

Optimal text-to-image synthesis model for generating portrait images using generative adversarial network techniques

Mohammed Berrahal, Mostafa Azizi

MATSI Research Lab, ESTO, Mohammed First University, Oujda, Morocco

Article Info

Article history:

Received Sep 7, 2021

Revised Nov 25, 2021

Accepted Dec 6, 2021

Keywords:

Deep fusion-GAN

Deep learning

Generative adversarial network

Portrait generation

Text-to-image synthesis

ABSTRACT

The advancements in artificial intelligence research, particularly in computer vision, have led to the development of previously unimaginable applications, such as generating new contents based on text description. In our work we focused on the text-to-image synthesis applications (TIS) field, to transform descriptive sentences into a real image. To tackle this issue, we use unsupervised deep learning networks that can generate high quality images from text descriptions, provided by eyewitnesses to assist law enforcement in their investigations, for the purpose of generating probable human faces. We analyzed a number of existing approaches and chose the best one. Deep fusion generative adversarial networks (DF-GAN) is the network that performs better than its peers, at multiple levels, like the generated image quality or the respect of the giving descriptive text. Our model is trained on the CelebA dataset and text descriptions (generated by our algorithm using existing attributes in the dataset). The obtained results from our implementation show that the learned generative model makes excellent quantitative and visual performances, the model is capable of generating realistic and diverse samples for human faces and create a complete portrait with respect of given text description.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Mohammed Berrahal

MATSI Research Lab, ESTO

Mohammed First University, Oujda, Morocco

Email: m.berrahal@ump.ac.ma

1. INTRODUCTION

With the advent of new technologies, deep learning (DL) had seen a tremendous success [1], helping in many fields like law enforcement applications [2], [3], to solve cases based on transforming text description (given by eyewitnesses) to real facial images in order to get closer to the portraits of potential suspects [4]. This technique known as text-to-face image generation synthesis has recently attracted the attention of many researchers and became an active research topic. There are a few works directly related to this area compared to text-to-image synthesis [5].

In this paper, we present an overview about the recent literature of text-to-image synthesis based on variants of generative adversarial network (GAN) [6]. Then, our main focus is directed towards the framework to generate object images from high-level descriptions or at least portrait images that match all the descriptions [7]. Our methodology consists of the following steps.

First, we define the architecture and strategy of every method then critically examine their main strengths and limitations. Secondly, we train a text encoder, char-convolutional neural network recurrent neural network (CNN-RNN) text embeddings [8], on a dataset of text captions to read the sentences and extract the relevant attributes. Then, we give a full implementation and adaptation on bird [9] and common object in context (COCO) [10] datasets, we compare their efficiencies on multiple levels like generated

image quality, consumed resources, accuracy for both generator and discriminator, number of epochs, steps and training time. We also use evaluation metrics like Inception Score [11] and frechet inception distance [12] for trained models. We deduce the best optimal model and apply-it on the CelebA [13] dataset.

Our main contribution in this work is:

- Propose an algorithm to transform facial attributes into sentences (text description),
- Build, evaluate each model, and make a comparative study, of existing models,
- Adapt the optimal method of text-to-image synthesis to portrait generation,
- Assess the whole model using specific metrics and discuss generated images.

The rest of this paper is organized: In the second section, we review the background of text-to-image synthesis using GAN networks, including a description of each method and the improvements made over GAN models for generating images. We prepare and preprocess our data in the third section. Before concluding, we analyze in the fourth section our findings about several text-to-image approaches used to construct an image and apply the optimal model on facial images to generate a realistic portrait generation.

2. BACKGROUND

Our research will focus on GAN models and their improvements. The early stages of text-to-image synthesis research are largely dependent on supervised learning [14], which relies on the connection between keywords and images to find a way to connect the written description to photos. We identify the following reasons for not qualifying these methods for our case: no generation of new visual contents, no model adaptation to fresh datasets, no use of unsupervised methods, and no image optimization during training.

