Semantics-enhanced Adversarial Nets for Text-to-Image Synthesis

Hongchen Tan[1]*, Xiuping Liu[1]†, Xin Li[2]*, Yi Zhang[1], Baocai Yin[1,3]
Dalian University of Technology[1], Louisiana State University[2], Peng Cheng Laboratory[3]
{tanhongchenphd, dlutzy}@mail.dlut.edu.cn, {xinli}@cct.lsu.edu, {xpliu, ybc}@dlut.edu.cn

Abstract

This paper presents a new model, Semantics-enhanced Generative Adversarial Network (SEGAN), for fine-grained text-to-image generation. We introduce two modules, a Semantic Consistency Module (SCM) and an Attention Competition Module (ACM), to our SEGAN. The SCM incorporates image-level semantic consistency into the training of the Generative Adversarial Network (GAN), and can diversify the generated images and improve their structural coherence. A Siamese network and two types of semantic similarities are designed to map the synthesized image and the groundtruth image to nearby points in the latent semantic feature space. The ACM constructs adaptive attention weights to differentiate keywords from unimportant words, and improves the stability and accuracy of SEGAN. Extensive experiments demonstrate that our SEGAN significantly outperforms existing state-of-the-art methods in generating photo-realistic images. All source codes and models will be released for comparative study.

1. Introduction

Synthesizing photographic images from text description has tremendous applications such as photo editing and computer-aided design. However, it remains as a challenging vision problem. Recently, based on Generative Adversarial Networks (GANs) [6], many effective approaches [28, 42, 9, 44, 40, 16] have been developed and achieved promising results. One issue that hinders the synthesis of realistic images with good resolutions is the big arbitrariness of the image contents to generate from limited text semantics. Most existing text-to-image synthesis approaches condition the training of content generation only using semantic information from textual data. However, from these limited words, it is sometimes difficult for a generator to learn rich-enough contents to form realistic images. The generated images, hence, are often prone to semantic structure ambiguity and class information confusion.

The training of image synthesizer is usually performed on data with both descriptive texts and corresponding groundtruth images. We believe these groundtruth images provide highly valuable content semantics to help train the image synthesizer. Therefore, our basic idea is to train an image synthesizer using both text semantics and image contents. We expect images generated by such a synthesizer would contain more structured context information and be more vivid.

In addition, many existing text-to-image synthesizers create contents based on the global sentence feature, which may miss important fine-grained information at the word level and impact the quality of synthesized images. Recently, AttnGAN [40] incorporate word-level and sentence-level attention mechanism to produce fine-grained image generation. However, in the word-level attention mechanism of AttnGAN [40], attention is paid to every word in the sentence. Our second technical development in this paper is to generate adaptive attention mechanism so that important words will gain plenty attention and unimportant words can be ignored. This could improve the description ability of detail to SEGAN, and further improve the stability and accuracy of the trained synthesizer.

Based on the above observations, we specifically design two new modules in our image synthesis GAN. The first module is the Semantic Consistency Module (SCM), in which we propose a Siamese network (SiaNet) to pull a synthesized image $I_{a'}$ towards its corresponding groundtruth image I_a, and pushes the image of $I_{a'}$ away from another image I_b that is associated with a different text description. SCM improves the semantic consistency by mapping the synthesized image and its corresponding groundtruth to nearby points on the output manifold in the latent semantic feature space. Furthermore, we observe the training of SCM involves both easy and difficult samples. We further revise the Contrastive Loss of SiaNets according to difficulty of samples to tackle this imbalance. The second module is the Attention Competition Module (ACM), which constructs attention weight for key words and suppresses the influence of unimportant words. We achieve this through an attention regularization term in the training of ACM.

*indicates equal contributions
†indicates corresponding author

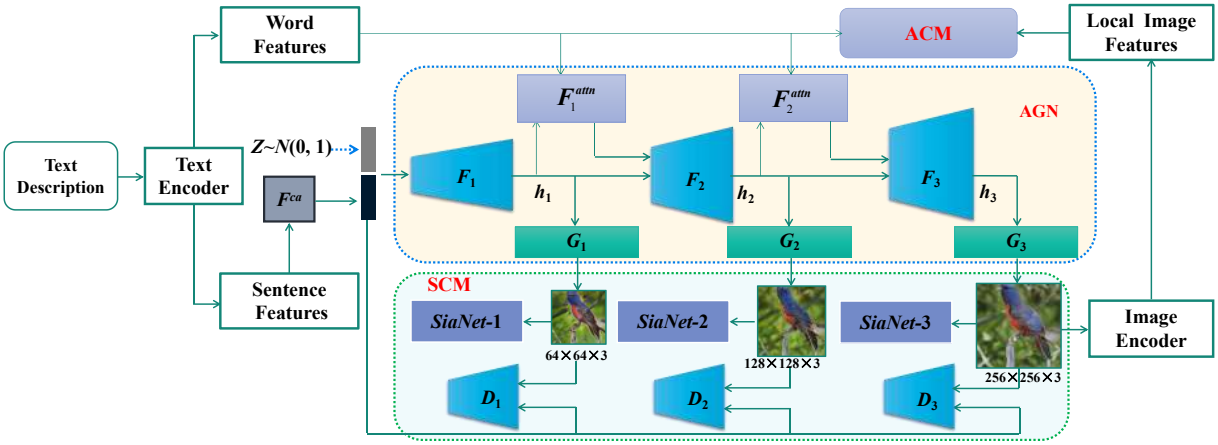

Figure 1. The architecture of the proposed SEGAN. SEGAN consists of text and image encoders, a Attention Generative Network (AGN) in [40], and two new components: the ACM and SCM. The ACM suppresses the word-level attention weights of visually unimportant words. The SCM improves the semantic consistency between the synthesized image and its corresponding groundtruth image.

