AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks

Tao Xu*[1], Pengchuan Zhang[2], Qiuyuan Huang[2],
Han Zhang[3], Zhe Gan[4], Xiaolei Huang[1], Xiaodong He[5]

[1]Lehigh University [2]Microsoft Research [3]Rutgers University [4]Duke University [5]JD AI Research
{tax313, xih206}@lehigh.edu, {penzhan, qihua, xiaohe}@microsoft.com
han.zhang@cs.rutgers.edu, zhe.gan@duke.edu, xiaodong.he@jd.com

Abstract

In this paper, we propose an Attentional Generative Adversarial Network (AttnGAN) that allows attention-driven, multi-stage refinement for fine-grained text-to-image generation. With a novel attentional generative network, the AttnGAN can synthesize fine-grained details at different subregions of the image by paying attentions to the relevant words in the natural language description. In addition, a deep attentional multimodal similarity model is proposed to compute a fine-grained image-text matching loss for training the generator. The proposed AttnGAN significantly outperforms the previous state of the art, boosting the best reported inception score by 14.14% on the CUB dataset and 170.25% on the more challenging COCO dataset. A detailed analysis is also performed by visualizing the attention layers of the AttnGAN. It for the first time shows that the layered attentional GAN is able to automatically select the condition at the word level for generating different parts of the image.

1. Introduction

Automatically generating images according to natural language descriptions is a fundamental problem in many applications, such as art generation and computer-aided design. It also drives research progress in multimodal learning and inference across vision and language, which is one of the most active research areas in recent years [20, 18, 36, 19, 41, 4, 30, 5, 1, 31, 33, 32]

Most recently proposed text-to-image synthesis methods are based on Generative Adversarial Networks (GANs) [6]. A commonly used approach is to encode the whole text description into a global sentence vector as the condition for GAN-based image generation [20, 18, 36, 37]. Although impressive results have been presented, conditioning GAN only on the global sentence vector lacks important fine-grained information at the word level, and prevents the generation of high quality images. This problem becomes even more severe when generating complex scenes such as those in the COCO dataset [14].

Figure 1. Example results of the proposed AttnGAN. The first row gives the low-to-high resolution images generated by G_0, G_1 and G_2 of the AttnGAN; the second and third row shows the top-5 most attended words by F_1^{attn} and F_2^{attn} of the AttnGAN, respectively. Here, images of G_0 and G_1 are bilinearly upsampled to have the same size as that of G_2 for better visualization.

To address this issue, we propose an Attentional Generative Adversarial Network (AttnGAN) that allows attention-driven, multi-stage refinement for fine-grained text-to-image generation. The overall architecture of the AttnGAN is illustrated in Figure 2. The model consists of two novel components. The first component is an attentional generative network, in which an attention mechanism is developed for the generator to draw different sub-regions of the

*work was performed when was an intern with Microsoft Research

image by focusing on words that are most relevant to the sub-region being drawn (see Figure 1). More specifically, besides encoding the natural language description into a global sentence vector, each word in the sentence is also encoded into a word vector. The generative network utilizes the global sentence vector to generate a low-resolution image in the first stage. In the following stages, it uses the image vector in each sub-region to query word vectors by using an attention layer to form a word-context vector. It then combines the regional image vector and the corresponding word-context vector to form a multimodal context vector, based on which the model generates new image features in the surrounding sub-regions. This effectively yields a higher resolution picture with more details at each stage. The other component in the AttnGAN is a Deep Attentional Multimodal Similarity Model (DAMSM). With an attention mechanism, the DAMSM is able to compute the similarity between the generated image and the sentence using both the global sentence level information and the fine-grained word level information. Thus, the DAMSM provides an additional fine-grained image-text matching loss for training the generator.

The contribution of our method is threefold. (*i*) An Attentional Generative Adversarial Network is proposed for synthesizing images from text descriptions. Specifically, two novel components are proposed in the AttnGAN, including the attentional generative network and the DAMSM. (*ii*) Comprehensive study is carried out to empirically evaluate the proposed AttnGAN. Experimental results show that the AttnGAN significantly outperforms previous state-of-the-art GAN models. (*iii*) A detailed analysis is performed through visualizing the attention layers of the AttnGAN. For the first time, it is demonstrated that the layered conditional GAN is able to automatically attend to relevant words to form the condition for image generation. Our code is available at https://github.com/taoxugit/AttnGAN.

2. Related Work

Generating high resolution images from text descriptions, though very challenging, is important for many practical applications such as art generation and computer-aided design. Recently, great progress has been achieved in this direction with the emergence of deep generative models [12, 27, 6]. Mansimov *et al*. [15] built the alignDRAW model, extending the Deep Recurrent Attention Writer (DRAW) [7] to iteratively draw image patches while attending to the relevant words in the caption. Nguyen *et al*. [16] proposed an approximate Langevin approach to generate images from captions. Reed *et al*. [21] used conditional PixelCNN [27] to synthesize images from text with a multi-scale model structure. Compared with other deep generative models, Generative Adversarial Networks (GANs) [6] have shown great performance for generating sharper samples [17, 3, 23, 13, 10, 35, 24, 34, 39, 40]. Reed *et al*. [20] first showed that the conditional GAN was capable of synthesizing plausible images from text descriptions. Their follow-up work [18] also demonstrated that GAN was able to generate better samples by incorporating additional conditions (*e.g*., object locations). Zhang *et al*. [36, 37] stacked several GANs for text-to-image synthesis and used different GANs to generate images of different sizes. However, all of their GANs are conditioned on the global sentence vector, missing fine-grained word level information for image generation.