2.1. Generative adversarial network (GAN)

In the last recent years, one of the most resulting and efficient deep learning methods is the GAN, GAN combine two neural networks, a generator and a discriminator. First the generator gets a randomly sampled noise vector as an input; in most cases, we sample, from a Gaussian distribution, a noise vector to feed the neural network with multiple convolutional layers, to generate an image, this latter will be directed to the discriminator who needs to decide whether that image comes from the actual dataset (the real images that were training on) or from the generator. The learning signal will be propagated through this entire model pipeline, the generator that has been learning from the feedback of this discriminator network and can eventually manage to generate images that look very similar to the dataset [6].

2.2. Deep fusion generative adversarial networks (DF-GAN)

It is a framework that has resolved several issues in text-to-image synthesis by using three major improvements: 1) Using one pair of generator discriminator to synthesize high-quality images. Therefore, using one stage training backbone, inspired by the unconditional image generation, for a better stabilization of the model, we use multiple loss function but the best performance was by using the hinge loss function. 2) To fuse sentence caption with graphical representation the approach uses a module deep text-image fusion block (DFBlock). This process is depending on staking multiple affine transformations and ReLU layers in fusion block, so that the text information can be fully exploited, 3) To improve the generator to synthesize realistic contents without using extra networks, the method proposes a target-aware discriminator composed of one-way output and matching-aware gradient penalty [15].

2.3. Stacked generative adversarial networks (StackGAN-v1/StackGAN-v2)

StackGAN-v1: This framework is capable of generating realistic images, conditioned by given text descriptions. The training is based on two stages, the first one is to generate low-resolution images based on primitive shapes and colors, the second is to generate high quality realistic resolution images, taking the result of the first stage and text description as an input to rectify the result of stage 1. Then to defect the results from both stages, the framework uses conditioning augmentation technique [16].

StackGAN-v2: This framework uses the same logic of the earlier version. The difference between the two frameworks is that StackGAN-v2 doesn't depend on only two stages, but exceeds it to multistage using a tree structure of multiple generative (Gs) and discriminators (Ds), in every stage a scene is generated from different branches of the tree to form the final generated image respecting every detail of the captions. That is why this version is more efficient than the old one, and it gives better results especially if the text is long and has a lot of details [17].

2.4. Attentional generative adversarial network (Attn-GAN)

This framework is based on two main components. The attentional generative network is the first one, it transforms text to vector, in order to generate low-resolution images in the first stage, then combine

image vector with the corresponding text vector to generate new images. The second component is deep attentional multimodal similarity model (DAMSM) that determines the degree of resemblance between the generated images and the text captions, providing an additional fine-grained image-text to match the loss of the trained generator [18].

2.5. Dynamic memory generative adversarial network (DM-GAN)

This framework proposes first place a refinement on fuzzy generated image contents by a dynamic memory module, its main work is to select the important text description based on the initial image content, to generate accurate images from text. The second process is working with a response gate to fuse the information read from the memories and the image features. By both qualitative and quantitative criteria, DM-GAN exceeds the current state-of-the-art in two real-world bird and COCO datasets [19].

2.6. Method categorization of GAN-based text-to-image synthesis

Up-to now, methods of GAN-based text-to-image synthesis has taken a big step into generating the images. However, each one is able to develop different aspects. We can regroup these methods into four categories, as shown in Table 1. Each category has its own characterization [36].

- Semantic enhancement: This category represents the GANs methods which focus on generating images semantically related to the input texts, by using the neural network to encode texts as dense features, and fed-it to the second network to generate images matching to the text’s description.
- Resolution enhancement: This category uses multistage training process where the output on the earlier stage is the input for the next stage, to guarantee the generation of better and high-quality images on every stage and capable of matching the text’s description.
- Diversity enhancement: This category adds an additional component to generate models capable of estimating semantic relevance between generated images and text description, in order to maximize the output diversity of generated images by multiple types and visual appearance.
- Motion enhancement: This category is based on adding a temporal dimension to ensure a meaningful action with respect to the text descriptions. It consists of two steps, the first one creates pictures based on the text’s actions, and the second is a mapping or alignment process used to guarantee that pictures are coherent in time.

