International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

Generative Adversarial Network Architectures for Text to Image Generation: A Comparative Study

Rida Malik Mubeen[1], Sai Annanya Sree Vedala[2]

[1]Student, Dept. of Computer Science and Engineering, Muffakham Jah College of Engineering and Technology, Telangana, India
[2]Student, Dept. of Computer Science and Engineering, Chaitanya Bharathi Institute of Technology, Telangana, India

---

Abstract - GANs, Generative Adversarial Networks [17] which are conditioned on textual descriptions, are capable of generating images that are very realistic and can fool the mind into believing that these images are genuine. But in reality, the image has been generated by a model using what is described to it. The multimodal task of generating an image from a text description is a very challenging task. In this study, we discuss the different methods and approaches to create realistic images that fit the corresponding textual descriptions. The techniques include various state of the art models as well as modified models that use neural networks, and the analysis of the performance of each method is shared and compared.

1. INTRODUCTION

The automatic synthesis of images from the text can be very useful and interesting to work on. The application is both practical and creative. The basic idea of this is to give a textual description to the computer, the computer will then generate an image that is meaningful and that contains what is textually described. It is difficult to include every feature, detail and object that is described, but the evolution of GANs over time has managed to prove the aforementioned wrong and has resulted in creating unbelievably realistic and meaningful results.

The text description is in the form of a sentence that contains what needs to be present in the image that is to be synthesised. For example, the sentence " A yellow sunset on the beach" gives the description of a beach during the sunset and the colour of the picture is yellow. The computer produces a realistic image which is that of a beach image of the said kind. Now, this beach most probably does not exist in the same manner that is present in the produced image. It is merely a creation of the computer using the description and by learning from the images from the used dataset.

Generally, the quality of the produced image largely depends on the quality of the images that are used to train the model. In most cases, if the dataset has a large number of images and each type of instance or object has multiple samples, the computer has a greater chance at a realistic and high-quality image. Learning from many images of a particular topic helps the model to create images of that topic with better performance. However, we also see that by changing the architectures and working on the same dataset, we obtain a vast range of varied metric values, proving that different architectures can majorly impact the results produced by GANs. Current AI models have not yet mastered the task of automatic synthesis of realistic images but neural network architectures and GANs have been successful in creating compelling images of different categories [16]. Much work is still left to be done to have a more general approach on this front.

2. GENERATIVE ADVERSARIAL NETWORKS

GANs Generative Adversarial Networks are an approach for generative modelling that uses deep learning methods. Generative modelling is a type of unsupervised learning task which includes discovering and learning the patterns in the data to generate output. GANs are a great way of dealing with this task, it has two models, the generator model and the discriminator model. The generator model is trained to create new examples and the discriminator model classifies the examples as real or fake. Real examples are from the dataset and the fake examples are the created ones. They are trained till the generator is fooled for half the time, this means that the generator is generating new and realistic examples. Simply put, GANs are a deep learning-based generative model.

Generative Modelling is an unsupervised task but using GANs, the architecture allows for training the generator model as a supervised learning problem. It is trained together with the discriminator model. The generator generates a batch of samples and the discriminator classifies all the samples as real or fake. It is then updated to perform better at classifying the samples and the generator is updated to produce images that are better at fooling the discriminator.

2.1 Generator Model

The generator takes an example from the domain and predicts the output as a binary class label of real or fake. The real example is from the training dataset and the fake examples are the ones generated by the generator model. The Generator Model is discarded after the training is completed. It can be used for other purposes later on like feature selection and extraction.

2.2 Discriminator Model

The Discriminator is primarily a binary classifier that discriminates whether the input sample is real or fake. It gives the scalar value 1 as the output if the input image is real and a scalar value 0 as output if the input image is fake. Samples that are generated by Generator are termed fake samples. The samples from the domain or the training dataset are termed real samples. The main objective of the whole process is to fool the discriminator. The generator should be able to fool the discriminator with the images that it generates as output. The role of the discriminator is no longer needed after training and it is discarded after completing training or can be updated to perform tasks for the model.

3. DATASETS

In this study, we have analysed the performance of multiple Generative Adversarial Network architectures on the Common Objects in Context dataset and the Caltech UCSD Birds dataset.

