Text to Image Generation with Semantic-Spatial Aware GAN

Wentong Liao[1*], Kai Hu[1*†], Michael Ying Yang[2], Bodo Rosenhahn[1]
[1]TNT, Leibniz University Hannover, Germany, [2]SUG, University of Twente, The Netherlands

Abstract

Text-to-image synthesis (T2I) aims to generate photo-realistic images which are semantically consistent with the text descriptions. Existing methods are usually built upon conditional generative adversarial networks (GANs) and initialize an image from noise with sentence embedding, and then refine the features with fine-grained word embedding iteratively. A close inspection of their generated images reveals a major limitation: even though the generated image holistically matches the description, individual image regions or parts of somethings are often not recognizable or consistent with words in the sentence, e.g. *"a white crown". To address this problem, we propose a novel framework Semantic-Spatial Aware GAN for synthesizing images from input text. Concretely, we introduce a simple and effective Semantic-Spatial Aware block, which (1) learns semantic-adaptive transformation conditioned on text to effectively fuse text features and image features, and (2) learns a semantic mask in a weakly-supervised way that depends on the current text-image fusion process in order to guide the transformation spatially. Experiments on the challenging COCO and CUB bird datasets demonstrate the advantage of our method over the recent state-of-the-art approaches, regarding both visual fidelity and alignment with input text description. Code available at* https://github.com/wtliao/text2image.

1. Introduction

The great advances made in Generative Adversarial Networks (GANs) [7, 20, 22, 38, 11, 35, 2, 13] boost a remarkable evolution in synthesizing photo-realistic images with diverse conditions, such as layout [19, 8], text [34, 30] and scene graph [12, 1, 5]. Particularly, generating images conditioned on text descriptions (see Fig. 1) has been catching increasing attention in computer vision and natural language processing communities because: (1) it bridges the gap between these two domains, and (2) linguistic description (text) is the most natural and convenient medium for human being to describe a visual scene. Nonetheless, T2I remains a challenging task because of the cross-modal problem (text to image transformation) and the ability to keep the generated image holistically as well as locally semantically consistent with the given text.

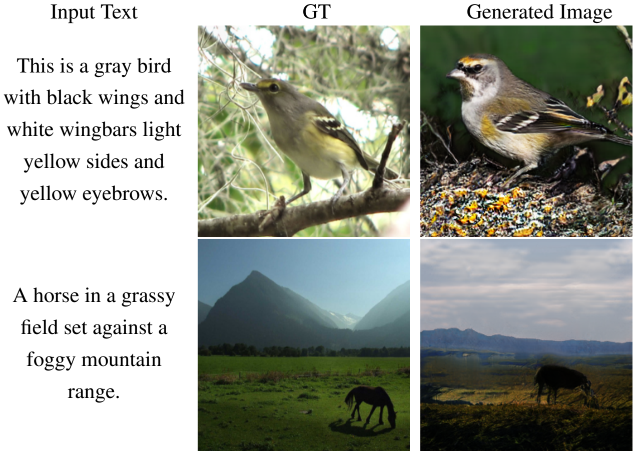
Input Text
GT
Generated Image
This is a gray bird with black wings and white wingbars light yellow sides and yellow eyebrows.
A horse in a grassy field set against a foggy mountain range.

Figure 1: Examples of images generated by our method (3rd column) conditioned on the given text descriptions.

The most recent T2I methods are usually multi-stage refinement frameworks which generates an initial image from noise with sentence embedding and refines the details with fine-grained word embedding in each following stage [34, 35, 10, 30, 16, 31]. In each stage, there is a pair of generator and discriminator to synthesize higher-resolution image and decide whether the generated image is real enough, respectively. This method has proved effective in synthesizing high-resolution images. However, multiple generator-discriminator pairs lead to higher computation and more unstable training processes. Moreover, the quality of the image generated by the earlier generator decides the final output. If the early generated image is poor, the later generators can not improve its quality. To address this problem, the one-stage generator is introduced in [28] which has one generator-discriminator pair. In this work, we also follow this one-stage structure.

On the other hand, the generated image should be holis-

[*] Equal contribution
[†] Kai Hu contributed to this work when he was doing his master thesis supervised by Wentong Liao at TNT.

tically consistent with the description and locally consistent with words in the sentence. For this purpose, the multi-stage refinement framework is used to fuse text and image information in each stage of the generation process to encourage the generated image to be semantically consistent with the corresponding text. AttGAN [30] plays a role in this task. It uses sentence embedding to initialize an image from noise, and judge whether the generated image matches the corresponding text in each stage. This helps the generated images holistically consistent with the description. In parallel, the attention mechanisms are used to select the important words in the text to complement the details in the sub-regions of images in each refinement stage. In this way, the generated image is encouraged to match the words in text semantically. Most of the recent T2I methods follow this framework [16, 21, 4, 24, 33]. Despite the remarkable performance that has been made with these methods, there still exists an important but unsolved limitation: local semantic are not well explored during the synthesis process due to the limited and abstractive textual information. Usually, a text description only describes part of a scene or an object (*e.g*. "a white crown"), and lacks explicit spatial information. To address this problem, previous methods normally utilize cross-modal attention mechanisms to attend word-level features to the image sub-regions [30, 16, 33]. However, the computation cost increases rapidly with larger image size. Moreover, the natural language description is in high-level semantics, while a sub-region of the image is relatively low-level [3, 32]. Last but also important, image sub-regions are still to coarse for complementing the details of somethings. Therefore, the high-level textual semantics cannot be explored well to control the image generation process, especially for complex image with multiple objects, such as in the COCO [18] dataset. Some methods [10, 17, 15] propose object-driven T2I approaches, which first predict object bounding box from text description, and then infer the corresponding segmentation masks. Finally, images are generated from the segmentation masks using PixelGAN [11]. However, such approaches convert T2I task to segmentation to image generation in practice, and the local features of objects are lost completely.

