Bulletin of Electrical Engineering and Informatics
Vol. 12, No. 2, April 2023, pp. 1180~1188
ISSN: 2302-9285, DOI: 10.11591/eei.v12i2.4199

Text-to-image generation based on AttnDM-GAN and DMAttn-GAN: applications and challenges

Razan Bayoumi[1], Marco Alfonse[1,2], Mohamed Roushdy[3], Abdel-Badeeh M. Salem[1]
[1]Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
[2]Laboratoire Interdisciplinaire de l'Université Française d'Égypte (UFEID LAB), Université Française d'Égypte, Cairo, Egypt
[3]Faculty of Computers and Information Technology, Future University in Egypt, New Cairo, Egypt

Article Info

Article history:

Received Jun 3, 2022
Revised Aug 14, 2022
Accepted Oct 2, 2022

Keywords:

Computer vision
Conditional image synthesis
Generative adversarial networks
Image generation
Text-to-face

ABSTRACT

The deep fake faces generation using generative adversarial networks (GANs) has reached an incredible level of realism where people can't differentiate the real from the fake. Text-to-face is a very challenging task compared to other text-to-image syntheses because of the detailed, precise, and complex nature of the human faces in addition to the textual description details. Providing an accurate realistic text-to-image model can be useful for many applications such as criminal identification where the model will be acting as the forensic artist. This paper presents text-to-image generation based on attention dynamic memory (AttnDM-GAN) and dynamic memory attention (DMAttn-GAN) that are applied to different datasets with an analysis that shows the different complexity of different datasets' categories, the quality of the datasets, and their effect on the results of the resolution and consistency of the generated images.

Corresponding Author:

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

1. INTRODUCTION

The deepfakes have been used widely in different domains as generating fake images that doesn't really exists but looks realistic as a real photographed image [1], or replacing faces of people in videos [2] which is considered incredible development yet dangerous. In the last few years, when the generative adversarial networks (GANs) have first introduced, they have shown very exciting results in the classification and generation fields [3], and they are considered a leap in this area. Text-to-image synthesis (the reverse of image captioning [4]) is considered an emerging domain where the generation process aims mainly to generate output images as real as possible which are consistent with the textual input description [5]–[10]. GAN is a deep neural network that consists of two competitive networks called generator and discriminator. The generative network is responsible for generating candidates as close as possible to the training dataset samples, while the other network is the discriminative network that evaluates the generated candidates and classifies them as real or fake images. In the training phase, both networks compete against each other, so each network becomes better in its task [11]. The aim of text-to-image is to generate high-quality images that match the input text description. It's needed to have a comprehensive understanding of the relationship between the latent spaces of text and the image to construct the image pixels from the input textual descriptions. This transformation is considered challenging because of the difference in the nature of both features. Even though text-to-image synthesis generally is a difficult field, text-to-face is more challenging and not well studied, and investigated because of the complexity of the human faces which leads to more obstacles and limitations in the generation process. Generating realistic images conditioned on text description has tremendous applications such as photo editing, computer designing, and crime scene investigation to recognize criminals.

The GANs are considered a leap in generating realistic images. The goal of the generation is that the images look so realistic to a level that people can't decide if the images are generated or real. Goodfellow *et al.* [11] are considered the pioneer who introduced the GAN architecture that consists of 2 networks called generator and discriminator. They used this architecture to generate images that have as close statistical distribution as possible to the training dataset. After that more attention has been directed to the image generation domain using GAN but with adding more conditions and categorization (such as label and text) to the generation process. There are many works that have been proposed that generate images based on English text descriptions with remarkable results. Generally, we can sum up the goals that the proposed works aim to achieve in the conditional generation process as: i) generating realistic high-resolution images; ii) generating images that are consistent with the input text description; and iii) generating manifold consistent images.

