2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)

2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE) | 978-1-7281-9111-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCEEE49695.2021.9429602

SemGAN: Text to Image Synthesis from Text Semantics using Attentional Generative Adversarial Networks

Ammar Nasr
Department of Electrical and Electronic Engineering
University of Khartoum
Khartoum, Sudan
Email: ammarnasraza@gmail.com

Ruba Mutasim
Department of Electrical and Electronic Engineering
University of Khartoum
Khartoum, Sudan
Email: rubamutasim335@gmail.com

Hiba Imam
Department of Electrical and Electronic Engineering
University of Khartoum
Khartoum, Sudan
Email: hiba.im@gmail.com

***Abstract*—Text to Image Synthesis is the procedure of automatically creating a realistic image from a particular text description. There are numerous innovative and practical applications for text to image synthesis, including image processing and computer-aided design. Using Generative Adversarial Networks (GANs) alongside the Attention mechanism has led to huge improvements lately. The fine-grained attention mechanism, although powerful, does not preserve the general description information well in the generator since it only attends to the text description at word-level (fine-grained). We propose incorporating the whole sentence semantics when generating images from captions to enhance the attention mechanism outputs. According to experiments, on our model produces more robust images with a better semantic layout. We use the Caltech birds dataset to run experiments on both models and validate the effectiveness of our proposal. Our model boosts the original AttnGAN Inception score by +4.13% and the Fréchet Inception Distance score by +13.93%. Moreover, an empirical analysis is carried out on the objective and subjective measures to: (i) address and overcome the limitations of these metrics (ii) verify that performance improvements are due to fundamental algorithmic changes rather than initialization and fine-tuning as with GANs models.**

I. Introduction

The automatic synthesis of images from descriptors has raised much interest in the research fields of Computer Vision and Natural Language Processing. Despite the complexity of the process, it is revolutionizing many real-world applications in marketing, e-commerce, games, advertisements, and many other industries, most notably this task underlies a significant step towards full machine intelligence. Recently the devolvement of generic and robust Deep Recurrent Neural Networks architectures has lead to learning discriminative feature representations. Meanwhile, generative adversarial networks (GANs) have begun to generate highly compelling images of specific categories under two major challenges of achieving visual realism in the sense of the extent to which an image appears to people like a photo rather than computer-generated; the other is fulfilling high-level semantic consistency and low-level semantic diversity because of the complexity of linguistic expressions.

II. Review

Based on the approach taken, efforts registered in the field can be categorized into three main categories :

Direct Generative Adversarial Text to Image Synthesis. The featured algorithms learn a text feature description that captures the important visual details and then use these features to synthesize an image without any intermediate representation.

Reed, Scott and Akata [12] trains a deep convolutional generative adversarial network (DC-GAN) conditioned on text features. In attempts to alleviate the mode collapse problem Cha et al. [1] proposed modification of the sampling procedure when learning to generate a matrix of mismatching text and image pairs which an extra discriminator uses to regress relevance between the text and the image.

To stabilize the GANs training Zhang et al. [20] were the first to generate realistic image with high fidelity at resolutions of 256×256 using multi-stage GAN architecture. Huang, Xin, Mingjie and Minglun[21] introduced accompanying hierarchical-nested adversarial objectives in the network with a single-stream generator architecture and multiple discriminators.

The Attentional Generative Adversarial Network (AttnGAN) Was first propose in [19], it allowed attention-driven, multi-stage text-to-image generation. The introduction of a

dynamic memory module by Zhu et al. [23] was to refine ambiguous image contents, when the initial images are not well generated.

Text-to-Image Synthesis with Layouts. Includes approaches that utilised meta-data about the image and extra information, such as object detection and labeling with bounding boxes or intermediate representations of the images in the form of scene graphs or scene layouts When generating complex images that contain scenes with multiple objects (MS-COCO dataset)

Li, Jianan and Yang [8] synthesizes layouts by first taking a collection of arbitrarily placed 2D graphic objects it then Uses attention modules to refine the labels of those objects to produce the layout.

