09$, which significantly outperforms the previous best method with an inception score $4.36 \pm .03$. More impressively, our SD-GAN boosts the best reported inception score on the MS-COCO dataset from $25.89 \pm .47$ to $35.69 \pm .50$. [2] The excellent performances on the datasets demonstrate the effectiveness of our proposed SD-GAN, thanks to the semantics-disentangling generation and visual-semantic embedding.

The results of subjective test are shown in Tab. 2. We compared the proposed SD-GAN with the previous methods, *i.e.*, StackGAN [40] and AttnGAN [37]. When users are asked to rank images based on their relevance to input text, they choose the generated images by SD-GAN as the best mostly, wining about 70% of the presented texts, much higher than others. This is consistent with the improvements of inception score listed in Table 1. Furthermore, the qualitative results are shown in Fig. 5. For each example, we compare the generation results from the descriptions of the same ground-truth image. Due to the lacking of the word-level details, StackGAN fails to predict the important semantic structure of object and scene. Although AttnGAN adopts the attention mechanism to extract details from the text, it is difficult to generate the corresponding visual concepts under linguistic expression variants, *e.g.*, *gray wings of white bird* in Fig. 5(b), *sheep on the grass* in Fig. 5(c), and *etc*. Comparing to them, the proposed SD-GAN generates more recognizable and semantically meaningful images based on the input texts.

Methods	CUB	MS-COCO
GAN-INT-CLS [29]	$2.88 \pm .04$	$7.88 \pm .07$
GAWWN [30]	$3.62 \pm .07$	-
StackGAN [40]	$3.70 \pm .04$	$8.45 \pm .03$
StackGAN++ [41]	$4.04 \pm .05$	-
PPGN [24]	-	$9.58 \pm .21$
AttnGAN [37]	$\underline{4.36 \pm .03}$	$\underline{25.89 \pm .47}$
HDGAN [42]	$4.15 \pm .05$	$11.86 \pm .18$
Cascaded C4Synth [19]	$3.92 \pm .04$	-
Recurrent C4Synth [19]	$4.07 \pm .13$	-
LayoutSynthesis [14]	-	$11.46 \pm .09$
SceneGraph [18]	-	$6.70 \pm .01$
SD-GAN	$\mathbf{4.67 \pm .09}$	$\mathbf{35.69 \pm .50}$

Table 1. Quantitative results of the proposed method comparing with the state-of-the-arts on CUB and MS-COCO test sets. The bold results are the highest and the underline ones are the second highest.

Methods	CUB	MS-COCO
StackGAN [40]	10.70%	6.53%
AttnGAN [37]	20.54%	17.69%
SD-GAN	68.76%	75.78%

Table 2. Human evaluation results (ratio of 1st by human ranking) of SD-GAN comparing with StackGAN [40] and AttnGAN [37].

Transferable Siamese structure and SCBN. Furthermore, we demonstrate the benefits of the proposed Siamese struc-

[2] The inception score of CUB dataset is much lower than that of MS-COCO because the CUB dataset consists of fine-grained bird images while MS-COCO consists of images from more diverse scenarios. The generated images in MS-COCO is more suitable to be classified by the Inception model.