The first textual conditional model was introduced by Reed *et al.* [12]. Reed's architecture is composed of two components: i) the character-level convolutional neural network-recurrent neural network (Char-CNN-RNN) [13] text encoder network and ii) the deep convolutional generative adversarial network (DCGAN) [14] as image decoder. There is an intense effort and works pointed to this area and new different architectures were designed to improve the generated results. The architecture starts to be multistage which means it consists of multiple generative and discriminative networks that work in an incremental fashion in order to improve the quality of the generated images. Zhang *et al*. took the lead in this idea and presented multistage models that were called StackGAN [15] and StackGAN-v2 [16]. StackGAN model is built on the idea of generating the images incrementally in two steps. The 1st step generates the primitive basic features of the image's objects, then at the 2nd step the construction of the image starts to be clearer, and more fine details are added to produce a higher resolution image. StackGAN-v2 has a tree-like structure that generates the images incrementally at 3 stages from low resolution to higher resolution. According to Xu *et al.* [17] AttnGAN was a revolution in conditional image generation because of the introduction of the 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 causes missing some fine details.

There are some researchers who have built architectures based on other different ideas. According to Dong *et al*. [18] and Qiao *et al*. [19] used the idea of bidirectional generation for building and training their models, in which the generated images are fed as input to another network that generates textual descriptions of the input image that represents their content. According to Singh *et al*. [20] and Qiao *et al*. [21] built their models as a simulation of the human learning process, at first people learn the colors, shapes, and textures then start drawing a picture incrementally by focusing on an object by object until the images are fully drawn and finished.

The two models; dynamic memory attention (DMAttn-GAN) and attention dynamic memory (AttnDM-GAN) that will be discussed and experimented in this paper concentrate on generalization by working on different datasets categories as well as generating diverse realistic images. This paper presents the application of different text-to-image generations models (DMAttn-GAN and AttnDM-GAN) to different datasets (representing different categories) in addition to discussing the limitations and the impact of different datasets' categories (faces and birds) on the quality of the generated images. DMAttn-GAN and AttnDM-GAN are multistage based architecture that consists of 3 stages. The first stage is the same at both in which it generates a low-level simple image. The second and third stages are switched together to construct the two different models. More details will be discussed at the next section. This paper is organized as follows, section 2 presents the text-to-image generation models and their architectures, section 3 shows the experiments and datasets' details, and section 4 presents the conclusion and future work.

2. METHOD

This section presents two text-to-image generation models proposed by the authors. These models are hybrid models of two other generation models that already exist and are used widely by researchers. These models are called AttnDM-GAN and DMAttn-GAN. The AttnDM-GAN is a multistage architecture model designed for text-to-image synthesis tasks. AttnDM-GAN is a hybrid model of the AttnGAN [17] and the DM-GAN [22] as shown in Figure 1. The images are generated in an incremental fashion design. The model is composed of 3 stages: initial image generation, attention image generation, and dynamic memory-based image refinement.

The 1st stage-initial image generation-synthesizes a low-level 64×64 image that contains the primitives features and the outline of the image's object based on the input description. It is considered the seed image for the final produced image. The 2nd stage-attention image generation-generates an improved image with 128×128, by computing a word-context vector for each sub-region of the image. The 3rd stage-dynamic memory-based image refinement-generates the final high-resolution 256×256 image through four components: memory writing, key addressing, value reading, and response.