The attention mechanism has recently become an integral part of sequence transduction models. It has been successfully used in modeling multi-level dependencies in image captioning [30, 38], image question answering [31] and machine translation [2]. Vaswani *et al*. [28] also demonstrated that machine translation models could achieve state-of-the-art results by solely using an attention model. In spite of these progress, the attention mechanism has not been explored in GANs for text-to-image synthesis yet. It is worth mentioning that the alignDRAW [15] also used LAPGAN [3] to scale the image to a higher resolution. However, the GAN in their framework was only utilized as a post-processing step without attention. To our knowledge, the proposed AttnGAN for the first time develops an attention mechanism that enables GANs to generate fine-grained high quality images via multi-level (*e.g*., word level and sentence level) conditioning.

3. Attentional Generative Adversarial Network

As shown in Figure 2, the proposed Attentional Generative Adversarial Network (AttnGAN) has two novel components: the attentional generative network and the deep attentional multimodal similarity model. We will elaborate each of them in the rest of this section.

3.1. Attentional Generative Network

Current GAN-based models for text-to-image generation [20, 18, 36, 37] typically encode the whole-sentence text description into a single vector as the condition for image generation, but lack fine-grained word level information. In this section, we propose a novel attention model that enables the generative network to draw different sub-regions of the image conditioned on words that are most relevant to those sub-regions.

As shown in Figure 2, the proposed attentional generative network has m generators $(G_0, G_1, ..., G_{m-1})$, which take the hidden states $(h_0, h_1, ..., h_{m-1})$ as input and generate images of small-to-large scales $(\hat{x}_0, \hat{x}_1, ..., \hat{x}_{m-1})$. Specifically,

$$
\begin{aligned}
&h_0 = F_0(z, F^{ca}(\bar{e}));\\
&h_i = F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, ..., m-1;\\
&\hat{x}_i = G_i(h_i).
\end{aligned} \tag{1}
$$

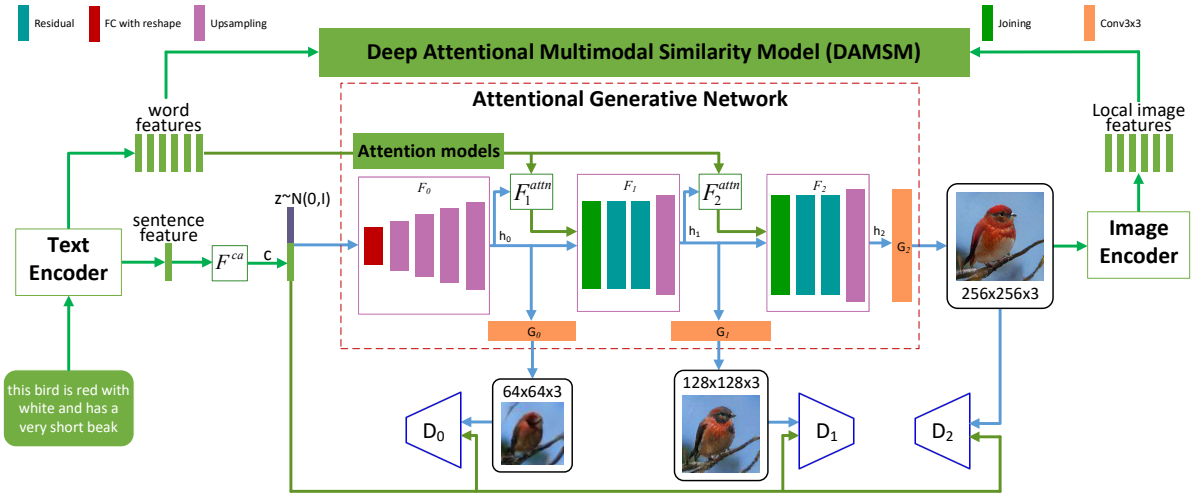

Figure 2. The architecture of the proposed AttnGAN. Each attention model automatically retrieves the conditions (*i.e.*, the most relevant word vectors) for generating different sub-regions of the image; the DAMSM provides the fine-grained image-text matching loss for the generative network.

Here, z is a noise vector usually sampled from a standard normal distribution. $\bar{e}$ is a global sentence vector, and e is the matrix of word vectors. F^{ca} represents the Conditioning Augmentation [36] that converts the sentence vector $\bar{e}$ to the conditioning vector. F_i^{attn} is the proposed attention model at the i^{th} stage of the AttnGAN. F^{ca}, F_i^{attn}, F_i, and G_i are modeled as neural networks.

The attention model $F^{attn}(e, h)$ has two inputs: the word features $e \in \mathbb{R}^{D \times T}$ and the image features from the previous hidden layer $h \in \mathbb{R}^{\hat{D} \times N}$. The word features are first converted into the common semantic space of the image features by adding a new perceptron layer, *i.e.*, $e' = Ue$, where $U \in \mathbb{R}^{\hat{D} \times D}$. Then, a word-context vector is computed for each sub-region of the image based on its hidden features h (query). Each column of h is a feature vector of a sub-region of the image. For the j^{th} sub-region, its word-context vector is a dynamic representation of word vectors relevant to h_j, which is calculated by

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i, \quad \text{where } \beta_{j,i} = \frac{\exp(s'_{j,i})}{\sum_{k=0}^{T-1} \exp(s'_{j,k})}, \qquad (2)$$

$s'_{j,i} = h_j^T e'_i$, and $\beta_{j,i}$ indicates the weight the model attends to the i^{th} word when generating the j^{th} sub-region of the image. We then donate the word-context matrix for image feature set h by $F^{attn}(e, h) = (c_0, c_1, ..., c_{N-1}) \in \mathbb{R}^{\hat{D} \times N}$. Finally, image features and the corresponding word-context features are combined to generate images at the next stage.