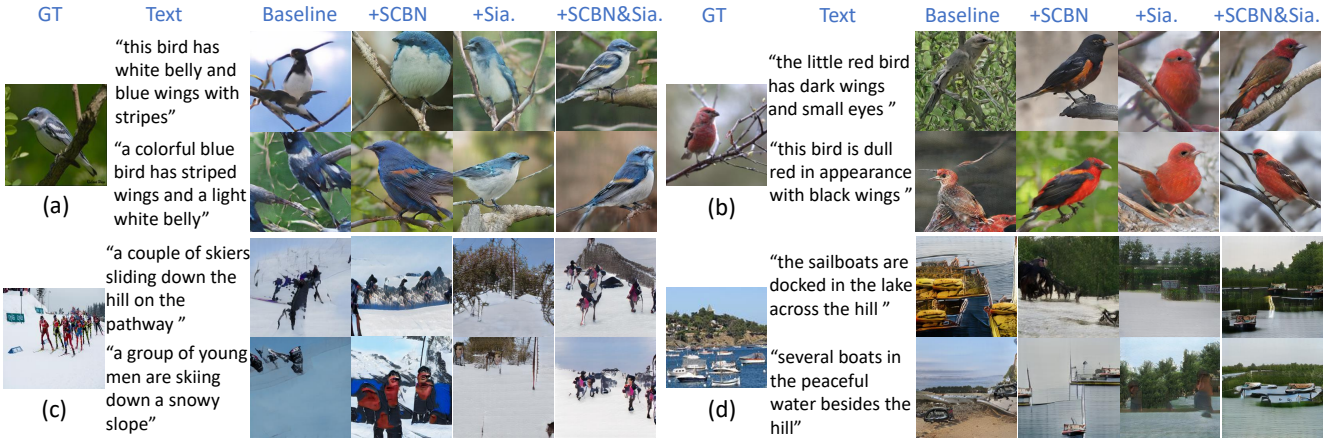

Figure 6. Image generation results of SD-GAN on CUB (top) and MS-COCO (bottom) test sets. For each sample, the images are generated by the methods based on two randomly-selected descriptions (Text) per ground-truth image (GT). The results of of baseline (SD-GAN without SCBN&Siamese) and its variants by adding the proposed SCBN and Siamese structure (Sia.) step by step.

Methods	CUB	MS-COCO
AttnGAN [37]	$4.36 \pm .03$	$25.89 \pm .47$
AttnGAN [37] + Siamese	$4.47 \pm .09$	$29.77 \pm .51$
AttnGAN [37] + SCBN	$4.48 \pm .08$	$29.42 \pm .45$
AttnGAN [37] + Siamese + SCBN	$4.62 \pm .09$	$35.50 \pm .56$

Table 3. Quantitative results of the combined models that incorporate the proposed Siamese structure and SCBN into the previous state-of-the-art architecture on CUB and MS-COCO test sets.

ID	Components		CUB	MS-COCO
	Siamese	SCBN		
1	√	√	$\mathbf{4.67 \pm .09}$	$\mathbf{35.69 \pm .50}$
2	√	-	$4.51 \pm .07$	$30.18 \pm .47$
3	-	√	$4.49 \pm .06$	$29.79 \pm .61$
4	-	-	$4.11 \pm .04$	$23.76 \pm .40$

Table 4. Component Analysis of the SD-GAN. *Siamese* indicates adopting the Siamese structure and *SCBN* indicates using the proposed SCBN layer. The bold results are the best.

ture and SCBN for image generation by plugging them into the existing works. Here we take the previous method, *i.e.*, AttnGAN [37], as the backbone because of its excellent performance. We compare three configurations, *i.e.*, *AttnGAN + Siamese*, *AttnGAN + SCBN* and *AttnGAN + Siamese + SCBN* under the same hyper-parameters for fair comparisons. As shown in Tab. 3, the performance of AttnGAN is improved by a considerable margin on the inception score after applying the Siamese structure (*i.e.*, *AttnGAN + Siamese*). The results again suggest the superiority of the proposed Siamese structure which is applied on AttnGAN. AttnGAN with SCBNs (*i.e.*, *AttnGAN + SCBN*) achieves a better performance than AttnGAN as well. Note that the overall performance by adding both Siamese structure and SCBN (*i.e.*, *AttnGAN + Siamese + SCBN*) surpasses that of AttnGAN itself and achieves the approximate results with our proposed SD-GAN.

4.3. Component Analysis

In this section, to evaluate the effectiveness of the proposed SCBN and Siamese structure with contrastive losses, we first quantitatively evaluate SD-GAN and its variants by removing each individual cue step by step, *i.e.*, 1) *SD-GAN w/o SCBN* (Model 2), SD-GAN without the proposed SCBNs, 2) *SD-GAN w/o Siamese* (Model 3), SD-GAN without Siamese structure, 3) *SD-GAN w/o SCBN & Siamese* (Model 4), SD-GAN without the proposed SCBNs and Siamese structure, regarded as the baseline of SD-GAN. The quantitative results are listed in Tab.4.

