CRITICAL REVIEW

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

Sachith M. Gunawardane

Contents

1. Introduction

The research paper delves into the pivotal challenge of translating textual descriptions into visual representations, a task with far-reaching implications in domains such as artistic creation, computer-aided design, and the advancement of multimodal learning at the intersection of vision and language. Traditionally, the prevalent method involves encoding the entire text description into a singular sentence vector, serving as a conditioning factor for Generative Adversarial Network (GAN)-based image synthesis. However, this approach overlooks crucial nuances at the level of individual words, thereby impeding the generation of high-fidelity images. In response to this limitation, the authors introduce the Attentional Generative Adversarial Network (AttnGAN), which aims to address the deficiency in word-level information processing and elevate the quality of image generation outcomes.

2. Comparison with Related Work

The comparative analysis of AttnGAN against prior leading GAN models for text-to-image synthesis, including GAN-INT-CLS, GAWWN, StackGAN, StackGAN-v2, and PPGN, provides valuable insights into the evolution of architectural advancements in this field.

GAN-INT-CLS Architecture

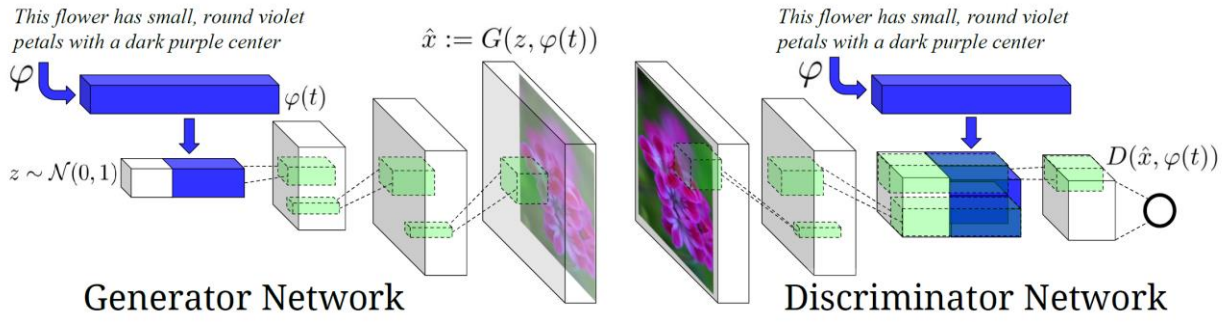

Figure 1: GAN-INT-CLS: RNN encoder with GAN decoder, as illustrated in [1]

The GAN-INT-CLS architecture, as proposed by Read et al. [1], sought to tackle the text-to-image challenge by incorporating an RNN encoder for textual input, which was then fused with a noise vector to form the input vector for the generator network. This approach aimed to train a deep convolutional generative adversarial network (DC-GAN) conditioned on text features encoded via

a hybrid character-level convolutional recurrent neural network. The generator and discriminator networks operated through feed-forward inference, with conditioning based on the extracted text features.

Generative Adversarial What-Where Network (GAWWN) Architecture

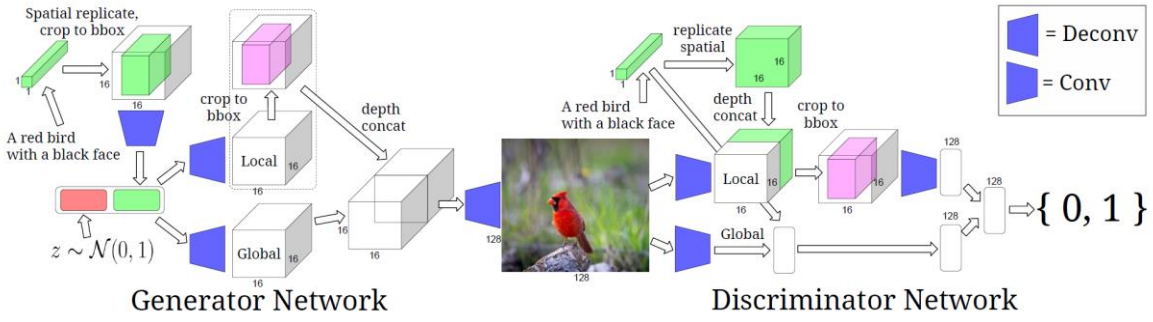

Figure 2: GAWWN with bounding box location control, as illustrated in [2]

The Generative Adversarial What-Where Network (GAWWN) architecture, developed by the same research group that introduced GAN-INT-CLS, represents a notable enhancement [2]. In this architecture, additional inputs such as bounding box and key point information were integrated to guide the network in determining the precise location of objects within the generated images. This enriched input data empowered the generator with more contextual information, resulting in improved performance compared to GAN-INT-CLS. Notably, GAWWN addressed the previous limitation of not having explicit control over the positioning of objects within the synthesised scenes.