To address the aforementioned issues, we propose a novel T2I framework dubbed as Semantic-Spatial Aware Generative Adversarial Network (SSA-GAN) (see Fig. 2). First, it has only one generator-discriminator pair and is trained in end-to-end fashion so that it can be trained more efficiently and stably compared to the multi-stage refinement framework. Second, only sentence embedding is used to control the image generation process. Compared to the previous methods that also use world-level features, our method requires lower computation. Last but important, our method complements the local details in pixel level rather than in sub-region level. Thus, the generated images are better consistent with the words in text semantically and locally. To realize the pixelwise control of image synthesis, we propose a novel Semantic-Spatial Aware (SSA) block (Fig. 3). On one hand, SSA block learns semantic-aware channel-wise affine parameters conditioned on the learned text feature vector (sentence embedding). On the other hand, a semantic mask is predicted depending on the current text-image fusion process (*i.e*. output of last SSA block). The semantic mask indicates where the generated images still need to be enhanced with the textual information in pixel level. This is how the name *Semantic-Spatial Aware* from. It is worth noting that the mask predictor is trained with weak supervision so that no additional mask annotation is required. Comprehensive experiments are conducted on the challenging benchmarks COCO [18] and CUB bird dataset [29] to validate the performance of SSA-GAN for T2I. The quantitative as well as qualitative experimental results show our superior performance over the previous methods. In summary, the **main contributions** of this paper are as follows:

- We propose a novel one-stage framework SSA-GAN for image synthesis from text. Compared to the popular multi-stage framework, one-stage framework requires less computation and can be trained more efficiently and stably.
- Our method only uses sentence embedding during the synthesis process. Compared to the methods which use world-level futures, our method is simple and has lower computation cost.
- A novel SSA block is introduced to fuse the text and image features effectively and deeply by predicting semantic mask to guide the learned text-adaptive affine transformation in pixel level.
- The semantic mask predictor is trained in a weakly-supervised way, such that no additional annotation is required and this block is potential to be applied on other T2I datasets.

2. Related Work

GANs for Text-to-image Synthesis T2I generation is becoming a hot topic in both CV and NLP communities. Generative Adversarial Networks (GANs) [7] is the most popular model for this task. Reed *et al*. [23] is the first to use conditional GANs (cGANs) to synthesize plausible images from text descriptions. To improve the resolution of generated images, the StackGAN structure is introduced in [34, 35], which stacks multiple generators in sequence in order to generate image from coarse to fine. For training, each generator has its own discriminator for adversarial training. Many recent works follow this structure [30, 37, 16, 39, 24, 33, 4] and have made advances. To overcome the training difficulties in the stacked structure, Ming

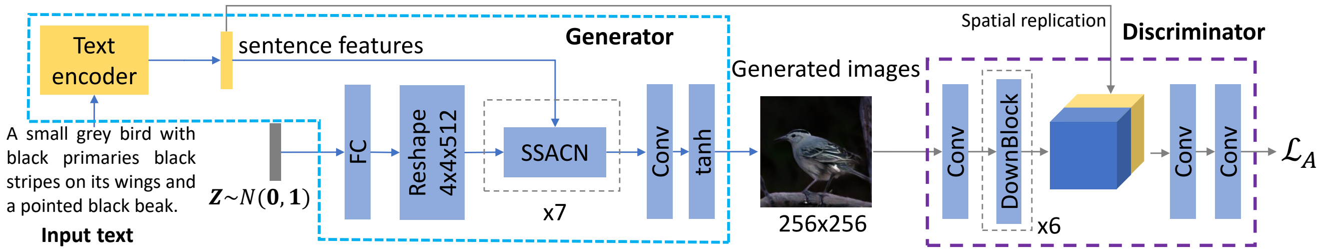

Figure 2: A schematic of our framework SSA-GAN. It has one generator-discriminator pair. The generator mainly consists of 7 proposed SSA blocks which fuse text and image features through the image generation process and guarantee the semantic text-image consistency. The gray lines indicate the data streams only for training.

et al. [28] propose a one-stage structure that has only one generator-discriminator pair for T2I generation. Their generator consists of a series of UPBlocks which is specifically designed to upsample the image features to generate high-resolution images. Our framework follows this one-stage structure to avoid the problems in the stacked structure.

Text-Image Fusion In the early T2I works [23, 34, 35], textual information is fused to the image features by naively concatenating text vector (sentence level) to the sampled noise and intermediate features. AttnGAN [30] utilizes cross-modal attention to repeatedly select important words in text for image sub-regions at each refinement stage for text-image fusion to capture better details. Moreover, it introduces Deep Attentional Multimodal Similarity Model (DAMSM) to measure the image-text similarity both at the word level and sentence level to compute a fine-grained loss for image generation. In this way, the generated image is forced to semantically consistent with the text. Control-GAN [16] further fuses text and image information with word-level spatial and channel-wise attention-driven generator which generates sub-regions features corresponding to the most relevant words during the generation process. Zhu *et al.* [39] proposes DM-GAN which uses memory network to adaptively select the important words to refine the image features iteratively. Yin *et al.* [31] introduces word-level conditioned batch normalization (CBN) in SD-GAN to better align text and image. DF-GAN [28] learns the affine transformation parameters from text vector at each stage. Then, multiple stacked affine transformations are operated on the image feature maps for text-image fusion.