Using sequence to sequence (Seq-to-Seq) learning, Boren et al. [7] generates semantic layout using from the scene graph. It derives sequence proxies for the two modalities, and a Transformer-based Seq-to-Seq model learns to transduce one into the other. Controllable object locations in [11] use an extension of Pixel Convolutional Neural Networks (PixelCNN) so the model can generate images conditioned on part keypoints and segmentation masks.

Semantic Image Manipulation This refers to generating a new image from a textual description conditioned on an already existing one such that the generated images contain the modifications described and also maintain text-irrelevant features of the source image.

[16] tried to mitigate the lack of image-caption paired datasets and the noisiness of captions by Adding a dialogue that further describes the scene. Hong, Seunghoon and Yan [4] employs structured semantic layout as an intermediate representation for manipulation at which the user can perform manipulation, e.g. adding or removing objects.

To enable object insertion and to facilitate image editing, sharma et al. [6] proposed a network that has two generative modules where the first one determines the location of the input object mask and the second one determines the pose and shape of the object. Instead of one step image generation [9] introduced a recurrent system that generates images iteratively.

A new task of Interactive Image Editing via conversational language suggested by El-Nouby, Alaaeldin and Sharma [9] in which the user can dictate what edits they wish via multi-turn dialogue sessions. Zhou et al. [22] presented a method to manipulate the visual appearance of a person image according to text descriptions.

III. Methodology

The Caltech CUB-200 birds dataset is a publicly accessible dataset used in this report [18]. The CUB-200 dataset comprises 11,788 images of 200 categories of birds. Each one of the images has ten captions. They are at a wide range of from 4 to 62 words in length, the context is not identified, and the bird species are not named. The dataset is split into train and test such that they contain disjoint classes of images as summarised in Table I.

The core architecture of the model is divided into two parts to a total of six modules, so is the training. We first train the DAMSM model, so we will have our text and image encoder ready to train the generative network.

The Text Encoder is modeled as an LSTM Recurrent Neural Network [15] which transform input caption into semantic representations. The representations are bidirectional and processed as cascaded hidden states to encapsulate the context of the word. All representations are then grouped into a matrix e of dimensions $D \times T$. Each column T of the matrix corresponds to a specific word semantic representation and the length of that column D is the representation vector dimension.

The Image Encoder is a Convolution Neural Network (CNN) that transfers images to semantic vectors. The image encoder is based on the ImageNet [13] pre-trained model Inception-v3 [17]. The input image is re-scaled to 299×299 pixels. The re-scaled image is divided into multiple sub-regions by the image encoder. Each one of the 289 sub-regions is further characterised by a vector of dimension 768 which are combined into a local feature matrix f. Finally, by adding a perceptron layer $(\underset{768\times289}{\mathrm{f}} \longrightarrow preceptorn \longrightarrow \underset{D\times289}{\mathrm{v}})$,we link the visual features (in the form of sub-regions) with context text features (in the form of semantic representations) to be processed later in the Multimodal (text and image modalities) model. Each column v_i is a sub-region but characterised by a common multimodal vector D with same dimension of the semantic representation vector D or the multimodal feature length. To enhance efficiency, only the parameters in separately introduced layers are trained simultaneously with the remainder of the network.

The Deep Attentional Multimodal Similarity Model DAMSM trains the RNN text encoder and the CNN image encoder in parallel to measure the similarity between each visual feature in the image and text feature in the caption. The similarity measure is then used to associate each word with the most relevant sub-regions, and to test the conformity of the generated image to the input text by calculating the fine-grained image generation loss. This model quantifies the matching of an image sub-region and a word pair by means of a multi-modal conversion between an image and a text resulting in an **attention-driven image-text matching score**.

The similarity matrix is calculated for all pairs of words and sub-regions by eq.1. The similarity matrix $\underset{T\times289}{\mathrm{s}}$ entries $s_{i,j}$ are the dot-product similarity between the i^{th} word and the j^{th} sub-region. For better performance, the similarity matrix is then normalized element-wise to $\bar{s}_{i,j}$

$$s = e^T v \tag{1}$$

TABLE I: Statistics Of Dataset

Dataset	CUB-200	
	train	test
#samples	8855	2933
#captions per image	10	10

When computing a word region-context vector, γ_1 is a smoothing factor deciding how much attention is paid to features of its respective sub-regions.