To generate realistic images with multiple levels (*i.e.*, sentence level and word level) of conditions, the final objective function of the attentional generative network is defined as

$$\mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_{DAMSM}, \quad \text{where } \mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}. \qquad (3)$$

Here, λ is a hyperparameter to balance the two terms of Eq. (3). The first term is the GAN loss that jointly approximates conditional and unconditional distributions [37]. At the i^{th} stage of the AttnGAN, the generator G_i has a corresponding discriminator D_i. The adversarial loss for G_i is defined as

$$\mathcal{L}_{G_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log(D_i(\hat{x}_i))]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log(D_i(\hat{x}_i, \bar{e})]}_{\text{conditional loss}}, \qquad (4)$$

where the unconditional loss determines whether the image is real or fake while the conditional loss determines whether the image and the sentence match or not.

Alternately to the training of G_i, each discriminator D_i is trained to classify the input into the class of real or fake by minimizing the cross-entropy loss defined by

$$\mathcal{L}_{D_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{x_i \sim p_{data_i}}[\log D_i(x_i)] - \frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log(1 - D_i(\hat{x}_i)] +}_{\text{unconditional loss}}$$
$$\underbrace{-\frac{1}{2}\mathbb{E}_{x_i \sim p_{data_i}}[\log D_i(x_i, \bar{e})] - \frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[\log(1 - D_i(\hat{x}_i, \bar{e})]}_{\text{conditional loss}}, \qquad (5)$$

where x_i is from the true image distribution p_{data_i} at the i^{th} scale, and $\hat{x}_i$ is from the model distribution p_{G_i} at the same scale. Discriminators of the AttnGAN are structurally disjoint, so they can be trained in parallel and each of them focuses on a single image scale.

The second term of Eq. (3), $\mathcal{L}_{DAMSM}$, is a word level fine-grained image-text matching loss computed by the DAMSM, which will be elaborated in Subsection 3.2.

3.2. Deep Attentional Multimodal Similarity Model

The DAMSM learns two neural networks that map sub-regions of the image and words of the sentence to a common semantic space, thus measures the image-text similarity at

the word level to compute a fine-grained loss for image generation.

The text encoder is a bi-directional Long Short-Term Memory (LSTM) [25] that extracts semantic vectors from the text description. In the bi-directional LSTM, each word corresponds to two hidden states, one for each direction. Thus, we concatenate its two hidden states to represent the semantic meaning of a word. The feature matrix of all words is indicated by $e \in \mathbb{R}^{D \times T}$. Its i^{th} column e_i is the feature vector for the i^{th} word. D is the dimension of the word vector and T is the number of words. Meanwhile, the last hidden states of the bi-directional LSTM are concatenated to be the global sentence vector, denoted by $\overline{e} \in \mathbb{R}^{D}$.

The image encoder is a Convolutional Neural Network (CNN) that maps images to semantic vectors. The intermediate layers of the CNN learn local features of different sub-regions of the image, while the later layers learn global features of the image. More specifically, our image encoder is built upon the Inception-v3 model [26] pretrained on ImageNet [22]. We first rescale the input image to be 299×299 pixels. And then, we extract the local feature matrix $f \in \mathbb{R}^{768 \times 289}$ (reshaped from 768×17×17) from the "$mixed\_6e$" layer of Inception-v3. Each column of f is the feature vector of a sub-region of the image. 768 is the dimension of the local feature vector, and 289 is the number of sub-regions in the image. Meanwhile, the global feature vector $\overline{f} \in \mathbb{R}^{2048}$ is extracted from the last average pooling layer of Inception-v3. Finally, we convert the image features to a common semantic space of text features by adding a perceptron layer:

$$v = W f\,, \qquad \overline{v} = \overline{W}\,\overline{f}, \tag{6}$$

where $v \in \mathbb{R}^{D \times 289}$ and its i^{th} column v_i is the visual feature vector for the i^{th} sub-region of the image; and $\overline{v} \in \mathbb{R}^{D}$ is the global vector for the whole image. D is the dimension of the multimodal (*i.e.*, image and text modalities) feature space. For efficiency, all parameters in layers built from the Inception-v3 model are fixed, and the parameters in newly added layers are jointly learned with the rest of the network.

The attention-driven image-text matching score is designed to measure the matching of an image-sentence pair based on an attention model between the image and the text.

We first calculate the similarity matrix for all possible pairs of words in the sentence and sub-regions in the image by

$$s = e^{T}\, v, \tag{7}$$

where $s \in \mathbb{R}^{T \times 289}$ and $s_{i,j}$ is the dot-product similarity between the i^{th} word of the sentence and the j^{th} sub-region of the image. We find that it is beneficial to normalize the similarity matrix as follows

$$\overline{s}_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=0}^{T-1} \exp(s_{k,j})}. \tag{8}$$

Then, we build an attention model to compute a region-context vector for each word (query). The region-context vector c_i is a dynamic representation of the image's sub-regions related to the i^{th} word of the sentence. It is computed as the weighted sum over all regional visual vectors, *i.e.*,

$$c_i = \sum_{j=0}^{288} \alpha_j v_j, \quad \text{where } \alpha_j = \frac{\exp(\gamma_1 \overline{s}_{i,j})}{\sum_{k=0}^{288} \exp(\gamma_1 \overline{s}_{i,k})}. \tag{9}$$

Here, γ_1 is a factor that determines how much attention is paid to features of its relevant sub-regions when computing the region-context vector for a word.