3.1 Common Objects in Context Dataset

The dataset, COCO common objects in context [9], was created by Microsoft with the goal of advancing the state of the art in object recognition. It focuses on the context of the broader question of scene understanding for object recognition. This was achieved by collecting images of everyday scenes that are complex and contained objects in a natural context. It contains photos of 91 object types and with a total of 2.5 million labelled instances in 328 thousand images, the dataset was contributed largely by crowd worker involvement. It is a large scale dataset that addresses the core research problems in scene understanding.

A large set of images containing contextual relationships and non-iconic object views was harvested by using a simple and effective technique which queries for pairs of objects in conjunction with images retrieved via scene-based queries. Each image is labelled as containing particular object categories using a hierarchical approach for labelling. For each of the categories, the individual instances were labelled, verified, and finally segmented. The dataset is significantly larger in the number of instances per category when compared to other datasets with similar objectives.

3.2 Caltech UCSD Birds Dataset

The CUB Caltech UCSD Birds dataset [11] is widely used for tasks such as visual categorisation. It contains 11,788 images. Out of the total images, 200 subcategories belong to birds, 5,994 for training and 5,794 for testing. Each of the images has detailed annotations. they are as: 1 subcategory label, 15 part locations, 312 binary attributes and 1 bounding box. 10 single sentence descriptions are collected for each image. The natural language descriptions are collected that have at least 10 words, without any information of subcategories and actions.

4. MODELS

In this study, we have included and analysed the following GAN architectures:

- Visual Information Captured Text Representation
- Mirror Generative Adversarial Network
- Dynamic Memory Generative Adversarial Network
- Control Generative Adversarial Network
- Attention Generative Adversarial Network + Object Pathway

We throw light on the aforementioned Generative Adversarial Network Architectures and highlight the importance of each of the architectures and finally, compare the different architectures on the basis of three metrics. Namely -

- Inception Score
- Frechet Inception Distance
- R-Precision

4.1 Visual Information Captured Text Representation

VICTR: Visual Information Captured Text Representation for Text-to-Image [17] - Multimodal Tasks captures rich visual semantic information [12] of the objects from the text input. The text description is used as the input and dependency parsing is conducted to extract synthetic structure and then analysed to obtain a scene graph. The extracted features are trained using Graph Convolutional network to generate text representation. There is aggregation with word level and sentence level embedding to generate visual and contextual word as well as semantic representation. The method includes five modules: Text to scene graph parsing, Scene graph embedding, Positional graph embedding, Visual semantic embedding, and Visual contextual text representation.

First, the scene graphs from the input description are extracted to define the object, attributes and relations from the image. Using dependency parsing and transformer-based object attribute relation classification, the scene graphs are generated. Then the extracted object, attribute, relation are trained using GCN to generate text representation that is visual contextual and then finally it is aggregated to word level and sentence level embedding to generate visual contextual word representation and along with it visual contextual sentence representation as well. Additionally, we see that StackGANs [14] can also be combined with VICTR to produce efficient results.

4.2 Mirror Generative Adversarial Network

MirrorGAN [6], The method focuses on learning text to image generation by redescription, which is a global-local attentive and semantic preserving text to image to text

framework. It includes three modules as (1)Semantic text embedding module that generates word and sentence level embeddings, (2)Global local collaborative attentive module for cascaded image generation and (3)Semantic text regeneration and alignment module.

Mirror structure is an integration of both T2I and I2T. After an image is generated, the model regenerates its description which aligns its underlying semantics with the given text description. During each stage of training, the generator and discriminator are trained alternately. The discriminator is trained alternately to avoid being fooled by the generators by distinguishing the inputs as real or fake. Two adversarial losses are employed: visual realism adversarial loss and text image paired semantic consistency adversarial loss. MirrorGAN has three generators and GLAM is employed over the last two. A pre-trained bidirectional LSTM is used to calculate semantic embedding from the text description. The model can generate more diverse images of better quality, along with semantic consistency with the input text description.

4.3 Dynamic Memory Generative Adversarial Network

In Generative Adversarial Networks, due to the process of multi-stage generation, the final resolution of images may sometimes be compromised. Two main hindrances can be observed in the multi-stage image generation [15] process. First, the quality of initial images ie., images that are generated in the first stage tend to highly influence the result of generation. Second, not every word in the given input sentence will carry the same weight ie., every word may have a unique level of content of the image it represents. Two major steps govern the functioning of a DMGAN [4]. These are Image Generation and Dynamic Memory Based Image Refinement.

Initial image generation is the first stage of the process that DMGAN undertakes. At first, a text encoder transforms the input text description into an internal representation. A deep conventional generator then predicts an initial image with a basic, rough shape and a couple of details according to the sentence feature and a random noise vector.