The attention-driven image-text matching score between the input image (Q) and the full caption (D) is formulated in eq.2 as featured in speech recognition by the minimum classification error formulation [5]. $R(c_i, e_i)$ is the relevance between the i^{th} word and the image using cosine similarity $\frac{c_i^T \cdot e_i}{||c_i|| \times ||e_i||}$

$$R(Q, D) = log(\sum_{i=1}^{T-1} exp(\gamma_2 R(c_i, e_i)))^{\frac{1}{\gamma_2}} \tag{2}$$

In the task of *automatically generating image descriptions* [2] developed a Deep Multimodal Similarity Model (DMSM) to model global similarity between images and text. The loss function minimized during training represents the negative log posterior probability of the caption given the corresponding image.Similarly, [19]'s DAMSM loss for a batch of image-sentence pairs $(Q_i, D_i)_{i=1}^{M}$ is computed as

$$P(D_i|Q_i) = \frac{exp(\gamma_3 R(Q_i, D_i))}{\sum_{j=1}^{M} exp(\gamma_3 R(Q_i, D_j))} \tag{3}$$

In eq.3 for a the batch of size M of sentences D_j only the sentence D_i matches image Q_i. Symmetrically, we define the equation for the posterior probability of image Q_j being matching with image D_i.

As proposed by [2], the loss function is defined as the negative log posterior probability that: (i) Each sub-region in all images is paired with the most relevant semantic word representations in the captions eq.4, (ii) the semantic word representations in the captions are matched with their corresponding images' sub-regions eq.5.

$$L_1^w = -\sum_{i=1}^{M} Log(P(Q_i|D_i)) \tag{4}$$

$$L_2^w = -\sum_{i=1}^{M} Log(P(D_i|Q_i)) \tag{5}$$

Meanwhile, using the bi-directional LSTM, we can obtain a *global sentence vector* $\bar{e}$ by concatenating its last hidden state. Redefining the word and region relevance equation eq.2 to be reflect the relevance between the sentence to the image using the sentence vector $\bar{e}$ and the image vector $\bar{v}$ we get:

$$R(Q, D) = \frac{\bar{v}^T \cdot \bar{e}}{||\bar{e}|| \times ||\bar{v}||} \tag{6}$$

Eventually,we substitute Eq.6 in Eq.3, Eq.4, and Eq.5 to compute L_1^s and L_2^s and the overall loss of DAMSM is therefore formulated as:

$$L_{DAMSM} = L_1^w + L_2^w + L_1^s + L_2^s \tag{7}$$

The Generative Adversarial Network. We introduced a multi-stage cascaded generator by successively grouping three Discriminator-Generator networks pairs in figure 1. Due to its consistent quality in producing realistic images we follow the underlying framework mentioned in [19]

Text Embedding Module uses bi-directional LSTM Text Encoder of to extract semantic embeddings from the given text description T eq.8, which include a word embedding e and a sentence embedding $\bar{\mathrm{e}}$ Similar to the STEM module of [10].

$$e, \bar{\mathrm{e}} = RNN(T) \tag{8}$$

where e , $\bar{\mathrm{e}}$ and T are the word embedding , sentence embedding and Text description. Due to the complexity of the text domain, it is likely that text with few permutations will share similar semantics. Therefore, we adopt the standard practise of conditioning the sentence embedding using conditioning augmentation [20]. In particular we 're using F^{ca} as the conditioning function and $\underset{C\times 1}{\bar{\mathrm{e}}_{ca}}$ as the augmented sentence vector where C is the conditioning dimension:

$$\bar{\mathrm{e}}_{ca} = F^{ca}(\bar{\mathrm{e}}) \tag{9}$$

Attentive Generative Module, for this part we used three image generators to generate images sequentially in three stages in rising quality from 64×64 to 128×128 and finally $256times256$. Theoretically, any number m of generators can be used denoted as $[G_0, G_1, ..., G_{m-1}]$.