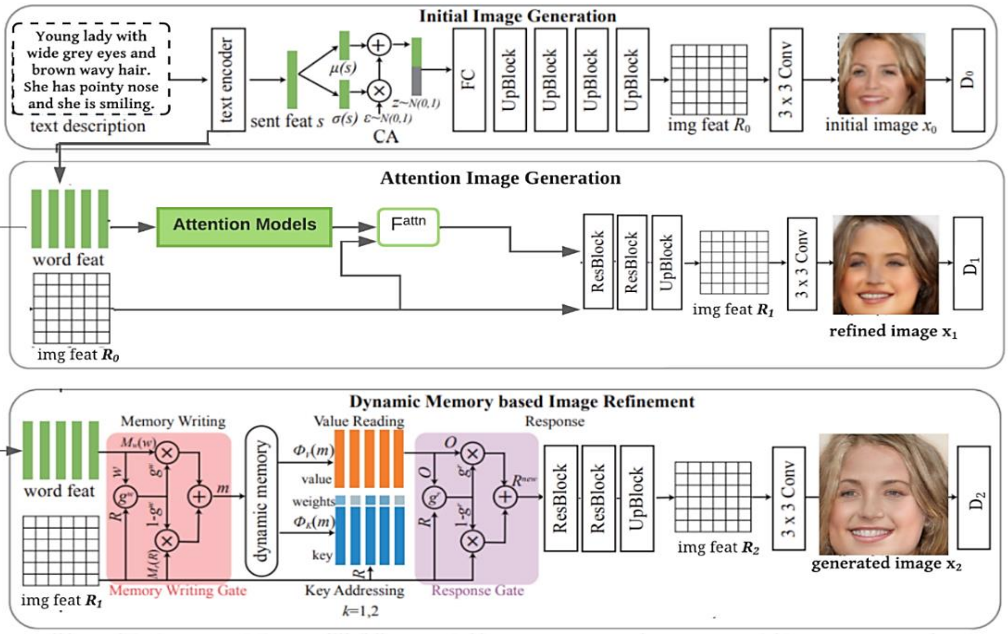


Figure 1. The AttnDM-GAN model architecture for text-to-image synthesis. It generates images through 3 incremental stages: initial image generation, attention image generation, and dynamic memory-based image refinement (adapted from [22])

The DMAttn-GAN is the same as the AttnDM-GAN but with the switch of the second stage (attention image generation) and the third stage (dynamic memory-based image refinement) as shown in Figure 2. In this model, the images are also generated through 3 incremental stages, from a low-level 64×64 image to the final high-resolution 256×256 image. The architectures' layers are discussed in more details in our previous work [23].

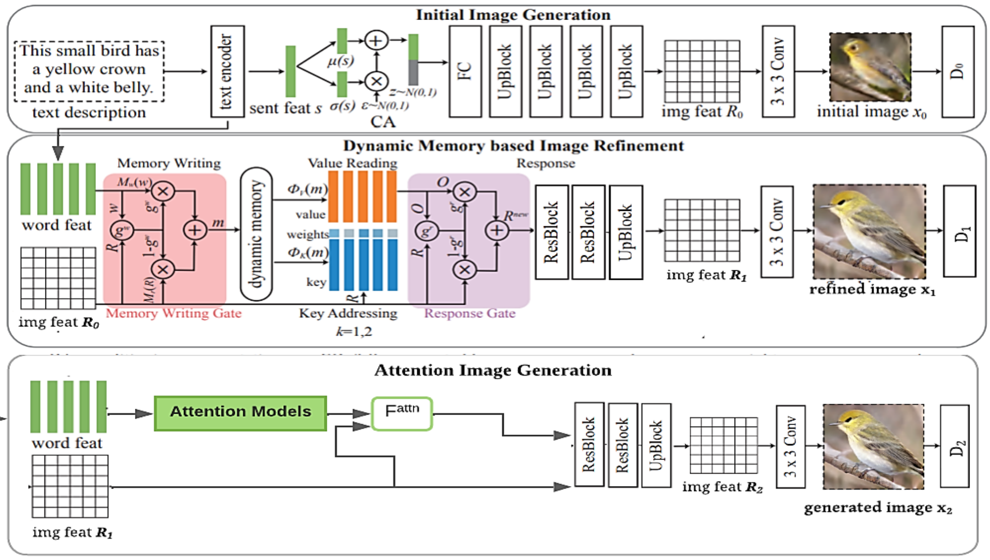


Figure 2. The DMAttn-GAN model architecture for text-to-image synthesis. It generates images through 3 incremental stages: initial image generation, dynamic memory-based image refinement, and attention image generation (adapted from [22])

3. RESULTS AND DISCUSSION

This section presents the datasets used in the experiments, evaluation metrics, and experimental setup, in addition to the analysis of the results obtained. The first subsection mentioned the datasets that were used in the experiments. The second subsection covers the used evaluation metrics. The last subsection presents the implementation details.

3.1. Datasets

In the experiments, three datasets were used which are Caltech-UCSD birds (CUB), CelebAText-HQ, and CelebAText. The CUB dataset [24] consists of different birds' categories. It has 200 different species, with 11,788 images, they are divided into 8,855 images for training and 2,933 images for testing. Each image in the dataset has 10 different descriptions.