After the fuzzy and rough initial images are created, finer visual contents are added to those initial images to generate better photo-realistic images. The refinement stage is reiterated over and over again to obtain more substantial and important information and generate a high-resolution image with finer details. This process consists of the following components: Memory Writing – It stores the text information into a key-value memory that is structured for further convenient retrieval. Key Addressing - It is employed to read features from the module of the memory to refine visual features of the images of low quality. Value Reading- It is employed to read features from the module of the memory to refine the visual features of the same images that are of low quality. Response - It is used to control the fusion of the image features and the reads of the memory. Gated Memory Writing - Instead of considering only partial text information, the memory writing gate allows the DM-GAN model to select the relevant word to refine the initial images.

4.4 Control Generative Adversarial Network

The ControlGAN [7] can synthesise extremely high-quality images, and additionally allow the user to manipulate different attributes of objects. All this, without affecting the generation of any other content that is a part of the image. The ControlGAN comprises three components mainly. It is essentially a channel-wise attention-driven generator and a word-level spatial generator, where an attention mechanism is used to permit the generator of the GAN to synthesize subregions that are related to the most relevant words. The backbone architecture is the multi-stage AttnGAN [10]. This model has a spatial attention module and a channel-wise attention module. The major job of the spatial attention module is to correlate words with individual spatial locations without taking any other channel-related information into perspective. The channel-wise attention module on the other hand uses the connection between words and channels.

The second part is a word-level discriminator, where the correlation between words and the subregions of images is utilised to understand and differentiate different visual attributes, which can provide the generator with fine training signals. The discriminator should provide the generator with good training feedback in order to encourage the generator to specifically modify parts of the image according to the text given. This can guide the generation of subregions corresponding to the most relevant words. A global average pooling layer is adopted by the text-adaptive discriminator to output a 1D vector as an image feature. It then calculates the correlation between image features and every word. Due to this, the image feature may lose important spatial information. Control GAN solves this issue.

The third component deals with the usage of a perceptual loss in text-to-image generation. This can enforce the generator to preserve the visual appearance and reduce the randomness involved in the generation of the text. The generated results can be highly random without adding any constraint on text-irrelevant regions. They may also fail to be semantically consistent with other content. To deal with the same, the perceptual loss is adopted based on a 16-layer VGG network pre-trained on the ImageNet dataset.

4.5 Attention Generative Adversarial Network + Object Pathway

The AttnGAN + Object Pathway [3] method allows to control the location of arbitrarily many objects within an image by adding something called an object pathway to both the generator and the discriminator in addition to the global pathway. It requires only bounding boxes and the respective labels of the desired objects. The main objective is to have objects generated at arbitrary locations within a scene while keeping the rest of the scene overall consistent. The generator constructs labels for the individual bounding boxes from the image caption ϕ and the provided labels of each bounding box. There are two parts to this model: Generator: Firstly, the generator consisting of the global pathway is responsible for creating a general layout of the global scene. It processes the previously generated local labels of the bounding boxes and replicates them spatially at the location of each of the bounding boxes. Secondly, the object pathway is responsible for generating features of the objects within the given bounding boxes. This pathway creates a feature map of predefined resolution using the convolutional layers which receive the previously generated label as the input. This feature map is further transformed with a Spatial Transformer Network (STN) to fit into the bounding box at the given location on an empty canvas. Discriminator: The discriminator also possesses both a global and an object pathway respectively. The global pathway takes the image and applies a collection of convolutional layers to obtain a representation of the whole image. An STN is first used by the object pathway in order to extract the objects from the given bounding boxes. It then concatenates the spatially replicated bounding box label with the extracted features. Finally, the outputs of both the object and global pathways are concatenated along the channel axis and we again apply convolutional layers to obtain a merged feature representation. This model can, however, also lead to suboptimal images if there are no bounding boxes for objects that in reality should be present within the image. This can often be the case if the object is too small (less than 2% of the total image) and is therefore not specifically labelled. Sometimes, here, the objects are not modelled in the image at all, despite being properly visible and vivid in the corresponding image caption, since features are not generated by the object pathway.