Table 1. Classification of existing GAN-based methods in four different categories

		Semantic Enhancement	Resolution Enhancement	Diversity Enhancement	Motion Enhancement
GAN-Based Methods					
2017	[20] DONG-GAN	✓	-	-	-
	[16] Stack-GAN	-	✓	-	-
	[21] AC-GAN	-	-	✓	-
	[22] TAC-GAN	-	-	✓	-
	[23] OBAMA-Net	-	-	-	✓
2018	[24] Paired-D GAN	✓	-	-	-
	[25] MC-GAN	✓	-	-	-
	[17] Stack-GAN++	-	✓	-	-
	[18] Attn-GAN	-	✓	-	-
	[26] HD-GAN	-	✓	-	-
	[27] Scene Graph GAN	-	-	✓	-
	[28] T2V	-	-	-	✓
2019	[29] T2S	-	-	-	✓
	[30] Obj-GAN	-	✓	-	-
	[19] DM-GAN	-	✓	-	-
	[31] Text-SeGAN	-	-	✓	-
	[32] MirrorGAN	-	-	✓	-
2020	[33] StoryGAN	-	-	-	✓
	[15] DF-GAN	✓	-	-	-
	[34] SAM-GAN	-	✓	-	-
	[35] PCCM-GAN	-	✓	-	-

3. PREPARING AND PREPROCESSING DATA

In this section, we update each existing code with the newest technologies, test and build a model for each method in order to generate images form text descriptions. As datasets for testing methods, we use the CUB-200 (Caltech-UCSD Birds 200) picture dataset that contains 6033 photographs of 200 different bird species (mostly from North American) and common object in context (COCO) is a large-scale object

detection, segmentation, and captioning dataset. For the implementation of such methods, we use celebrity facial images (CelebA) datasets with more than 200K facial images, each with 40 attributes.

Text description: To create a description of every facial image existing in CelebA datasets, we define descriptions or options from the attribute that helps us form sentences of the facial description (see some examples in Table 2). We use gender (option in random state) to begin the sentence, after that we focus randomly on other attributes that match the image. We connect the sentence by “and” or “which”, for instance, if the attributes ‘gender’ and ‘wearing glass’ are both valued is 1, then the sentence will be “The man is wearing glasses”. We create multiple sentences for each image between 2 and 5, even though the algorithm does most of the work, we must manually correct some errors, for a clean training.

Table 2. Attributes, descriptions and examples of generated sentence

Attribute	Description (Options)	Sentences formed examples
Gender	He / she - his / here	“The woman has high cheekbones.”
	Man / women	“The boy has a chubby face.”
	Girl / boy	
	Male / female	
Face shape	Oval / round / square / rectangular / triangle	“The man has an oval face and high cheekbones.”
Mouth	Small/ big / medium	“She has big lips with arched eyebrows and a slightly open mouth.”
	Open / slightly open	
Color	Hair [dark, white ...]	“The woman has wavy hair which is brown in color.”
	Beard [dark, white ...]	“His hair is black in color.”
	Eyes [brown, bleu ...]	
	Eyes-brow [dark, white ...]	
Glasses Necklace ...	Has/ wearing	“She's wearing eyeglasses, necklace.”

4. RESULTS AND DISCUSSION

4.1. Hardware characteristics

To test DL model, we use a remotely accessible high-performance computing (HPC) infrastructure is available to Moroccan researchers through the national center for scientific and technical research (CNRST). Parallel computing performance is maximized by using a low latency network with 100Gbps capacity to connect nodes. A 5Gbps link connects the infrastructure to MARWAN to ensure fast data transfer from MARWAN-connected universities and organizations:

- Compute Nodes: 2 * Intel Xeon Gold 6148(2.4GHz/20-core)/192 GB RAM
- GPU Node : 2 * NVIDIA Tesla P100/192 GB RAM
- Storage Node: 2 * Intel Xeon Silver 4114(2.2GHz/20-core)/18 * SATA 6 TB

4.2. Evaluation metrics

For evaluating our models, we use both quantitative and qualitative evaluations. The quantitative evaluation uses the following metrics,

- a) Inception scores (IS): is calculated by first predicting the class probabilities for each generated image using a pre-trained inception v3 model [11].
- b) Frechet inception distance (FID): is a distance measure that compares feature vectors calculated for real and generated images [12].

As the qualitative evaluation relays on observing visualization results, we compare the quality of the portrait images generated by different methods, and how much detail is similar to real images.