Finally, we define the relevance between the i^{th} word and the image using the cosine similarity between c_i and e_i, *i.e.*, $R(c_i, e_i) = (c_i^T e_i)/(||c_i||||e_i||)$. Inspired by the minimum classification error formulation in speech recognition (see, *e.g.*, [11, 8]), the *attention-driven image-text matching score* between the entire image (Q) and the whole text description (D) is defined as

$$R(Q, D) = \log\Big(\sum_{i=1}^{T-1} \exp(\gamma_2 R(c_i, e_i))\Big)^{\frac{1}{\gamma_2}}, \tag{10}$$

where γ_2 is a factor that determines how much to magnify the importance of the most relevant word-to-region-context pair. When $\gamma_2 \rightarrow \infty$, $R(Q, D)$ approximates to $\max_{i=1}^{T-1} R(c_i, e_i)$.

The DAMSM loss is designed to learn the attention model in a semi-supervised manner, in which the only supervision is the matching between entire images and whole sentences (a sequence of words). Similar to [4, 9], for a batch of image-sentence pairs $\{(Q_i, D_i)\}_{i=1}^{M}$, the posterior probability of sentence D_i being matching with image Q_i is computed as

$$P(D_i|Q_i) = \frac{\exp(\gamma_3 R(Q_i, D_i))}{\sum_{j=1}^{M} \exp(\gamma_3 R(Q_i, D_j))}, \tag{11}$$

where γ_3 is a smoothing factor determined by experiments. In this batch of sentences, only D_i matches the image Q_i, and treat all other $M-1$ sentences as mismatching descriptions. Following [4, 9], we define the loss function as the negative log posterior probability that the images are matched with their corresponding text descriptions (ground truth), *i.e.*,

$$\mathcal{L}_1^w = -\sum_{i=1}^{M} \log P(D_i|Q_i), \tag{12}$$

where 'w' stands for "word".

Symmetrically, we also minimize

$$\mathcal{L}_2^w = -\sum_{i=1}^{M} \log P(Q_i|D_i), \tag{13}$$

where $P(Q_i|D_i) = \frac{\exp(\gamma_3 R(Q_i,D_i))}{\sum_{j=1}^{M}\exp(\gamma_3 R(Q_j,D_i))}$ is the posterior probability that sentence D_i is matched with its corresponding image Q_i. If we redefine Eq. (10) by $R(Q,D) = \left(\overline{v}^T\overline{e}\right)/\left(||\overline{v}||||\overline{e}||\right)$ and substitute it to Eq. (11), (12) and (13), we can obtain loss functions $\mathcal{L}_1^s$ and $\mathcal{L}_2^s$ (where 's' stands for "sentence") using the sentence vector $\overline{e}$ and the global image vector $\overline{v}$.

Finally, the DAMSM loss is defined as

$$\mathcal{L}_{DAMSM} = \mathcal{L}_1^w + \mathcal{L}_2^w + \mathcal{L}_1^s + \mathcal{L}_2^s. \quad (14)$$

Based on experiments on a held-out validation set, we set the hyperparameters in this section as: $\gamma_1 = 5$, $\gamma_2 = 5$, $\gamma_3 = 10$ and $M = 50$. Our DAMSM is pretrained [1] by minimizing $\mathcal{L}_{DAMSM}$ using real image-text pairs. Since the size of images for pretraining DAMSM is not limited by the size of images that can be generated, real images of size 299×299 are utilized. In addition, the pretrained text-encoder in the DAMSM provides visually-discriminative word vectors learned from image-text paired data for the attentional generative network. In comparison, conventional word vectors pretrained on pure text data are often not visually-discriminative, *e.g.*, word vectors of different colors, such as red, blue, yellow, etc., are often clustered together in the vector space, due to the lack of grounding them to the actual visual signals.

In sum, we propose two novel attention models, the attentional generative network and the DAMSM, which play different roles in the AttnGAN. (*i*) The attention mechanism in the generative network (see Eq. 2) enables the AttnGAN to automatically select word level condition for generating different sub-regions of the image. (*ii*) With an attention mechanism (see Eq. 9), the DAMSM is able to compute the fine-grained text-image matching loss $\mathcal{L}_{DAMSM}$. It is worth mentioning that, $\mathcal{L}_{DAMSM}$ is applied only on the output of the last generator G_{m-1}, because the eventual goal of the AttnGAN is to generate large images by the last generator. We tried to apply $\mathcal{L}_{DAMSM}$ on images of all resolutions generated by $(G_0, G_1, ..., G_{m-1})$. However, the performance was not improved but the computational cost was increased.

4. Experiments

Extensive experimentation is carried out to evaluate the proposed AttnGAN. We first study the important components of the AttnGAN, including the attentional generative network and the DAMSM. Then, we compare our AttnGAN with previous state-of-the-art GAN models for text-to-image synthesis [36, 37, 20, 18, 16].