5. EVALUATION METRICS AND ANALYSIS

5.1 Inception Score

An Inception v3 Network is used by Inception Score [18] which is pre-trained on an ImageNet and it calculates a statistic of the outputs of the network when it is applied to generated images. This metric is utilized in order to automatically evaluate the quality of image generative models. This Inception-v3 Network is a 48 layer-deep convolutional neural network. A pre-trained version of the network can be trained on 1.2 million RGB images from the ImageNet database and it can classify images into 1000 categories, these categories include objects such as keyboards, cats, pencil, mouse and many animals. Consequently, the network has achieved the learning of rich feature representations for a wide variety of images. The image input size of the network is 299-by-299. Being one of the most widely used networks for transfer learning, the pre-trained models of the Inception v3 network are available in most deep learning software libraries. This model mainly looks for two desirable qualities. First, the images generated should contain distinct and clear objects. Secondly, it should be able to output a high diversity of images. The IS is seen to correlate very well with the human judgment of image quality. A higher Inception score is considered more desirable.

5.1.1 Analysis on the CUB dataset

The best IS score for the generation of text to image using the CUB dataset is for the DMGAN model with a score of 4.75 ± 0.07. The DMGAN solves two major issues in this domain. Firstly, the generation result depends heavily on the quality of the initial images. Secondly, each word in an input sentence depicts a different level of information of the content of the image. The DMGAN solves the first issue by adding the key-value memory structure to the GAN framework. The fuzzy image features of the images generated in the initial stage are taken as queries in order to read features from the memory module. The memory reads are then used for the purpose of refining the initial fuzzy images. To solve the second issue, a memory writing gate is introduced to dynamically select the words that are relevant to the generated image. This makes the generated image well conditioned on the text description. Additionally, a response gate is used instead of directly concatenating the image and the memory. This is used to adaptively receive information from images and memory. The architecture followed in order to receive such a high IS score within the DMGAN basically comprises the concepts of dynamic memory, memory writing gate, and response gate. Followed by this, with an IS score of 4.58 ± 0.09 is the Control GAN. It is the word-level spatial and channel-wise attention-driven generator, where an attention mechanism [8] is exploited to allow the generator to synthesize subregions corresponding to the most relevant words. The correlation between the image subregions and the words is used to identify different visual attributes. This helps to provide the generator with fine training signals related to the visual attributes. This adoption of the perceptual loss in text-to-image generation helps to reduce the randomness that comes with the generation, and this causes an enforcing to the generator

which preserves the visual appearance that corresponds to the unmodified text.

5.1.2 Analysis on the COCO dataset

The highest IS for the COCO dataset is the DM-GAN with VICTR [1] with a score of 32.37±0.31. DMGAN alone gives an IS of 30.49±0.57 which proves to be higher than the IS given by Control GAN and Attn GAN ie., 25.89±0.47 . The Control GAN however has a lower IS than MirrorGAN with an IS of 26.47±0.41. Above these, the AttnGAN with VICTR [1] framework has a score of 28.18±0.51 which stands next to the DM-GAN with VICTR.

5.2 Frechet Inception Distance

The Frechet Inception Distance (FID) [13] has become a standard measure due to its simplicity. It is also used frequently in the analysis of conditional generators. It is one of the simplest metrics and it is based on an Inception embedding and it is a particularization of Wasserstein distance to the simple case of multivariate normal distributions. The Wasserstein metric is considered a true probability metric, it considers the probability of various outcome events and also the distance between them. Unlike other distance metrics like KL-divergence, Wasserstein distance provides a meaningful and smooth representation of the distance between distributions. The above-mentioned properties tend to make the Wasserstein suited to domains where there is higher importance given to underlying similarity than exactly matching likelihoods in the outcome. In Frechet Inception Distance, the Inception network is used to extract features from a layer that is intermediate. Then the data distribution is modelled using a multivariate Gaussian distribution with mean μ and covariance Σ. FID is more robust to noise than IS. If one image is generated by the model per class, the distance will tend to be on the higher side. Thus, FID is considered a better measurement for the diversity of images. The FID score denotes the similarity of two considered groups in terms of the statistics on computer vision features of the raw images which the inception v3 model calculates for image classification. So naturally, similarity will be denoted by lower scores ie., lower scores indicate the two groups of images are more similar or have similar statistics. A perfect score is considered to be 0.0 indicating that the two groups of images are identical.