5. IMPLEMENTATION AND RESULTS ON CUB-200 AND COCO

After an implementation which took a lot of time and resources at the graphics processing unit (GPU) level, we succeeded to train our models of the seven methods, applying each of them on CUB-200 and COCO datasets, summing up, with 14 models capable of generating birds and context object common form text description. Each method has its own parameter and its own environment. As shown in Table 3, our trained models managed to get better scores at multiple levels. The highest score of is on both datasets achieved by DF-GAN, and the lowest score of FID on both datasets is also made by DF-GAN (we remind that a model M1 whose FID is lower than those of another model M2, mean that M1 performs better than M2).

So far, we could say that the best model in terms of performance is DF-GAN. However before concluding whether it is the best model, we need to compare the quality of generated images, and how much it matches text description. For this purpose, we generate images using three models DF-GAN, DM-GAN and Attn-GAN, and the description of the bird shown in the Figures 1 and 2. From these figures, we can clearly see the potential of DF-GAN, it generates an accurate image with high quality, and full respect to text

description, like for example “a bright yellow and black bird perched on the end of a branch”, the DF-GAN model manages to respect every detail in the text especially at the end of the sentence. However, the other model manages to draw the color and shape but fails at the end. The other example “The small bird has a red head with feathers that fade from red to gray from head to tail” where the DF-GAN model succeeds at fading color from red to gray, when the other models fail this process even if Attn-GAN succeeded the fading part it failed the realism and high quality of the image.

Even with the other tests on CUB-200 and COCO, the DF-GAN model succeeds over 80% of generating a better result in 50 tests. Compared with the other models, despite that some models on tests surpassed the DF-GAN, but we must remember that this model respects all the text and surpassed the others in terms of the realism of generated images. From the results, we deduced that the best method through many levels is DF-GAN, so we use this method for implementation on CelebA datasets to create a model capable of generating portrait images from text description.



Figure 1. Generated yellow bird’s images from text description using three models: DF-GAN, DM-GAN and Attn-GAN



Figure 2. Generated red bird’s images from text description using three models: DF-GAN, DM-GAN and Attn-GAN

Table 3. A summary of performance of different methods on CUB and COCO datasets, using two performance metrics: Inception score (IS), frechet inception distance (FID)

Methods	Datasets and Metrics			
	CUB		COCO	
	IS	FID	IS	FID
DF-GAN	5.10	14.81	42.05	21.41
StackGAN-v1	3.74	51.89	8.60	--
StackGAN-v2	4.09	15.30	8.40	--
Attn-GAN	4.39	23.98	26.36	35.49
DM-GAN	4.75	16.09	31.06	32.64
HD-GAN	4.20	--	12.04	--
DCGAN	2.88	68.79	7.88	--

6. IMPLEMENTATION AND RESULTS ON CELEBA DATASET

We adapt and train DF-GAN model to generate portrait images from description. As data we use images from CelebA datasets, and text captions generated from our algorithm. Implementation details: As shown in Figure 3. The generator G takes two inputs the sentence vector encoded by text encoders from text caption and noise vector, we feed them into a fully connected layer, to subsample the image features we apply a multiple UPBlocks on the output of the first layer, every UPBlock is divided into three components and upsample layer, a residual block and DFBlock, the last component has as a goal to combine text and picture elements during image generation. Lastly, we use a convolution layer to converts image features to the real images. The image output of the G will be direct to the Discriminator D as the first input, the second input is the image from the dataset, the generator applies on each image a convolution layers the output will be down sampled by multiple layers of DownBlocks to convert images into feature maps, every to assess the visual realism and semantic consistency of inputs, the sentence vector will be duplicated and concatenated, the role of D is to determine the generated images from the real data, by repenting this process, the discriminator encourages the generator to generate pictures of greater quality and text-image semantic coherence.