By comparing Model 3 (with SCBNs) and Model 4 (baseline) in Tab. 4, the proposed SCBN can help to enforce the visual-semantic embedding, which significantly improves the inception score from 4.11 to 4.49 on CUB and 23.76 to 29.79 on MS-COCO. When adopting the Siamese structure (Model 2) based on Model 4, the inception score can achieve 4.51 (versus 4.11) on CUB dataset. By combining the proposed SCBNs and Siamese structure, Model 1 obtains a significantly improvement and outperforms Model 3 by improving the inception score from 4.49 to 4.67 on CUB and 29.79 to 35.69 on MS-COCO. The Siamese structure makes it possible to maintain the generation consistency and handle the generation deviation because of the input expression variations. The comparisons demonstrate the superiority of the proposed SCBN and Siamese structure for text-to-image generation.

To better understand the effectiveness of the proposed modules, we visualize the generation results of SD-GAN and its variants. As shown in Fig. 6, the baseline without Siamese structure and SCBN just sketches the primitive shape of objects lacking the exact descriptions. By adding the proposed SCBN (+SCBN), the models learn to

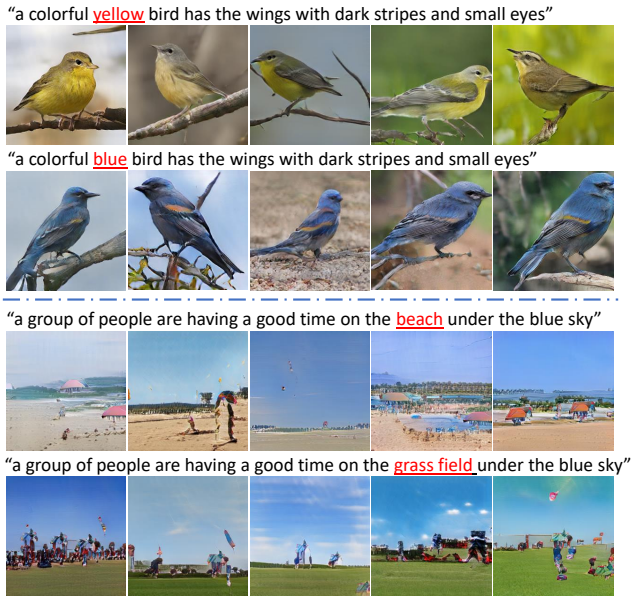

Figure 7. Examples of SD-GAN on the ability of catching subtle changes (underline word or phrase in red) of the text descriptions on CUB (top) and MS-COCO (bottom) test sets.

Methods		CUB	MS-COCO
α	0.01	$4.50 \pm .08$	$32.53 \pm .77$
	0.05	$4.55 \pm .10$	$33.18 \pm .62$
	0.1	$4.67 \pm .09$	$35.69 \pm .50$
	0.2	$4.49 \pm .07$	$31.74 \pm .91$
position	(D1, D2, D3)	$4.67 \pm .09$	$35.69 \pm .50$
	(D2, D3)	$4.59 \pm .10$	$33.13 \pm .74$
	(D3)	$4.56 \pm .09$	$32.88 \pm .82$

Table 5. Ablation study on the contrastive loss. We compare the variants of SD-GAN with different values of hyper-parameter α, *i.e.* $0.01, 0.05, 0.1, 0.2$. Then we compare the variants of SD-GAN by removing the contrastive loss at the individual stage.