StackGAN: multi-stage generation process

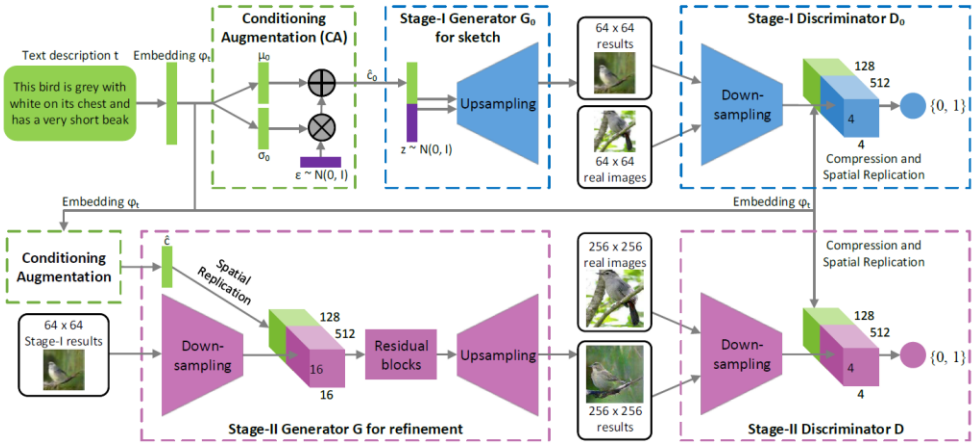

Figure 3: StackGAN Architecture, as illustrated in [3]

In their work on StackGAN, Zhang et al. [3] introduced a multi-stage generation process that forms the basis for the subsequent AttnGAN model. The conditional-GAN (GAN-INT-CLS) framework, when conditioned on textual descriptions, demonstrates the capability to produce images closely aligned with the semantics of the given text. However, achieving the generation of high-resolution, lifelike photographs poses a formidable challenge within the GAN framework. The conventional approach of augmenting upsampling layers in state-of-the-art GAN architectures to handle high-resolution image generation often leads to training instability and outputs of poor quality.

To surmount these hurdles, Zhang et al. [3] proposed the Stacked Generative Adversarial Networks (StackGAN) architecture, designed to generate 256 x 256 photo-realistic images conditioned on textual descriptions. The approach involves decomposing the complex problem into more manageable sub-problems through sketch-refinement. In the first stage, known as Stage-I GAN, the network sketches the rudimentary shape and colours of the object based on the provided text, resulting in low-resolution Stage-I images. Subsequently, Stage-II GAN refines these Stage-I outputs along with the text descriptions to generate high-resolution images imbued with photo-realistic details. This process rectifies any imperfections from Stage-I and enhances the visual fidelity through intricate refinements. The architectural layout of StackGAN comprises two sets of generators and discriminators, as depicted in Figure 3.

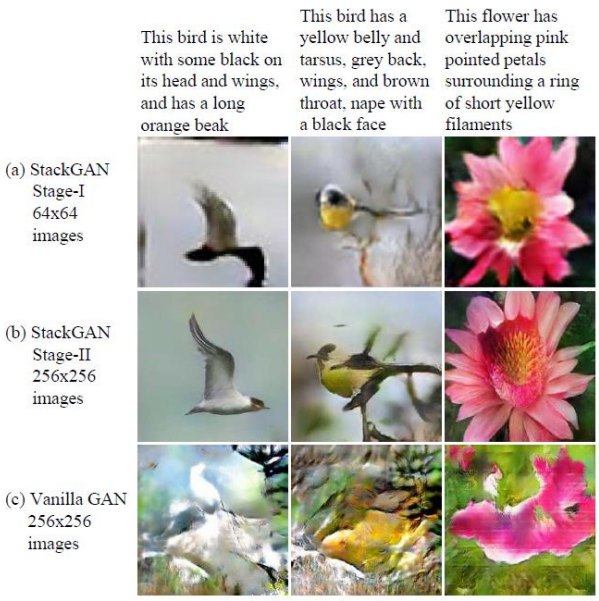

Figure 4: Comparison of the proposed StackGAN and a CAN-INT-CLS

Illustrated in Figure 4, the functionality of StackGAN unfolds as follows: (a) Given textual descriptions, Stage-I sketches preliminary shapes and basic colours, yielding low-resolution images. (b) Stage-II inputs Stage-I results and text descriptions to generate high-resolution images with nuanced photo-realistic details. (c) In contrast, a vanilla 256×256 GAN, which merely increases upsampling layers in GAN-INT-CLS, struggles to produce coherent images of 256×256 resolution.

