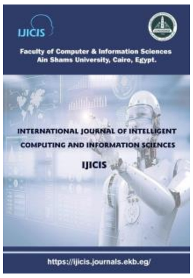
IJICIS, Vol.22, No.3, 82-91
DOI: 10.21608/ijicis.2022.117124.1157

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/


IJICIS
Faculty of Computer & Information Sciences Ain Shams University, Cairo, Egypt.
INTERNATIONAL JOURNAL OF INTELLIGENT COMPUTING AND INFORMATION SCIENCES
IJICIS
https://ijicis.journals.ekb.eg/

MULTI-STAGE HYBRID TEXT-TO-IMAGE GENERATION MODEL

Razan Bayoumi*
Computer Science Department,
Faculty of Computer and
Information Sciences, Ain Shams
University,
Cairo, Egypt
razan.bayoumi@cis.asu.edu.eg

Marco Alfonse
Computer Science Department,
Faculty of Computer and
Information Sciences, Ain Shams
University,
Cairo, Egypt
marco_alfonse@cis.asu.edu.eg

Abdel-Badeeh M. Salem
Computer Science Department,
Faculty of Computer and
Information Sciences,Ain Shams
University,
Cairo, Egypt
abSalem@cis.asu.edu.eg

Received 2021-01-18; Revised 2022-06-02; Accepted 2022-06-09

Abstract

Generative Adversarial Networks (GANs) have proven their outstanding potential in creating realistic images that can't differentiate between them and the real images, but text-to-image (conditional generation) still faces some challenges. In this paper, we propose a new model called (AttnDM GAN) stands for Attentional Dynamic Memory Generative Adversarial Memory, which seeks to generate realistic output semantically harmonious with an input text description. AttnDM GAN is a three-stage hybrid model of the Attentional Generative Adversarial Network (AttnGAN) and the Dynamic Memory Generative Adversarial Network (DM-GAN), the 1st stage is called the Initial Image Generation, in which low resolution 64x64 images are generated conditioned on the encoded input textual description. The 2nd stage is the Attention Image Generation stage that generates higher-resolution images 128x128, and the last stage is Dynamic Memory Based Image Refinement that refines the images to 256x256 resolution images. We conduct an experiment on our model the AttnDM GAN using the Caltech-UCSD Birds 200 dataset and evaluate it using the Frechet Inception Distance (FID) with a value of 19.78. We also proposed another model called Dynamic Memory Attention Generative Adversarial Networks (DMAttn-GAN) which considered a variation of the AttnDM GAN model, where the second and the third stages are switched together, its FID value is 17.04.

Keywords*— *Generative Adversarial Networks, Image Generation, Text-to-image, Computer Vision, Conditional Image Synthesis

* Corresponding author: Razan Bayoumi
Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
E-mail address: razan.bayoumi@cis.asu.edu.eg

1. Introduction

In the last few years, there has been great growth in the utilization of Generative Adversarial Networks (GANs) in different applications, and it has been proved the pioneering of the quality and accuracy of their output in generating photos and videos. The GAN is a neural network that can be internally considered as two neural networks, one of them is the generative network that is responsible for generating the new samples while the other one is the discriminative network. They compete with each other in the training phase, so each network becomes better in its prediction [1]. The generative network is trained to generate realistic visual output that the discriminative network can't distinguish from the real images (dataset) until they reach the nash equilibrium point. They are trained to generate new images that have the statistics close as possible to the real training dataset samples. Text-to-Image refers to generating high-quality images that match the input text description. Generating realistic images conditioned on text description has tremendous applications just as photos editing, computer designing, crime scene investigation to recognize criminals, etc. Transforming the textual descriptions into image pixels which requires a comprehensive understanding of the relationship between the two latent spaces is considered challenging because of the difference between the nature of both features. In this paper, we proposed a new model called AttnDM-GAN which is a hybrid model of the AttnGAN [2] and the DM-GAN [3]. The AttnGAN [2] is a multistage architecture that uses an attention layer to pay attention to the relevance between the input description words and sub-regions of the generated images. The DM-GAN [3] consist of 2 phases; initial image generation and dynamic memory-based image refinement that can be any number of layers, but the authors in [3] introduced their architecture with two stages of this layer. The proposed AttnDM-GAN model is a multi-stage model, where the **attentional image generation stage** is from the AttnGAN**,** and the **dynamic memory-based image refinement stage** is from the DM-GAN. There is also another proposed model called DMAttn-GAN which is a variation of AttnDM-GAN. AttnDM-GAN and DMAttn-GAN aim to generate high-resolution images based on input description.