The main contributions of this paper are as follows: (i) We propose a new Semantics-enhanced Generative Adversarial Network (SEGAN) for text to image generation. Its effectiveness is supported by two novel components: a Semantic Consistency Module (SCM) and an Attention Competition Module (ACM). (ii) We design a new Sliding Loss to replace the commonly adopted Contrastive Loss, so that easy and difficult data samples can be handled with better balance in SCM. (iii) We validate our proposed method on two datasets: CUB birds [38] and large-scale MSCOCO [20]. Extensive experiments demonstrate that our algorithm clearly outperforms the existing state-of-the-art.

2. Related Work

Generating photo-realistic images from text descriptions is a fundamental but challenging problem in computer vision. Recently, significant progress has been made in this research. Various approaches, such as variational inference [7, 24], approximate Langevin process [29], conditional PixelCNN via maximal likelihood estimation [29, 26], and conditional generative adversarial networks (GANs) [30, 28, 42, 9], have been developed to tackle this problem. Among all these approaches, GAN-based algorithms has produced the state-of-the-art results.

Semantic Consistency. Recent GAN-based text-to-image synthesis [28, 42, 9, 44, 40, 16, 10] design the generative procedure only using semantics from the input text. Such textual description, as discussion in Section 1, is prone to semantic structure ambiguity and class information confusion. On the other hand, structural and spatial information from realistic groundtruth images that match with the given texts may provide abundant valuable extra information to better guide the image synthesis.

In several other computer vision tasks, such as unsupervised domain person Re-identification [3, 39, 46] and image-to-image style transfer (CycleGAN) [15], such guidance from groundtruth images has been demonstrated effective in enhancing structural coherence of synthesized images. For example, CycleGAN [15] introduces the perceptual consistency to improve quality of semantic structure for translated image. Based on CycleGAN [15], SPGAN [3] introduces semantic similarity between translated image and its counterpart in the source dataset to better preserve person ID label and produce cross-domain person re-identification. Inspired by SPGAN [3], we believe introducing image-level semantic consistency into text-to-image synthesis could also desirably enrich the semantic structure information in synthesized images.

Visual attention mechanism. Earlier GAN-based text-to-image synthesis algorithms such as [28, 42, 9, 44] use the whole-sentence features, which sometimes overlook important fine-grained information at the word level and result in lower-quality synthesized images. Recently, AttnGAN [40] introduces a combined sentence-level and word-level visual attention mechanism for text-to-image synthesis, and by paying attentions to the relevant words in the text description, it enhances the synthesis of fine-grained details at different image regions. Based on AttnGAN [40], obj-GAN [18] proposes a object-driven attention mechanism to further improve the detail synthesis, and produces finer images. However, obj-GAN [18] requires the training images to contain bounding box and shape information of each interested object. But generating such labels are non-trivial: it is time consuming and expensive when dealing with large image datasets. This limits its scalability and usability in synthesizing images from more general text description. A limitation of both AttnGAN [40] and obj-GAN [18] is that their attention weight is defined for every word, even an unimportant one, from the sentence. This unnecessary focus

could impact the stability and quality of the trained generative model (See Figure 3 for example). Intuitively, only focusing on keywords could benefit the efficiency, accuracy, and stability of the generator's training.

In other computer vision tasks, such as person search [35], person Re-identification [17, 2], object tracking [11, 45] and Image Captioning [41, 22], visual attention mechanism been studied. GLIA [1] extracts noun phrases by chunking, and then considers local association based upon the implicit correspondences between image regions and noun phrases for person Re-ID. [34] employs a diversity regularization term in the attention mechanism to ensure multiple models do not discover the same body part in the task of video person re-identification. Inspired by [34], based on the word-level fine-grained attention Module in [40], we employ an attention regularization term into the attention mechanism to improve the quality of semantic attention from key words and try to supress the attention weight from non-key words.

3. Semantics-enhanced Generative Adversarial Network

We illustrate the design of our Semantics-enhanced Generative Adversarial Network (SEGAN) architecture in Figure 1. Our SEGAN has four main components: text and image encoders, Attention Generative Network (AGN), a Semantic Consistency Module (SCM), and an Attention Competition Module (ACM). The two encoders prepare text features and image features for the ACM and AGN. The ACM includes a new attention regularization term and *DAMSM loss* [40], which helps the text encoder extract *visually important keywords* for AGN. In the AGN, the text-encoder pre-trained by the ACM provides semantic vectors encoding visually important words. The SEGAN generator then synthesizes image subregions according to these keywords. In the SCM, multi-scale SiaNets are used to produce semantic consistency constraints for the generator. They help enrich semantic structural information and image ID information in image synthesis.

3.1. Text and Image Encoders

We first describe the design of text and image encoders that prepare text and image features for SEGAN's other modules. This design follows the widely adopted design from state-of-the-art text-to-image synthesizers [18, 40, 14].