Methods	CUB	MS-COCO
SCBN - sent	$4.39 \pm .06$	28.81 ± 0.53
SCBN - word	$4.45 \pm .06$	29.79 ± 0.61
BN - sent	$4.19 \pm .05$	$24.18 \pm .56$
BN - word	$4.23 \pm .05$	$25.34 \pm .79$

Table 6. Ablation study on SCBN. *SCBN-sent* indicates using the SCBN layers conditioned on the sentence-level cues; *SCBN-word* indicates using the SCBN layers conditioned on the word-level cues; *BN-sent* indicates using BN layers and then concatenating sentence-level cues by channel-wise; *BN-word* indicates using BN layers and then concatenating word-level cues by channel-wise.

rectify defects by embedding more linguistic details into the generation procedure, *e.g.* "blue wings" in Fig. 6(a), but the generated birds belong to different categories in appearance due to the expression variants. The model with Siamese structure (+Sia.) can generate similar images from different descriptions of the same image, but might lose the detailed semantic informations, *e.g.*, "black wings" in Fig. 6(b). By combining the Siamese structure and SCBN (+SCBN&Sia.), the models can achieve visibly significant improvements. On the challenging MS-COCO dataset, we have similar observations. Although the generation is far from perfection, the generated images can still be recognized from the text semantics as shown in the bottom of Fig. 6. Those observations demonstrate that the SD-GAN not only maintain the generation consistency but also contains the detailed semantics.

Furthermore, to evaluate the sensitivity of the proposed SD-GAN, we change just one word or phrase in the input text descriptions. As shown in Fig. 7, the generated images are modified according to the changes of the input texts, *e.g.*, bird color (*yellow* versus *blue*) and image scene (*beach* versus *grass field*). It demonstrates the proposed SD-GAN retains the semantic diversities & details from text and has the ability to catch subtle changes of the text descriptions. On the other hand, there are no collapsed nonsensical mode in the visualization of the generated images.

Contrastive Losses. The value of α in Eq. (7) is worth investigating because it can be used to find a trade-off between effectiveness of distilling semantic commons and retaining the semantic diversities from the descriptions of the same image. We validate the value of α among $0.01, 0.05, 0.1$ and 0.2 of SD-GAN. By comparing the results listed in Tab. 5, we adopt α as 0.1 for further experiments as it has the best performances on both CUB and MS-COCO datasets.

Furthermore, we explore the effectiveness of contrastive losses at each stage by removing the contrastive loss stage by stage, *i.e.*, 1) (D1, D2, D3) indicates the contrastive losses are implemented at all the stages as shown in Fig. 2, 2) (D2, D3) indicates only at the last two stages and 3) (D3) indicates only at the last stage. By comparing (D1, D2, D3) with (D2, D3) and (D3) in Tab. 5, the model with contrastive loss implemented at each stage (D1, D2, D3) achieves the best performances.

Semantic-Conditioned Batch Normalization (SCBN). To evaluate the benefits of the proposed SCBN layer, we compare the variants of the SCBN layers. We conduct the experiments with the architecture of SD-GAN without Siamese structure due to the less computational cost during the training. As introduced in Sec. 3.2, the linguistic cues are from sentence-level and word-level. Firstly, we compare the model with SCBN layer on sentence-level linguistic cues, *i.e.*, *SCBN - sent*, and that with word-level cues, *i.e.*, *SCBN - word*. By comparing the results listed in Tab. 6, the SCBN layer with word-level cues outperforms that with sentence-level cues, *i.e.*, 4.45 versus 4.39 on CUB dataset. The word-level features provide more details than the coarse sentence-level features and the visual-semantic embedding defined in Eq. (6) enables the visual modulation in the spatial config-

urations by the linguistic cues.

In addition, we replace the proposed SCBN layer with the general BN layer. The linguistic cues are embedded into the visual feature maps as well by concatenating in channels directly after BN. The BN layers with sentence-level and word-level cues are represented by *BN - sent* and *BN - word* respectively. By comparing the results of *SCBN - sent* versus *BN - sent* and *SCBN - word* versus *BN - word* in Tab. 6, both of the SCBN layers outperform the corresponding BN layers in the experiments. No doubt that the proposed SCBN is more efficient and powerful for embedding the linguistic cues into the generated vision.