StackGAN-v2

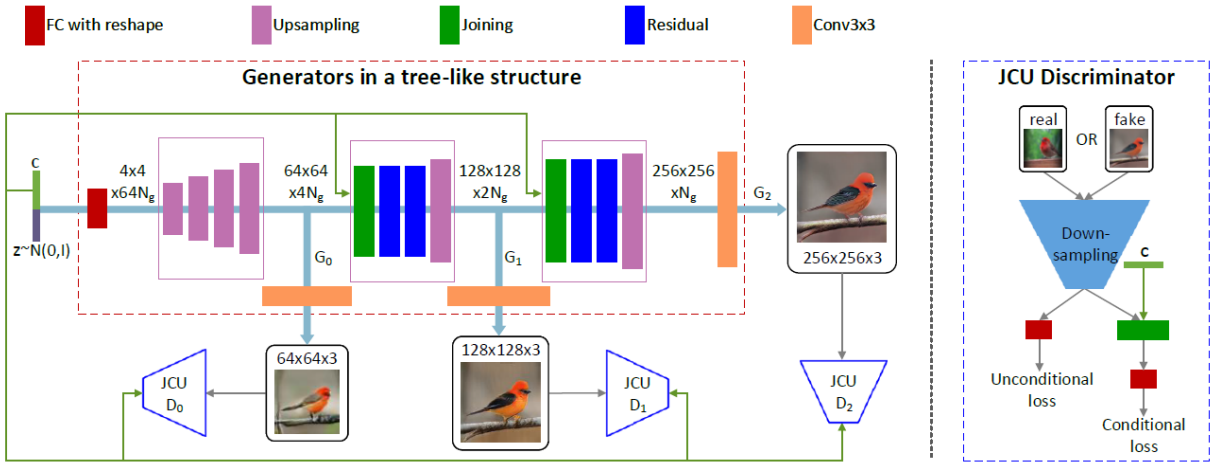

Figure 5: StackGAN-v2 Architecture, as illustrated in [4]

Building upon StackGAN, Zhang et al. [4] introduced StackGAN-v2, presenting minor enhancements over the original model. StackGAN-v2 incorporates multiple generators and discriminators organised in a hierarchical tree-like structure. This design facilitates the generation of images at various scales corresponding to the same scene through distinct branches of the tree. The improved stability in training exhibited by StackGAN-v2 stems from jointly approximating multiple distributions within its architecture.

Plug & Play Generative Network (PPGN) Architecture

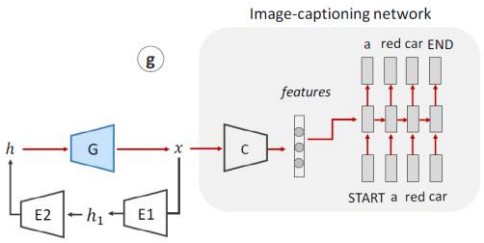

Figure 6: PPGN Architecture - Sampling conditioning on captions, as illustrated in [5]

The Plug & Play Generative Network (PPGN) architecture introduces a unique setting where image generation is conditioned not on classes but on captions. This is achieved by integrating a recurrent image-captioning network with the output layer of the generator network G, followed by iterative sampling. This approach significantly enhances the reverse process of describing an image, showcasing the innovative strategies within the PPGN framework.

The authors of AttnGAN emphasise its distinctive contributions, particularly the incorporation of attention mechanisms that significantly elevate the quality of generated images, especially in complex scenes. This addresses previous limitations observed in methods that solely rely on global sentence vectors for conditioning image generation.

The motivations behind the development of this novel architecture are multifaceted. Firstly, the text-to-image problem is akin to an Encoder-Decoder paradigm, where the given text descriptor needs to be encoded and decoded to correspond to an image. However, the authors sought to approach this challenge from a Sequence-to-Sequence perspective, focusing on processing image sequences. This departure from conventional methods is evident in StackGAN, which adopts a multi-stage generation process instead of a single final stage.

