



Faculty of Informatics and Computer Science

Artificial Intelligence

Image Synthesis from Text using Generative Adversarial Network

By: Ahmed Tammaa

Supervised by: Associate Professor. Nahla Barakat

December 2021

Contents

Faculty of Informatics and Computer Science	1
1 Abstract	3
2 Introduction	3
2.1 Overview	3
2.2 Problem Statement	4
2.3 Scope and Objectives	4
2.4 Work Methodology	4
3 Preliminaries of Adversarial Text-To-Image	4
3.1 Background	4
3.2 Pure GAN Architecture	5
3.2 Conditional GAN (cGAN)	6
3.3 Text Encoding	7
4 Related Work (State-of-Art)	7
4.1 First Adversarial Text-To-Image Work	8
4.2 Stacked Generative Adversarial Network	9
4.4 Brief Cover of Interactive Text-to-image Approach	17
4.5 The Evaluation Challenge	18
5 Gannt Chart	20
Reference	21

1 Abstract

The Text-to-Image generation is an open research field aims to generate a photo-realistic image from a text description. The text input can vary from a single statement, the generic approach, and/or an interactive dialogue which makes the human and computer interact to modify and generate the image. The current results are far from perfect since the research has started in 2016 and the task is complex with various challenges, such as, a lack of a good metric. Furthermore, there is no discovered metric can give a single score to measure it. However, the current metrics can give higher scores to a synthetic image more than the ground truth image even if the synthetic image quality is remarkably low. Moreover, in the early stages of the research the generator was sacrificing the relevance and the quality to deceive the discriminator. Fortunately, the steps of the research solved this problem by enforcing the generator to generate a text-relevant image. The semantic and colours of the image are improved by introducing the attention-based GAN architectures.

2 Introduction

Humans are naturally visual; furthermore, listening to phrases, stories, or reading a text immediately an image is visualized in our head. This ability is proven to be key factor on humans' cognitive abilities [1]. On contrary, the computers are fast calculators. However, with the modern advances in artificial intelligence specifically the Generative Adversarial Network (GAN), the computers can form a form of visualization from a given text. This project aims to generate photo-realistic images that are consistent with a given text input. This process is the reverse of image captioning, moreover, the image captioning describes a given image while this project generates an image from text. This project is helpful in image editing and contributes to the early stages of creativity and imagination. Furthermore, an artist can describe what is in the mind and get a close image to describe what is in the mind or the early design stages. Besides, it can help designers to make quick modifications and facilitate communication with their customers. Finally, the text-to-image is used in image retrieval application.

2.1 Overview

The image generation is using Generative Adversarial Network (GAN). The GANs architecture is comprised of two deep learning models. The First one is the discriminator which is implicitly a classifier to differentiate between the real and fake image. The second is the

generator which tries to generate a realistic output to deceive the discriminator. Thus, the images generated must be as realistic as possible. However, this architecture generally has some undesired features, such as, there is no agreed metric for evaluation of GANs. In 2019, Barua et al [2] proposed cross local intrinsic dimensionality (CrossLID) as metric that is based on the manifold degree of coincidence between two data distributions. The CrossLID is proposed as sensitive to mode collapse image transformation, and robust to small-scale noise. However, this new evaluation metric is critiqued for scalability problems and clarity of the paper is questioned by the researchers. Hence, this measure still under research and it is not certain if it should be used or not. From other challenges of GANs is it takes very long time to train and at some case the GANs do not converges.

2.2 Problem Statement

Given a textual description for a scene generate a photo-realistic image that aligns with the given description with diversity of the output.

2.3 Scope and Objectives

Improve the quality of the generated image by improving preprocessing approach and/or by improving a network architecture.

2.4 Work Methodology

This paper attempts to generate a higher quality image by modifying the GANs architecture and/or improving the pre-processing.

3 Preliminaries of Adversarial Text-To-Image

3.1 Background

A lot of research has been conducted on the text-to-image synthesis using GANs. As its motivation is to synthesis a realistic input that can deceive a classifier to predict it as real. At some cases, the GAN networks can deceive humans as it has various applications like image super resolution, human face synthesis and many other applications. On the text-to-image context, the first research conducted on text-to-image was by Reed et al [3] by developing a deep convolutional GAN (DC-GAN) where the encoded text is the conditional factor. Furthermore,

both deep learning models, generator, and discriminator, are trained feed-forward with inference conditioned on the given text description [3].