In our work, the semantic-aware batch normalization is conditioned on text vector which requires much less computation compared to the word-level CBN-based methods and word-level cross-modal attention-based methods. Compared to the existing methods, our affine transformation is spatially guided by the semantic mask predicted based on the current text-image fusion process.

3. Method

The architecture of our SSA-GAN is shown in Fig. 2. SSA-GAN has a text encoder that learns text representations, a generator that has 7 SSA blocks for deepening text-image fusion and improving resolution, and a discriminator that is used to judge whether the generated image is semantically consistent to the given text. SSA-GAN takes a text description and a normal-distributed noise vector $z \in \mathbb{R}^{100}$ as input, and outputs an RGB image in size of 256×256. We elaborate each part of our model as follows.

3.1. Text Encoder

We adopt the pre-trained text encoder provided by [30] that has been used in many existing works [16, 28, 39]. The text encoder is a bidirectional LSTM [26] and pre-trained using real image-text pairs by minimizing the Deep Attentional Multimodal Similarity Model (DAMSM) loss [30]. It encodes the given text description into a text vector $\bar{e} \in \mathbb{R}^{256}$, and word features with length 18 $\mathbf{e} \in \mathbb{R}^{256 \times 18}$. The i-th column e_i of $\mathbf{e}$ is the feature vector of the i-th word.

3.2. Semantic-Spatial Aware Block

The core of SSA-GAN is the SSA block as shown in Fig. 3. It takes the encoded text feature vector $\bar{e}$ and image feature maps $f_{i-1} \in \mathbb{R}^{ch_{i-1} \times \frac{h_i}{2} \times \frac{w_i}{2}}$ from last SSA block as input, and outputs the image feature maps $f_i \in \mathbb{R}^{ch_i \times h_i \times w_i}$ which are further fused with the text features. w_i, h_i, ch_i are the width, height and number of channels of the image feature maps generated by the i-th SSA block. The input image feature maps of the first SSA block (no upsampling) are in shape of $4 \times 4 \times 512$ which are achieved by projecting the noise vector z to visual domain using a fully-connected (FC) layer and then reshaping it. Therefore, after 6 times upsampling by SSA blocks, the image feature maps have 256×256 resolution. Each SSA block consists of an upsample block, a semantic mask predictor, a Semantic-Spatial Condition Batch Normalization block with a residual connection. The upsample block is used to double the resolution of image feature maps by bilinear interpolation operation. The residual connection is used to maintain the main contents of the image features to prevent text-irrelevant parts from being changed and the image information being overwhelmed by the text information. More details are introduced as follows.

Weakly-supervised Semantic Mask Predictor The structure of the semantic mask predictor is shown in Fig. 3, as highlighted by the gray dash box. It takes the upsampled image feature maps as input and predicts a semantic mask map $m_i \in \mathbb{R}^{h_i \times w_i}$. The value of its elements $m_{i,(h,w)}$ ranges between $[0, 1]$. Each value decides how much the following affine transformation should be operated on location (h, w). This semantic mask is predicted based on the current generated image feature maps. Thus, it intuitively indicates which parts of the current image feature maps still need to be reinforced with text information so that the refined image feature maps are more semantically consistent to the given text. The semantic mask predictor is trained jointly with the whole network without specific loss function to guide its learning process nor additional mask annotation. The only supervision is from the adversarial loss given by the discriminator which will be discussed in Sec. 3.4. Therefore, it is a weakly-supervised learning process. In the experiments, we will demonstrate at different stages of SSA blocks, how the semantic mask indicates the text-image fusion spatially.

Semantic Condition Batch Normalization We first give a brief review on standard BN and CBN. Given an input batch $x \in \mathbb{R}^{N \times C \times H \times W}$, where N is the batch size, BN first normalizes it into zero mean and unit deviation for each feature channel:

$$\begin{aligned} \hat{x}_{nchw} &= \frac{x_{nchw} - \mu_c(x)}{\sigma_c(x)}, \\ \mu_c(x) &= \frac{1}{NHW}\Sigma_{n,h,w}x_{nchw}, \\ \sigma_c(x) &= \sqrt{\frac{1}{NHW}\Sigma_{n,h,w}(x_{nchw} - \mu_c)^2 + \epsilon}, \end{aligned} \tag{1}$$

where ϵ is a small positive constant for numeric stability. Then, a channel-wise affine transformation is operated:

$$\tilde{x}_{nchw} = \gamma_c \hat{x}_{nchw} + \beta_c, \tag{2}$$

where γ_c and β_c are learned parameters that work on all spatial locations of all samples in a batch equally. During the test, the learned γ_c and β_c are fixed. Apart from using a fixed set of γ and β learned from training data, Dumoulin *et al.* [6] proposed the CBN which learns the modulation parameters γ and β adaptive to the given condition for the affine transformation. Then, Eq. (2) can be reformulated as:

$$\tilde{x}_{nchw} = \gamma(con)\hat{x}_{nchw} + \beta(con). \tag{3}$$

To fuse the text and image features, the modulation parameters γ and β are learned from the text vector $\bar{e}$:

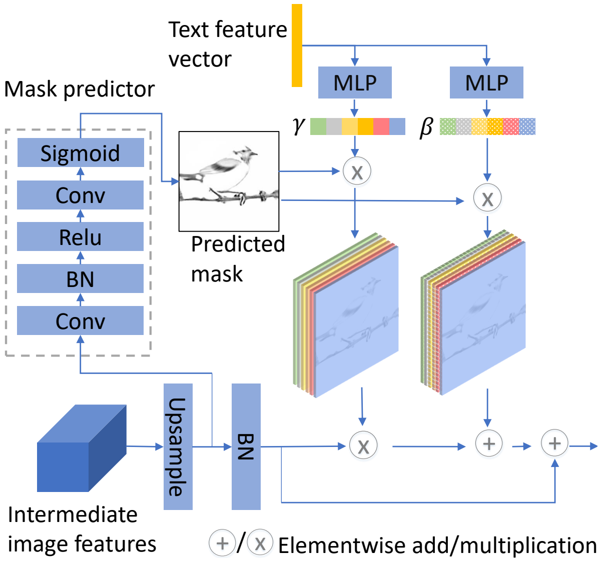

Figure 3: Structure of the SSA block. The text-aware affine parameters are learned and semantic mask is predicted from current image features in order for Semantic-Spatial Condition Batch Normalization.

$$\gamma_c = P_\gamma(\bar{e}), \quad \beta_c = P_\beta(\bar{e}) \tag{4}$$

$P_\gamma(\cdot)$ and $P_\beta(\cdot)$ represent the MLPs for γ_c and β_c, respectively. Here, semantic CBN is realized.

Semantic-Spatial Aware Batch Normalization The semantic aware BN from the last step would work on the image feature maps spatial equally. Ideally, we expect the modulation only works on the text-relevant parts of the feature maps. For this purpose, the predicted semantic mask is added to Eq. (3) as the spatial condition:

$$\tilde{x}_{nchw} = m_{i,(h,w)}(\gamma_c(\bar{e})\hat{x}_{nchw} + \beta_c(\bar{e})). \tag{5}$$

One can see from the formulation that $m_{i,(h,w)}$ does not only decide where to add the text information but also decides how much text information needs to be reinforced on the image feature maps in pixel level.

Summary The modulation parameters γ and β are learned conditioned on the text information, and the predicted semantic mask control the affine transformation spatially. Thus, the text-image fusion is semantic-spatial aware.

3.3. Discriminator

We adopt the one-way discriminator proposed in [28] because of its effectiveness and simplicity, as shown in Fig. 2 (in the violet dashed box). It concatenates the features extracted from generated image and the text vector for computing the adversarial loss through two convolution layers. Associated with the Matching-Aware zero-centered Gradient Penalty (MA-GP) [28], it guides the generator to synthesize more realistic images with better text-image semantic consistency. Because the Discriminator is not the contribution of this work, we will not extend its details here and please refer to the paper for more information.

To further improve the quality of generated images and the text-image consistency, and help train the text encoder jointly with the generator, we add the widely applied DAMSM [30] to our framework. Note that, even without the DAMSM, our method already reports the state-of-the-art performance (see Table 2 in Sec.4).

3.4. Objective Functions

Discriminator Objective The adversarial loss associated with the MA-GP loss is used to train our network.

$$\begin{aligned}\mathcal{L}_{adv}^{D} =& E_{x \backsim p_{data}}[max(0, 1 - D(x, s))] \\ &+ \frac{1}{2} E_{x \backsim p_G}[max(0, 1 + D(\hat{x}, s))] \\ &+ \frac{1}{2} E_{x \backsim p_{data}}[max(0, 1 + D(x, \hat{s}))] \\ &+ \lambda_{MA} E_{x \backsim p_{data}}[(\|\nabla_x D(x, s)\|_2 \\ &+ \|\nabla_s D(x, s)\|_2)^p],\end{aligned} \tag{6}$$

where s is the given text description while $\hat{s}$ is a mismatched text description. x is the real image corresponding to s, and $\hat{x}$ is the generated image. $D(\cdot)$ is the decision given by the discriminator that whether the input image matches the input sentence. The variables λ_{MA} and p are the hyperparameters for MA-GP loss.

Generator Objective The total loss for the generator is composed of an adversarial loss and a DAMSM loss [30]:

$$\begin{aligned}\mathcal{L}_G &= \mathcal{L}_{adv}^{G} + \lambda_{DA} \mathcal{L}_{DAMSM} \\ \mathcal{L}_{adv}^{G} &= -E_{x \backsim p_G}[D(\hat{x}, s)],\end{aligned} \tag{7}$$

where $\mathcal{L}_{DAMSM}$ is a word level fine-grained image-text matching loss, and λ_{DA} is the weight of DAMSM loss.

4. Experiments

We evaluated our method on the COCO [18] and CUB bird [29] benchmark datasets, and compared the performance with the recent state-of-the-art GAN methods on T2I generation, StackGAN++ [35], AttnGAN [30], ControlGAN [16], SD-GAN [31], DM-GAN [39], DF-GAN [28], and DAE-GAN [24]. Series of ablation studies are conducted to get insight of how each proposed module works.

