Exploring Multimodal Learning: Text Conditioned Image Generation

Priyanka, Shailesh D Kamble, Amita Dev

Indira Gandhi Delhi Technical University for Women, Delhi, India

Corresponding author: Priyanka, Email: priyanka014mtaids22@igdtuw.ac.in

Over the years, with the advancement of technologies, Artificial Intelligence has played a huge role. Text to image-based conversion has taken up the market when the user looks to make their tasks simpler and easier. With plain text commands, one may obtain an image without wasting time in searching for that image. With the use GAN (generative adversarial network) and through the intersection of Natural language processing in decoding the texts through tokens, deep learning, and Artificial intelligence and with the help of image datasets, we would be able to generate images by preprocessing the text and understanding it.

Keywords: Artificial Intelligence, Natural Language processing, Machine learning, Generative adversarial network.

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 387–396. DOI: https://doi.org/10.56155/978-81-955020-7-3-34

1. Introduction

To understand what multimodal approach is, here is the explanation, it is a method in artificial intelligence that combines multiple types of data and modes which creates more exact, understandable insights and conclusions to real-world problems. It uses various modes such as text, images, speech, or numerical datasets. GPT 5 can be one such example of being a multimodal AI. Here, we are using a multimodal approach where the user will input text which will help in synthesis of image. The user will write a description for the images which he wants to fetch, and the model will generate the images accordingly. All these processes are enabled by Generative adversarial networks (GAN). The text to image-based generation, a visual synthesis has gained many interests of users in the market. The rapid progress of GAN brings a remarkable evolution in natural image generation with diverse conditions [1]. The text to image generation is still a challenging process in the market to decode the text and describe the meaning.

There are various issues in text to image generation, so to resolve these issues, this paper introduces a new way of image generation using GAN (Generative adversarial network), which will generate a fine grained, noise free image, according to the description provided by the user.

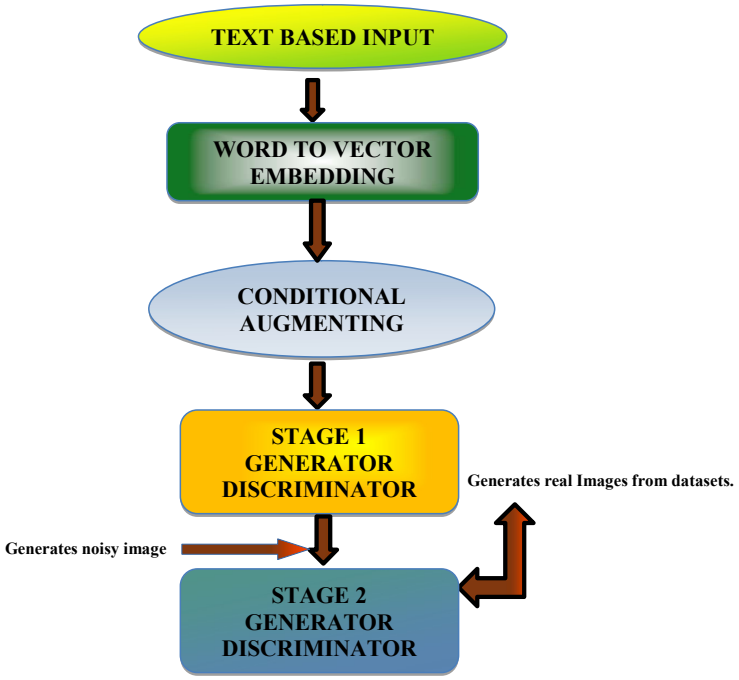

Figure 1. Structure of GAN

In the above figure 1, GAN consists of two independent networks which are Generator and Discriminator. Generator is meant to generate synthetic samples with random noise and discriminator is used to classify in the form of binary sample, which classify in the form of 0 and 1 in which 0 is for fake images and 1 is for real image sample [2].