The second dataset is CelebAText-HQ dataset [25]. It is a human faces dataset; each image is manually annotated with 10 different descriptions that cover the human different features without redundancy between the descriptions. The advantage of manual annotating is that the descriptions are diverse and natural. The third used dataset is the proposed dataset in [26], it is a human faces dataset and could be called as *CelebAText*. It consists of 10,000 images each with one text description. It is divided into 8,000 training and 2,000 testing. The descriptions are generated from the celebA binary annotations using a written script [27]. A sample of each dataset is shown in Figure 3.

Figure 3. An example of the 3 used datasets: CelebAText-HQ (upper), CUB (middle), and CelebAText (lower)

3.2. Evaluation metrics

The GANs, as mentioned before, consist of 2 models: the generator and the discriminator. They both train each other, in other words, there is no objective model to evaluate the generator model. Generally, we can evaluate the output of the generator by its quality but it's still an open issue and a point of research. One of the most common evaluation metrics for evaluating the quality of the generated images is the frechet inception distance (FID) [28]. The FID uses the Inception v3 network coding layer [29] to capture the image features of the real (training dataset) and the synthesized images (generated from the model). The FID evaluates the performance of the model by calculating the differences between the generated images and the ground truths. The distance is calculated based on their statistics (mean and covariance), instead of depending only on the generated images. The lower FID, the closer the generated and real images are.

3.3. Experimental setup

The AttnDM-GAN and DMAttn-GAN models were trained on the CelebAText-HQ, CelebAText, and CUB datasets over 600 epochs with the hyperparameters shown in Table 1. The Adam optimizer [30] is used to update network weights iteratively based in training data. The source code is available at https://github.com/RazanBayoumi/AttnDMGAN.

Table 1. The training hyper parameters' values

Hyper parameters	Value AttnDM-GAN/DMAttn-GAN
Learning rate	0.0002
λ_1	1
λ_2	50
β_1	0.5
β_2	0.999
Text vector (N_w)	256
Image vector (N_r)	64
Memory feature vector (N_m)	128

4. RESULTS AND ANALYSIS

We evaluated the AttnDM-GAN and DMAttn-GAN models using the FID metric [28] on the 3 previously mentioned datasets and the results are recorded in Table 2. From Table 2 we can conduct which model is better. If we check the values, it is obvious that the values of the two models on CelebAText dataset are lower than the values of the CelebAText-HQ dataset but, as shown in Figure 4, the output of CelebAText-HQ dataset is much better. To be able to use the value as a comparative value, it needs to be applied to the same dataset as it is mainly calculated by using the mean and covariance of the testing dataset, so to use the FID evaluation metrics and have a meaningful conclusion, we need to use it for comparison over the same dataset. Figure 4 shows some of the output samples of our two models over two different datasets, and as illustrated the output of the CelebAText-HQ dataset is much better qualitatively.

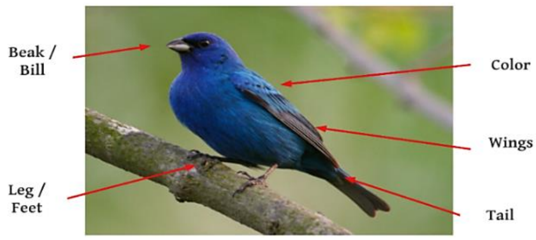
Table 2. The FID values of the AttnGAN, AttnDM-GAN and DMAttn-GAN models applied to different datasets

Dataset	AttnGAN	DMAttn-GAN	AttnDM-GAN
CUB	23.98 [22]	17.04	19.78
CelebAText	-	43.815	35.056
CelebAText-HQ	35.494	52.384	51.195

Figure 4. Qualitative results on CelebAText dataset (first two rows) and CelebAText-HQ (last two rows)

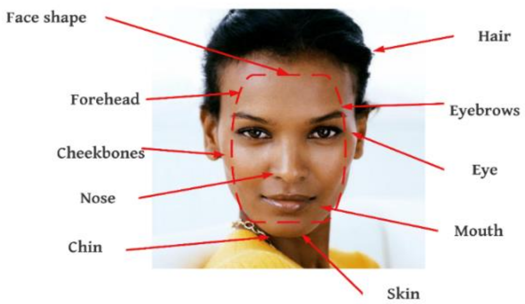
In the text-to-image synthesis domain, there are mainly different categories of datasets, this study focuses on two of them: birds and faces. The human faces are more detailed, fine-grained, and have vague properties that need to be covered in the descriptions unlike the descriptions of birds and flowers that are mainly related to color. Figures 5 and 6 present an example of bird attributes covered in the description in the CUB dataset and an example of human face attributes covered in the description in the CelebAText-HQ dataset respectively.

