Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 8, 2020

# GENERATING IMAGES USING DESCRIPTIVE CAPTIONS VIA ADVERSARIAL TRAINING

**Pranjal Jain[1], Tanmay Jayaswal[2]**

[1,2]Computer Science and Engineering Department,
SRM Institute of Science and Technology KTR, Chennai, India

*Received: 20.05.2020* *Revised: 17.06.2020* *Accepted: 04.07.2020*

**Abstract**

We present an approach to tackle one of the most popular problems in the deep learning community of Text to Image synthesis by using the highly effective generative adversarial networks (GANs). We modified the original GAN architecture to take inputs of descriptive caption and render images that are visualized form of the caption. The architecture utilizes multiple stages of deep attention networks to get high detailing for the various sub-regions of the image based on the usage of words and its association with other parts of the image. The system also uses a multistage communication methodology that helps in increasing structural coherence and render better word association. Our approach trains on and experiments on the Caltech USCD Birds-200 dataset and gets an inception score of 4.21 ± 0.05 with faster training and execution time.

**Keywords--** Generative adversarial networks, Text to image synthesis


DOI: http://dx.doi.org/10.31838/jcr.07.08.328

## INTRODUCTION

The conversion of descriptive caption or sentence to an image is a very useful technique that can have a wide variety of use cases in many industries.

The system can be utilized but isn't limited to, CAD applications, Fashion Designing, designing equipment, and art generation. The visualization of the image based on the caption using this system can reduce the work of the developer by automating the task of synthesizing the target via deep generative models.

Most of the recently proposed systems for Text to image synthesis utilize Generative adversarial networks or GANs for achieving significant results. This is due to the unique nature of GANs in how they excel in image generation tasks and the proficiency in the generation of realistic output which seems indistinguishable from real samples.

GANs achieve this level of accuracy by contesting two neural networks, a generator, and a discriminator, against each other. The generator renders images and the discriminator verifies if the image is real or generated. When they are trained, the generator is never exposed to the dataset, making it less likely to overfit the dataset (as seen in various auto-encoders) and also learns features and attributes that might be lost while training using existing systems.

Initially, the generator isn't able to synthesize images of target very well, but over time and after several steps of training they become proficient, till they beat the discriminator and generate an image with very high accuracy and high detailing.

This precise nature of realistic target output via this adversarial training also makes it versatile enough for learning a range of targets with very high accuracy, from image synthesis to style transfer and more, making it suitable for text to image generation. The original GANs architecture uses a noise vector as input, this has been modified to generate specific outputs by using a conditional input based GANs called conditional GANs [9].

Though conditional GANs are very powerful, the modified versions of conditional GANs for text to image synthesis are not sufficient enough for generating a high resolution and highly detailed image. This is due to the complex associations of text, which is in natural language, and the visualized image to which it corresponds. The image can be enhanced via attention layers on each generator and make the image much more realistic and reduce defects or inconsistencies in synthesis as we have attempted to do in this paper.

## RELATED WORK

The conversion of text description to image is a very challenging task, it requires a multi-disciplinary utilization of deep learning. First, the text needs to be understood via natural language processing.

Then there has to be an association of the words in context from the caption and how to visualize them into image sub-portions. The image sub-portions have to be converted into one image where each attribute is placed in just the right position to give perfect meaning to the image, while at the same time minimizing loss.

The current methodologies of text to image generation are proficient in generating high-quality images and of significant resolution. The utilization of a multistage GAN approach, where they stack multiple GAN pairs in a sequence to increase the size and detailing as seen in [2][3], has proven effective in increasing resolution, but still lacks structural coherence and utilization of multi-attribute association.

The use of multiple GANs leads to an increase in ways a defect can creep into the image. The methods like [2][3] use two stages of GANs (which can be modified for more stages) to render various defects due to the multi-stage architecture.

The lack of communication between the stages has made it oblivious to the higher layers of what the various attributes are and how should they be detailed. This results in attributes that might be necessary for image visualization to become defects that reduce the image's accuracy concerning the caption.

These insufficiencies are tackled by [4][5][7]by adding communication between the stages but still aren't able to give properly minimize the structural defects.

**PROBLEM DEFINITION AND CHALLENGES**

The task of "text to image" bases upon the generation of an image that successfully consists of what the text is defining accurately with full fidelity. This system needs to understand what the text is describing and at the same time understand where all the features have to be placed according to each other.