Text encoder extracts the semantics of the whole sentence and each word in the sentence. Following [18, 40, 14], we use a bi-directional Long Short-Term Memory [33] to construct this encoder. It takes a sentence (i.e., word sequence) as input, and outputs a *sentence feature vector* $\bar{e} \in \mathbb{R}^D$ and a *word feature matrix* $e \in \mathbb{R}^{D \times T}$, where its i^{th} column e_i is the feature vector of the i^{th} word, D is the dimension of the word vector, and T is the number of different words in the given sentence.

Image encoder. Following [18, 40, 14], our image encoder uses the Inception-v3 model [36] pretrained on ImageNet [31]. We first rescale the input image to 299×299 pixels before feeding it to the encoder. Then, a local feature matrix $f \in \mathbb{R}^{768 \times 289}$ (reshaped from $768 \times 17 \times 17$) is extracted from the "$mixed\_6e$" layer of the Inception-V3 model. Each column of f is a feature vector of a subregion in the image. The dimension of a local feature vector and the number of sub-regions of an image are 768 and 289, respectively. Meanwhile, a global feature vector $\bar{f} \in \mathbb{R}^{2048}$ is extracted from the last average pooling layer of the Inception-v3. Finally, we map image features to a canonical semantic space of text features by adding a perceptron layer: $v = Wf, \bar{v} = \bar{W}\bar{f}$, where $v \in \mathbb{R}^{D \times 289}$, $\bar{v} \in \mathbb{R}^D$, and D is the dimension of this semantic space.

3.2. Attention Competition Module

We design a new mechanism, the Attention Competition Mechanism (ACM) in Figure 1, to help text encoder identify visually important keywords. To fulfill this goal, we design a new *attention regularization term* and reuse *DAMSM loss* [40] to filter out unimportant words. The DAMSM [40] is to measure the matching degree between images and text descriptions, which makes generated images better conditioned on text descriptions.

Proposed Attention Regularization Term. Firstly, we define an image-text similarity matrix $S = (s_{i,j}) = e^T v \in \mathbb{R}^{T \times 289}$ to encode the dot-product similarity between the i^{th} word in the sentence and j^{th} sub-region in the image. This S is normalized to $\hat{S} = (\hat{s}_{i,j}) = \frac{exp(s_{i,j})}{\sum_{k=1}^{T} exp(s_{k,j})}$, and then normalized to $R = (r_{i,j}) = \frac{exp(\hat{s}_{i,j})}{\sum_{k=1}^{289} exp(\hat{s}_{i,k})}$.

Secondly, in AttnGAN [40], attention is paid to every word in the sentence. But unnecessary emphasis on non-visual words such as "is", "the", and "has" could negatively impact the stability and quality of the generative model (See Figure 3 for example). Our observation is that only focusing on visual keywords could benefit the efficiency, accuracy, and stability of the generator's training. Inspired by the attention regularization employed in recent text embedding networks [21, 34], we propose a new attention regularization term to construct attention weights of words through finding "survival of the fittest". Specifically, we define

$$\mathcal{L}_c = \sum_{i,j} (\min(r_{i,j}, \alpha))^2, \tag{1}$$

where the subscript "c" stands for "*competition*", and $\alpha > 0$ is a threshold. In the training process, *visually important words* are those whose attention weights with respect to certain image's subregions exceed the α. With $\mathcal{L}_c$, the Cross-modal Similarity Matching Loss $\mathcal{L}_W$ and $\mathcal{L}_S$ will push the

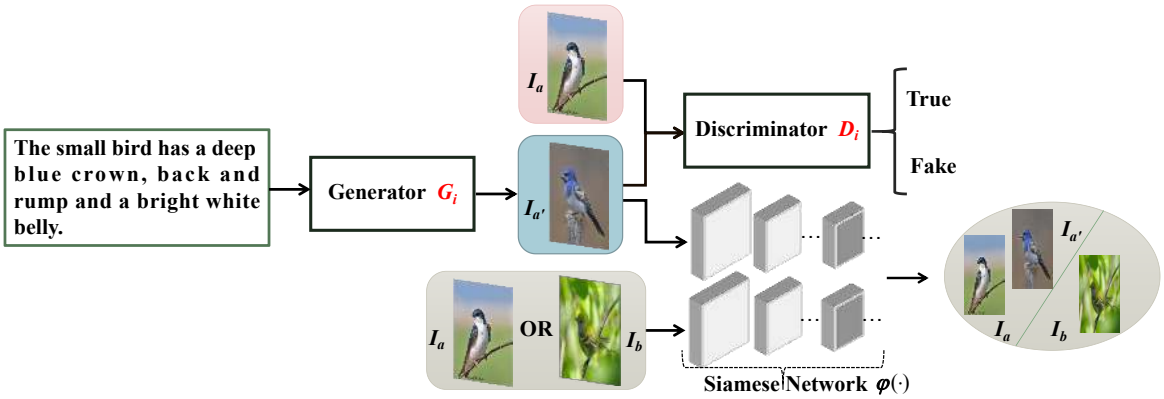

Figure 2. The architecture of the Semantic Consistency Module (SCM) in SEGAN.

attention weights of *visually important words* to exceed the threshold α. And their attention weights will be preserved. In contrast, *visually unimportant words* are those whose attention weights with respect to all the image's subregions are lower than α. Their attention weights will decrease and move towards 0. Thus, these words will be suppressed. More analysis refer to Section A in supplementary material.

Finally, the combined loss in ACM is formulated as

$$\mathcal{L}_{ACM} = \mathcal{L}_{DAMSM} + \lambda_1 \mathcal{L}_c. \tag{2}$$

Through experiments on a hold-out validation set, we set the hyperparameter $\lambda_1 = 2$ in this subsection.