During the period of this research, attention models gained momentum within the research community, prompting exploration into leveraging Attention models (Seq2Seq) to further refine solutions to the Text-to-Image problem. Lastly, there is an exploration into mechanisms for validating the coherence between the generated text and images, ensuring alignment and agreement between the two modalities.

3. Methodology

The AttnGAN introduces two innovative elements: an attentional generative network and a deep attentional multimodal similarity model (DAMSM). The attentional generative network incorporates an attention mechanism enabling the generator to selectively focus on specific sub-regions of the image based on the salient words in the input text. On the other hand, the DAMSM comprises two neural networks that map subregions of the image and words of the sentence into a shared semantic space. This process facilitates the measurement of image-text similarity at the word level, enabling the computation of a fine-grained loss for the generated image. The

architecture of AttnGAN, as depicted in Figure 7, visually encapsulates these components and their interconnected functionalities.

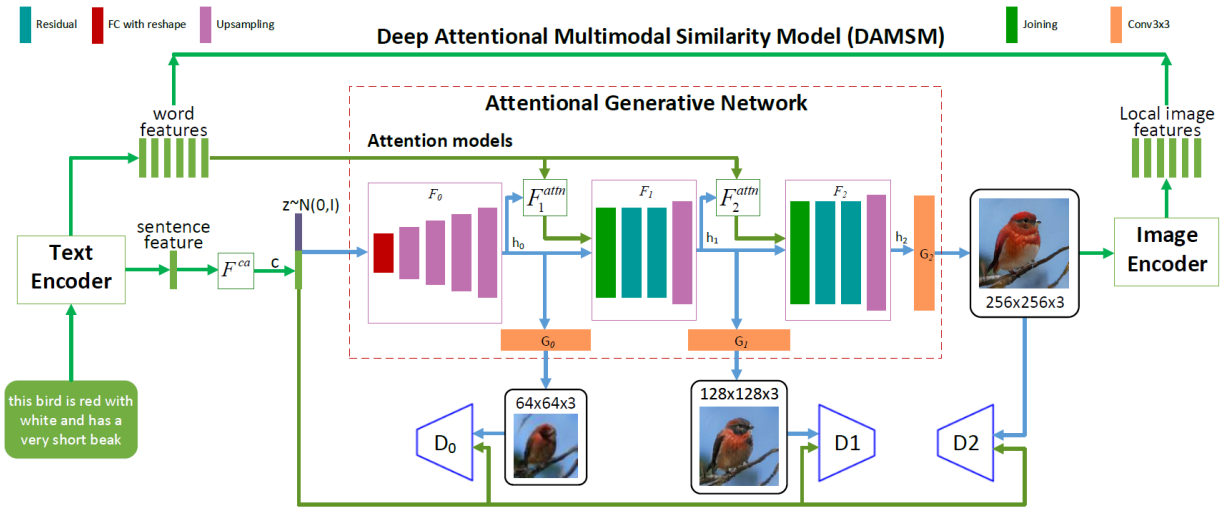

Figure 7: Architecture of AttnGAN, as illustrated in [6]

The architecture of AttnGAN, as depicted in Figure 7, closely mirrors the StackGAN-V2 framework. In addition, AttnGAN incorporates an attention network (F_{attn}) within the generator and the deep attentional multimodal similarity model (DAMSM). AttnGAN's design allows for incorporating "m" generators, denoted by ($G_0;G_1;..;G_{m-1}$) notation in the diagram. Text descriptions are processed in two distinct ways: as a unified global sentence vector (represented as a sentence feature in the diagram) and a matrix of word vectors (word features).

In the initial stage, the first generator inputs the sentence features concatenated with noise. Subsequent generators, however, use inputs created by combining the output from the preceding generator with the output from the attention network. The attention network (F_{attn}) takes in word features and the input from previous generators, directing the generator's focus to relevant words, thus generating word-context vectors. These vectors, combined with image features, enable the generation of higher-resolution details in subsequent stages.

The research conducted by the authors explores two configurations: one with a single attention network (equivalent to two generators) and another with two attention networks (equivalent to three generators, as illustrated in the diagram). Upsampling within the generator is achieved using nearest neighbour interpolation with a scaling factor of 2. The components F_{ca}, F_{attn}, F_i, and G_i are implemented as neural networks.