That means that if we are generating a bird, the tail cannot be on the head and the wings cannot be more than two in number and more such trivial aspects that are difficult for the generator to understand. The text has to be understood via natural language processing and has to generate high-quality images with no defects that would render the image incomprehensible.

The various challenges posed by this problem are:

1. GANs are very unstable to train, which causes it to be unpredictable. At any time the loss/accuracy can increase and decrease, which makes it uncertain if we are training it correctly.

2. The loss of dimension is very common in GANs, where they suffer in the inadequate representation of the depth in the image. The attributes seem to be flattened out. This is due to the lack of dimensional understanding from the 2-D image dataset.

3. The loss of frequency, or lack of association of the number of attributes. Here the generator doesn't understand how many features should be generated, i.e. four legs of a dog or six legs, two wings, or four wings.

4. The structural coherence, or loss of association between the image sub-regions and the coherence between the various attributes. The tail should be on the back of the bird, the wings on its side, and more should be properly rendered.

5. The proper utilization and visualization of the caption and rendering everything that it describes with maximum comprehension.

**PROPOSED METHODS**

**Dataset Evaluation**

The dataset that we use is the Caltech-UCSD Birds 200 dataset that encompasses a variety of bird species, majorly North-American.

It has over 11,788 images and 200 categories and each image have captions attributed to it for easily determining the image's elements and descriptions that explain the image given in a sentence form.

This makes it the perfect dataset for our needs. This Dataset is comprehensive and extensive for training a GAN as being extensive makes the discriminator very strict and the generator more accurate in our tasks.

The dataset is sanctioned by developers of the University of California San-Diego and Caltech University, making it quite reliable and easy to use for developers.

The dataset is small in size and has various files that are easy to utilize and compatible with the majority of computer environments.

The dataset is also updated regularly over 2-3 years and is popular for deep learning, computer vision, and natural language processing projects due to its known bias and defects.

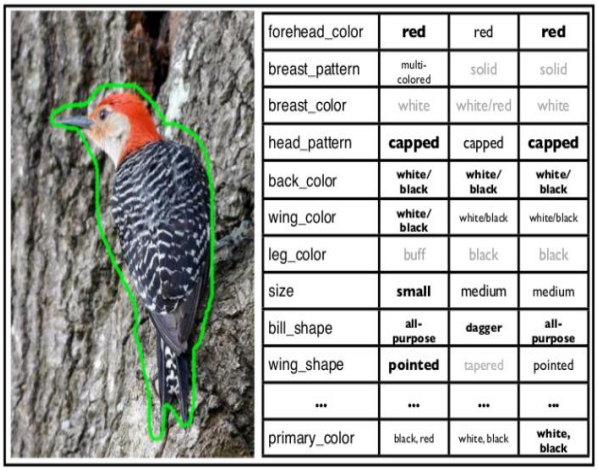



**Figure 1.** Image containing in the dataset and how it has been used to categorise. Source [12]

The data analysis on this has made it much easier to use and discovered a variety of feature extraction making the analysis useful for more efficient results.

As we can see the image has been segmented into a variety of layer filters and using the dataset's captions we can correctly determine the various instances of words it is describing and attributing to.

The words corresponding to the image's attributes help the attention layer in synthesizing realistic details with ease and the bounding box and rough segmentation provided by the dataset make it suitable for rendering.

**Architecture**

The architecture works by taking the caption or description as an input and then giving it to the text encoder. The text encoder then forms the caption into a word-level feature and a sentence-level context feature.

This feature is a conversion of the natural language into an N-dimension vector that helps associates the various words with each other. The word-level features help in identifying the most important words concerning the sentence feature vector and augments the attention layer to make better elements with high detailing.

The sentence feature is also provided to each generator so that it can form the image. Once the features are passed on to the attention layer which then subsequently passes it to the generator, the image of 64x64 is synthesized.

The image is then sent to the discriminator that identifies if the image is real or not. This process is repeated another two times to get an image of 256x256 resolution with high detailing. The higher layers communicate with all the layers below it and increase structural coherence.

The evaluation metric helps the system to understand how the image is associated with the caption and how many defects it has by giving a score to it.