This paper is organized as follows, section 2 overviews the related work, section 3 presents the proposed methodology and architecture, section 4 mentions the implementation details, section 5 shows the experiments and datasets details, and section 6 presents the conclusion and future work.

2. Related work

The GANs have made a revolution in generating images that look so realistic to a level that people can't differentiate between the real ones and the generated ones. Goodfellow et al. [1] introduced the GAN that consists of 2 networks and used it to generate random images that have the same statistics as the training dataset. Then great progress has been done in the direction of the conditional generation, whether the generation was based on label, text, etc. There is a variety of proposed work that achieved remarkable results focusing specifically text-to-image generation. Text-to-image synthesis models aim to achieve one or more of these goals: to generate high-resolution images, to generate images that are highly consist with the input textual description, or to generate diverse images consistent with the input text. The first text-to-image model was introduced in 2016 by Reed at al. [4], who is considered the pioneer in this domain. The model architecture mainly consists of the text encoder which is the Char-CNN-RNN network, and the image decoder is like the Deep Convolutional Generative Adversarial Network (DCGAN). There were lots of progress after that in this area and new models and architectures have been designed, the architecture starts to be multistage which means it consists of multiple generative and discriminative networks work in incremental fashion in order to improve the quality of the generated images. Zhang et al. [5] introduced this idea and proposed two multistage models StackGAN and StackGAN-v2. StackGAN generates the images in two incremental stages, the 1st stage generates an image with the outline and basic shape and color of the object, then in the 2nd stage more fine details are added and higher resolution image is produced. StackGAN-v2 has tree-like structure that generates the images in incremental fashion of 3 stages from low resolution to high resolution. Xu et al. [2] proposed AttnGAN which was a leap because it introduced an attention mechanism that pays attention to the relation between the input description words and the most relevant image subregions instead of generating the image based on the global embedding sentence which cause missing some fine details. There are some researchers have built architectures based on other different ideas. Dong et al. [6] and Qiao et al. [7] used the idea of bidirectional generation for building and training their model, in which the generated images are fed as input to another network that generate textual descriptions of the input image that represents their content. Singh et al. [8] and Qiao at al. [9] built their models as a simulation to the human learning process, at first people learn the colors, shapes and textures then start drawing any pictures incrementally by focusing on object by object until the images are fully drawn and finished.

3. Methodology

The proposed Attentional Dynamic Memory model (AttnDM-GAN) is shown in Figure. 1. The model consists of two modules: the generation network and the Deep Attentional Multimodal Similarity Model (DAMSM) [2]. The images are generated in a multi-stage fashion, where the generative network consists of 3 incremental stages. The first stage is the **initial image generation** that synthesis 64x64 images, the second stage is the **attentional image generation** that improves the images' resolution and generates 128x128 images, and the final stage is the **dynamic memory-based image refinement** the generates high-resolution 256x256 images. DAMSM is trained to map each sentence's word to the corresponding sub-image into a common semantic space in order to increase the correlation between the input text and the generated image. DAMSM consists of 2 networks: text encoder and image encoder; the text encoder is a bi-directional Long Short-Term Memory (LSTM) [10] that extracts text features into a semantic vector, the image encoder is a Convolutional Neural Network (CNN) that extracts the sub-region images' local features and the images' global features as well.