3.3. Attention Generative Network [40]

We adopt the Attention Generative Network (AGN) in [40] as our basic generator, due to its good performance in generating realistic images. Thus, we revisit the AGN in this subsection. As shown in Figure 1, firstly text-encoder pretrained by ACM provides *visually important keywords* for the Attention Generative Network(AGN). Then AGN synthesizes different subregions of the image following their most relevant words.

The AGN has Ω generators (G_1, G_2, $\cdots$, G_Ω), which take the hidden states (h_1, h_2, $\cdots$, h_Ω) as input and generate images of small-to-large scales ($\hat{I}_1$, $\hat{I}_2$, $\cdots$, $\hat{I}_\Omega$):

$$\begin{aligned} &h_1 = F_1(z, F^{ca}(\bar{e})); \\ &h_i = F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})), \quad \text{for } i = 1, 2, \cdots, \Omega; \\ &\hat{I}_i = G_i(h_i). \end{aligned} \tag{3}$$

Here, $z \sim N(0, 1)$. F^{ca} is a conditioning augmentation module [42] that converts a sentence feature $\bar{e}$ to a conditioning feature for the generator. F^{ca}, F_i^{attn}, F_i, and G_i are modeled as neural networks. F_i^{attn} represents the i^{th} stage key attentional model.

The attention model $F^{attn}(e, h)$ has two inputs: the word features $e \in \mathbb{R}^{D \times T}$ from pretrained text encoder and the image features from the previous hidden layer $h \in \mathbb{R}^{\hat{D} \times N}$. Firstly, the word features are mapped to the same latent semantic space as the image features, i.e., $e' = Ue$, where $U \in \mathbb{R}^{\hat{D} \times D}$ is corresponds to a perceptral layer. Each column of h (hidden features) is a feature vector of an image subregion. Then, for the j^{th} sub-region, its *dynamic representation* of word vectors w.r.t. h_j is

$$q_j = \sum_{i=1}^{T} \theta_{j,i} e'_i, \quad \text{where } \theta_{j,i} = \frac{exp(S'_{j,i})}{\sum_{k=1}^{T} exp(S'_{j,k})}. \tag{4}$$

Here $S'_{j,i} = h_j^T e'_i$, and $\theta_{j,i}$ indicates the weight the model assigned to the i^{th} word when generating the j^{th} sub-region of the image. With the text encoder pretrain by the ACM, attention weights on unimportant words in θ are lower than visually important words. Then, *text-vision matrix* for image feature set h is denoted by $F^{attn}(e, h) = (q_1, q_2, ..., q_N) \in \mathbb{R}^{\hat{D} \times N}$, which provide key-word information for next stage. Finally, image features h and corresponding text-vision features are combined to generate images at next stage.

3.4. Semantic Consistency Module

As discussed in Section 1, we want to align image ID information and semantic information of our synthesized image with the groundtruth image during training.

Contrastive Loss. Given a synthesized image $I_{a'}$, its corresponding groundtruth image I_a, and another random image I_b described by a different sentence, suppose their **normalized** feature vectors are $\varphi(I_a)$, $\varphi(I_{a'})$, and $\varphi(I_b)$, respectively. As shown in Figure 2, we use a Siamese Network (SiaNet) to push a positive pair, $\varphi(I_a)$ and $\varphi(I_{a'})$, towards each other, and pull a negative pair, $\varphi(I_a)$ and $\varphi(I_b)$, away from each other. This SiaNet can be trained using a Contrastive Loss [8]. At the i^{th} stage of the SEGAN,

$$\mathcal{L}_{con_i} = \begin{cases} d(\varphi(I'_a), \varphi(I_a))^2 & \text{(positive pair)} \\ max^2\left(0, m_i - d(\varphi(I'_a), \varphi(I_b))\right) & \text{(negative pair)} \end{cases} \tag{5}$$

where $m_i > 0$ is separability margin and $d(\mathbf{x}, \mathbf{y})$ gives the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$.

Slidling Loss. In the SiaNet training process, some sample pairs are easy to differentiate but some are hard. In many computer vision tasks [23, 19, 4], effectively balancing easy versus hard samples is important to the training quality. Inspired by the Focal Loss [19], in the aforementioned Contrastive Loss function, we further add two modulating factors, $(\frac{d}{1+d})^\gamma$ and $(1-\frac{d}{1+d})^\gamma$, to positive and negative pairs, respectively, and call this a Sliding Loss (SL):

$$\mathcal{L}_{SL_i} = \begin{cases} (\frac{d}{1+d})^\gamma d^2 & \text{(positive pair)} \\ (1-\frac{d}{1+d})^\gamma max^2(0, m_i - d) & \text{(negative pair)} \end{cases} \tag{6}$$

Here $\gamma \geq 0$ is a tunable sliding parameter. When $\gamma = 0$, the Sliding Loss becomes the Contrastive Loss. The two modulating factors $(\frac{d}{1+d})^\gamma$ and $(1-\frac{d}{1+d})^\gamma$ could adaptively adjust the weights of positive and negative sample pairs. Taking a positive sample as an example, When $\gamma > 0$, the bigger d is, the greater the punishment weight $(\frac{d}{1+d})^\gamma$ is, and vice versa. Therefore, setting $\gamma > 0$ reduces the relative loss for well-trained pairs, and we focus more on hard sample pairs. With an increasing γ, the effect of the modulating factor increases too. We found $\gamma = 1/2$ works best in our experiments.