Each generator is accompanied by a corresponding discriminator, denoted as D_0, D_1, and D_2 in the diagram, which is validated against sentence features rather than word features or the noise vector. Discriminators operate in two modes: unconditional and conditional. The unconditional form follows the typical adversarial network setup, classifying images as real or fake. In contrast, the conditional form assesses whether an image and its caption form a matched pair or not.

Given that each generator produces images at different resolutions across multiple levels, the objective function of the attentional generative network is formulated to incorporate outputs from all generators. This objective function is visually represented in Figure 8.

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

Figure 8: AttnGAN Objective Function

The hyperparameter λ plays a crucial role in balancing the terms within the equation, contributing to the loss function L_G of each generator. While a generator may excel in generating high-resolution images, it may not necessarily align with the provided text description. Hence, the Deep Attentional Multimodal Similarity Model (DAMSM) plays a pivotal role in ensuring the coherence between the generated image and the text input. DAMSM achieves this by computing a fine-grained image-text matching loss, employing an attention mechanism to map image sub-regions and words into a shared semantic space and measuring their similarity.

This additional loss term provides word-level supervision during the generator training, enhancing the model's ability to generate images that accurately correspond to the given text. DAMSM receives inputs from two sources: word features extracted from the text and the final image generated. The text encoder utilised in DAMSM is a bi-directional Long Short-Term Memory (LSTM) network, adept at extracting semantic vectors from textual descriptions. Within the bi-directional LSTM, each word corresponds to two hidden states, one for each direction. Concatenating these two hidden states forms a representation of the word's semantic meaning.

On the other hand, the image encoder in DAMSM is a Convolutional Neural Network (CNN) responsible for mapping images to semantic vectors. The CNN's intermediate layers specialise in learning local features of different sub-regions within the image, while the later layers focus on capturing global features. Specifically, the image encoder is constructed based on the Inception-

v3 model, pre-trained on the ImageNet dataset, ensuring a robust representation of image semantics.

The experimental setup entailed utilising two distinct datasets, namely CUB and COCO while employing appropriate evaluation metrics such as inception score and R-precision to gauge the efficacy of the proposed methodology. Notably, the inclusion of R-precision as an evaluation metric is unique to this research, distinguishing it from prior studies within this domain. The comparative analysis of these datasets is summarised in Table 1.

Dataset	CUB		COCO	
	Train	Test	Train	Test
No of Samples	8,855	2,933	80K	40K
Captions per image	10	10	5	5

Table 1: Statistics of datasets

Table 1 reveals that while the CUB dataset comprises a smaller number of images compared to COCO, it excels in text classification for each image, achieving a score of up to 10, whereas COCO achieves a score of 5. This distinction underscores the strengths and nuances of each dataset.

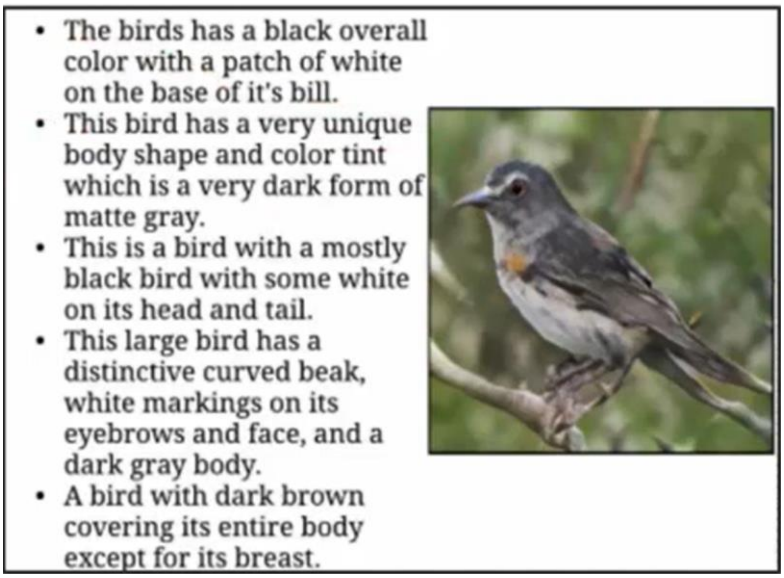

Figure 9: Sample from CUB dataset

Figure 9 showcases a sample of the data from the CUB dataset, highlighting its detailed and lengthy descriptions coupled with focused images devoid of clutter. The CUB dataset is inherently object-specific, revolving around birds, despite the overarching goal of this research being the development of a model capable of generating various types of objects. The rationale behind starting with a specific dataset like CUB is to establish foundational learning and subsequently transition towards more intricate datasets.