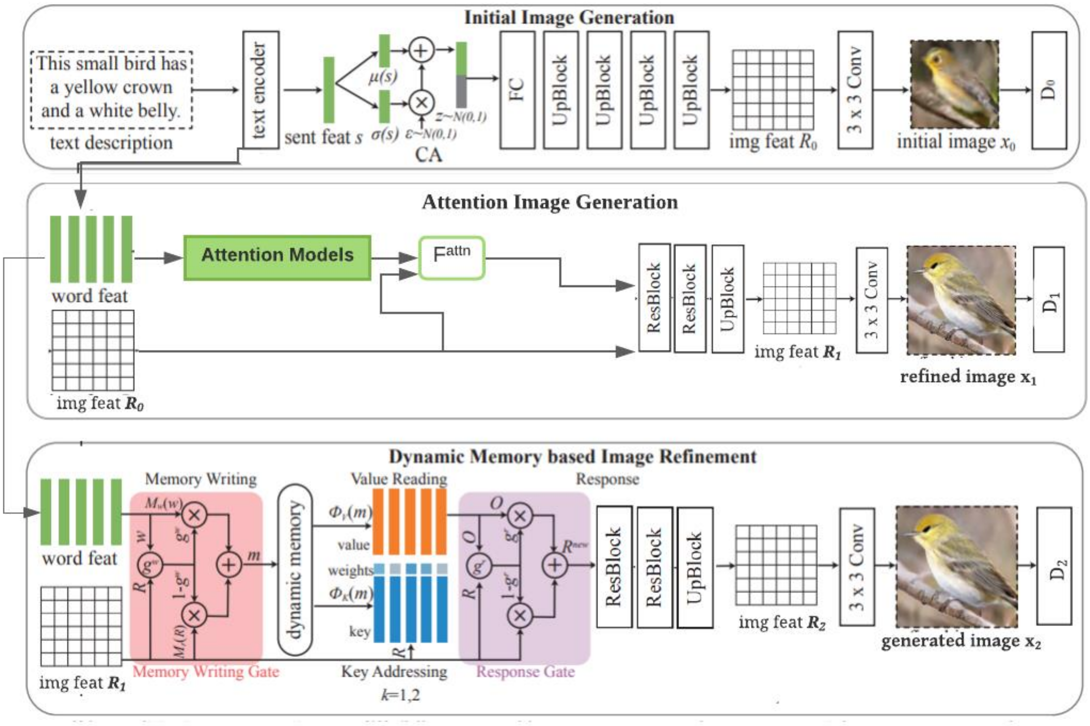


Figure. 1 The model architecture for text-to-image synthesis. It generates images through 3 incremental stages: initial image generation, attention image generation, and dynamic memory-based image refinement (adapted from *[3]*).

At the initial image generation stage, the input text is encoded to extract the global sentence features vector s and the local word features matrix. The global sentence features vector s is resampled by applying Conditional Augmentation (CA) and concatenated to a generated normal distribution noise vector to generate a low resolution 64x64 image. At the attention image generation stage, the input is the words features matrix w and the image features R_0 that is extracted from the previous stage's generated image. The word features matrix is converted to a common semantic of the image features then passed to the attention model $F^{attn}(w, R_0)$. Each column in matrix R_0 represents a sub-region of the image, then for each sub-region, a word context vector is computed. To generate the 128x128 images, both the image features and their corresponding context word vectors are combined together. At the third stage, dynamic memory-based image refinement, the input is the same as the previous stage; the word features matrix w and the previous image features R_1. This stage is based on the dynamic memory that consists of 4 modules: Memory Writing, Key Addressing, Value Reading, and Response. The **Memory Writing** operations store the text information into the memory in a key-value mapping fashion. In dynamic memory, encoding the previous knowledge is an important step to allow recovering high-quality images. In the **Key Addressing**, we retrieve the most relevant information from the memory based on the probability between a memory slot m_i and an image feature r_j. At **Value Reading**, the output is calculated according to the similarity probability and based on the weighted summation of the memory values. At the **Response**, the output is the memory representation output concatenated to the current image to generate the new image features.