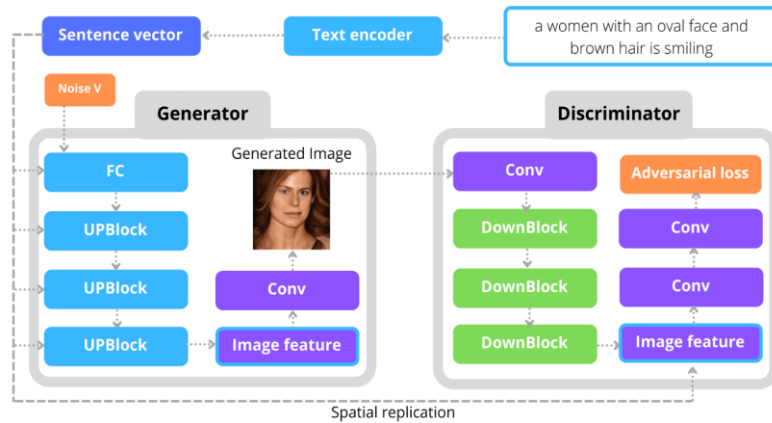


Figure 3. The architecture of the method DF-GAN used for portrait generations

Training details: we train our model over 60 epochs. We choose a batch size of 24, for optimal training we use hinge loss. We also use as the dimension of the model a generator with 128 and discriminator with 64. The training process took one week of training.

We generate for each 20 epochs an image to follow the evolution of the model, as shown in Figure 4. The generator starts training from noise vectors. After 20 epochs the model begins to reshape faces, we can see it in first parts of the picture, the form of faces with noise, we can also distinguish the gender and the other attributes, but still in an early stage.

After 60 epochs the model can generate an accurate face with high resolution as shown in part 2 of the picture, we can distinguish the male’s face from the females and it is Capable of respecting other attributes with high precision. A remark which shows us that the model can generate images by itself is visible in the 5th and 6th photos which were quite similar the only obvious difference is the hair color, instead of generating the image from the beginning, the generator just changes the color of the hair and the pose of the face. After the training process, we test our DF-GAN trained model, on some outside text description, as shown in Figure 5. The model succeeds most of the time at respecting the contents of the text and generate more accurately high-quality images. However, we observe that when we stick with the 40 attributes that exist in CelebA datasets, the model generates accurate images, but once we stray a way from these attributes the model fails the generation process.

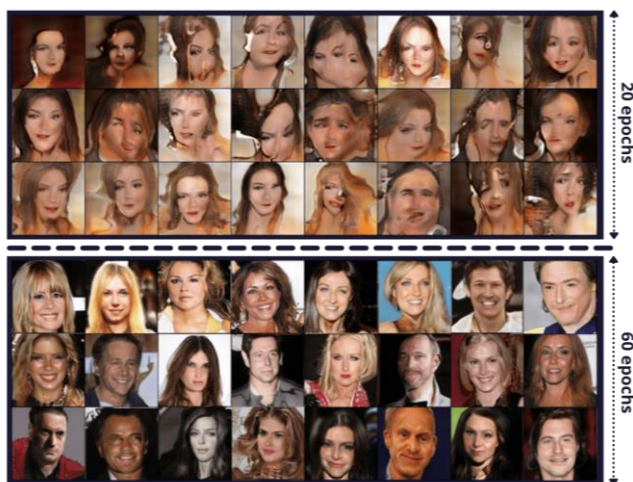


Figure 4. Generated images from training model using DF-GAN method on 20 and 60 epochs



Figure 5. Examples of generated facial images from text using our DF-GAN trained models

7. CONCLUSION

Through this work, we have implemented the most recent approach DL-based for text-to-image synthesis, to compare their efficiency in terms of quantitative and qualitative assessment of generated images. On the light of this experiment, the optimal model that performs better than its peers, is DF-GAN. This latter is further implemented on the human face, to generate high-quality facial images from text descriptions. Indeed, the implementation and testing of several methods over CUB-200 and COCO datasets, we found that DF-GAN was the most accurate one compared to the other studied methods. To implement DF-GAN on portraits, we have used CelebA dataset of face images. However, the gap of text descriptions led us to create an algorithm to randomly create sentence descriptions using existing attributes in CelebA datasets. The experiment conducted over our implementation shows relevant quantitative and visual results. The resulting model is not only capable of generating realistic and diverse samples for human faces, but also it makes a complete portrait with the respect of text descriptions.

ACKNOWLEDGEMENTS

This research was supported through computational resources of HPC-MARWAN (www.marwan.ma/hpc) provided by the National Center for Scientific and Technical Research (CNRST), Rabat, Morocco.