As shown in Figures 5 and 6, there is a difference between the number and the values of the attributes between the faces category and the birds' category. The nature of the face structure is more complex and has more features that have to be mentioned and described, unlike the birds which have much fewer details needed to be covered to have a fully completed description. Not only the number of attributes in the face structure is higher, but also the single component has much more features that should be mentioned, for example, the eyes: they have size, shape, color, eyebrows, and eyelashes. Based on this, the different proposed models applied to different categories, don’t achieve close performance and there is a significant difference as illustrated in Table 2. Another important point to consider is that the dataset’s properties, quality, and diversity affect the convergence process as well as the quality and realism of the generated images. In Figure 3, there is an example of CelebAText dataset and CelebAText-HQ dataset, and the details of the datasets are shown in Table 3.



Medium , bright blue bird body throughout, black wing edges and tail feather, black eyes.
This bird has a wide curved bill, a dark blue crown, and a medium blue throat, breast, and belly.
A medium sized bird with a blue belly, and a bill that curves downwards
A big blue bird with a large breast and a silver bill.
A medium sized bird that has tones of blue and purple with a medium sized bill
A bright blue bird with black wings and tail feathers and a black bill with a sharp point.
Gorgeous deep blue coloring with darker, almost black secondaries with a grayish colored pointed beak, an enlongated tail, and legs with a tint of blue.
The colorful bird has a blue back and small bill.
This bird has blue crown with a blue throat and black feet.
A small blue bird with black wings and a sharp bill.

Figure 5. Example of bird attributes covered in the description in the CUB dataset



She is a young lady with a melon seed face.
This lady has a narrow chin, and she has a broad forehead.
The bridge of this lady's nose is high, and she has a long nose.
This woman has double eyelids, and she has slightly curved eyebrows and long eyelashes.
The woman with black eyebrows and big eyes has a pair of brown eyes.
This woman has wheat skin, she has deep eye sockets and a small mouth.
She has big eyes and a long broad nose, and her short black hair is straight.
The woman has thin lips and her small mouth is closed.
The woman wears a silver necklace around her neck, and her hair is pulled back to reveal a wide forehead.
The woman has bags under her eyes and her cheekbones are a bit prominent.

Figure 6. Example of human face attributes covered in the description in the CelebAText-HQ dataset

Table 3. Datasets of text-to-faces synthesis

	CelebAText	CelebAText-HQ
Images	10,000	15,010
Captions	10,000	150,100
Manually annotated	No (generated)	Yes

As illustrated in Figure 3 and Table 3, the CelebAText dataset has only one description for each image but a very long one, while CelebAText-HQ has 10 different descriptions for every single image (with a medium number of words) that together fully describe the image. The CelebAText is constrained in a small corpus as the descriptions are generated from the attribute labels of celebA dataset [22]. Figure 7 shows the training of the DMAttn-GAN model on two different datasets: the CelebAText and CelebAText-HQ. From Figure 7, it's clear that the convergence of the model on CelebAText dataset is very slow and doesn't show great potential in achieving high-resolution images and it needs at least another 100 epochs (as mentioned in [21]) in order to generate a real human face structure, while the other experiment (CelebAText-HQ) outputs a reasonable human face structure with only 100 epochs.