The objective function L to generate realistic images is defined as:

$$L = L_G + \lambda_1 L_{CA} + \lambda_2 L_{DAMSM}, \; where \; L_G = \sum_{i=0}^{3} L_{G_i}, \qquad (1)$$

Here, λ_1 and λ_2 are the hyperparameters that are corresponding to the conditioning augmentation loss and DAMSM loss respectively. The discriminators D$_i$ are trained to categorize the generated images as real or fake by minimizing the cross-entropy loss L_{D_i}:

$$L_{D_i} = -\frac{1}{2}[E_{x_i \sim P\,data_i} \, logD_i\,(x_i) + E_{\hat{x}_i \sim PG_i} \, log\big(1 - D_i\,(\hat{x}_i)\big) + \, E_{x_i \sim P\,data_i} \, logD_i(x_i, s)\big] + E_{\hat{x}_i \sim PG_i} \, log\big(1 - D_i(\hat{x}_i, s)\big), \qquad (2)$$

where $\hat{x}_i$ is from the model generated distribution PG_i, and x_i is from the training image distribution $P\,data_i$ at the i^{th} level. The first half of the equation represents the unconditional loss that distinguishes between the real images and the generated images, while the second half is the conditional loss that distinguishes between the matched text-image pairs and the unmatched ones. On the other hand, the adversarial loss L_{G_i} for G_i is defined as:

$$L_{G_i} = -\frac{1}{2}[E_{\hat{x}_i \sim PG_i} \, logD_i\,(\hat{x}_i) + E_{\hat{x}_i \sim PG_i} \, logD_i(\hat{x}_i, s)], \qquad (3)$$

where the first term is the unconditional term that tries to make the generated images as close as possible like the real images, while the second term is the conditional term that makes the generated images match the input text description.

We also proposed another model which is a variation of AttnDM-GAN, we switched the second stage (Attention Image Generation) and the third stage (Dynamic Memory-based Image Refinement) as shown in Figure. 2. We called this proposed model as DMAttn-GAN stands to Dynamic Memory Attention Generative Adversarial Networks. It has been proved qualitatively and quantitively that DMAttn-GAN is an improved version of AttnDM-GAN.