3.2 Pure GAN Architecture

The GAN on the pure form is a deep learning model consisting of two neural networks which are the Generator and Discriminator. Where the discriminator is a binary classifier to differentiate the real vs fake inputs, while the generator tries to generate image that deceive the classifier by generating a realistic input. Formally, both neural networks can be represented by a zero-sum game. Moreover, each neural network tries to maximize its chance of winning the game and minimize their counterpart [4]. More formally, this architecture is represented mathematically by this equation.

$$\min_G \max_D V(D, G) = E_{x \sim p_{\text{data}}(x)} [\log(D(x))] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

The ‘D’ and ‘G’ Represent the Discriminator and Generator respectively. ‘E’ represents the mathematical expectation of the real data distribution $p_{\text{data}}(x)$ of the variation of binary cross entropy of ground truth 1,0 (real - fake) $p_z(z)$ are the noise variable. $G(Z; \theta_g)$ is denoting the differential function of the generator that is represented by Multi-Layer Perceptron (MLP) $D(X)$ is the discrimination function that denotes the probability of X were from the original dataset not from the generated dataset. Moreover, the generator tries to minimize $\log(1 - D(G(z)))$ and the discriminator tries maximizing it. Hence, the mathematical representation shows the idea of GAN clearly, moreover, other GAN architectures will be discussed is using a variation of this basic idea. Diagram 1 shows the architecture of the pure generative adversarial network

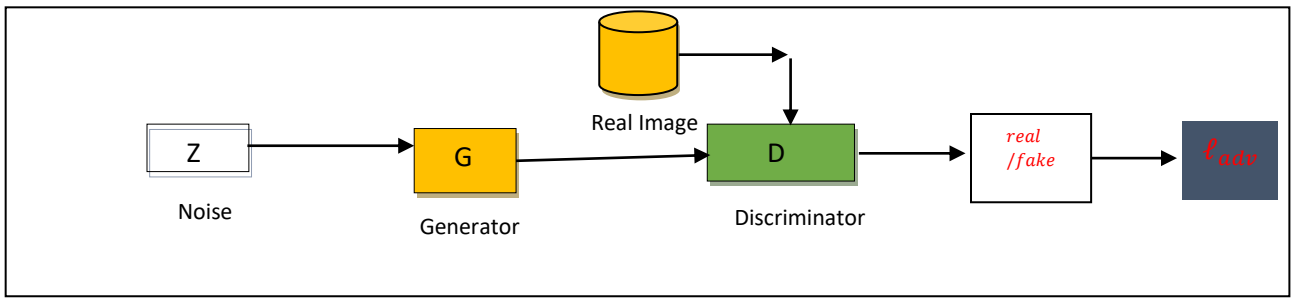


Diagram 1. Shows pure GAN

3.2 Conditional GAN (cGAN)

The GAN generation of image is very powerful approach, however, adding a control of the generated image is a great enhancer for the quality of the image. Mirza et al [5] introduced the conditional GAN (cGAN) which added new variable (y) that is class labels to have the generated image be conditioned on it. Formally, it can be represented in this formula.

$$\min_G \max_D V(D, G) = E_{x \sim p_{\text{data}}(x)} [\log(D(x|y))] + E_{z \sim p_z(z)} [\log(1 - D(G(z|y)))] \quad (2)$$

This Architecture is later extended with Auxiliary Classifier that is changes the discriminator to predict the class label of the classifier to check it instead of being given as an input [6]. This improvement has improved the stability and enhanced the quality of the generated image. This idea will enhance the relation between the generated image to the text description. Diagram 2 shows cGAN and Diagram 3 shows ACGAN.