5. Conclusion

In this paper, we propose an innovative text-to-image generation framework, named as Semantics Disentangling Generative Adversarial Networks (SD-GAN), that effectively exploit the semantics of the input text within the generation procedure. The proposed SD-GAN adopts a Siamese structure to distills semantic commons from the linguistic descriptions, based on which the generated images can keep generation consistency under expression variants. Furthermore, to complement the Siamese mechanism that may lose unique semantic diversities, we design an enhanced visual-semantic embedding method by reformulating the batch normalization layer with the instance linguistic cues. Extensive experiments demonstrate the respective effectiveness and significance of the proposed SD-GAN on the CUB dataset and the challenging large-scale MS-COCO dataset.

Acknowledgment This work is supported in part by the National Natural Science Foundation of China (Grant No. 61371192), the Key Laboratory Foundation of the Chinese Academy of Sciences (CXJJ-17S044) and the Fundamental Research Funds for the Central Universities (WK2100330002, WK3480000005), in part by SenseTime Group Limited, the General Research Fund sponsored by the Research Grants Council of Hong Kong (Nos. CUHK14213616, CUHK14206114, CUHK14205615, CUHK14203015, CUHK14239816, CUHK419412, CUHK14207-814, CUHK14208417, CUHK14202217), the Hong Kong Innovation and Technology Support Program (No.ITS/121/15FX).

References

[1] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

[4] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *ICCV*, 2017.

[5] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *NIPS*, pages 6594–6604, 2017.

[6] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. 2017.

[7] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[9] Hao Ge, Yin Xia, Xu Chen, Randall Berry, and Ying Wu. Fictitious gan: Training gans with historical models. In *ECCV*, September 2018.

[10] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Xiaogang Wang, and Hongsheng Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NIPS*, 2018.

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

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, pages 1735–1742. IEEE, 2006.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, pages 1735–1780, 1997.

[14] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *CVPR*, 2018.

[15] Xun Huang and Serge J Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[16] Xun Huang, Yixuan Li, Omid Poursaeed, John E Hopcroft, and Serge J Belongie. Stacked generative adversarial networks. In *CVPR*, 2017.

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *JMLR*, 2015.

[18] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. *CVPR*, 2018.

[19] KJ Joseph, Arghya Pal, Sailaja Rajanala, and Vineeth N Balasubramanian. C4synth: Cross-caption cycle-consistent text-to-image synthesis. *WACV*, 2019.

[20] Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino. Generative attribute controller with conditional filtered generative adversarial networks. In *CVPR*, 2017.

[21] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *CVPR*, 2017.

[22] 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 *ECCV*, 2014.

[23] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual gan. In *ECCV*, September 2018.

[24] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *CVPR*, 2017.

[25] Vu Nguyen, Tomas F Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, 2017.

[26] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *NIPS*, 2016.

[27] Ethan Perez, Harm De Vries, Florian Strub, Vincent Dumoulin, and Aaron Courville. Learning visual reasoning without strong priors. In *ICML 2017's Machine Learning in Speech and Language Processing Workshop*, 2017.

[28] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[29] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *ICML*, 2016.

[30] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *NIPS*, 2016.

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, pages 2234–2242, 2016.

[32] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, pages 791–808. Springer, 2016.

[33] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer, 2016.

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

[35] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *ECCV*, September 2018.

[36] Jiqing Wu, Zhiwu Huang, Janine Thoma, Dinesh Acharya, and Luc Van Gool. Wasserstein divergence for gans. In *ECCV*, September 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. *CVPR*, 2018.

[38] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *CVPR*, 2018.

[39] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *CVPR*, pages 4187–4195. IEEE, 2017.

[40] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *ICCV*, 2017.

[41] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1710.10916*, 2017.

[42] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, 2018.

[43] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, September 2018.