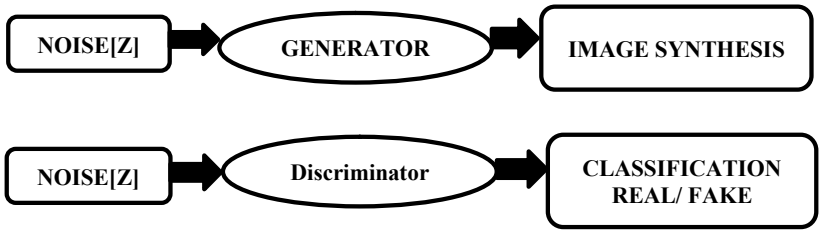

Figure 2. Structure of GAN

In the above figure 2, generator and discriminator do its job in a best way, it labels the image as fake image or real image. Generator generates the samples in a best possible way so that it can fool the discriminator [2], as it is sometimes unable to identify real image and mistakenly it classifies real image as fake image. Therefore, two-waystack GAN is sometimes a good option where two generators and two discriminators are used to obtain a high quality and better image.

2. Background

As discussed above, text to image generation has been evolutionary and by the time it has shown changes. A lot of models, for text to image generation such as autoencoders, diffusion model, GAN has evolved. All of these models are compared on the basis of diversity, visually reality, semantic alignment to understand how and which model is working fine to generate images.

The image datasets are being trained and fake or real images are obtained. There have been a number of research work based on this domain, where different authors have used different models. With the development of GAN, text to image synthesis have gained significant attentions [3].

There is use of datasets which are processed with a large number of samples. Each sample of this training data is referred to as observations. And these observations are made up of various types of features. The model will be able to produce more collections of images and features.
It is to be noted that the model performance depends on the task for which it is used for. In this research, a conclusion will be drawn by using the Specific type of GAN [3], where the images will be produced and compared with the earlier image.

This helps in analyzing how the model is performing and able to understand clearly the text description and generate clear images.

Text to image conversion is the one which has made our task simpler and easier. Now, with just one click a person is able to synthesize an image which he wishes to do. This not only make our lives simpler and easier but also helped in various presentations, academic and corporate sector. There have been a large number of works related to this in the past, just because of its best uses in the market. With the advent of AI and obviously ChatGPT, google bard, image bind tool and many more, these areas have taken up the path for more research.

With this development of ChatGPT and many other AI devices, many authors have worked in this area. This area is the one which is in the scope of continuous improvement. By the time, new features have been deployed, and Multimodal AI have come into picture from a single modal AI. This is how the evolution has taken place in GPT and AI based devices. The Multimodal AI is one which can accept different forms of data such as image, video, text, voice and give the same in output. This type of modal came

into picture when GPT 5 based devices have been heard to come in the market. But these Software are not free, and the cost must be paid.

What if a user designs its own method of developing the text to Image based system where with just small text description, an image can be synthesized.

3. Literature Review

There are number of research work related to this domain, the authors have worked in various areas and implemented using various models, as there is large number of scopes in the implementation of this project which is text to image synthesis. The author in the paper related to text to image generation using textual inversion, has implemented latent diffusion models (LDMs) and introduced class of denoising diffusion probabilistic models [1], also they used LAION-400M datasets, and used word embeddings. The experiment was conducted by them using 2xV100 GPUs having batch size as 4 and the base learning was set to 0.005. They further scaled the base learning and produced results through 5,000 optimization steps. They obtained various results of images which are much clear, included long captions, short captions and modified images [1].

Another research is all about building a generating image from text through controllable text [2]. In this paper the author has talked about converting text to images using Control GAN which has the ability for high quality images synthesis effectively, they also used Natural language processing to control the parts of image generation. There are 3 stages used in this. They have used three scales where 64x64, 128x128 and 256x256 are used. Stage 2 had Spatial and channel wise attentions. The loss in the content was computed at the relu2 of VGG-16, and the whole network was trained using the Adam optimizer with learning rate of 0.0002 [2]. They visualized different results and corresponding attention maps on various stages. They experimentally found that the channel wise attention correlates closely with semantic parts of objects, while the spatial attention mainly focuses on color descriptions [2].

According to the author in another research work where they have done work in the area of Mirror GAN, learning text to image generation by redescription [3]. They have proposed a novel global local attentive and semantic preserving text to image to text framework which is called as Mirror GAN. This Mirror GAN consists of 3 modules which are semantic text embedding module (STEM), a global local collaborative attentive module for cascaded image generation (GLAM) and a semantic text regeneration and alignment module (STREAM) [3]. They have presented a qualitative and quantative comparisons where they verified the Mirror GAN. They have listed inception scores and observed that Mirror GAN obtained high quality images. They observed that GLAM has produced the best results. According to author in another research work [4], It has been mentioned about zero shot text to Image generation, wherein the author has defined a simple transformer instead of using complex models to process the image. Here he has used transformer that autoregressively models the text and image tokens as a single stream of data.