Datasets. Same as previous text-to-image methods [36, 37, 20, 18], our method is evaluated on CUB [29] and COCO [14] datasets. We preprocess the CUB dataset according to the method in [36]. Table 1 lists the statistics of datasets.

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

Dataset	CUB [29]		COCO [14]	
	train	test	train	test
#samples	8,855	2,933	80k	40k
caption/image	10	10	5	5

Table 1. Statistics of datasets.

Evaluation. Following Zhang *et al.* [36], we use the inception score [23] as the quantitative evaluation measure. Since the inception score cannot reflect whether the generated image is well conditioned on the given text description, we propose to use R-precision, a common evaluation metric for ranking retrieval results, as a complementary evaluation metric for the text-to-image synthesis task. If there are R relevant documents for a query, we examine the top R ranked retrieval results of a system, and find that r are relevant, and then by definition, the R-precision is r/R. More specifically, we conduct a retrieval experiment, *i.e.*, we use generated images to query their corresponding text descriptions. First, the image and text encoders learned in our pretrained DAMSM are utilized to extract global feature vectors of the generated images and the given text descriptions. And then, we compute cosine similarities between the global image vectors and the global text vectors. Finally, we rank candidate text descriptions for each image in descending similarity and find the top r relevant descriptions for computing the R-precision. To compute the inception score and the R-precision, each model generates 30,000 images from randomly selected unseen text descriptions. The candidate text descriptions for each query image consist of one ground truth (*i.e.*, $R = 1$) and 99 randomly selected mismatching descriptions.

Besides quantitative evaluation, we also qualitatively examine the samples generated by our models. Specifically, we visualize the intermediate results with attention learned by the attention models F^{attn}. As defined in Eq. (2), weights $\beta_{j,i}$ indicates which words the model attends to when generating a sub-region of the image, and $\sum_{i=0}^{T-1}\beta_{j,i} = 1$. We suppress the less-relevant words for an image's sub-region via

$$\hat{\beta}_{j,i} = \begin{cases} \beta_{j,i}, & \text{if } \beta_{j,i} > 1/T, \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

For better visualization, we fix the word and compute its attention weights with N different sub-regions of an image, $\hat{\beta}_{0,i}, \hat{\beta}_{1,i}, ..., \hat{\beta}_{N-1,i}$. We reshape the N attention weights to $\sqrt{N} \times \sqrt{N}$ pixels, which are then upsampled with Gaussian filters to have the same size as the generated images. Limited by the length of the paper, we only visualize the top-5 most attended words (*i.e.*, words with top-5 highest $\sum_{j=0}^{N-1}\hat{\beta}_{j,i}$ values) for each attention model.

4.1. Component analysis

In this section, we first quantitatively evaluate the AttnGAN and its variants. The results are shown in Table 2

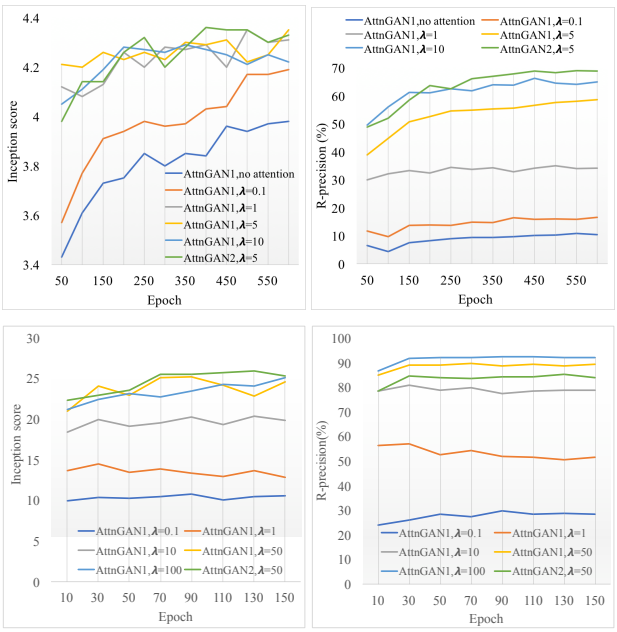

Figure 3. Inception scores and R-precision rates by our AttnGAN and its variants at different epochs on CUB (top) and COCO (bottom) test sets. For the text-to-image synthesis task, $R = 1$.

Method	inception score	R-precision(%)
AttnGAN1, no attention	3.98 $\pm$ .04	10.37$\pm$ 5.88
AttnGAN1, $\lambda = 0.1$	4.19 $\pm$ .06	16.55$\pm$ 4.83
AttnGAN1, $\lambda = 1$	4.35 $\pm$ .05	34.96$\pm$ 4.02
AttnGAN1, $\lambda = 5$	4.35 $\pm$ .04	58.65$\pm$ 5.41
AttnGAN1, $\lambda = 10$	4.29 $\pm$ .05	63.87$\pm$ 4.85
AttnGAN2, $\lambda = 5$	**4.36** $\pm$ **.03**	**67.82** $\pm$ **4.43**
AttnGAN2, $\lambda = 50$ **(COCO)**	**25.89** $\pm$ **.47**	**85.47** $\pm$ **3.69**

Table 2. The best inception score and the corresponding R-precision rate of each AttnGAN model on CUB (top six rows) and COCO (the last row) test sets. More results in Figure 3.

and Figure 3. Our "AttnGAN1" architecture has one attention model and two generators, while the "AttnGAN2" architecture has two attention models stacked with three generators (see Figure 2). In addition, as illustrated in Figure 4, Figure 5, Figure 6, and Figure 7, we qualitatively examine the images generated by our AttnGAN.