Datasets The CUB bird dataset [29] has 8,855 training images (150 species) and 2,933 test images (50 species). Each bird has 10 text descriptions. The COCO dataset [18] contains 80k training images and 40k test images. Each image has 5 text descriptions. Compared with the CUB dataset, the images in COCO show complex visual scenes, making it more challenging for T2I generation tasks.

Evaluation Metric We follow the previous works to adopt the widely used Inception Score (IS) [25], Fréchet Inception Distance (FID) [9] and R-precision [30] to quantify the performance. For the IS scores, a pre-trained Inception v3 network [27] is used to compute the KL-divergence between the conditional class distribution (generated images) and the marginal class distribution (real images). A large IS indicates that the generated images are of high quality, and each image clearly belongs to a specific class. The FID computes the Fréchet Distance between the features distribution of the generated and real-world images. The features are extracted by a pre-trained Inception v3 network. A lower FID implies the generated images are more realistic. The R-precision is used to evaluate the image-text semantic consistency. The cosine distance between the global image vector and the global sentence vectors of 100 candidates (one ground truth, *i.e.* $R = 1$, and 99 randomly selected mismatching descriptions). The generated image is considered as semantically consistent with the ground truth if their distance is the shortest. To evaluate the IS, FID scores and R-precision, 30k images in resolution 256×256 are generated from each model by randomly selecting text descriptions from the test dataset. For COCO dataset, previous works [28, 36, 17] reported that the IS metric completely fails in evaluating the synthesized images. Hence, we do not compare the IS on the COCO dataset. The FID is more robust and aligns manually evaluation on the COCO dataset.

Implementation details Our model is implemented in Pytorch. The batch size is set to 24 distributed on 4 Nvidia RTX 2080-Ti GPUs. The Adam optimizer [14] with $\beta_1 = 0.0$ and $\beta_2 = 0.9$ is used in the training. The learning rates of the generator and the discriminator are $1e^{-4}$ and $4e^{-4}$, respectively. The hyper-parameters $p = 6$, $\lambda_{MA} = 2$ and $\lambda_{DA} = 0.1$ are adopted. The model is trained for 600 epochs on CUB dataset and 120 epochs on COCO dataset.

4.1. Quantitative Results

Table 1 shows the quantitative results of SSA-GAN and several recent state-of-the-art GAN models for T2I. From the second column of the table we can see that, SSA-GAN reports the significant improvements in IS (from 4.86 to 5.17) on CUB dataset compared to the most recent state-of-the-art method DF-GAN [28]. Higher IS means higher quality and text-image semantic consistency. Our method remarkably decreases the FID score from 28.12 to 19.37 on COCO dataset compared to the state-of-the-art performance. On CUB dataset, our FID score is slightly inferior to the ones given by StackGAN++ [35] and DAE-GAN [24] (15.61 *v.s.* 15.30 and 15.19) but much lower than the other recent methods: 19.24 in DF-GAN [28] and 16.09 in DM-GAN [39]. Our R-precision scores are better than most of the previous methods but inferior to DAE-GAN. The overall

Table 1: Performance of IS, FID and R-precision scores of different state-of-the-art methods, and our method on the CUB and COCO test set. The results are taken from the authors' own papers. Note that the numbers reported in DF-GAN‡ [28] is the updated results from DF-GAN [28]. Best results are in bold.

Methods	IS ↑	FID ↓		R-precision ↑	
	CUB	CUB	COCO	CUB	COCO
StackGAN++ [35]	4.04±0.06	15.30	81.59	-	-
AttnGAN [30]	4.36±0.03	23.98	35.49	67.82±4.43	85.47±3.69
ControlGAN [16]	4.58±0.09	-	-	69.33±3.23	82.43±2.43
SD-GAN [31]	4.67±0.09	-	-	-	-
DM-GAN [39]	4.75±0.07	16.09	32.64	72.31±0.91	88.56±0.28
DF-GAN [28]	4.86±0.04	19.24	28.92	-	-
DF-GAN‡ [28]	5.10	14.81	21.42	-	-
DAE-GAN [24]	4.42±0.04	**15.19**	28.12	**85.4±0.57**	**92.6±0.50**
Ours	**5.17±0.08**	15.61	**19.37**	75.9±0.92	90.6±0.71