REFERENCES

- [1] M. Berrahal and M. Azizi, "Review of DL-Based Generation Techniques of Augmented Images using Portraits Specification," in *4th International Conference on Intelligent Computing in Data Sciences, ICDS 2020*, 2020, doi: 10.1109/ICDS50568.2020.9268710.
- [2] M. Boukabous and M. Azizi, "Review of Learning-Based Techniques of Sentiment Analysis for Security Purposes," in *Innovations in Smart Cities Applications Volume 4*, 2021, pp. 1-14, doi: 10.1007/978-3-030-66840-2_8.
- [3] I. Idrissi, M. Azizi, and O. Moussaoui, "Accelerating the update of a DL-based IDS for IoT using deep transfer learning," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 23, no. 2, pp. 1059-1067, Aug. 2021, doi: 10.11591/IJECS.V23.I2.PP1059-1067.
- [4] M. Berrahal and M. Azizi, "Augmented binary multi-labeled CNN for practical facial attribute classification," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 23, no. 2, pp. 973-979, Aug. 2021, doi: 10.11591/IJECS.V23.I2.PP973-979.
- [5] X. Wu, K. Xu, and P. Hall, "A survey of image synthesis and editing with generative adversarial networks," *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 660-674, Dec. 2017, doi: 10.23919/TST.2017.8195348.
- [6] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, doi: 10.3156/jsoft.29.5_177_2.
- [7] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional image generation from visual attributes," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9908 LNCS, pp. 776-791, doi: 10.1007/978-3-319-46493-0_47.
- [8] M. Boukabous and M. Azizi, "A comparative study of deep learning based language representation learning models," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 22, no. 2, pp. 1032-1040, May 2021, doi: 10.11591/IJECS.V22.I2.PP1032-1040.
- [9] "Caltech-UCSD Birds 200." Accessed: Jul. 1, 2021. [Online]. Available: <http://www.vision.caltech.edu/visipedia/CUB-200.html>
- [10] "COCO - Common Objects in Context." Accessed: Jul. 27, 2021. [Online]. Available: <https://cocodataset.org/>
- [11] S. Barratt and R. Sharma, "A Note on the Inception Score," *Proc. ICML 2018 Workshop on Theoretical Foundations and Applications of Deep Generative Models*, Jan. 2018.
- [12] A. Odena *et al.*, "Is Generator Conditioning Causally Related to GAN Performance?," *PMLR*, Jul. 2018.
- [13] "Large-scale CelebFaces Attributes (CelebA) Dataset." Accessed: May 7, 2020. [Online]. Available: <http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>
- [14] I. Idrissi, M. Boukabous, M. Azizi, O. Moussaoui, and H. El Fadili, "Toward a deep learning-based intrusion detection system for IoT against botnet attacks," *IAES Int. J. Artif. Intell.*, vol. 10, no. 1, Mar. 2021, doi: 10.11591/IJAI.V10.I1.PP%P.
- [15] M. Tao *et al.*, "DF-GAN: Deep Fusion GAN for Text-to-Image Synthesis," *IEEE Trans. Multimed.*, Aug. 2020.
- [16] H. Zhang *et al.*, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," *ICCV 2017 Oral Presentation*, vol. 2017-Oct, pp. 5908-5916, Dec. 2016.
- [17] 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, Oct. 2017, doi: 10.1109/TPAMI.2018.2856256.
- [18] T. Xu *et al.*, "AttnGAN: Fine-Grained Text to Image Generation with Attentional GAN," *Computer Vision and Pattern Recognition*, 2017.
- [19] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic Memory GAN for Text-to-Image Synthesis," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 5795-5803, Apr. 2019.
- [20] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic Image Synthesis via Adversarial Learning," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 5707-5715, Jul. 2017.
- [21] A. Odena, C. Olah, and J. Shlens, "Conditional Image Synthesis With Auxiliary Classifier GANs," *34th Int. Conf. Mach. Learn. ICML 2017*, vol. 6, pp. 4043-4055, Oct. 2016.
- [22] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, "TAC-GAN - Text Conditioned Auxiliary Classifier Generative Adversarial Network," Mar. 2017.
- [23] R. Kumar, J. Sotelo, K. Kumar, A. de B, and Y. B, "ObamaNet: Photo-realistic lip-sync from text," Dec. 2017.