According to the author in next research, it has been mentioned about the hierarchical text to image generation using CLIP (Contrastive Language Image pretraining) latent. They have proposed a 2-stage model where first They compared their model by using MS coco datasets and found that unclip produces realistic scenes by capturing the text prompts [5].

By their experiments using CLIP, it has been found that, more realistic images are obtained, with high quality photos [5]. According to the author in another research, it has been observed that the author had explained about AttnGAN where they produced fine grained images through attentional generative adversarial networks [6]. This attentional generative adversarial network allowed attention driven multi-grained refinement of images where through this they were able to generate the image by different sub regions only after focusing on a relevant word in the natural language descriptions.

According to next author, it has been observed that the he has researched about the semantics disentangling for image generation. There the author has observed about text to image generation which is able to disentangles semantics implicitly so that it can fulfill high- and low-level semantic consistency [7]. He visualized the images generated using various models. The model learnt to remove the defects after embedding.

According to another paper, the author has mentioned about cross modal contrastive learning where he generated image from text. The cross modal contrastive GAN addresses some challenges through increasing mutual information within text and image [8]. This is done through multiple contrastive losses. It is shown using challenging datasets such as MS COCO through XMC GAN. According to another paper, the author has brought a new approach of generating image through text without using text data. This method has aligned multimodal semantic space of the CLIP model [9]. Another research paper is based on the use of Deep learning in obtaining images from text. They have addressed the problem where they noticed that existing algorithms are not able to create images which match text description [10]. So, in order to address this issue, they have introduced the use of RC-GAN which is recurrent convolutional generative adversarial network. This RC GAN was able to bridge the gap of advancement in text and picture modelling, and thereby converting visual notions from words to pixels [10]. According to another research work which is a review paper based on text to image generation, here the author has examined the strategies for evaluation of the image synthesis through text as well as he highlighted shortcomings, identified new research areas of research design [11].

Here the author has come to the conclusion that there are lot more future scopes and improvements in this area. As the Quality and variety of datasets improves, the model will show good results. According to the research work which is based on semantic object accuracy for text to image generation [12], here the author has introduced a new model where the model explicitly models individual objects within an image and they have introduced a new semantic object accuracy that evaluates images with their given caption [12]. In another paper which is based on Random Forest regression for magnetic resonance image synthesis [13], which is based on REPLICA which is a supervised random forest image synthesis approach. Another research is a review paper based on the topic of image synthesis using generative adversarial networks [14]. They have introduced recent research on GANs and summarized the methods of these applications. They have also discussed future scopes and challenges faced by this GAN [14]. They have also reviewed likely future research directions in the field of GANs, such as video generations, facial animation synthesis, and 3D face reconstruction [14].

4. Methodology

The Methodology of this project is based on using a GAN and Glove vectors. Here the libraries are imported which are glob, pandas, NumPy, TensorFlow, Keras, Sklearn, PIL, matplotlib. The glove function is being loaded so that the glove vectors can be loaded. Glove vectors are vector representation of the words, using which training is performed on global words, and this results into a linear substructures of word spaces.

Glove vectors basically makes the task easier and simpler, using these the words are automatically represented to vectors. After this the datasets used here are oxford flower datasets, where text caption file, images file and data file are being loaded.

Then the data preprocessing stage is being started, where are images are preprocessed. The processed files are stored in a binary form, and hence it will enhance the quick use of data. After the Images are being preprocessed, then the image captions are being preprocessed, where the glove embedding is being performed. After this the embedding file is saved. A data frame is being created to store the captions. In this way the captions file is stored in the form of csv files.

After the following operations are being performed, then the loading and combining of NumPy takes place. These files are then trained, the trained data is finally proceeded to data modelling.

Data modelling is the main part where we apply the process of GAN, in this way, we deploy generator and discriminator, and do its further process of up sampling and down sampling. After the initialization of generator, the image has been obtained which is not clear and noisy image, so it is passed through discriminator by calling this function. Both the generator and discriminator used ADAM and same learning rate and momentum.

After the above process, the model is again trained, the datasets are being trained with 500 epochs performed. After the training of model, the testing phase starts, and clear images are obtained. So, after defining the descriptions of flower in the text, the flowers images are obtained. This is how, this project has been proceeded with. The below figure 3 explains the methodology used in implementation of this project.