We trained the system in the same way we train most GANs, by contesting them against each other and achieve convergence. All three pairs of GANs are contested together in one step so that the three stages can communicate with each other while generating an image.

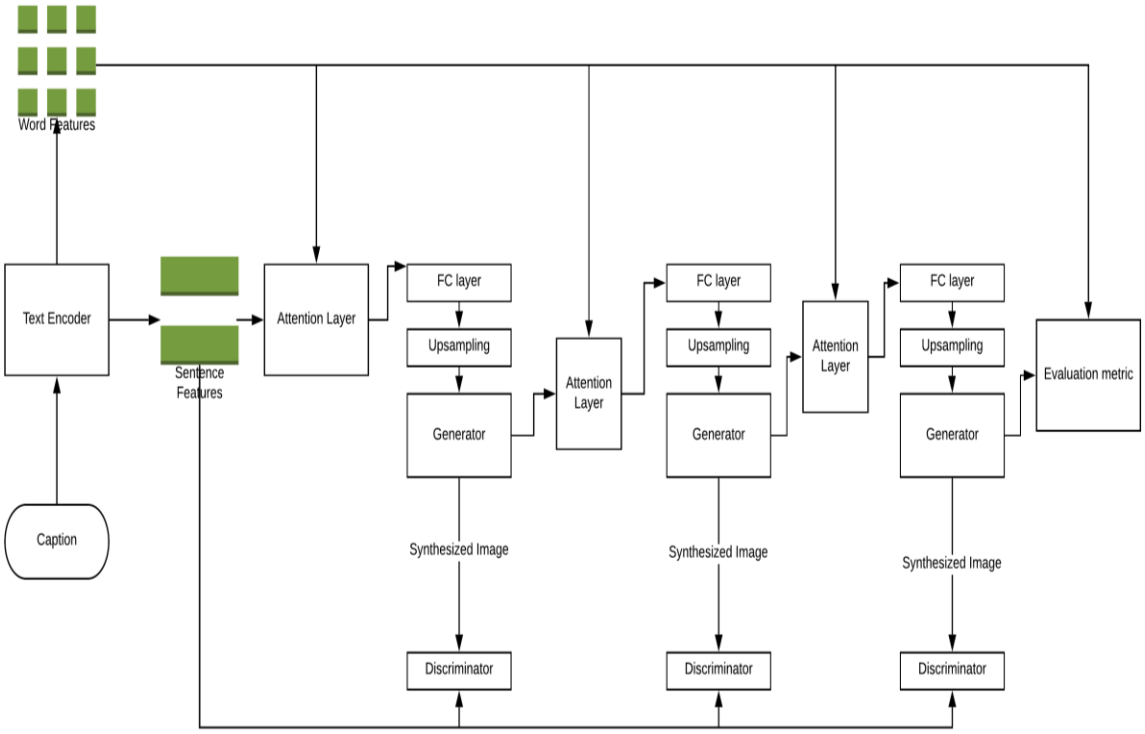



**Figure 2.** Architecture of the System to generate images from captions.

**Advantages of Proposed System**

The architecture we use is highly suitable as it breaks down the generation process into stages for better utility. The input in most conditional GANs is converted into images directly. This reduces the possibilities of using a wide variety of vocabulary and hinders the use of synonym words that is important as not everyone uses the same words to describe a particular attribute of the target. Users can describe a word like "Short" as "small", "bright" as "vivid" and more. To overcome this we convert the text into a word embedding which is an N-dimensional Space for word associations. This can then be used to provide for the generator. Stage one concentrates on feature generation rather than detailing and high-resolution generation and successfully generates images with a high frequency of features. This is then up-scaled using the later stages where detailing is focused and generates a higher quality image. The various stages have communication in them so they train all at once in each step and generate an output image. The communication helps in reducing defects caused by an incomplete understanding of the work done in previous layers and eliminates many of the defects. The defects are not eliminated, but we see increased accuracy.

With the use of an attention layer, we can generate multiple elements with great accuracy. Since it focuses on each and individual layer we can get more detailed sub-portions of the image and have better structural coherence. Using attention we focus on each image sub-portion and generate high-quality images. The system also uses a newer evaluation metric that can help get better accuracy by informing the generator how to amend its functioning while also considering caption.