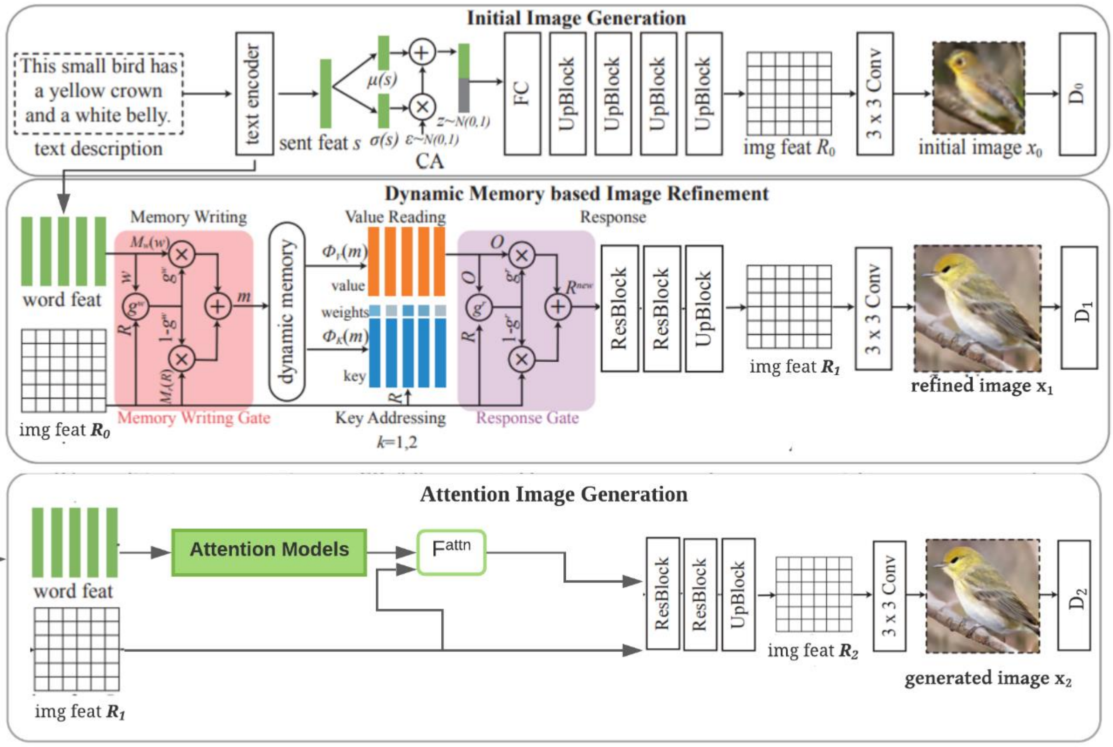


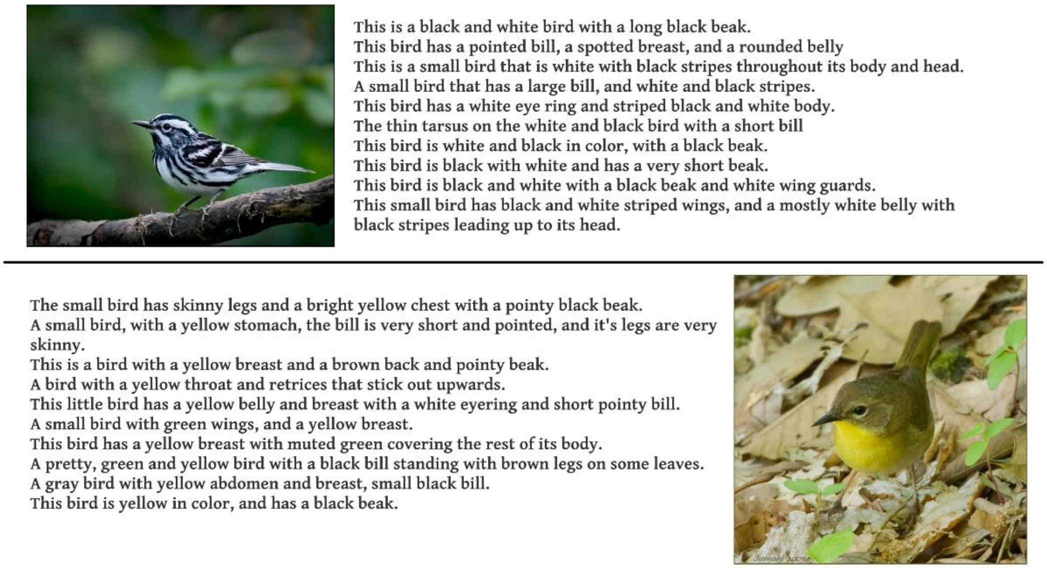
Figure. 2 The model architecture for text-to-image synthesis. It generates images through 3 incremental stages: initial image generation, dynamic memory-based image refinement, and attention image generation (adapted from *[3]*).

4. Implementation and Experiments

We used a pre-trained bidirectional LSTM text encoder for the text embedding [2]. The input text is encoded as words (vector for each word) and as a whole sentence. The first stage generates low-resolution 64x64 images. Then the attention image generation stage upgrades the images to 128x128. Then the final stage, dynamic memory-based image refinement, refine the images to generate high-quality images of 256x256 resolution. We trained AttnDM-GAN and DMAttn-GAN model with 504 epochs on CUB dataset with a learning rate of 0.0002. We set the hyperparameter $\lambda_1 = 1$ and $\lambda_2 = 5$ and used the ADAM optimizer [11] with a batch size of 10, β1 = 0.5 and β2 = 0.999. The dimension of the text vector (N_w) is 256, while the image vector (N_r) is 64 and the memory feature vector (N_m) is 128.

4.1 Dataset

We evaluated our proposed models on the CUB [12] dataset to demonstrate their capability. The CUB dataset has 200 different species, with 11,788 images, they are split into 8,855 images for training and 2,933 images for testing. Each image in the dataset has 10 different descriptions. A sample example of the CUB dataset is shown in Figure. 3.