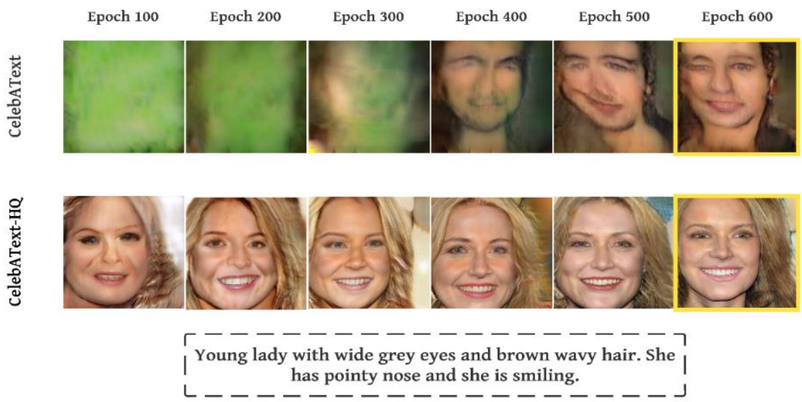


Figure 7. Training progress of DMAttn-GAN on CelebAText (upper) and CelebAText-HQ (lower)

5. CONCLUSION

This paper presented two text-to-image generation models: AttnDM-GAN and DMAttn-GAN. These models 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 64×64 images are generated based on the encoded input text. The second stage is the attention image generation stage, which generates higher-resolution images 128×128, and the last stage is dynamic memory based image Refinement which refines the images to 256×256 resolution images. DMAttn-GAN is the same as AttnDM-GAN but with switching the second and the third stage. These models are applied to different datasets considering two categories: birds and faces. The experiments show the effect of the different datasets' categories and quality on the realism and consistency of the output images and how challenging, complex, and vague the faces' datasets are. In the future, more work will be directed to measure both the quality and the consistency between the input textual data and output images which minimize the human intervention.

REFERENCE

[1] M. Berrahal and M. Azizi, "Optimal text-to-image synthesis model for generating portrait images using generative adversarial network techniques," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 2, pp. 972–979, Feb. 2022, doi: 10.11591/ijeecs.v25.i2.pp972-979.

[2] W. J. Hadi, S. M. Kadhem, and A. R. Abbas, "Fast discrimination of fake video manipulation," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 3, pp. 2582–2587, Jun. 2022, doi: 10.11591/ijece.v12i3.pp2582-2587.

[3] C. Kim, H. Lee, and H. Jung, "Fruit tree disease classification system using generative adversarial networks," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, pp. 2508–2515, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2508-2515.

[4] M. A. Jishan, K. R. Mahmud, A. K. Al Azad, M. R. A. Rashid, B. Paul, and M. S. Alam, "Bangla language textual image description by hybrid neural network model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 2, pp. 757–767, Feb. 2021, doi: 10.11591/ijeecs.v21.i2.pp757-767.