**EXPERIMENT**

**Generation of images**

The GAN, once trained, can generate multiple images each trying to render the text description given from the user. Not all these images are going to be perfect so we handpick the best ones we require. This is since many of the images generated will have defects due to the black-box nature of the network and the limitations of the training conditions. This can however be improved with deeper networks but will require more training, more data, and much more sophisticated hardware. By working on the discriminators and making them stricter we can get many accurate results [11]. We do not see any permanent defects, and some of the images generated are of very high quality.

**Generation and Testing**

We gave the system the input *"A red bird with a short break on the tree"* and got the following results that were very promising, and successfully depicted the variety of features that were needed to be visualized from the input. We also saw various defects that had crept in the system i.e. depth perception, loss of frequency of features, and incorrect attribute visualization, but were found to be not permanent to all the images rendered. The system was tested multiple times on the same input and we found it to be not attenuated or suffering from mode collapse. [1]


**Figure 3.** Render from system for the input *"A red bird with short beak on tree"*

The image generated has adequately described the various features in the caption. We can observe that the object generated is in-fact a bird and is one of red colour. The beak is small though

this is assumed as the bird is facing us so we don't evaluate that. The tree is not generated properly but we can still notice tree-like features under the legs of the bird.

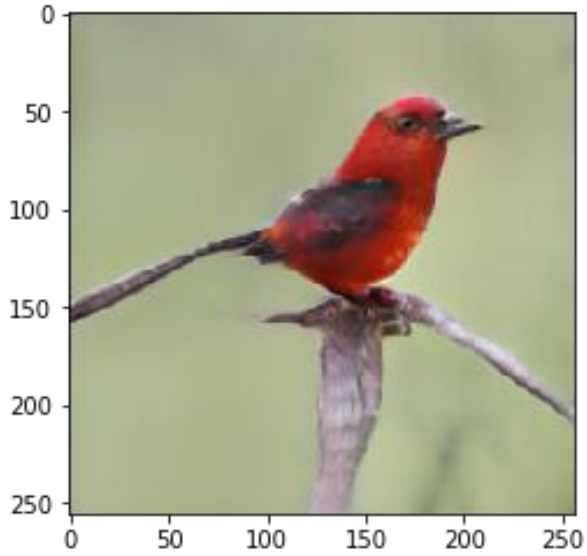



**Figure 4.** Render from system for the input *"A red bird with short beak on tree"*

The next render as we can see is more accurate as the bird is rendered with almost no defects. The eyes are prominent and the beak is short as demanded in the input. The colour of the bird is red and we can observe the various features necessary for it to be classified as a bird. The bird appears to be sitting on a tree and hence successfully checks off all criteria for it to be the perfect image as required by the caption.

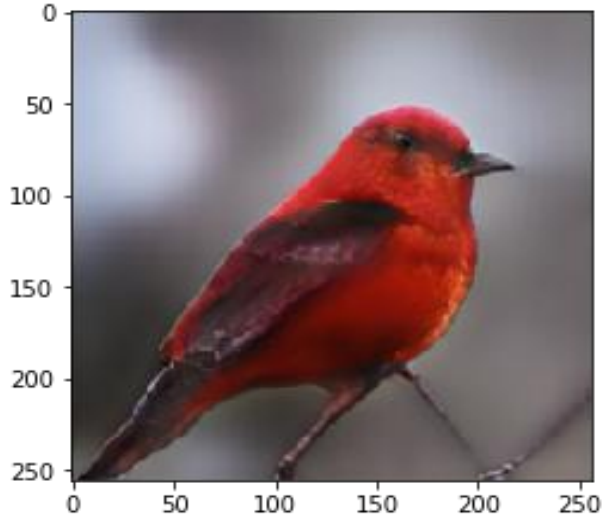



**Figure 5.** Render from system for the input *"A red bird with short beak on tree"*

This is also a perfect render of the input text. This has flawless detailing of the chest and the wings where we can see patterns of feathers of red colour. The eyes and beak are synthesized with perfection and the bird is sitting on the tree as demanded in the input. The same goes for the render below as well.