This is a black and white bird with a long black beak.
This bird has a pointed bill, a spotted breast, and a rounded belly
This is a small bird that is white with black stripes throughout its body and head.
A small bird that has a large bill, and white and black stripes.
This bird has a white eye ring and striped black and white body.
The thin tarsus on the white and black bird with a short bill
This bird is white and black in color, with a black beak.
This bird is black with white and has a very short beak.
This bird is black and white with a black beak and white wing guards.
This small bird has black and white striped wings, and a mostly white belly with black stripes leading up to its head.

The small bird has skinny legs and a bright yellow chest with a pointy black beak.
A small bird, with a yellow stomach, the bill is very short and pointed, and it's legs are very skinny.
This is a bird with a yellow breast and a brown back and pointy beak.
A bird with a yellow throat and retrices that stick out upwards.
This little bird has a yellow belly and breast with a white eyering and short pointy bill.
A small bird with green wings, and a yellow breast.
This bird has a yellow breast with muted green covering the rest of its body.
A pretty, green and yellow bird with a black bill standing with brown legs on some leaves.
A gray bird with yellow abdomen and breast, small black bill.
This bird is yellow in color, and has a black beak.

Figure. 3 Sample of CUB dataset.

4.2 Evaluation

We evaluated AttnDM-GAN and DMAttn-GAN quantitively and qualitatively. We quantify our models quantitively in terms of Frechet Inception Distance (FID) [13] which is considered an upgrade version of the Inception Score (IS) [14]. The IS evaluates the model by using the pre-trained Inception v3 network [15] to measure how distinct each generated image is and the variety of the generated images. The higher IS, the better and more distinct the generated images are. The FID measures the performance of the model by calculating the distance between the generated images and the real images depending on the statistics (mean and covariance) of them, instead of depending only on the generated images. The lower FID, the closer the generated and real images are. The FID equation is defined as:

$$d^2(x,g) = ||\mu_x - \mu_g||_2^2 + Tr\left(\Sigma_x + \Sigma_g - 2\left(\Sigma_x\Sigma_g\right)^{\frac{1}{2}}\right), \qquad (4)$$

We compare AttnDM-GAN and DMAttn-GAN to the state-of-the-art models; AttnGAN and DM-GAN on the CUB dataset and the results are shown in Figure 4.

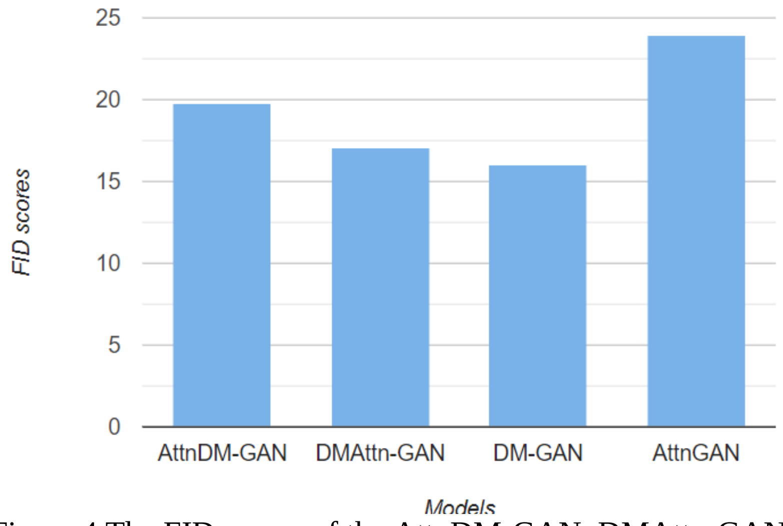


Figure 4 The FID scores of the AttnDM-GAN, DMAttn-GAN, DM-GAN and AttnGAN.

According to the FID scores presented in table 1, the AttnDM-GAN and DMAttn-GAN are better than the AttnGAN.