3.5. Generative and Discriminative Loss

Combining the above modules together, at the i^{th} stage of the SEGAN, the Generative loss $\mathcal{L}_{G_i}$ and Discriminative loss $\mathcal{L}_{D_i}$ are defined as

$$\mathcal{L}_{G_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{I}_i \sim P_{G_i}}[logD_i(\hat{I}_i)]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{I}_i \sim P_{G_i}}[logD_i(\hat{I}_i, \bar{e})]}_{\text{conditional loss}}, \tag{7}$$

where the unconditional loss is trained to generate images towards the true data distribution to fool the discriminator, and the conditional loss is trained to generate samples to match text descriptions.

The discriminator D_i is trained to classify the input into the class of real or fake images by minimizing the cross-entropy loss

$$\begin{aligned} \mathcal{L}_{D_i} = & \underbrace{-\frac{1}{2}\mathbb{E}_{I_i \sim P_{data_i}}[logD_i(I_i)] - \frac{1}{2}\mathbb{E}_{\hat{I}_i \sim P_{G_i}}[log(1 - D_i(\hat{I}_i)]}_{\text{unconditional loss}} + \\ & \underbrace{-\frac{1}{2}\mathbb{E}_{I_i \sim P_{data_i}}[logD_i(I_i, \bar{e})] - \frac{1}{2}\mathbb{E}_{\hat{I}_i \sim P_{G_i}}[log(1 - D_i(\hat{I}_i, \bar{e})]}_{\text{conditional loss}}, \end{aligned} \tag{8}$$

where I_i is from the realistic image distribution p_{data} at the i^{th} scale, and $\hat{I}_i$ is from distribution p_{G_i} of the generative images at the same scale.

To generate realistic images, the final objective function of the generative network and discriminative network are defined as

$$\mathcal{L}_{G^*} = \mathcal{L}_G + \lambda_2 \mathcal{L}_{ACM} + \mathcal{L}_{SL}, \mathcal{L}_{D^*} = \sum_{i=1}^{\Omega} \mathcal{L}_{D_i}. \tag{9}$$

Here, $\mathcal{L}_G = \sum_{i=1}^{\Omega} \mathcal{L}_{G_i}, \mathcal{L}_{SL} = \sum_{i=1}^{\Omega} \eta_i \mathcal{L}_{SL_i}$. In our current implementation, SEGAN has three stage generators ($\Omega = 3$). And $\lambda_2 = 100$ is to balance these terms of Eq. 9. In $\mathcal{L}_{SL}$, $\eta_1 = 1$, $\eta_2 = 5$ and $\eta_3 = 10$. And through experiments on a hold-out validation set, we set the hyper-parameters $m_1 = 1$, $m_2 = 2$, and $m_3 = 3$ in this paper.

4. Experimental Results

We perform extensive experiments to evaluate the proposed SEGAN. Firstly, we discuss the effectiveness of each new module introduced in SEGAN: ACM and SCM. Then, we compare our SEGAN with other state-of-the-art GAN models [28, 30, 25, 42, 9, 13, 44, 43, 13, 40, 14, 18, 27].

Datasets. Two widely used datasets are used. The CUB dataset [38] contains $11,788$ bird images belonging to 200 categories, and 10 visual description sentences for each image. We pre-process and split the images following the same pipeline in [28, 42]. The COCO dataset [20] contains $80k$ training images and $40k$ test images, and each image has 5 text annotations.

Evaluation. We use the Inception score [32], Fréchet Inception Distance (FID) [12] and Visual-semantic similarity in [44, 5] as the quantitative evaluation measures. Synthesized images are also visually compared for qualitative evaluation.

Inception score [32] is a measure for both objectiveness and diversity of generated images. The FID computes the Frechet distance between synthetic and real-world images based on the extracted features from a pre-trained Inception-V3 network [36]. A lower FID implies a closer distance between generated image distribution and real-world image distribution.

The aforementioned metrics are widely used for evaluating standard GANs. However, they can not measure the semantic consistency between synthesized image and corresponding text description. Thus, the same as [44], we also use Visual-semantic similarity as our third evaluation metric. We use the trained model in [5] to evaluate the semantic consistency, and select the Rank-1 as our evaluation score.

Network Settings. We use SiaNets with Sliding Loss (SL) or Contrastive Loss (CL) on each stage of SEGAN. Because AttnGAN and other state-of-the-art methods could synthesize high-resolution images. Thus, all synthesized images discussed in this paper are 256×256. Our baseline model is AttnGAN [40] due to its excellent performance.

Training Details. In SEGAN, the generator and the discriminator losses of the proposed SEGAN follow those in [40] due to its excellent performance. The text encoder and

inception model for visual features used in visual-semantic embedding are pretrained by [40] and fixed during the end-to-end training. The network parameters of the generator and discriminator are initialized randomly.

4.1. Ablation Study

Table 1. Inception scores produced by combining different components of the SEGAN. The final SEGAN=baseline+AC+SL.