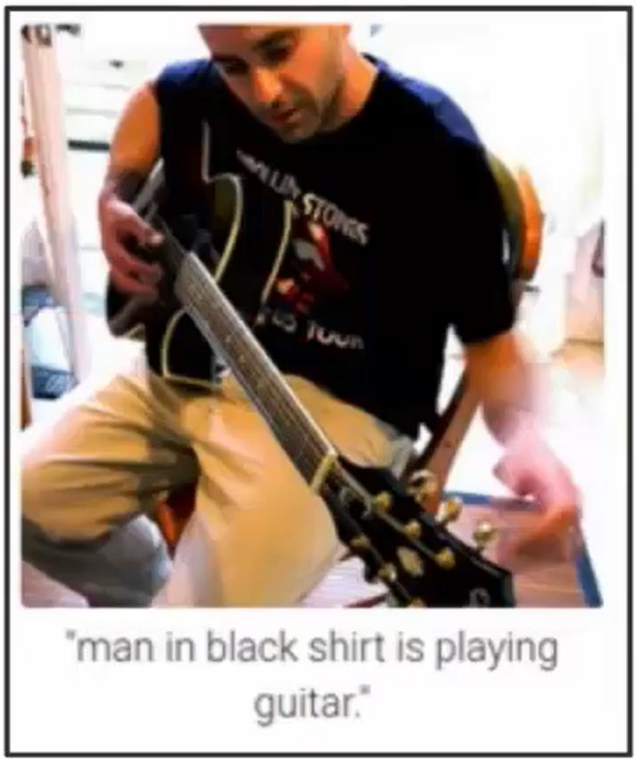

Figure 10: Sample of COCO dataset

In contrast, Figure 10 illustrates examples from the COCO dataset, characterised by complex object settings and comparatively shorter captions when compared with the CUB dataset. This contrast emphasises the varying complexities and characteristics inherent in different datasets, informing the approach taken in this research to progressively navigate through increasingly complex datasets.

4. Results and Analysis

The performance evaluation of AttnGAN spans multiple levels, encompassing a component-level quantitative analysis summarised in Table 2. The DAMSM aims to keep generated image relevant

to description and aim of Attntional GAN is to add details to the image in each stage. AttnGAN1 features a single attention network, whereas AttnGAN2 incorporates two attention models, aligning with the architecture diagram Figure 7. AttnGAN1 go through testing with varying λ values, as depicted in the table, with the optimal value identified as 5 for evaluating AttnGAN2. The inception score indicates how crisp the images are, and R-precision provides a matrix showing how similar the caption and image are.

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

Table 2: The best inception score and the corresponding R precision

This comparative analysis underscores the significance of appropriately weighting the L_{DAMSM} component, which contributes to generating higher-quality images that are more closely aligned with the given text descriptions. This improvement is attributed to the fine-grained image-text matching loss L_{DAMSM}, which provides additional supervision, specifically at the word level, during the generator training.