Each generator takes the hidden states $[h_0, h_1, ..., h_{m-1}]$ as input and generate images of small-to-large scales $[I_0, I_1, ..., I_{m-1}]$. The first hidden state h_0 is expressed as :

$$h_0 = F_0(z, \bar{\mathrm{e}}_{ca})) \tag{10}$$

where, $z \sim N(0, 1)$ denotes random noises y sampled from a standard normal distribution, F_i is the visual feature transformers modeled as neural network. Now, to get $h_i i \in [1, 2, ..., m-1]$ we introduce the global-local collaborative attention model [10] with two components Att_{i-1}^w and Att_{i-1}^s which are then concatenated to generate F_i^{att}

$$h_i = F_i(h_{i-1}, F_i^{att}(e, \bar{e_{ca}}, h_{i-1})) \tag{11}$$

The word-level attention model $Att_{i-1}^w(e, h)$ generate attentive word-context feature from the matrix of the semantic word representation and local feature image sub-regions matrix after being converted to the common multimodal dimension with the semantic representation using the perceptron layer. Next we obtain the attention score which indicates the attention paid to each word in the caption text during the process of generating local feature of every sub-region in the image as :

$$\beta_{j,i} = softmax(h_{i-1}^T \cdot e') \tag{12}$$

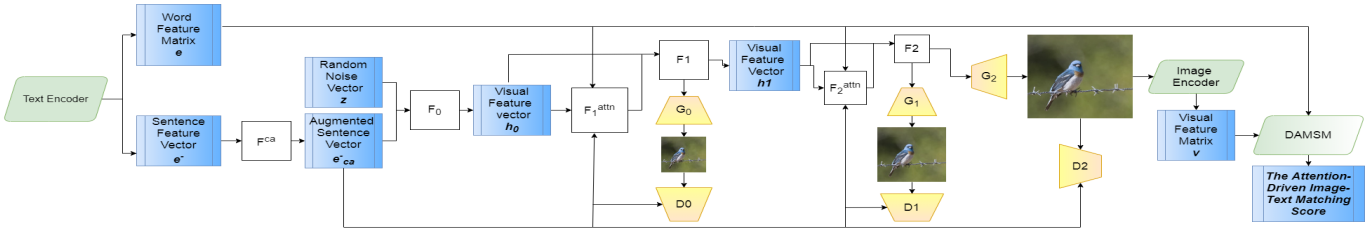


Fig. 1: The Model

Then, a word-context vector [19] is computed for each sub-region of d by calculating the inner product between the attention score and e':

$$c_j = \sum_{i=0}^{T-1} e'_i \cdot \beta_{j,i} \tag{13}$$

The attentive word-context feature is donated $Attn^w_{i-1}(e,h) = (c_0, c_1, ..., c_{N-1})$ where it has the same dimension as $\underset{D' \times N}{h}$ i.e., $\underset{D' \times N}{Attn^w_{i-1}}$.

Add the sentence level attention model Att^s_{i-1} to place a global constraint on the image regions and the text description semantics to force the generator to produce a coherent layout of the entire image that matches the overall meaning of its caption. Analogous to word level attention model, the augmented sentence embedding $\bar{e}_{ca}$ is Transformed to the shared semantic dimension of the image feature by introducing a new perceptron layer $\underset{D' \times C}{V}$. The transformed vector is element-wise multiplied with the attention score to obtain the final *attentive sentence-context vector*.

$$Attn^s_{i-1} = (V \cdot \bar{e}_{ca}) \odot (softmax(h_{i-1}) \odot (V \cdot \bar{e}_{ca})) \tag{14}$$

The sentence attention score in eq.14 is the element-wise product of the image feature h_{i-1} and the augmented sentence vector after being transformed to the common semantic space $V \cdot \bar{e}_{ca}$. The attentive sentence-context of dimensions $\underset{D \times N}{Att^s_{i-1}}$ is finally concatenated with the attentive word-context Att^w_{i-1} and image feature h_{i-1} to be passed to the feature visual transformer F_i in eq.11.The resulting image feature h_i is further passed to the generators from each of the m stages to produce the respective image I_i.