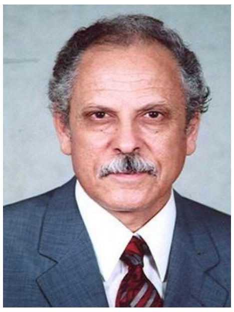
[5] X. Chen, L. Qing, X. He, X. Luo, and Y. Xu, "FTGAN: A fully-trained generative adversarial networks for text to face generation," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.05729.
[6] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, "Semantics disentangling for text-to-image generation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 2322–2331, doi: 10.1109/CVPR.2019.00243.
[7] J. Liang, W. Pei, and F. Lu, "CPGAN: Content-parsing generative adversarial networks for text-to-image synthesis," in *European Conference on Computer Vision*, 2020, pp. 491–508, doi: 10.1007/978-3-030-58548-8_29.
[8] L. Gao, D. Chen, J. Song, X. Xu, D. Zhang, and H. T. Shen, "Perceptual pyramid adversarial networks for text-to-image synthesis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 8312–8319, Jul. 2019, doi: 10.1609/aaai.v33i01.33018312.
[9] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, "ManiGAN: Text-guided image manipulation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 7877–7886, doi: 10.1109/CVPR42600.2020.00790.
[10] S. Yu, H. Dong, F. Liang, Y. Mo, C. Wu, and Y. Guo, "SIMGAN: Photo-realistic semantic image manipulation using generative adversarial networks," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 734–738, doi: 10.1109/ICIP.2019.8804285.
[11] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
[12] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *33rd International Conference on Machine Learning, ICML 2016*, 2016, pp. 1681–1690.
[13] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 49–58, doi: 10.1109/CVPR.2016.13.
[14] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *4th International Conference on Learning Representations, ICLR 2016-Conference Track Proceedings*, 2016, pp. 1–15.
[15] H. Zhang *et al.*, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5908–5916, doi: 10.1109/ICCV.2017.629.
[16] H. Zhang *et al.*, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019, doi: 10.1109/TPAMI.2018.2856256.
[17] T. Xu *et al.*, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 1316–1324, doi: 10.1109/CVPR.2018.00143.
[18] H. Dong, 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)*, Sep. 2017, pp. 2015–2019, doi: 10.1109/ICIP.2017.8296635.
[19] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 1505–1514, doi: 10.1109/CVPR.2019.00160.
[20] A. Singh and S. Agrawal, "CanvasGAN: A simple baseline for text to image generation by incrementally patching a canvas," in *Science and Information Conference*, 2020, pp. 86–98, doi: 10.1007/978-3-030-17795-9_7.
[21] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Learn, imagine and create: Text-to-image generation from prior knowledge," in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019, vol. 32, pp. 887–897.
[22] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 5795–5803, doi: 10.1109/CVPR.2019.00595.
[23] R. Bayoumi, M. Alfonse, and A.-B. Salem, "Multi-stage hybrid text-to-image generation models," *International Journal of Intelligent Computing and Information Sciences*, vol. 22, no. 3, pp. 82–91, Aug. 2022, doi: 10.21608/ijicis.2022.117124.1157.
[24] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.
[25] J. Sun, Q. Li, W. Wang, J. Zhao, and Z. Sun, "Multi-caption text-to-face synthesis: Dataset and algorithm," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2290–2298, doi: 10.1145/3474085.3475391.
[26] R. Bayoumi, M. Alfonse, and A.-B. M. Salem, "An intelligent hybrid text-to-image synthesis model for generating realistic human faces," in *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, Dec. 2021, pp. 172–176, doi: 10.1109/ICICIS52592.2021.9694194.
[27] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 3730–3738, doi: 10.1109/ICCV.2015.425.
[28] M. Heusel, 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 *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 6627–6638.
[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
[30] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–15.

BIOGRAPHIES OF AUTHORS

Razan Bayoumi is a teaching assistant at Department of Computer Science, Faculty of Computer and Information Science, Ain Shams University, Abbasia, Cairo, Egypt. She received the B.Sc degree in computer sciences from Computer and Information Sciences, Ain Shams University, Cairo, Egypt in 2018. Her research interests include deep neural networks, AI, and generative networks. She has published 4 publications in refereed international conferences and journals. She can be contacted at email: razan.bayoumi@cis.asu.edu.eg.

Marco Alfonse is an Associate Professor at Department of Computer Science, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt. He got Ph.D. of Computer Science since August 2014, University of Ain Shams. His research interests include semantic web, ontological engineering, medical informatics, machine learning, deep learning, and artificial intelligence. He has published 78 publications in refereed international journals and conferences. He can be contacted at email: marco_alfonse@cis.asu.edu.eg.

Mohamed Roushdy received his Ph.D. in 1993, M.Sc degree in 1984 and B.Sc. degree in 1979 from the Faculty of Science, Ain Shams University. His experimental doctoral research work was conducted at Bochum University in Germany during 1989-1991. He is a Professor of Computer Science since 2007 till now and Dean of the Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt since 2010 till 2018. He is a Dean of Faculty of Computers and Information Technology, Future University in Egypt from Sept. 2019 till now. His areas of research are artificial intelligence, knowledge engineering, machine learning, expert systems, intelligent medical, health informatics, image processing, pattern recognition, biometrics, and medical data mining. He published more than 130 scientific papers in International Journals, International Conferences and Book Chapter. He got Ain Shams University Appreciation Award in Technological Sciences (Year 2018). He can be contacted at email: mohamed.roushdy@fue.edu.eg.



Abdel-Badeeh M. Salem he is a Professor of Computer Science since 1989 at Ain Shams University, Egypt. His research includes intelligent computing, knowledge-based systems, biomedical informatics, and intelligent e-leaming. He has published around 250 papers in refereed journals and conferences. He has been involved in more than 400 Confs and workshops as a keynote speaker, scientific program committee, organizer, and session chair. He is a member of many national and international informatics associations. He can be contacted at email: abSalem@cis.asu.edu.eg.