5.2.1 Analysis on the CUB dataset

The reduction in the FID from 23.98 to 16.09 is seen from AttnGAN [5] to DM-GAN. DM-GAN clearly provides a lesser distance or variation between the actual and the generated images. The baseline DM-GAN architecture has an FID of 23.32; the baseline architecture with dynamic memory has an FID of 21.41. When a memory writing gate is added, the FID falls to 20.83 and when a response gate is added, the FID drops to 20.83. This shows that in terms of the FID, even the very baseline architecture portrays a better performance than an AttnGAN. This is the case because, in DM-GAN, more fine-grained visual contents are added to the fuzzy initial images to generate a photo-realistic image. The refinement stage can be repeated over and over again to get more information and generate a high-resolution image with finer details.

5.2.2 Analysis on the COCO dataset

The reduction in the FID from 35.49 to 32.64 is seen from AttnGAN to DM-GAN. DM-GAN clearly provides a lesser distance or variation between the actual and the generated images. However, the FIDs of modified AttnGANS are also not as satisfactory as a DM-GAN. The FID of AttnGAN with VICTR is 29.26 and that of DM-GAN with VICTR also surprisingly is higher with a score of 32.37.

5.3 R-Precision

Consider a given query topic *Q*, the *R*-precision [2], as the name suggests, is the precision at *R*. Here, *R* refers to the number of relevant documents for *Q*. In other words, if there are *r* relevant documents among the top-*R* retrieved documents, then *R*-precision is r/R. Considering a ranked list of documents that are returned in response to a given query, the average precision is the average of the precisions of all the relevant documents. Approximately, this is the area under the precision-recall curve. The R-precision of that list is the precision at the rank of R, where R is the number of documents that are considered relevant to the query.

5.3.1 Analysis on the CUB and the COCO datasets

Just like in the case of IS and FID, we see that DM-GAN tends to have an upper hand even in the case of R-precision. The difference between the R-precision of an AttnGAN and a DM-GAN in the case of the CUB dataset is around 5 and the difference in the case of COCO is around 3. In both cases, the score of the DM-GAN is preferred over the AttnGAN. The above is the case of DM-GAN with dynamic memory, memory writing gate and response gate. However, the baseline model without the above three additions to the architecture still has a higher R-precision than an AttnGAN. The baseline score stands at around 68.64, followed by the baseline with the dynamic memory at 70.66 and the architecture of the baseline with dynamic memory and the memory writing gate with a score of around 71.40.

Table -1: Evaluation on the CUB dataset.

MODEL	INCEPTION SCORE	FRECHET INCEPTION DISTANCE	R-PRECISION
DMGAN	4.75±0.07	16.09	72.31±0.91
CONTROL GAN	4.58±0.09	-	-
MIRROR GAN	4.56±0.05	-	-

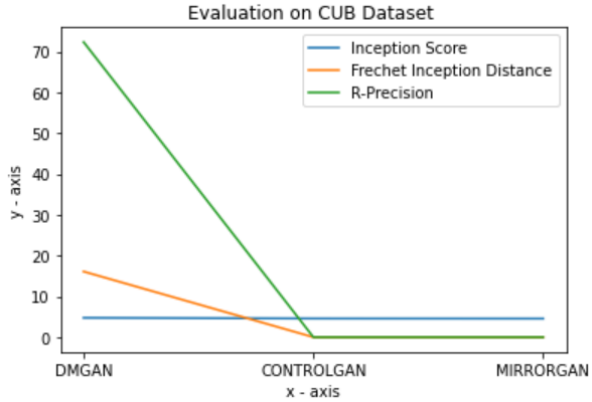

Fig-1: Visual representation of the performance of the architectures on the CUB dataset.

Table -2: Evaluation on the COCO dataset.

MODEL	INCEPTION SCORE	FRECHET INCEPTION DISTANCE	R-PRECISION
DMGAN	30.49±0.57	32.64	88.56 ±0.28
CONTROL GAN	24.06±0.60	-	-
AttnGAN+OP	24.76±0.43	33.35±1.15	-
StackGAN+OP	12.12±0.31	55.30±1.78	-
AttnGAN+VICTR	28.18±0.51	29.26	86.39±0.0039
StackGAN+VICTR	10.38±0.20	-	-
DM-GAN	32.37±0.31	32.37	90.37±0.00
with VICTR			63
AttnGAN	25.89±0.47	35.49	85.47±3.69
MirrorGAN	26.47±0.41	-	80.21