$$I_i = G_i(h_{i-1}), i \in [1, 2, ..., m-1] \tag{15}$$

IV. Results

A. Evaluation Metric

The Inception Score IS is an objective metric for assessing the realism of the images generated, particularly the synthetic image produced by the GANs models. The inception score was proposed by [14] As an alternative to the intuitive metric of using human annotators evaluate the visual quality of the images. The writers proved that their scores are well correlated with the subjective assessment.

$$IS = exp(\mathbb{E}_x \, KL(p(y|x)||p(y))) \tag{16}$$

The Fréchet Inception Distance FID was proposed by Heusel, Martin and Ramsauer [3] to provide an alternative for quantifying the quality of samples generated from GANs.FID embeds a set of generated samples into a feature space given by a specific layer of Inception Net (activations from the penultimate layer of the Inception network). The embedded layer -viewed as a continuous multivariate Gaussian- is parameterized by the $mean\,\mu$ and $covariance\,\Sigma$ is for both the generated data P_g and the real data P_r.The Fréchet distance between these two Gaussians (a.k.a Wasserstein-2 distance) is then used to quantify the quality of generated samples.

$$FID(r,g) = ||\mu_r - \mu_g||^2 + T_r(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}) \tag{17}$$

B. Experiments

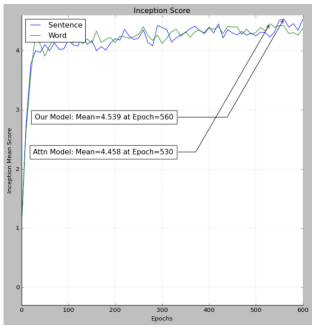
We perform comprehensive quantitative and qualitative analyses to validate the effectiveness of our approach over the then state of the art methods for text to image synthesis [19]. The performance of the two methods of comparison is evaluated using the code released by their authors. In addition, we test a variety of simple generator network models to analyse the overall design and function of each of the components of our proposed architecture. For the first baseline, we directly generate 64x64 images from the Stage-I generator, then the Stage-II and stage-III to generate 128x128 and 256x256 images, respectively. Although all three stages are trained simultaneously, we generate images from each stage at the evaluation to investigate whether if the stacked architecture is beneficial or not.

We calculate The Inception Score means and standard deviations in addition to the FID distance for both models on the CUB birds dataset in table II. These findings concretely indicate an improvement in performance by **+4.13% in the inception score** and **+13.93% in the FID score**. To further understand and analyse these results, we calculated each of the scores 60 times for both models and plotted them side by side in figure 2. It can be seen that our model take 5% more time to reach it is minimum due to the extra processing of the whole sentence semantic but in return produce more semantically robust images as evident by the boost in both scores. Finally, the larger increase in the FID compared the IS metric could be interpreted as our model is improving in two ways: (i)

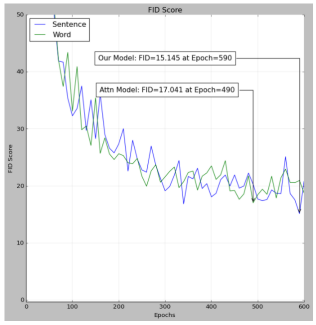
TABLE II: Comparison Of Our Improved Model Vs The Base AttnGAN Model

Model	Inception Score	FID Score
AttnGAN	*4.36*	*17.04*
Our Model	*4.54*	*15.15*

Generating slightly better objects for discrimination and (ii) Generating an overall more realistic layout of the whole image.



(a) *The Inception score every 10 epochs for both Models.*



(b) *The FID score every 10 epochs for both Models*

Fig. 2: FID and IS vs number of epochs for AttnGAN and our model

Figures 3, 4 and 5 show a side by side comparison of the models. The images are generated from random test labels captions, while Figure 6 include a side by side comparison of the models to real images from the training dataset. All images are of size 256x256.