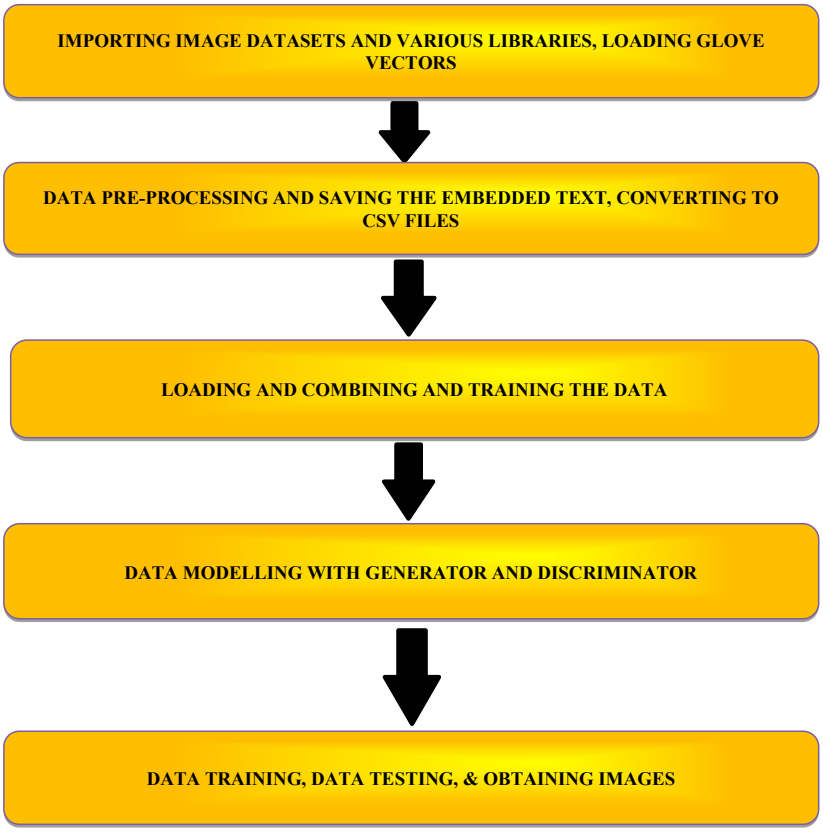

Figure 3. flow of methodology used

5. Experiment Design & Results

The following results are obtained while working on Generator, a noisy image was obtained after up sampling. The below figure explains how the experiment result worked.

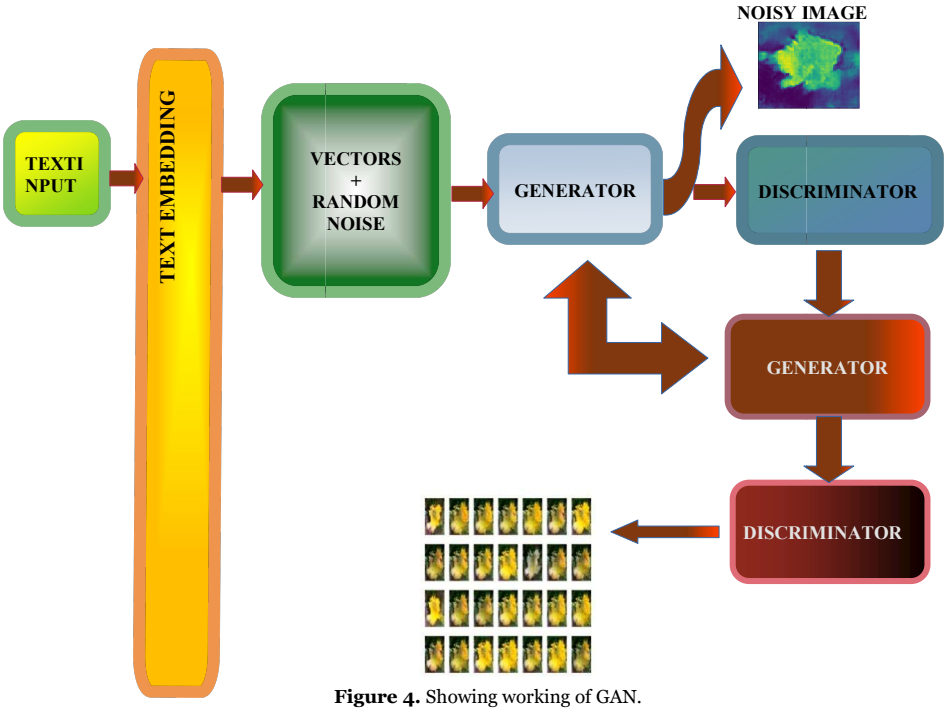

Figure 4. Showing working of GAN.