6. Appendix

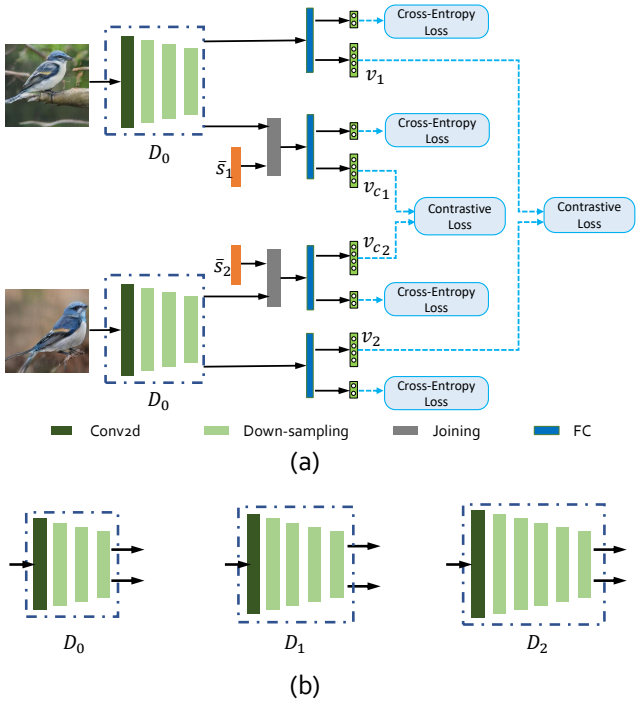

Figure 8. Network architecture of discriminators.

6.1. Architecture of Discriminators

As described in Sec.3 in the main paper, at each branch of the Siamese structure, we adopt hierarchical stages from low-resolution to high-resolution for the photo-realistic image generation. At each stage, the generator is followed by a discriminator that distinguishes whether the image is real or fake. As shown in Fig. 8 (a), the input image is processed by several convolutional layers (in dot-line bounding box) for extracting the visual features. The visual features are fed into two branches, where each branch has two outputs, *i.e.*, a classification vector for cross-entropy loss (1 for real, 0 for fake) and a feature vector for contrastive loss. Differs to the first branch, the second branch has its input as a concatenation of the sentence-level feature $\bar{s}$ and the visual feature map, following [40, 37, 41]. The contrastive loss (Eq.2 in the main paper) is calculated as follows,

$$L_c = \frac{1}{2N}\sum_{n=1}^{N} y \max(d,\alpha)^2 + (1-y)\max(\varepsilon - d, 0)^2 + \frac{1}{2N}\sum_{n=1}^{N} y \max(d_c,\alpha)^2 + (1-y)\max(\varepsilon - d_c, 0)^2, \tag{7}$$

where $d = \|v_1 - v_2\|_2$, $d_c = \|v_{c_1} - v_{c_2}\|_2$ is the distance between the visual feature vectors from the two Siamese branches respectively.

The discriminators D_0, D_1, D_2 in Fig.8 (b) have similar structures. To obtain the output of each discriminator with the same size, the discriminator is constructed with different number of down-sampling layers. These discriminators are independent and will not share the parameters.

6.2. More Results

Additional qualitative results of SD-GAN. The additional qualitative comparisons are visualized in Fig. 9: (1) we compare the results between different module configurations of the proposed SD-GAN, and (2) we show the excellent performance of SD-GAN, compared with the state-of-the-art methods, *i.e.*, StackGAN [40] and AttnGAN [37]. The details are depicted in Sec.4.2 and Sec.4.3 in the main paper.

More generated results. When visualizing a large number of generated images by the proposed SD-GAN, we do not observe obvious nonsensical modes on both CUB and MS-COCO datasets. Since the limited size of the supplementary material, here we only show 400 images for each dataset, as shown Fig. 10 and Fig.11 respectively. These images are randomly selected, and the original resolution is 256×256 (Please zoom in to view more details).

Figure 9. Additional qualitative resutls of the proposed SD-GAN. For each example, the images are generated by the methods based on two randomly-selected descriptions (Text) per ground-truth image (GT). (a) The results of baseline (SD-GAN without SCBN & Siamese) and its variants by adding the proposed SCBN and Siamese structure step by step. (b) The results of SD-GAN comparing with the state-of-the-art methods, *i.e.*, StackGAN [40] and AttnGAN [37].