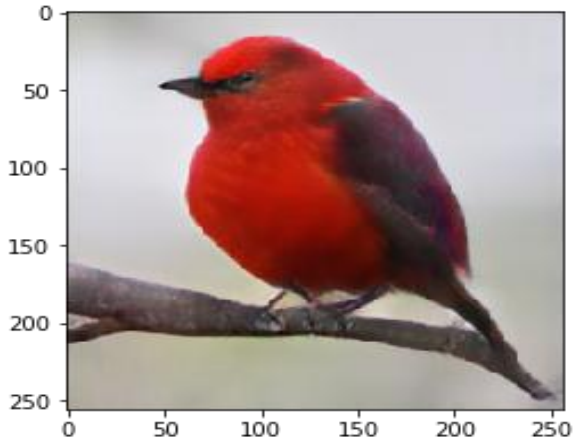



**Figure 6.** Render from system for the input *"A red bird with short beak on tree"*

Even though we got many successful renders we got many defected ones as well. Figure 7 is a failed render. We do not see permanent results like this one, which means the system still works properly.

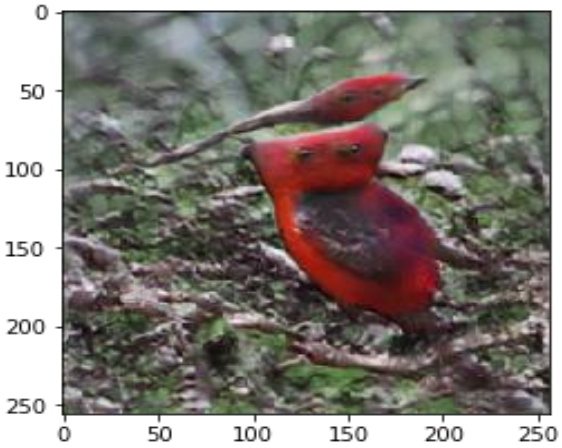



**Figure 7.** Defective Render from system for the input *"A red bird with short beak on tree"*

The next input we fed the generator was "A bird with blue wings". This is a simple input and the results were satisfactory.

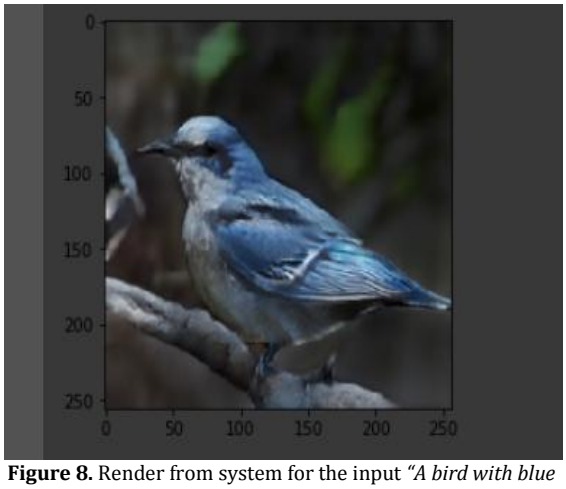



**Figure 8.** Render from system for the input *"A bird with blue wings"*

As we observe the image has perfect structural coherence and we can see all the elements in perfect harmony with the other i.e. all attributes are well placed and complement each other. The bird's wings are a blue colour and are properly painted making it very realistic. This image looks as if it was photographed in the real world as we can see various features like feathers and patterns on the bird making it very realistic.

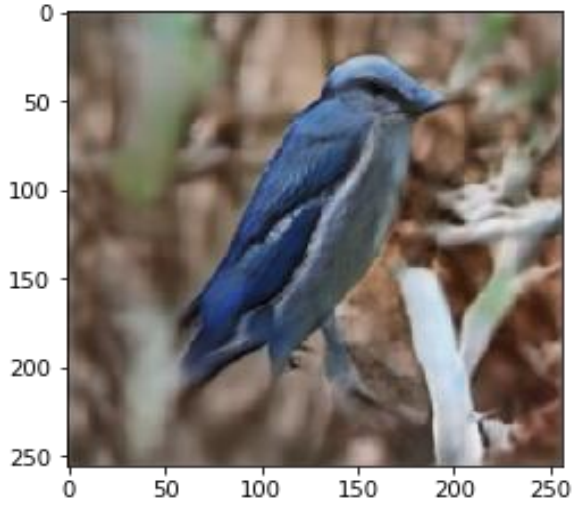
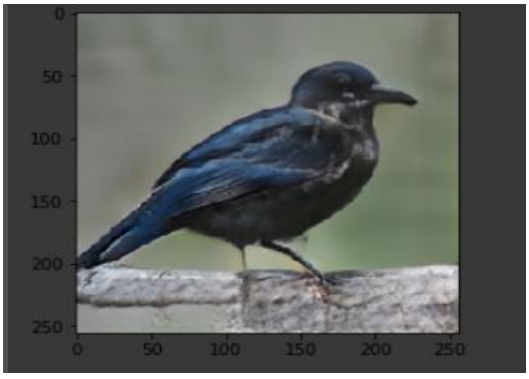


**Figure 9.** Render from system for the input *"A bird with blue wings"*

This is also a satisfactory render of the caption, though we can notice minor defects. The image still successfully renders the properly.