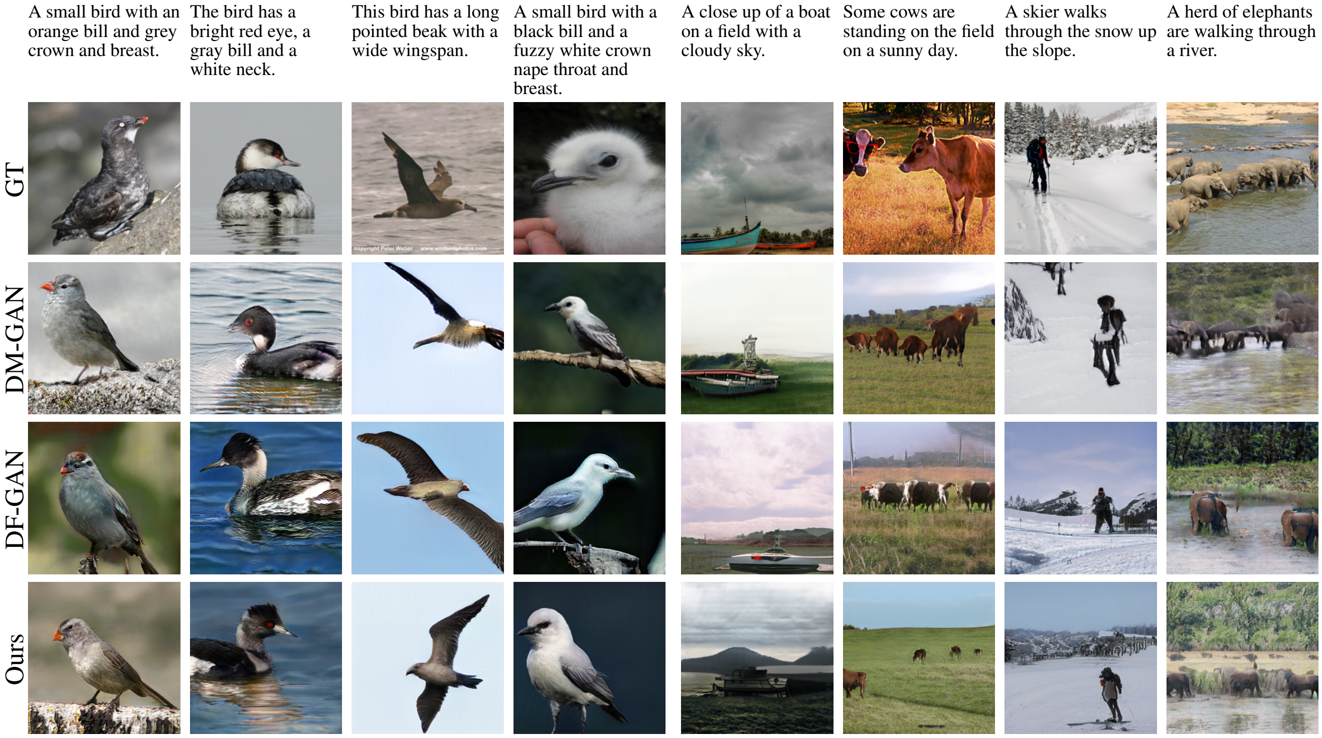
A small bird with an orange bill and grey crown and breast.
The bird has a bright red eye, a gray bill and a white neck.
This bird has a long pointed beak with a wide wingspan.
A small bird with a black bill and a fuzzy white crown nape throat and breast.
A close up of a boat on a field with a cloudy sky.
Some cows are standing on the field on a sunny day.
A skier walks through the snow up the slope.
A herd of elephants are walking through a river.
GT
DM-GAN
DF-GAN
Ours

Figure 4: Qualitative comparison between our method and DM-GAN [39], DF-GAN [28] on the test set of CUB bird dataset (1st - 4th columns) and COCO dataset (5th - 8th columns). The input text descriptions are given in the first row and the corresponding generated images from different methods are shown in the same column. Best view in color and zoom in.

superiority and effectiveness of our SSA-GAN are demonstrated by the extensive quantitative evaluation results that SSA-GAN is able to generate high-quality images with better holistic and local semantic consistency, both for the images with many detailed attributes and more complex images with multiple objects.

Compared with the CUB dataset, the COCO dataset is more challenging because there are always multiple objects in images and the background is more complex. Our superior performances indicate that SSA-GAN is able to synthesize complex images in high quality.

4.2. Qualitative Results

We qualitatively compare the generated images from our method and three recent state-of-the-art GAN models for T2I, *i.e.* DM-GAN [39], DF-GAN [28] and DAE [24].

For the CUB Bird dataset, shown in the first 4 columns in Fig. 4, our SSA-GAN generates images with more vivid de-

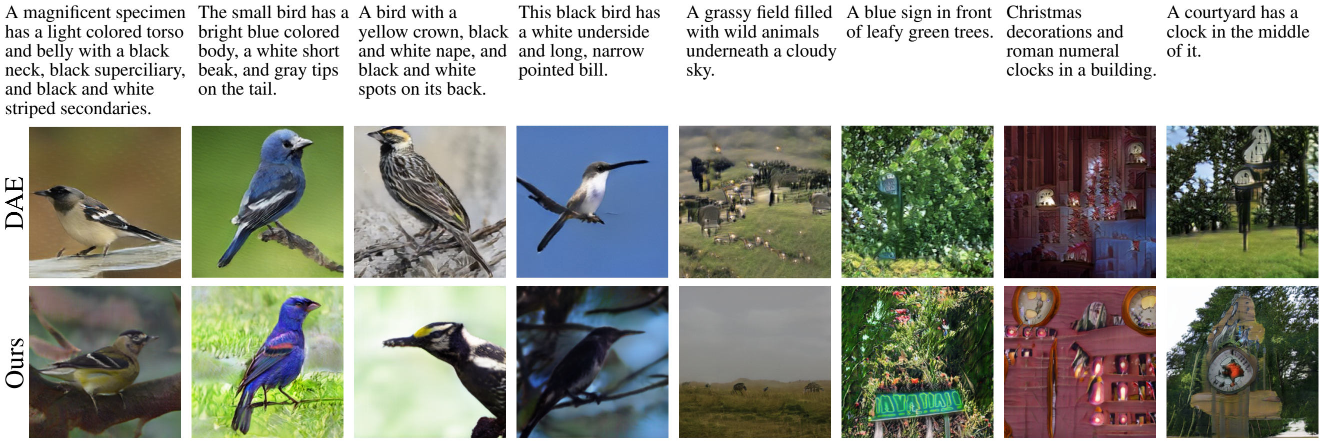
A magnificent specimen has a light colored torso and belly with a black neck, black superciliary, and black and white striped secondaries.
The small bird has a bright blue colored body, a white short beak, and gray tips on the tail.
A bird with a yellow crown, black and white nape, and black and white spots on its back.
This black bird has a white underside and long, narrow pointed bill.
A grassy field filled with wild animals underneath a cloudy sky.
A blue sign in front of leafy green trees.
Christmas decorations and roman numeral clocks in a building.
A courtyard has a clock in the middle of it.
DAE
Ours