Table 1 The FID scores of the AttnDM-GAN, DMAttn-GAN, DM-GAN and AttnGAN

	FID ↓
AttnDM-GAN	19.78
DMAttn-GAN	17.04
DM-GAN [3]	**16.09**
AttnGAN [2]	23.98

For the qualitative evaluation, Figure 5 shows examples of the generated images from AttnDM-GAN and DMAttn-GAN and the state-of-the-art models AttnGAN and DM-GAN. The images are generated from new unseen text descriptions. Generally, all the synthesized images are high quality and meet the input description, but AttnDM GAN is clearer with vivid details. Our models learn to understand the textual and image features and bridge the gap between them. It learns to understand the semantic details in the input text description and reflects it in the output by generating accurate images with a main highlighted object (bird) that represents the textual mentioned features.

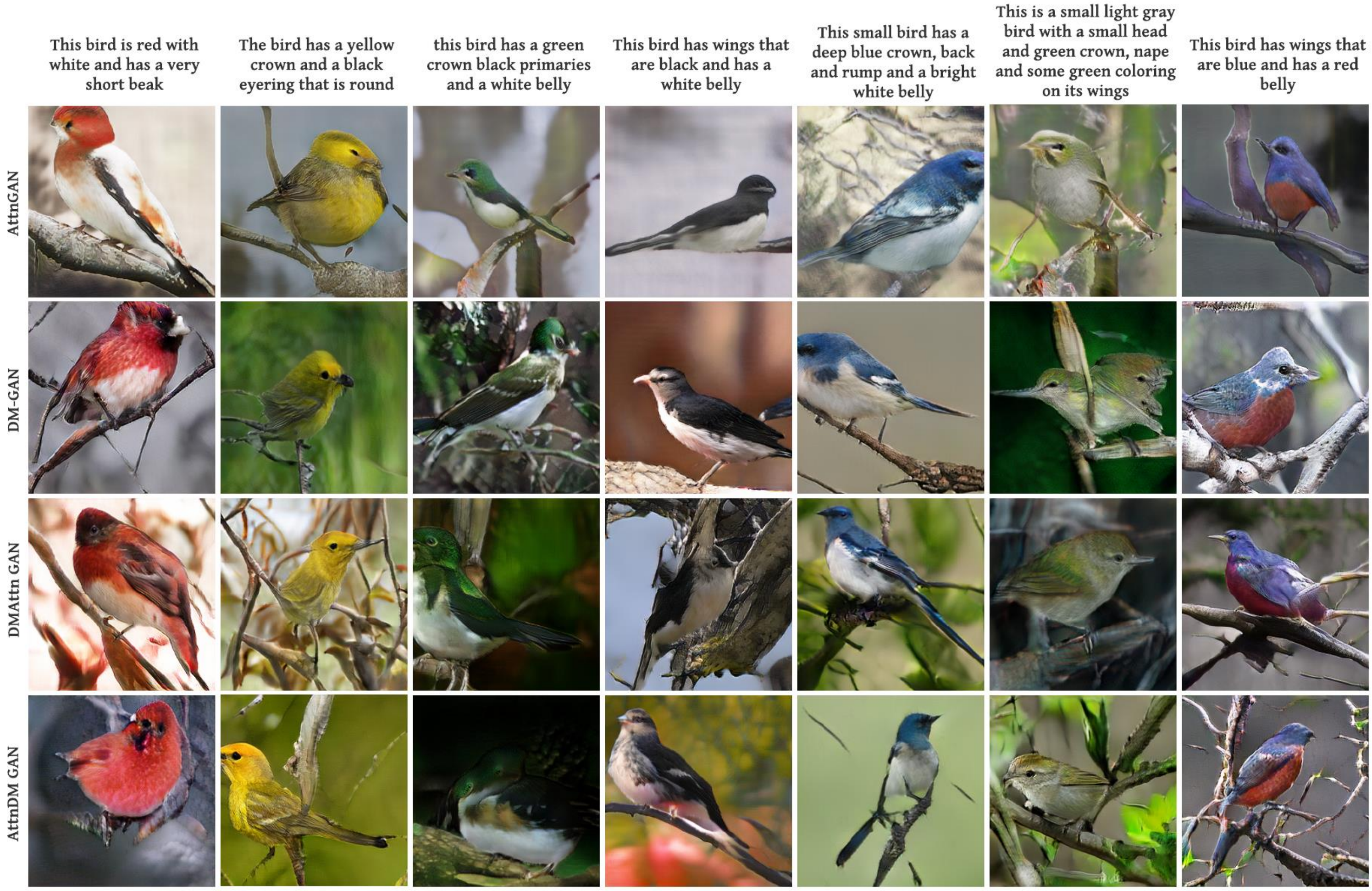

This bird is red with white and has a very short beak
The bird has a yellow crown and a black eyering that is round
this bird has a green crown black primaries and a white belly
This bird has wings that are black and has a white belly
This small bird has a deep blue crown, back and rump and a bright white belly
This is a small light gray bird with a small head and green crown, nape and some green coloring on its wings
This bird has wings that are blue and has a red belly
AttnGAN
DM-GAN
DMAttn GAN
AttnDM GAN

Figure 5 Generated bird images by conditioning on text from CUB dataset.

By observing the samples in Figure 5, the output of the 5th column is better in AttnDM-GAN and DMAttn-GAN than the others, as there is part of the bird is cut out of the image. The output of the 6th column is better too as the images are clearer and the bird is bolder and more correctly structured.

5. Conclusion and future work

In this paper, we proposed two architectures; AttnDM-GAN and DMAttn-GAN, which are hybrid models of both AttnGAN and DM-GAN. AttnDM-GAN consists of 3 stages, the first stage is called the Initial Image Generation, in which a low resolution 64x64 images are generated based on the encoded input text. The second stage is the Attention Image Generation stage, which generates higher-resolution images 128x128, and the last stage is Dynamic Memory Based Image Refinement that refines the images to 256x256 resolution images. DMAttn-GAN is the same as AttnDM-GAN but the second and the third stages are switched. The qualitative and quantitative evaluation demonstrates the effectiveness of the model. In the future work, we will work on generalizing the model use and try to apply it to more complex datasets to generate better and more accurate images with a wide variety.