Method	CUB	COCO
baseline [40]	4.31 ± 0.02	25.56 ± 0.19
baseline+AC, $\alpha = 0.005$	4.61 ± 0.04	26.83 ± 0.33
baseline+CL	4.44 ± 0.03	-
baseline+SL, $\gamma = 1/2$	4.58 ± 0.03	27.13 ± 0.26
SEGAN, $\gamma = 1/2$, $\alpha = 0.005$	4.67 ± 0.04	27.86 ± 0.31

Effectiveness of New Modules. We evaluate the effectiveness of two new components, ACM and SCM, and document the results in Table 1. (1) We introduce the SiaNets with Contrastive Loss (baseline + CL) into the baseline model, which leads to 3.0% improvement of Inception score over the baseline on CUB test dataset. (2) Replacing the Contrastive Loss by Sliding Loss (baseline + SL) further leads to 6.3% and 6.1% improvement of Inception score, on CUB and COCO test datasets, respectively. (3) If we introduce ACM into the baseline (baseline + AC), we obtain 7.0% and 5.0% improvement over the baseline in Inception scores on CUB and COCO datasets. (4) When adding both ACM and SCM into baseline, we get the SEGAN and it leads to 8.4% and 9.0% improvement over baseline in Inception scores on CUB and COCO test datasets. This shows that both components contribute to the SEGAN's performance improvement. The Inception scores of SEGAN are 4.67 on CUB and 27.86 on COCO test dataset.

Table 2. Results on CUB testing Data from Different Hyperparameters γ in Sliding Loss (SL).

Method	Inception Score
baseline [40]	4.31 ± 0.02
baseline+SL ($\gamma = 0$)	4.44 ± 0.03
baseline+SL ($\gamma = 1/5$)	4.47 ± 0.02
baseline+SL ($\gamma = 1/3$)	4.51 ± 0.04
baseline+SL ($\gamma = 1/2$)	4.58 ± 0.03
baseline+SL ($\gamma = 1$)	4.41 ± 0.01
baseline+SL ($\gamma = 2$)	4.41 ± 0.02

Semantic Consistency Module. We first discuss the effect of parameter γ in Sliding Loss (SL) to the Inception score, then show some results produced through the SCM module.

The hyperparameter γ in the Sliding Loss (LS) controls the strength of the weight term in Eq. 6. When $\gamma = 0$, Sliding Loss degenerates to the Contrastive Loss. Setting $\gamma > 0$ reduces the relative loss for easy examples, putting more focus on hard examples, which is beneficial to the training of SCM model. To find the suitable γ, we perform through experiments to evaluate the SCM's performance under different γ values. As shown in Table 2, when $\gamma \in (0, 1/2]$, baseline+SL has better performance than baseline+SL ($\gamma = 0$), the Inception score of baseline+SL ($\gamma = 1/2$) reaches 4.58, significantly better than $\gamma = 0$. When γ is too big, the performance decreases. With bigger and bigger γ, the weights of most sample pairs become lower and lower. If we do not adjust the weight of $\mathcal{L}_{SL}$ in Eq. 9, a very big γ makes the weights of most sample pairs to be much lower than 1, which is not conducive to the training of SCM model. In summary, while baseline+SL is usually better than baseline model in Inception scores, we find $\gamma = 1/2$ works the best.

Table 3. Discussion of hyperparameter α in ACM on CUB dataset.

Method	Inception Score
baseline [40]	4.31 ± 0.02
baseline+AC ($\alpha = 0.001$)	4.41 ± 0.03
baseline+AC ($\alpha = 0.003$)	4.53 ± 0.03
baseline+AC ($\alpha = 0.005$)	4.61 ± 0.02
baseline+AC ($\alpha = 0.008$)	4.55 ± 0.01
baseline+AC ($\alpha = 0.05$)	4.51 ± 0.02
baseline+AC ($\alpha = 0.5$)	4.46 ± 0.01
baseline+AC ($\alpha = 1$)	4.25 ± 0.01

Visualization results are shown in second row (AttnGAN+SCM) of Figure 4. Compared with AttnGAN (baseline), images synthesized by AttnGAN+SCM contain more semantic structural information and is more realistic.

Attention Competition Module. For ACM, firstly we discuss the effect of hyperparameter α in Eq. 1, and its selection for SEGAN. Secondly, we demonstrate that attention weights of non-key words could be supressed by ACM.

We document the overall Inception scores under different α values. As shown in Table 3, we found $\alpha = 0.005$ works best in our experiments. With $\alpha = 0.005$, the Inception score of SEGAN obtains 4.61 which is much higher than baseline model (4.31). When $\alpha = 1$, attention regularization term (Eq. 1) is equivalent to $\mathcal{L}_c = \|R\|_F^2$, which means that attention weight of all words in sentence should be supressed. In this case, the Inception score of baseline + AC drops to 4.25.

Figure 3 visualizes the attention weight maps on synthesized images. For sub-regions whose semantic meaning are expressed in the description text description, the attentions are allocated to their most relevant words (bright regions in Figure 3). AttnGAN pays attention on all the words including unimportant ones. But such kind of attention may result in strange synthesized subparts (the left example) or chaotic structures (the right example). In contrast, AttnGAN + ACM could better focus on visually important words and synthesize higher-quality images.