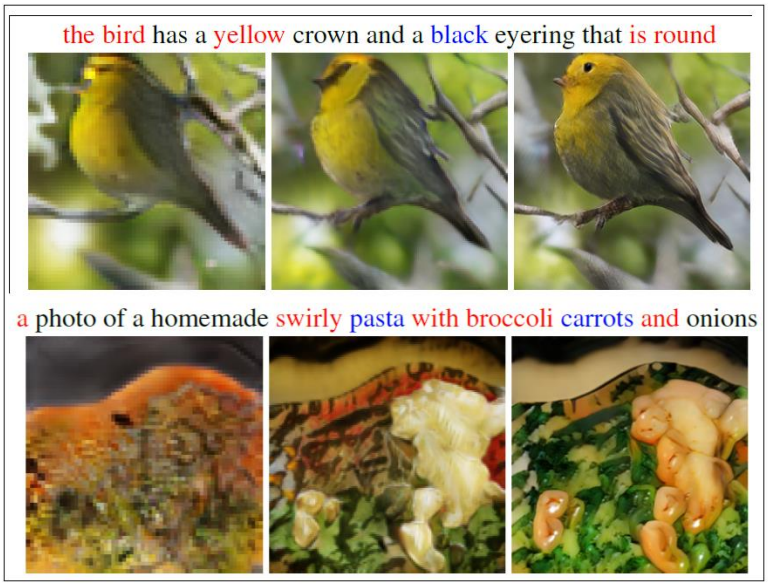

Figure 11: Intermediate results of our AttnGAN

The finalised model configurations for CUB and COCO datasets are denoted as "AttnGAN2, λ =5" and "AttnGAN2, λ =50," respectively. The higher λ value for the COCO dataset compared to CUB suggests that L_{DAMSM} is particularly crucial for generating complex scenarios, as observed in the COCO dataset. To gain deeper insights into the learning process of AttnGAN, refer to Figure 11, which illustrates how the initial generator sketches primitive shapes and colours, generating low-resolution images due to the utilisation of global sentence vectors in the first stage. Subsequent stages, specifically generators 2 and 3, refine these results by leveraging word vectors to rectify any shortcomings from the previous stage.

Dataset	GAN-INT-CLS	GAWWN	StackGAN	StackGAN-V2	PPGN	AttnGAN
CUB	2.88 ± 0.04	3.62 ± 0.04	3.70 ± 0.04	3.82 ± 0.06	/	**4.36 ± 0.03**
COCO	7.88 ± 0.07	/	8.45 ± 0.03	/	9.58 ± 0.21	**25.89 ± 0.49**

Table 3: Inception scores by state-of-the-art GAN models

Comparative assessments against previous methods reveal that AttnGAN significantly outperforms previous state-of-the-art GAN models. Table 3 visualises inception scores captured for each model under the CUB and COCO datasets, yielding a substantial boost in the best-reported inception score by 14.14% on the CUB dataset and an impressive 170.25% on the more challenging COCO dataset. Increasing the weight of the DAMSM loss further enhances R-precision rates, underscoring the efficacy of the fine-grained image-text matching loss in improving the alignment between generated images and text descriptions.

5. Conclusion and Future Works

The paper presents a significant contribution to the realm of text-to-image synthesis and multimodal learning through the introduction of AttnGAN. The method's contribution can be delineated into three key aspects. Firstly, it proposes a novel Attentional Generative Adversarial Network comprising two innovative components: the attentional generative network and DAMSM. Secondly, it empirically evaluates the proposed AttnGAN architecture, showcasing its substantial improvement over previous state-of-the-art GAN models. Thirdly, it conducts a detailed analysis of the attention layers within AttnGAN, illustrating how the layered conditional GAN autonomously attends to relevant words to shape the conditions for image generation.

To address the identified limitations, such as resource constraints, future work could explore leveraging online platforms like Google Colab to introduce additional generators with Attention networks. Moreover, alternative architectures or techniques could be explored to enhance the model's capacity to capture cohesive global structures while retaining finesse in generating fine-grained details. This avenue of research could lead to further advancements in the field of text-to-image synthesis and multimodal learning.

6. Strength and Weaknesses

The strengths of the paper include the novel attention mechanism, the improved performance of complex datasets (especially COCO), and the ability to capture fine-grained details through attention-driven image-text matching loss. Potential weaknesses include the lack of global coherence in some generated images and the computational complexity of the model, which limited the exploration of architectures with more than two attention models.

7. References

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran REEDSCOT, B. Schiele, and H. Lee SCHIELE, "Generative Adversarial Text to Image Synthesis." PMLR, pp. 1060–1069, Jun. 11, 2016. Accessed: Apr. 06, 2024. [Online]. Available: https://proceedings.mlr.press/v48/reed16.html

[2] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning What and Where to Draw," *Adv Neural Inf Process Syst*, vol. 29, 2016.

[3] H. Zhang *et al.*, "StackGAN: Text to Photo-Realistic Image Synthesis With Stacked Generative Adversarial Networks." pp. 5907–5915, 2017. Accessed: Apr. 06, 2024. [Online]. Available: https://github.com/hanzhanggit/StackGAN.

[4] H. Zhang *et al.*, "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019, doi: 10.1109/TPAMI.2018.2856256.

[5] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space." pp. 4467–4477, 2017. Accessed: Apr. 06, 2024. [Online]. Available: http://EvolvingAI.org/ppgn.

[6] T. Xu *et al.*, "AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks." pp. 1316–1324, 2018. Accessed: Apr. 07, 2024. [Online]. Available: https://github.com/taoxugit/AttnGAN.