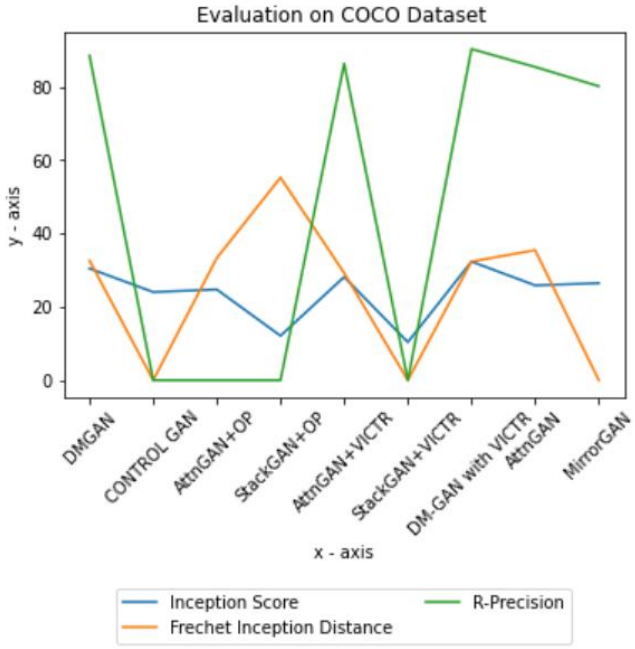

Fig-2: Visual representation of the performance of the architectures on the COCO dataset.

6. CONCLUSION

This study tries to present the current methods for text to image generation using Generative Adversarial Networks. The methods have been evaluated on two different datasets, the COCO dataset and the CUB dataset. The benchmark has been highlighted for its performance compared to various other methods using metrics Inception score, Frechet Inception Distance and the R-Precision. The study can be extended further to evaluation metrics other than the traditional metrics that can enhance the performance. The models can also be evaluated using other datasets from different domains to better understand the degree of their performance.

REFERENCES

[1] Soyeon Caren Han, Siqu Long, Siwen Luo and Kunze Wang, 'VICTR: Visual Information Captured Text Representation for Text-to-Image Multimodal Tasks' Retrieved from https://arxiv.org/pdf/2010.03182v3.pdf

[2] Aslam, Javed & Yilmaz, Emine & Pavlu, Virgil. (2005). A geometric interpretation of r-precision and its correlation

with average precision. 573-574. 10.1145/1076034.1076134.

[3] Tobias Hinz, Stefan Heinrich, Stefan Wermter, 'Generating Multiple Objects at Spatially Distinct Locations', Conference Paper at ICLR 2019.

[4] Minfeng Zhu, Pingbo Pan, Wei Chen, Yi Yang, 'DMGAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image-Synthesis' , Conference Paper at ICLR 2019

[5] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He, AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks, Conference Paper at CVPR 2018.

[6] Tingting Qiao, Jing Zhang, Duanqing Xu, DaCheng Tao, MirrorGAN: Learning Text-to-image Generation by Rediscription, Conference Paper at CVPR 2019.

[7] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, Phillip H.S. Torr, Controllable Text-to-Image Generation, Conference Paper at NeurIPS 2019

[8] Y. Kataoka, T. Matsubara and K. Uehara, "Image generation using generative adversarial networks and attention mechanism," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 2016, pp. 1-6, doi: 10.1109/ICIS.2016.7550880.

[9] Lin TY. et al. (2014) Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48

[10] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1316-1324

[11] Welinder P., Branson S., Mita T., Wah C., Schroff F., Belongie S., Perona, P. "Caltech-UCSD Birds 200". California Institute of Technology. CNS-TR-2010-001. 2010.

[12] Tobiaz Hinz, Stephan Heeinich, Stefan Wermterr, Semantic Object Accuracy for Generative Text-to-Image Synthesis, Conference Paper at CVPR 2018.

[13] Soloveitchik, Michael & Diskin, Tzvi & Morin, Efrat & Wiesel, Ami. (2021). Conditional Frechet Inception Distance. DOI:https://doi.org/10.1145/3446374

[14] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris Metaxas, StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks,

[15] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, Improved Techniques for GANS, Retrieved from https://arxiv.org/abs/1606.03498

[16] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, Yoshua Bengio, ChatPainter: Improving Text to Image Generation using Dialogue, Retrieved from https://arxiv.org/abs/1802.08216

[17] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng and F. -Y. Wang, "Generative adversarial networks: introduction and outlook," in IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 4, pp. 588-598, 2017, doi: 10.1109/JAS.2017.7510583.

[18] Shane Barratt, Rishi Sharma, A Note on the Inception Score, Retrieved from https://arxiv.org/abs/1801.01973