Visualization Figures 4 and 5 show some more qualitative comparisons. Description in the left five columns in

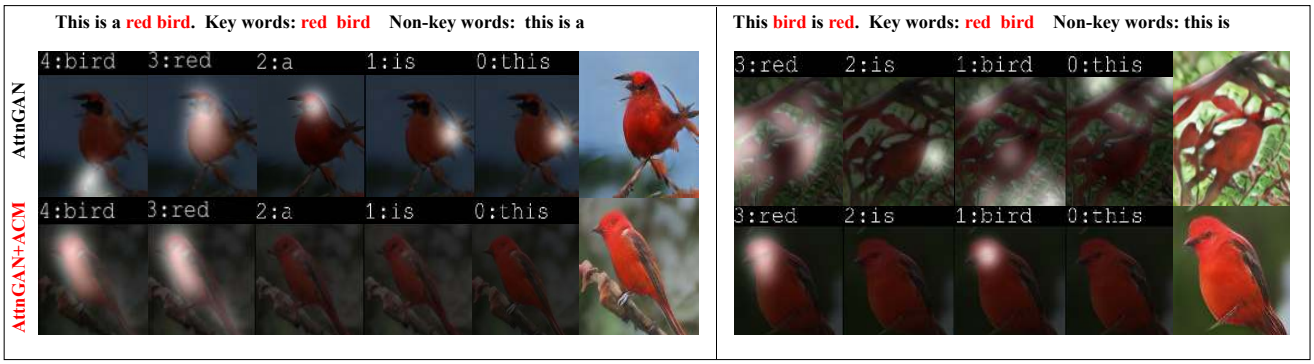

Figure 3. Word-level Attention Weight Maps generated from AttnGAN (first row) and our AttnGAN + ACM (second row).

Figure 4 are from the CUB test dataset. Images synthesized by AttnGAN are prone to semantic structure ambiguity. In contrast, images synthesized by SEGAN contain more semantic classification details and structured information. Besides using the testing sentences from the benchmark for evaluation, we also compose new descriptive sentences with the same meaning to test the stability and generalizability of different synthesizers. The sentences in blue on the right three columns in Figure 4 are such composed sentences. We can see that our SEGAN still synthesize realistic and accurate images. Descriptions in Figure 5 are from the COCO dataset. They describe much more complex scenarios. All recent synthesizers cannot properly handle this dataset. Some SEGAN generated images are illustrated. While they are still far away from perfect, they contain reasonable subparts. To sum up, Figure 4 and Figure 5 further demonstrate the generalization ability of the SEGAN. More synthesized images on CUB and COCO datasets are given in Section B of supplementary material.

Figure 5. Images of 256 × 256 Resolution Generated by SEGAN using Texts from COCO testing dataset [20].

Table 4. Inception scores by state-of-the-art GAN models and our SEGAN on CUB and COCO test datasets. The best scores for text-to-image methods are shown in bold. AttnGAN+O.P.*: Inception score of AttnGAN in paper [37] (AttnGAN+Object Pathways) is 23.61 ± 0.21 on COCO dataset.

Method	**CUB**	**COCO**	Reference
GAN-INT-CLS [28]	2.88 ± 0.04	7.88 ± 0.07	ICML 2016
GAWWN [30]	3.62 ± 0.07	-	NIPS 2016
PPGAN [25]	-	9.58 ± 0.21	CVPR 2017
StackGAN [42]	3.70 ± 0.04	8.45 ± 0.03	ICCV 2017
mvGAN [43]	-	9.94 ± 0.12	PCM 2018
StackGAN-V2 [9]	3.84 ± 0.06	-	TPAMI 2018
ISL-GAN [13]	-	11.46 ± 0.09	CVPR 2018
HDGAN [44]	4.15 ± 0.05	11.86 ± 0.18	CVPR 2018
Infer [13]	-	12.40 ± 0.08	CVPR 2018
AttnGAN [40]	4.31 ± 0.02	25.56 ± 0.19	CVPR 2018
AttnGAN+O.P.*[37]	-	24.76 ± 0.43	ICLR 2019
RAGAN [14]	-	23.74 ± 0.36	arXiv 2019
MirrorGAN [27]	4.56 ± 0.05	26.47 ± 0.41	CVPR 2019
Obj-GAN [18]	-	**30.29 ± 0.33**	CVPR 2019
baseline [40]+AC	4.61 ± 0.04	26.83 ± 0.33	**Our**
baseline [40]+SL	4.58 ± 0.03	27.13 ± 0.26	**Our**
SEGAN	**4.67 ± 0.04**	27.86 ± 0.31	**Our**

Table 5. AttnGAN versus SEGAN in FID. A lower FID implies a closer distance between generated image distribution and real-world image distribution.

Method	Bird	COCO
AttnGAN (Baseline)	22.504	34.398
SEGAN (Ours)	**18.167**	**32.276**

Table 6. The Visual-Semantic Similarity evaluation (Rank-1). **A higher score indicates higher semantic consistency between the generated images and conditioned text.** The groundtruth score is shown in the first row.

Method	Bird	COCO
GroundTruth	46.3%	21.2%
AttnGAN (Baseline)	27.9%	7.1%
SEGAN (Ours)	**30.2%**	**8.9%**

Figure 4. Images of 256×256 resolution are generated by our SEGAN and AttnGAN [40] conditioned on text descriptions. Texts in the left five columns are from CUB [38] test datasets. Texts in the right three columns are composed by us to test the generators' stability and generalizability.