We gave many more inputs and got the following renders:



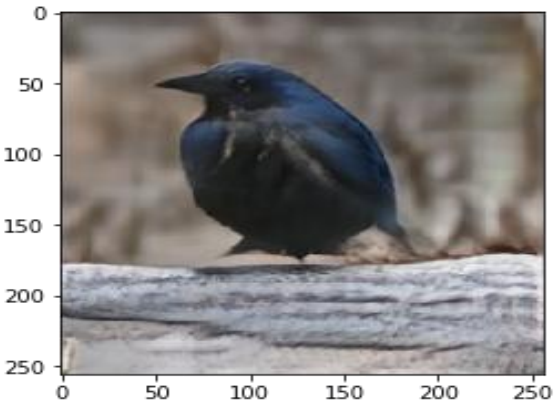


**Figure 10.** Render from system for the input *"A black bird with blue wings"*

**Figure 11.** Render from system for the input *"A red bird with blue beak and white wings"*

**RESULT**

The following table helps us understand our system as compared to the existing systems in terms of inception score.

| Dataset | GAN-INT-CLS [9] | GAWWN [8] | StackGAN [2] | StackGAN-v2 [3] | Our |
|---|---|---|---|---|---|
| **CUB** | 2.88 ± .04 | 3.62 ± .07 | 3.70 ± .04 | 3.82 ± .06 | 4.21 ± 0.05 |

**Figure 12.** Comparison of renders from other systems based on inception score

As we can see the system is proficient in detailing and creating a more accurate image compared to other models in the CUB dataset.

**FUTURE WORK**

The system can be modified to generate targets of a wide variety. Because it uses adversarial networks, it can be suitable for any kind of image, even with multi-object generation. The system can be used to generate designs for apparels and CAD applications and more, with only needing to edit the hyper-parameters and the dataset. The system can also be attached with more pairs of the GANs that can increase the resolution and detailing with further training as well. The system can still have increased accuracy by updating the evaluation metric based on the dataset used.

**CONCLUSION**

In this paper, we proposed a system that can synthesize images from descriptive captions using multistage generative adversarial networks. The system we utilize has three stages of GANs and can successfully depict the caption in a 256x256 image with minimal and non-recurring defects. The system was tested on the CUBs dataset and got an inception score of 4.21 ± 0.05 which is significantly better than many systems and trains faster as well. The extensive testing and experimentation have clearly demonstrated that the system is highly effective and can be modified for other tasks as well.

**REFERENCES**

1. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680.
2. H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in

Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5907– 5915.

3. Zhang, Han & Xu, Tao & Li, Hongsheng & Zhang, Shaoting & Wang, Xiaogang & Huang, Xiaolei & Metaxas, Dimitris. (2017). StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. PP. 10.1109/TPAMI.2018.2856256
4. T. Xu, 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, pp. 1316–1324.
5. Hinz, Tobias, Stefan Heinrich, and Stefan Wermter. "Semantic Object Accuracy for Generative Text-to-Image Synthesis." arXiv preprint arXiv:1910.13321 (2019).
6. Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset. http://www.vision.caltech.edu/visipedia/CUB-200.html
7. Zhu, M., Pan, P., Chen, W., & Yang, Y. (2019). Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5802-5810).
8. Reed, S. E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., & Lee, H. (2016). Learning what and where to draw. In Advances in neural information processing systems (pp. 217-225).
9. S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In ICML, 2016. 1, 2, 5, 7W
10. M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014
11. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems* (pp. 2234-2242).
12. https://vision.cornell.edu/se3/caltech-ucsd-birds-200/#:~:text=Caltech%2DUCSD%20Birds%20200%20(CUB,Caltech%2D101%2C%20etc).