The above figure 4 explains the working of GAN and how the results are obtained withthis methodology. Below are some results of flower images obtained from writing the text descriptions as shown in figure 5.

Figure 5. showing results of flower images obtained after text input.

The above process includes 2 stages: -

Stage 1: - Text embedding :
A piece of text which is projected in a high dimensional space. To calculate embeddingof the target and store it an indexed database. Each word in a text is assigned some vector space, and this vector space is clearly understood by our model. Using word2vec, we canobtain the embedding vector, and we will get integer encoded data.During the stage of conditioning augmentation, some Random noise is added in GAN, with the embedding vector and a d-dimensional array will be obtained.

Stage 2: - Generator :
The generator generates image datasets by taking random noise and embedded vector as input. The generator generates output which contains noisy image which mimics the text description. The generated samples contain losses and it is classified as: -

$$J = -1/m \sum_{i=1}^{m} \log D(G(z)) \quad (1)$$

where J = is used to measure how the generator fools the discriminator.
Log D(G(z) represents probability of discriminator.

Stage 3: -Discriminator :
The discriminator in GAN plays an important role in classifying the generated image as real or fake in the binary from 0 to 1. By the time, discriminator is trained in such a way that it is able to differentiate easily between the real image as per textual description of fake image. There are losses associated with the discriminator, and these losses helps to categorize the samples as fake (log(1-D(G(z)) close to 1 and samples which are real as (log D(x)) close to 1. The ability of discriminator to differentiate between fake and real samples are given: -

$$J = -1/m \sum_{i=1}^{m} \log D(x) - 1/m \sum_{i=1}^{m} \log(1 - D(G(z)) \quad (2)$$

where J = ability of discriminator to differentiate the image generated
log D(x) means the real image as per text input
log(1- D(G(z)) means the fake image not as per text input.

6. Challenges

There are various challenges, which is offered in this area, such as accuracy challenge, able to obtain more diverse and desired images, user friendliness, able to obtain the results according to the text and extract the exact meanings out of the text descriptions on the image as well.

6.1 Accuracy
This is one such challenge that exits in this kind of project, where the image is sometimes not much accurate, and according to the text descriptions. Therefore, sometimes the desired results are not able to be obtained here.

6.2 Image Quality
This is another issue that exist here, the pixels and quality of image are sometimes not clear and desired ones. Therefore, sometimes the images which are obtained are blurred.

6.3 Less Diverse Datasets
This is another issue where due to less diverse datasets and low-quality datasets; the users are not able to get the desired results, and sometimes the images are not obtained as desired.

6.4 Semantically Consistent
Semantically consistency is another issue, wherein, the images that are obtained are according to the descriptions. It is really important to understand the semantics behind the text descriptions, and understand each meaning of the texts, so as to obtain accurate and desired images.

6.5 Model Accuracy
In this type of issue, the model is sometimes not able to obtain the good results. Therefore, it is important to deploy a good model. And due to large varieties of models such as types of GANs, stable diffusion, pytorch and other methods, this sometimes leads to a lot of confusions.

7. Conclusion and Future Scope

7.1 Use of Machine Learning and Deep Learning
Machine learning algorithms offer incredible learning potential. These features can be used to make the text to image synthesis more exact and with reliable results. Machine learning and deep learning models can help in deploying more varieties of models, and thereby it gives us many options to apply various models to have a good modification in the function of this project.

By using more diverse data and processing them, one can obtain good results. Machine learning algorithms also offer best results for analyze the performance of the model and datasets.

Deep learning model plays an important role in synthesis of image from text. Through the use of Various Deep learning models such as stable diffusion, Generative adversarial networks, one can able to obtain various types of images.

7.2 Accuracy
Through the above model of GAN, it can be concluded that, GAN is used to produce the best results. But the model depends upon the variety datasets it has been trained to, therefore the more diverse datasets and accurate models, the model will produce an accurate and best results. There are number of Models used based on this method, where results are quite good such as stable diffusion method, Style GAN, Attn GAN etc.