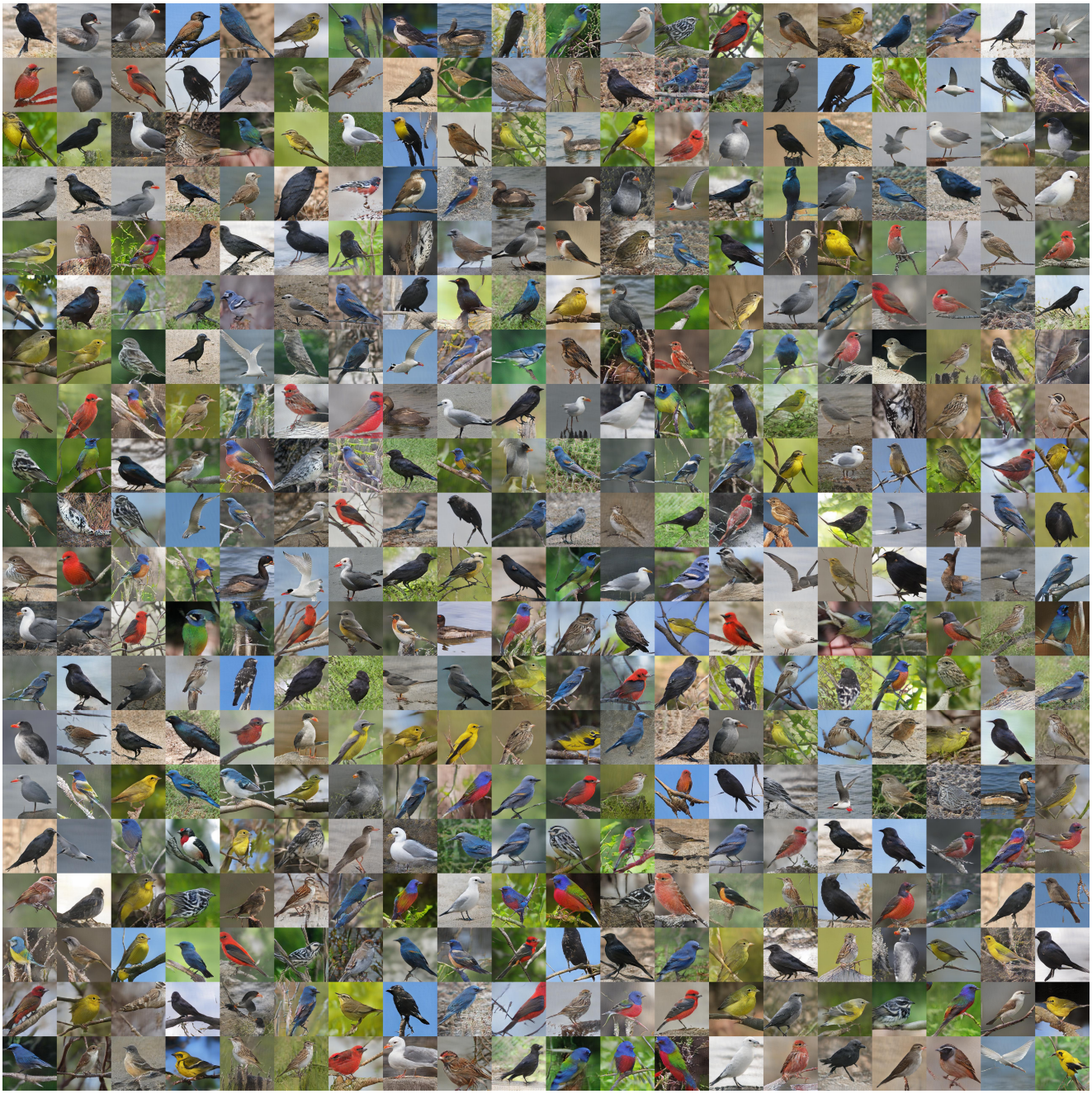

Figure 10. Generated images randomly-sampled from CUB dataset (Please zoom in to view more details).

Figure 11. Generated images randomly-sampled from MS-COCO dataset (Please zoom in to view more details).