- [24] D. M. Vo and A. Sugimoto, "Paired-D GAN for Semantic Image Synthesis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11364 LNCS, pp. 468-484, 2019, doi: 10.1007/978-3-030-20870-7_29.
- [25] H. Park, Y. Yoo, and N. Kwak, "MC-GAN: Multi-conditional Generative Adversarial Network for Image Synthesis," *Br. Mach. Vis. Conf. 2018, BMVC 2018*, May 2018.
- [26] Z. Zhang, Y. Xie, and L. Yang, "Photographic Text-to-Image Synthesis with a Hierarchically-nested Adv-Net," *CVPR2018 Spotlight*, 2018.
- [27] J. Johnson, A. Gupta, and L. Fei-Fei, "Image Generation from Scene Graphs," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1219-1228, Apr. 2018.
- [28] Y. Li, M. R. Min, D. Shen, D. Carlson, and L. Carin, "Video Generation From Text," *32nd AAAI Conf. Artif. Intell. AAAI 2018*, pp. 7065-7072, Oct. 2017.
- [29] S. Stoll, N. Cihan Camgoz, S. Hadfield, and R. Bowden, "Sign Language Production using Neural Machine Translation and Generative Adversarial Networks," *SMILE: Scalable Multimodal sign language Technology for sign language Learning and assessmEnt*, 2018.
- [30] W. Li *et al.*, "Object-driven Text-to-Image Synthesis via Adversarial Training," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 12166-12174, Feb. 2019.
- [31] H. Tan, X. Liu, X. Li, Y. Zhang, and B. Yin, "Semantics-enhanced adversarial nets for text-to-image synthesis," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-October, pp. 10500-10509, Oct. 2019, doi: 10.1109/ICCV.2019.01060.
- [32] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning Text-to-image Generation by Redescription," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 1505-1514, Mar. 2019.
- [33] Y. Li *et al.*, "StoryGAN: A Sequential Conditional GAN for Story Visualization," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 6322-6331, Dec. 2018.
- [34] D. Peng, W. Yang, C. Liu, and S. Lü, "SAM-GAN: Self-Attention supporting Multi-stage Generative Adversarial Networks for text-to-image synthesis," *Neural Networks*, vol. 138, pp. 57-67, Jun. 2021, doi: 10.1016/J.NEUNET.2021.01.023.
- [35] Z. Qi, J. Sun, J. Qian, J. Xu, and S. Zhan, "PCCM-GAN: Photographic Text-to-Image Generation with Pyramid Contrastive Consistency Model," *Neurocomputing*, vol. 449, pp. 330-341, Aug. 2021, doi: 10.1016/J.NEUCOM.2021.03.059.
- [36] J. Agnese, J. Herrera, H. Tao, and X. Zhu, "A Survey and Taxonomy of Adversarial Neural Networks for Text-to-Image Synthesis," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 4, Oct. 2019.

BIOGRAPHIES OF AUTHORS



Mohammed Berrahal    is a Ph.D. student in computer engineering at Mohammed First University in Oujda, Morocco, where he is conducting research on security and law enforcement applications utilizing Deep Learning. He holds an M.Sc. in internet of things from ENSIAS, Mohammed 5 University in Rabat, Morocco (2018) and a B.Sc. in computer engineering from ESTO, Mohammed First University in Oujda, Morocco (2016). Furthermore, he is certified in artificial intelligence, 3D modeling, and programming. Additionally, he has served as a reviewer for a number of international conferences and journals. And is currently employed at Mohammed First University as an administrative assistant. He can be contacted at email: m.berrahal@ump.ac.ma.



Prof. Dr. Mostafa Azizi    received a State Engineer degree in Automation and Industrial Computing from the Engineering School EMI of Rabat, Morocco in 1993, then a Master degree in Automation and Industrial Computing from the Faculty of Sciences of Oujda, Morocco in 1995, and a Ph.D. degree in Computer Science from the University of Montreal, Canada in 2001. He earned also tens of online certifications in Programming, Networking, AI, and Computer Security. He is currently a Professor at the ESTO, University Mohammed 1st of Oujda. His research interests include Security and Networking, AI, Software Engineering, IoT, and Embedded Systems. His research findings with his team are published in over 100 peer-reviewed communications and papers. He also served as PC member and reviewer in several international conferences and journals. He can be contacted at email: azizi.mos@ump.ac.ma.