4.2. Comparison with state-of-the-art GAN models

We compare our SEGAN with state-of-the-art GAN models for text-to-image synthesis on CUB and COCO test datasets. Table 4 lists all the Inception scores. Our AttnGAN achieves 4.67, which is 8.4% higher than AttnGAN on the CUB test dataset. On the COCO test dataset, the SEGAN achieves 27.86 Inception score, 9.0% better than AttnGAN. Besides, combining baseline with each module, baseline+AC and baseline+SL, which also could have better performance than most state-of-the-art GAN models in Inception score. In Table 4, obj-GAN [18] is better than our SEGAN in Inception score. However, they require additional information, including the interested object's bouding box and shape, for training synthesizing. This additional information, although available in the COCO dataset, is often unavailable for other datasets such as CUB. Hence, Obj-GAN cannot produce images on the CUB dataset. In general, producing such information on new database to train the generator is expensive. This limits its scalability and usability in more general text and image datasets.

In Tabel 5, we compare the performance between AttnGAN and our SEGAN with respect to FID on the CUB and MS-COCO datasets. Our SEGAN decreases the FID from 22.504 to 18.167 on the CUB dataset and from 34.398 to 32.276 on the COCO dataset. It demonstrates that SEGAN can learn a better data distribution.

In order to demonstrate that our SEGAN has better performance on semantic consistency between synthesized image and corresponding text description. The results of Visual-Semantic Similarity evaluation (Text-to-Image Retrieval) are shown in Table 6. We use the Rank-1 score to evaluate the performance of our SEGAN and AttnGAN. The scores of the groundtruth image-text pair are also shown for reference. On CUB dataset, our SEGAN achieves the Rank-1 30.2%, which significantly outperforms AttnGAN 27.9%. On MS-COCO dataset, our SEGAN achieves the Rank-1 8.9%, which also significantly outperforms AttnGAN 7.1%. These results demonstrate that SEGAN can better capture the visual semantic information from textual data in generated images.

5. Conclusions

We present a novel model, Semantics-enhanced Generative Adversarial Network (SEGAN), to tackle the problem of generating images from text descriptions. We develop two novel components for SEGAN, the Attention Competition Module (ACM) and Semantic Consistency Module (SCM). Extensive visual experimental results demonstrate that SEGAN outperforms recent state-of-the-art approaches in text-to-image synthesis. In the future, we will explore the adding of object sketch constraint and style constraint to GAN model and explore their, to further improve the quality of synthesized images.

Acknowledgment

This work has been supported by National Natural Science Foundation of China under Grant (No. U1811463, No. 61728206, No. 61632006, No. 61562062).

References

[1] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*, 2018.

[2] Liang Wang Wanli Ouyang Chunfeng Song, Yan Huang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018.

[3] Weijian Deng, Zheng Liang, Guoliang Kang, Yang Yi, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018.

[4] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *ECCV*, 2018.

[5] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improved visual-semantic embeddings. In *BMVC*, 2018.

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

[7] Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *ICML*, 2015.

[8] R. Hadsell, S. Chopra, and Y. Lecun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[9] Zhang Han, Xu Tao, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[10] Dong Hao, Simiao Yu, Wu Chao, Yike Guo, Dong Hao, Simiao Yu, Wu Chao, Yike Guo, Dong Hao, and Simiao Yu. Semantic image synthesis via adversarial learning. In *CVPR*, 2017.

[11] Anfeng He, Luo Chong, Xinmei Tian, and Wenjun Zeng. A twofold siamese network for real-time object tracking. In *CVPR*, 2018.

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

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

[14] Wanming Huang, Yida Xu, and Ian Oppermann. Realistic image generation using region-phrase attention. In *https://arxiv.org/abs/1902.05395*, 2019.

[15] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

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

[17] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gokmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018.

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

[19] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):2999–3007, 2017.

[20] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[21] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *http://arxiv.org/abs/1703.03130*.

[22] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017.

[23] Xiankai Lu, Chao Ma, Bingbing Ni, Xiaokang Yang, Ian Reid, , and Ming-Hsuan Yang. Deep regression tracking with shrinkage loss. In *ECCV*, 2018.

[24] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In *ICLR*, 2016.

[25] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune. Plug and play generative networks: Conditional iterative generation of images in latent space. In *CVPR*, 2017.

[26] Aaron Van Den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixel cnn decoders. In *NeurIPS*, 2016.

[27] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *CVPR*, 2019.

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

[29] Scott Reed, Aron Van Den Oord, Nal Kalchbrenner, Sergio Gmez Colmenarejo, Ziyu Wang, Belov Dan, and Nando De Freitas. Parallel multiscale autoregressive density estimation. In *ICML*, 2017.

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

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Chen Xi. Improved techniques for training gans. In *NeurIPS*, 2016.

[33] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *Image and Vision Computing*, 45(11):32673–2681, 1997.

[34] Li Shuang, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018.

[35] Li Shuang, Xiao Tong, Hongsheng Li, Yang Wei, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *CVPR*, 2017.

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[37] Hinz Tobias, Heinrich Stefan, and Wermter Stefan. Generating multiple objects at spatially distinct locations. In *ICLR*, 2019.

[38] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[39] Longhui Wei, Shiliang Zhang, Gao Wen, and Tian Qi. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[40] 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*, 2018.

[41] Quanzeng You, Hailin Jin, Zhaowen Wang, Fang Chen, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.

[42] Han Zhang, Tao Xu, and Li Hongsheng. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[43] Shengyu Zhang, Hao Dong, Wei Hu, Yike Guo, Chao Wu, Di Xie, and Fei Wu. Identity-aware textual-visual matching with latent co-attention. In *PCM*, 2018.

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

[45] Zhu Zheng, Wu Wei, Zou Wei, and Junjie Yan. End-to-end flow correlation tracking with spatial-temporal attention. In *CVPR*, 2018.

[46] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018.