The DAMSM loss. To test the proposed $\mathcal{L}_{DAMSM}$, we adjust the value of λ (see Eq. (3)). As shown in Figure 3, a larger λ leads to a significantly higher R-precision rate on both CUB and COCO datasets. On the CUB dataset, when the value of λ is increased from 0.1 to 5, the inception score of the AttnGAN1 is improved from 4.19 to 4.35 and the corresponding R-precision rate is increased from 16.55% to 58.65% (see Table 2). On the COCO dataset, by increasing the value of λ from 0.1 to 50, the AttnGAN1 achieves both high inception score and R-precision rate (see Figure 3). This comparison demonstrates that properly increasing the weight of $\mathcal{L}_{DAMSM}$ helps to generate higher quality images that are better conditioned on given text descriptions. The reason is that the proposed fine-grained image-text matching loss $\mathcal{L}_{DAMSM}$ provides additional supervision (*i.e.*, word level matching information) for training the generator. Moreover, in our experiments, we do not observe any collapsed nonsensical mode in the visualization of AttnGAN-generated images. It indicates that, with extra supervision, the fine-grained image-text matching loss also helps to stabilize the training process of the AttnGAN. In addition, a baseline model, 'AttnGAN1, no attention", with the text encoder used in [19], is trained on the CUB dataset. Without using attention, its inception score and R-precision drops to 3.98 and 10.37%, respectively, which further demonstrates the effectiveness of the proposed $\mathcal{L}_{DAMSM}$.

The attentional generative network. As shown in Table 2 and Figure 3, stacking two attention models in the generative networks not only generates images of a higher resolution (from 128$\times$128 to 256$\times$256 resolution), but also yields higher inception scores on both CUB and COCO datasets. In order to guarantee the image quality, we find the best value of λ for each dataset by increasing the value of λ until the overall inception score is starting to drop on a held-out validation set. "AttnGAN1" models are built for searching the best λ, based on which a "AttnGAN2" model is built to generate higher resolution images. Due to GPU memory constraints, we did not try the AttnGAN with three attention models. As the result, our final model for CUB and COCO is "AttnGAN2, λ=5" and "AttnGAN2, λ=50", respectively. The final λ of the COCO dataset turns out to be much larger than that of the CUB dataset, indicating that the proposed $\mathcal{L}_{DAMSM}$ is especially important for generating complex scenarios like those in the COCO dataset.

To better understand what has been learned by the AttnGAN, we visualize its intermediate results with attention. As shown in Figure 4, the first stage of the AttnGAN (G_0) just sketches the primitive shape and colors of objects and generates low resolution images. Since only the global sentence vectors are utilized in this stage, the generated images lack details described by exact words, *e.g.*, the beak and eyes of a bird. Based on word vectors, the following stages (G_1 and G_2) learn to rectify defects in results of the previous stage and add more details to generate higher-resolution images. Some sub-regions/pixels of G_1 or G_2 images can be inferred directly from images generated by the previous stage. For those sub-regions, the attention is equally allocated to all words and shown to be black in the attention map (see Figure 4). For other sub-regions, which usually have semantic meaning expressed in the text description such as the attributes of objects, the attention is allocated to their most relevant words (bright regions in Figure 4). Thus, those regions are inferred from both word-context features and previous image features of those regions. As shown in Figure 4, on the CUB dataset, the words *the, this, bird* are usually attended by the F^{attn} models for locating the ob-