Figure 5: Qualitative comparison between our method and DAE [24]. The DAE images are taken from their paper, and we generate the images using the same descriptions as theirs for fairness purpose. Best view in color and zoom in.

tails that are semantically consistent with the given text descriptions as well as clearer backgrounds. For example, in the 1st column, given text "A small bird with an orange bill and grey crown and breast", our method generates an image that has all the mentioned attributes. However, the image generated by DM-GAN does not reflect "small" while the image generated by DF-GAN does not have "grey crown and breast". More limitations of other methods can be observed in other examples. DF-GAN can neither generate the "red eye" in the 2nd column nor the "black bill" in the 4th column. The birds generated by DM-GAN in the 2nd and 3rd columns are not natural or photo-realistic. The qualitative results demonstrate that our SSA-GAN is more effectively and deeply to fuse text and image features and has higher text-image consistency. Particularly, the better generated details of a bird demonstrate that pixelwise text-image fusion of SSA block performs better than the sub-region-based methods in capturing details.

For the COCO dataset, shown in the last 4 columns in Fig. 4, one can observe that SSA-GAN is able to generate complex images with multiple objects with different backgrounds. In the 5th column, our image is more realistic than the ones generated by DM-GAN and DF-GAN. In 6th column, each of the generated cows can be clearly recognized and separated, while the cows are mixed together generated by DF-GAN. The images in the 6th - 8th columns are poorly synthesized by DM-GAN: the objects cannot be recognized and the backgrounds are fuzzy. In the 7th and 8th columns, the "skier" and "elephants" generated by DF-GAN do not seem as a natural part in the corresponding image. These qualitative examples on the more challenging COCO dataset demonstrate that SSA-GAN is able to generate a complex image with multiple objects as well as the corresponding background. The images generated by SSA-GAN are better holistically semantically consistent with given text as well as locally semantically consistent with important words in the text.

We compare the qualitative results of our method and DAE [24] in Fig. 5. For fair comparison, we generate images using the same captions as in [24] and compare with the images taken from their paper. We can see that, the birds generated by our method are in comparable quality as DAE but have better local semantic consistency with the given text: the "black and white nape" is more visible in our image (3st column), and our "long, narrow pointed bill" is more natural and realistic (4nd column). DAE generates better image in the 2nd column. In the complex scenes, our images have better quality. In the 5th column, our image is natural and visually recognizable while the image of DAE is abstract and chaos. In the 8th column, our method generates "a clock in the middle" while DAE generates a clock with ghost image. DAE has difficulty in generating multiple objects which is confirmed by the authors. Both methods failed to generate images in the 7th column because the description is too abstract.

4.3. Ablation Studies

In this subsection, we verify the effectiveness of each component in SSA-GAN by conducting extensive ablation studies on the testing set of the CUB dataset [29].

SSA Block and DAMSM Firstly, we verify how the proposed SSA block and the additional DAMSM affect the performance of the network. The results of using different components are given in Table 2. We treat the DF-GAN as the baseline denoted (ID0). Replacing the UPBlocks in DF-GAN with our SSA blocks, both the IS and FID performance are improved (ID1), which shows that our SSA block is able to fuse text and image features better. When DAMSM is added to our network (ID2), the overall performance is improved. It indicates that DAMSM helps im-

This small bird has a short beak, a light gray breast, a darker gray and black wing tips.

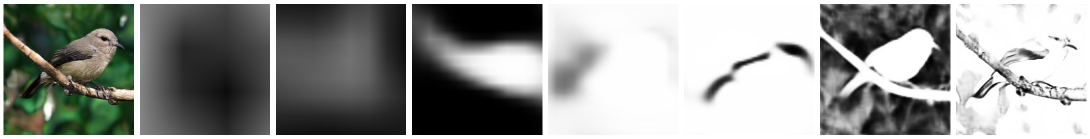

Figure 6: Example of semantic masks predicted in different SSA blocks. From left to right: input text, generated image and the 7 predicted semantic masks (from shallower to deeper layer). Best view in color and zoom in.

Table 2: Ablation study of evaluating the impact of SSA block and DAMSM in our framework on the CUB test set.

ID	Components		IS ↑	FID ↓
	SSA	DAMSM		
0	-	-	4.86 ± 0.04	19.24
1	✓	-	4.97 ± 0.09	18.54
2	✓	✓	5.07 ± 0.04	**15.61**
3	✓	✓(fine-tune)	$\mathbf{5.17 \pm 0.08}$	16.58

Table 3: Ablation study of evaluating how the performance is affected by different numbers of semantic masks used in the SSA-GAN. Note that, text encoder is not fine tuned here.