7.3 Quality of Images and Approach
The Quality of image and also the variety that we want to bring into the image depends on the type of models we are using. Obtaining complex scenes with interacting objects is still very difficult task. Quality of text embeddings is an important part. Also, the diverse datasets, the variety in the datasets can be a good part, for having better representations.

The goal of this research Exploring multimodal Leaning: text conditioned image generation and summarization is to explore and evaluate recent research based on this and propose a unique way of preprocessing the images and deploying easiest model and by using text embeddings and glove vectors, obtaining a best variety of images.

The report is all about an approach taken in the development of the above text to image synthesis-based model, wherein using GAN, we are able to obtain good results at the end, obtaining a fine-grained image. At the end, considering the scope of the text to image-based synthesis, there are lots of improvement that can be done in this domain. There are still some challenges presents in this area, where variety of images, clarity in the images, and accuracy in the images according to the text descriptions is still lagging.

Acknowledgment

The authors express the gratitude towards Indira Gandhi Delhi Technical University for Women for the opportunity for this research.

References

[1] Gal2022AnII, An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, Rinon Gal and Yuval Alaluf and Yuval Atzmon and Or Patashnik and Amit H. Bermano and Gal Chechik and Daniel Cohen-Or (2022).

[2] li2019controllable, Controllable Text-to-Image Generation, Bowen Li and Xiaojuan Qi and Thomas Lukasiewicz and Philip H. S. Torr},(2019).

[3] qiao2019mirrorgan, MirrorGAN: Learning Text-to-image Generation by Redescription, Tingting Qiao and Jing Zhang and Duanqing Xu and Dacheng Tao (2019).

[4] ramesh2021zeroshot, Zero-Shot Text-to-Image Generation, Aditya Ramesh and Mikhail Pavlov and Gabriel Goh and Scott Gray and Chelsea Voss and Alec Radford and Mark Chen and Ilya Sutskever (2021).

[5] ramesh2022hierarchical, Hierarchical Text-Conditional Image Generation with CLIP Latents, Aditya Ramesh and Prafulla Dhariwal and Alex Nichol and Casey Chu and Mark Chen (2022).

[6] AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks, Tao Xu and Pengchuan Zhang and Qiuyuan Huang and Han Zhang and Zhe Gan and Xiaolei Huang and Xiaodong He (2017).

[7] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang and J. Shao, "Semantics Disentangling for Text-To-Image Generation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 2322-2331, doi: 10.1109/CVPR.2019.00243 (2019).

[8] H. Zhang, J. Koh, J. Baldridge, H. Lee and Y. Yang, "Cross-Modal Contrastive Learning for Text-to-Image Generation," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021 pp. 833-842 (2021).

[9] Y. Zhou et al., "Towards Language-Free Training for Text-to-Image Generation," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 17886-17896, doi: 10.1109/CVPR52688.2022.01738 (2022).

[10] Ramzan, S.; Iqbal, M.M.; Kalsum, T. Text-to-Image Generation Using Deep Learning. Eng. Proc. 2022, 20, 16 (2022).

[11] Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, Andreas Dengel, Adversarial text-to-image synthesis: A review, Neural Networks (2021).

[12] T. Hinz, S. Heinrich and S. Wermter, "Semantic Object Accuracy for Generative Text-to-Image Synthesis," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 3, pp. 1552-1565, 1 March 2022, doi: 10.1109/TPAMI.2020.3021209(2022).

[13] Amod Jog, Aaron Carass, Snehashis Roy, Dzung L. Pham, Jerry L. Prince, Random Forest regression for magnetic resonance image synthesis, Medical Image Analysis.

[14] L. Wang, W. Chen, W. Yang, F. Bi and F. R. Yu, "A State-of-the-Art Review on Image Synthesis With Generative Adversarial Networks," in IEEE Access, vol. 8, pp. 63514-63537, 2020, doi: 10.1109/ACCESS.2020.2982224 (2020).

[15] Santiago Gonzalez, Mohak Kant, Risto Miikkulainen,15 - Evolving GAN formulations for higher-quality image synthesis, Robert Kozma, Cesare Alippi, Yoonsuck Choe, Francesco Carlo Morabito, Artificial Intelligence in the Age of Neural Networks and Brain Computing (Second Edition), Academic Press (2023).