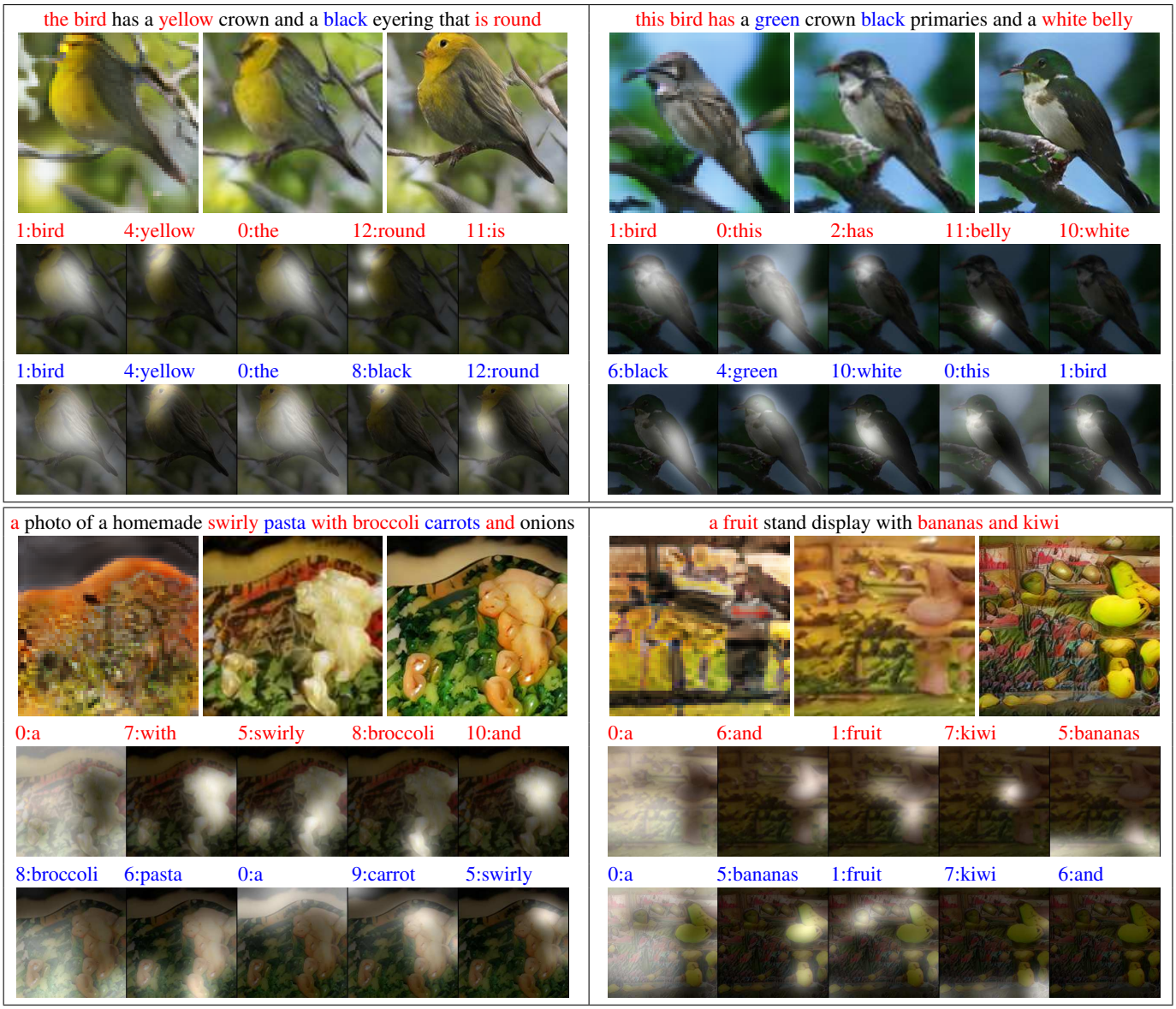

Figure 4. Intermediate results of our AttnGAN on CUB (top) and COCO (bottom) test sets. In each block, the first row gives 64×64 images by G_0, 128×128 images by G_1 and 256×256 images by G_2 of the AttnGAN; the second and third row shows the top-5 most attended words by F_1^{attn} and F_2^{attn} of the AttnGAN, respectively. Refer to the supplementary material for more examples.

Dataset	GAN-INT-CLS [20]	GAWWN [18]	StackGAN [36]	StackGAN-v2 [37]	PPGN [16]	Our AttnGAN
CUB	2.88 ± .04	3.62 ± .07	3.70 ± .04	3.84 ± .06	/	**4.36 ± .03**
COCO	7.88 ± .07	/	8.45 ± .03	/	9.58 ± .21	**25.89 ± .47**

Table 3. Inception scores by state-of-the-art GAN models [20, 18, 36, 37, 16] and our AttnGAN on CUB and COCO test sets.

ject; the words describing object attributes, such as colors and parts of birds, are also attended for correcting defects and drawing details. On the COCO dataset, we have similar observations. Since there are usually more than one object in each COCO image, it is more visible that the words describing different objects are attended by different sub-regions of the image, *e.g.*, *bananas, kiwi* in the bottom-right block of Figure 4. Those observations demonstrate that the AttnGAN learns to understand the detailed semantic meaning expressed in the text description of an image. Another observation is that our second attention model F_2^{attn} is able to attend to some new words that were omitted by the first attention model F_1^{attn} (see Figure 4). It demonstrates that, to provide richer information for generating higher resolu-

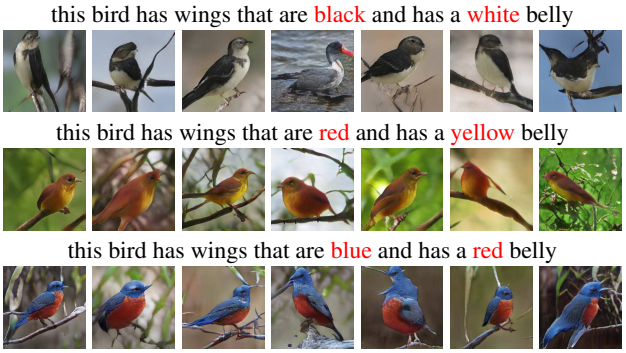

Figure 5. Example results of our AttnGAN model trained on CUB while changing some most attended words in the text descriptions.

Figure 6. 256×256 images generated from descriptions of novel scenarios using the AttnGAN model trained on COCO. (Intermediate results are given in the supplementary material.)

Figure 7. Novel images by our AttnGAN on the CUB test set.

tion images at latter stages of the AttnGAN, the corresponding attention models learn to recover objects and attributes omitted at previous stages.

Generalization ability. Our experimental results above have quantitatively and qualitatively shown the generalization ability of the AttnGAN by generating images from unseen text descriptions. Here we further test how sensitive the outputs are to changes in the input sentences by changing some most attended words in the text descriptions. Some examples are shown in Figure 5. It illustrates that the generated images are modified according to the changes in the input sentences, showing that the model can catch subtle semantic differences in the text description. Moreover, as shown in Figure 6, our AttnGAN can generate images to reflect the semantic meaning of descriptions of novel scenarios that are not likely to happen in the real world, *e.g.*, *a stop sign is floating on top of a lake*. On the other hand, we also observe that the AttnGAN sometimes generates images which are sharp and detailed, but are not likely realistic. As examples shown in Figure 7, the AttnGAN creates birds with multiple heads, eyes or tails, which only exist in fairy tales. This indicates that our current method is still not perfect in capturing global coherent structures, which leaves room to improve. To sum up, observations shown in Figure 5, Figure 6 and Figure 7 further demonstrate the generalization ability of the AttnGAN.