Parameter	Stages	IS ↑	FID ↓
#masks	2	4.98 ± 0.09	19.69
	3	5.04 ± 0.07	18.40
	4	5.05 ± 0.05	**15.03**
	5	5.02 ± 0.07	17.64
	6	4.97 ± 0.04	16.62
	7	$\mathbf{5.07 \pm 0.04}$	15.61

prove the text-image consistency. Then, we train the whole framework in order to fine tune the text encoder (ID3). Our method achieves further improvements in IS but inferior performance in FID. The reason is that fine tuning the text encoder helps text-image fusion and improves the text-image consistency so that the IS score is improved. However, when the encoded text features become more adaptive to the image features, the diversity of generated images also increases (more deeply constrained by the diverse text descriptions). Thus, the FID performance decreases while it measures the KL divergence between the real images and generated images. It is worth noting that, without adding DAMSM, our method (ID1) achieves better performance compared to the most recent state-of-the-art method (ID0).

Semantic Mask The predicted semantic mask provide spatial information for the semantic CBN in each SSA block. To evaluate how the semantic masks affect the text-image fusion process, we add the mask predictor one by one from the last SSA block to the first one and observe how the performance varies. The results are given in Table 3. We can see that the performance increases constantly by increasing the semantic masks up to 4. However, the performance is marginally worse when adding the 5th and 6th semantic mask. When the framework uses 7 masks, it has the highest IS score and second best FID performance. This phenomenon demonstrates that more semantic masks help text-image fusion process and the generated images are more realistic and text-image consistent (higher IS scores). Meanwhile, deeper text-image fusion also makes the generated images be stronger controlled by the diverse text descriptions. Consequently, the generated images become more diverse which leads to higher FID. Note that, we use 7 semantic masks for all the rest experiments in this work.

To gain more insight, Fig. 6 shows the semantic masks predicted on different stages. One can see that, the semantic masks become more focused on the bird when the text-image fusion becomes deeper. Especially in the last two stages, the main attention is on the whole bird to generate the bird, then on the specific local parts of the bird to refine the details. It visually demonstrates that the masks are predicted based on the current generated image features and deepen the text-image fusion process.

5. Conclusion

In this paper, we proposed a novel framework of Semantic-Spatial Aware GAN (SSA-GAN) for T2I generation. It has one generator-discriminator pair and is trained in an end-to-end fashion. The core module is the Semantic-Spatial Aware (SSA) block which operates Semantic-Spatial Condition Batch Normalization by predicting the semantic mask based on the current generated image features, and learning the affine parameters from the encoded text vector. The SSA block deepens the text-image fusion through the image generation process, and guarantees the text-image consistency. In the experimental results and ablation studies, we demonstrated the effectiveness of our model and the significant improvement over previous state-of-the-art approaches in terms of T2I generation.

Acknowledgment

This work has been supported by the Federal Ministry of Education and Research (BMBF), Germany, under the project LeibnizKILabor (grant no. 01DD20003), the Center for Digital Innovations (ZDIN) and the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122).

References

[1] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation. In *ICCV*, pages 4561–4569, 2019. 1

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. 2019. 1

[3] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, pages 5659–5667, 2017. 2

[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 *CVPR*, pages 10911–10920, 2020. 2

[5] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16372–16382, 2021. 1

[6] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *ICLR*, 2017. 4

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 2

[8] Sen He, Wentong Liao, Michael Ying Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Context-aware layout to image generation with enhanced object appearance. In *CVPR*, 2021. 1

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. In *NeuIPS*, pages 6626–6637, 2017. 5

[10] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *CVPR*, pages 7986–7994, 2018. 1, 2

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 1, 2

[12] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *CVPR*, pages 1219–1228, 2018. 1

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[15] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Text-to-image generation grounded by fine-grained user attention. In *WACV*, pages 237–246, 2021. 2

[16] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Controllable text-to-image generation. In *NeurIPS*, 2019. 1, 2, 3, 5, 6

[17] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *CVPR*, pages 12174–12182, 2019. 2, 5

[18] 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*, pages 740–755, 2014. 2, 5

[19] Xihui Liu, Guojun Yin, Jing Shao, and Xiaogang Wang. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *NeurIPS*, 2019. 1

[20] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[21] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *CVPR*, pages 1505–1514, 2019. 2

[22] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. 2016. 1

[23] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, pages 1060–1069, 2016. 2, 3

[24] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *ICCV*, pages 13960–13969, 2021. 2, 5, 6, 7

[25] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeuIPS*, pages 2234–2242, 2016. 5

[26] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 3

[27] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 5

[28] 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. 1, 3, 4, 5, 6

[29] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 5, 7

[30] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, pages 1316–1324, 2018. 1, 2, 3, 5, 6

[31] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *CVPR*, pages 2327–2336, 2019. 1, 3, 5, 6

[32] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *CVPR*, pages 4709–4717, 2017. 2

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

[34] 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 *ICCV*, pages 5907–5915, 2017. 1, 2, 3

[35] 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. *Transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 1, 2, 3, 5, 6

[36] Zhenxing Zhang and Lambert Schomaker. Dtgan: Dual attention generative adversarial networks for text-to-image generation. *arXiv preprint arXiv:2011.02709*, 2020. 5

[37] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, pages 6199–6208, 2018. 2

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 1

[39] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, pages 5802–5810, 2019. 2, 3, 5, 6