Fig. 3: Images generated by AttnGan (top row) vs our model(bottom row) from the caption: this bird is red with white and has a very short

Fig. 4: Images generated by AttnGan (top row) vs our model(bottom row) from the caption: this is a small light grey bird with a small head and green crown, nape and some green colouring on its wings

Fig. 5: Images generated by AttnGan (top row) vs our model(bottom row) from the caption: the bird has a yellow crown and a black eyering that is round

Fig. 6: Comparison between real images , AttnGAN samples and samples from our model

C. Discussion

After close inspection of images and the corresponding captions from both models, we conclude the following points regarding the general types of captions that affect the generation of our model:

First, All images are generated from the random noise in the latent-space. Depending on this initialization, the same label/caption can produce a countless number of images, some of which may look like more noise if it is initialized far from the centre of the latent-space.

Second, both metrics have their limitations, and their correlation to the human perceived quality and image-text matching is not absolute. Both metrics show better correlation with bigger dataset sizes since they find the average.

Third, the output of our model varies extremely with the type of captions, generating very realistic images for one type and bad images for other types. For extended Captions, our model tries to make sense of the whole sentence semantic while paying attention to specific words, resulting in images with varying level of quality. The second type is captions with short sentences that clearly and specifically describe the image. This is the type at which our model excels and produce subjectively and objectively great images. The third type is when the captions s short and does not have much information, in this case, the model collapses and regenerate the same object multiple times, this mode collapse may be hard for

the evaluation metrics to detect and is an inherited problem in GANs that has been receiving a great deal of attention in the community. It was likely to attain even better results had we used bigger sample size but due to processing power limitations that was not possible at the time.

V. Conclusion

In this paper, we proposed attentional GAN for coarse-grained and fine-grained image generation. Because of the comprehensiveness provided by the new design of condition sets and as we use an excellent dataset in the sense that most of the captions describe the object of the image explicitly, our model, as provided by metrics, significantly improves the quality of images generated in the original attention framework boosting the inception score by +4.13% and the Fréchet Inception Distance score by +13.93%.

References

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

[2] Hao Fang et al. "From captions to visual concepts and back". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1473–1482.

[3] Martin Heusel et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems*. 2017, pp. 6626–6637.

[4] Seunghoon Hong et al. "Learning hierarchical semantic image manipulation through structured representations". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2708–2718.

[5] Biing-Hwang Juang, Wu Hou, and Chin-Hui Lee. "Minimum classification error rate methods for speech recognition". In: *IEEE Transactions on Speech and Audio processing* 5.3 (1997), pp. 257–265.

[6] Donghoon Lee et al. "Context-aware synthesis and placement of object instances". In: *Advances in neural information processing systems*. 2018, pp. 10393–10403.

[7] Boren Li et al. "Seq-sg2sl: Inferring semantic layout from scene graph through sequence to sequence learning". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 7435–7443.

[8] Jianan Li et al. "Layoutgan: Generating graphic layouts with wireframe discriminators". In: *arXiv preprint arXiv:1901.06767* (2019).

[9] Alaaeldin El-Nouby et al. "Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 10304–10312.

[10] Tingting Qiao et al. "Mirrorgan: Learning text-to-image generation by redescription". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1505–1514.

[11] Scott Reed et al. "Generating interpretable images with controllable structure". In: (2016).

[12] Scott Reed et al. "Generative adversarial text to image synthesis". In: *arXiv preprint arXiv:1605.05396* (2016).

[13] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[14] Tim Salimans et al. "Improved techniques for training gans". In: *Advances in neural information processing systems*. 2016, pp. 2234–2242.

[15] Mike Schuster and Kuldip K Paliwal. "Bidirectional recurrent neural networks". In: *IEEE transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[16] Shikhar Sharma et al. "Chatpainter: Improving text to image generation using dialogue". In: *arXiv preprint arXiv:1802.08216* (2018).

[17] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[18] Catherine Wah et al. "The caltech-ucsd birds-200-2011 dataset". In: (2011).

[19] Tao Xu et al. "Attngan: Fine-grained text to image generation with attentional generative adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1316–1324.

[20] Han Zhang et al. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5907–5915.

[21] Zizhao Zhang, Yuanpu Xie, and Lin Yang. "Photographic text-to-image synthesis with a hierarchically-nested adversarial network". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6199–6208.

[22] Xingran Zhou et al. "Text guided person image synthesis". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3663–3672.

[23] Minfeng Zhu et al. "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5802–5810.