References

[1] Goodfellow, Ian, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and a. Y. Bengio, "Generative adversarial nets.," in *Advances in neural information processing systems 27*, 2014.

[2] Xu, Tao, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang and X. He, "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.

[3] Zhu, Minfeng, P. Pan, W. Chen and Y. Yang., "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.

[4] Reed, S. E., Z. Akata, S. Mohan, S. Tenka, B. Schiele and H. Lee., "Learning what and where to draw," in *Advances in neural information processing systems*, 2016.

[5] Zhang, Han, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence 41, no. 8,* 2018.

[6] Dong, Hao, J. Zhang, D. McIlwraith and Y. Guo, "I2T2I: Learning text to image synthesis with textual data augmentation," in *2017 IEEE International Conference on Image Processing (ICIP),*

2017.

[7] Qiao, Tingting, J. Zhang, D. Xu and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription.," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[8] Singh, Amanpreet and S. Agrawal., "CanvasGAN: A simple baseline for text to image generation by incrementally patching a canvas," in *Science and Information Conference, Springer*, 2019.

[9] Qiao, Tingting, J. Zhang, D. Xu and D. Tao., "Learn, imagine and create: Text-to-image generation from prior knowledge.," in *Advances in Neural Information Processing Systems*, 2019.

[10] Schuster, Mike and K. K. Paliwal., "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing,* vol. 45, no. 11, pp. 2673-2681, 1997.

[11] Kingma, D. P. and J. Ba., "Adam: A method for stochastic optimization.," in *arXiv preprint arXiv:1412.6980*, 2014.

[12] Wah, Catherine, S. Branson, P. Welinder, P. Perona and S. Belongie, "The caltech-ucsd birds-200-2011 dataset.," 2011.

[13] Heusel, Martin, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter., "GANs trained by a two time-scale update rule converge to a local nash equilibrium.," in *Advances in neural information processing systems*, 2017.

[14] Salimans, Tim, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen., "Improved techniques for training GANs.," in *Advances in neural information processing systems*, 2016.

[15] Szegedy, Christian, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna., "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.