4.2. Comparison with previous methods

We compare our AttnGAN with previous state-of-the-art GAN models for text-to-image generation on CUB and COCO test sets. As shown in Table 3, on CUB, our AttnGAN achieves 4.36 inception score, which significantly outperforms the previous best inception score of 3.82. More impressively, our AttnGAN boosts the best reported inception score on COCO from 9.58 to 25.89, a 170.25% improvement relatively. The COCO dataset is known to be much more challenging than the CUB dataset because it consists of images with more complex scenarios. Existing methods struggle in generating realistic high-resolution images on this dataset. Examples in Figure 4 and Figure 6 illustrate that our AttnGAN succeeds in generating 256×256 images for various scenarios on the COCO dataset, although those generated images of the COCO dataset are not as photo-realistic as that of the CUB dataset. The experimental results show that, compared to previous state-of-the-art approaches, the AttnGAN is more effective for generating complex scenes due to its novel attention mechanism that catches fine-grained word level and sub-region level information in text-to-image generation.

Besides StackGAN-v2 [37], the proposed attention mechanisms can also be applied to the widely used DCGAN framework [17]. On the CUB dataset, we build an AttnDCGAN and a vanilla DCGAN. While the vanilla DCGAN conditioned only on the sentence vector (without the proposed attention mechanisms) is shown unable to generate plausible 256×256 images, our AttnDCGAN is able to generate realistic images. The AttnDCGAN achieves 4.12±.05 inception score and 38.45±4.26% R-precision. The vanilla DCGAN only achieves 2.47±.01 inception score and 3.69±1.82% R-precision because of severe mode collapse. The comparison result further demonstrates the effectiveness of the proposed attention mechanisms.

5. Conclusions

In this paper, an Attentional Generative Adversarial Network, named AttnGAN, is proposed for fine-grained text-to-image synthesis. We build a novel attentional generative network for the AttnGAN to generate high quality image through a multi-stage process. We present a deep attentional multimodal similarity model to compute the fine-grained image-text matching loss for training the generator of the AttnGAN. Our AttnGAN significantly outperforms state-of-the-art GAN models, boosting the best reported inception score by 14.14% on the CUB dataset and 170.25% on the more challenging COCO dataset. Extensive experimental results demonstrate the effectiveness of the proposed attention mechanisms in the AttnGAN, which is especially critical for text-to-image generation for complex scenes.

References

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. VQA: visual question answering. *IJCV*, 123(1):4–31, 2017. 1

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*, 2014. 2

[3] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015. 2

[4] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, 2015. 1, 4

[5] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng. Semantic compositional networks for visual captioning. In *CVPR*, 2017. 1

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2

[7] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. DRAW: A recurrent neural network for image generation. In *ICML*, 2015. 2

[8] X. He, L. Deng, and W. Chou. Discriminative learning in sequential pattern recognition. *IEEE Signal Processing Magazine*, 25(5):14–36, 2008. 4

[9] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*, 2013. 4

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[11] B.-H. Juang, W. Chou, and C.-H. Lee. Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 5(3):257–265, 1997. 4

[12] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[13] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 5

[15] E. Mansimov, E. Parisotto, L. J. Ba, and R. Salakhutdinov. Generating images from captions with attention. In *ICLR*, 2016. 2

[16] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune. Plug & play generative networks: Conditional iterative generation of images in latent space. In *CVPR*, 2017. 2, 5, 7

[17] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 2, 8

[18] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In *NIPS*, 2016. 1, 2, 5, 7

[19] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 1, 6

[20] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In *ICML*, 2016. 1, 2, 5, 7

[21] S. E. Reed, A. van den Oord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, Y. Chen, D. Belov, and N. de Freitas. Parallel multiscale autoregressive density estimation. In *ICML*, 2017. 2

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 4

[23] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, 2016. 2, 5

[24] T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving gans using optimal transport. In *ICLR*, 2018. 2

[25] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Processing*, 45(11):2673–2681, 1997. 4

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 4

[27] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional image generation with pixelcnn decoders. In *NIPS*, 2016. 2

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv:1706.03762*, 2017. 2

[29] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5

[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1, 2

[31] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 1, 2

[32] H. Zhang and K. Dana. Multi-style generative network for real-time transfer. *arXiv:1703.06953*, 2017. 1

[33] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018. 1

[34] H. Zhang and V. M. Patel. Densely connected pyramid dehazing network. In *CVPR*, 2018. 2

[35] H. Zhang, V. Sindagi, and V. M. Patel. Image de-raining using a conditional generative adversarial network. *arXiv:1701.05957*, 2017. 2

[36] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 1, 2, 3, 5, 7

[37] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *arXiv: 1710.10916*, 2017. 1, 2, 3, 5, 7, 8

[38] Z. Zhang, Y. Xie, F. Xing, M. Mcgough, and L. Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *CVPR*, 2017. 2

[39] Z. Zhang, Y. Xie, and L. Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, 2018. 2

[40] Z. Zhang, L. Yang, and Y. Zheng. Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In *CVPR*, 2018. 2

[41] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, 2018. 1