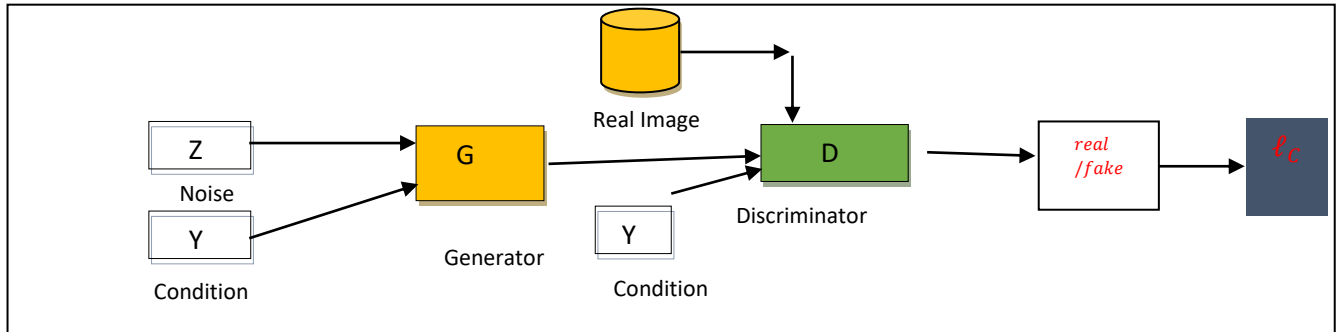


Diagram 2. Shows pure cGAN

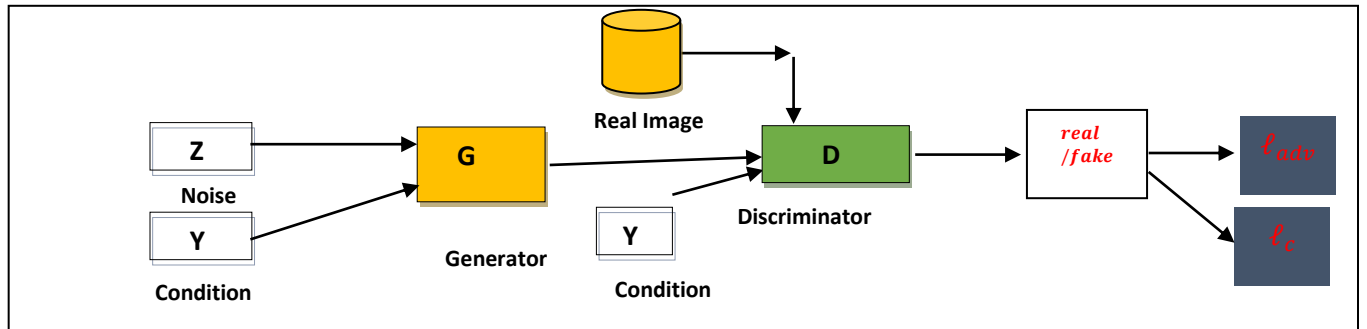


Diagram 3. Shows pure ACGAN with extra Auxiliary loss

3.3 Text Encoding

The first step in the text encoding of the text to image was proposed by Reed et al [3] on the first research by using character-level Convolutional Recurrent Neural Network (Char-CNN-RNN). Furthermore, it is trained neural network that can learn the relation between a text and image by a given set of labels. Moreover, the Reed et al [3] experimented other text encoding techniques such as Bag-of-Words and Word2Vec their experiment proved that Char-CNN-RNN was more effective in the text-to-image context. Zhang et al [7] proposed a better modeling way conditioning augmentation uses the covariance matrix and text embedding as functions for text embedding and samples latent variable from Gaussian distribution. This approach is regularized by the Kullback-Leibler Divergence (KL-Divergence) term among the training. This technique became used frequently for the text to image translation. Moreover, another approach was introduced Souza [8] by proposing the Sentence Interpolation similar to the conditioning augmentation it smoothen the text embedding throughout the training phase. The advances continue by proposing the bi-directional Long Short-Term Memory (BiLSTM) for forming feature matrix for each word using hidden states [8]. Recent research [9,10] is starting to use pre-trained transformers such as BERT to perform the text embedding task [11]. Diagram 4 shows the basic architecture.

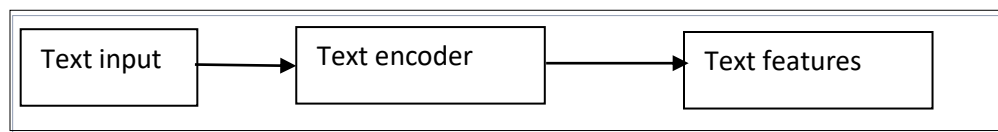


Diagram 4

4 Related Work (State-of-Art)

This section will review the related work and state-of-art of the text-to-image with different neural network architecture

4.1 First Adversarial Text-To-Image Work

As mentioned in the background section the first text-to-image translation was introduced by Reed et al [3]. The architecture proposed in Reed et al was a variation of cGAN by replacing the class label y by the text embedding ϕ . Their training approach forces both components of GAN to focus on text-alignment to the given textual description; Furthermore, the input to the discriminator is real image with mismatched text and generated image with a mismatched text. This approach is named “Matching-aware discriminator” GAN-CLS. The research added another architecture that learns based on manifold interpolation which is named GAN-INT. Furthermore, the idea is motivated by the property of deep learning models to learn from interpolation between embedding when it nears to the data manifold. The interpolation does not have to correspond to the input text; thus, it does not add labelling cost. Formally, it is an additional generator objective to minimize

$$E_{t_1, t_2, p_{data}} [\log(1 - D(G(z, \beta t_1 + (1 - \beta)t_2)))] \quad (3)$$

Where z is noise distribution and β interpolate the t_1 and t_2 embedding. Experiments showed that $\beta = 0.5$ showed good results [3]. Lastly, the final architecture is combination between GAN-INT and GAN-CLS to result into GAN-INT-CLS. Figure 1 shows a comparison of the three architectures.



Figure 1, shows comparison between GAN, GAN-INT- and GAN-INT-CLS [7]

Qualitatively, the results of GAN-INT-CLS outperform its counterparts. Notice that all generated images are 64x64.

4.2 Stacked Generative Adversarial Network

The stacked GAN approaches came to solve the problem of building higher resolution images than the GAN-INT-CLS which is limited to 64 by 64. Its mechanism works by first generate an initial 64x64 image from the given input text and random noise. The generated image is inserted into another generator with text embedding to generate 256 x 256. In the two-staged iterations there are two discriminators that check matching to non-matching text [7]. Zhang et al [11] proposed stackGAN++ that improved their previous architecture by training the three generator and the three discriminators jointly trained to simultaneously figure out the image distribution for multi-scale and conditional training. Besides, they introduced the dynamic text embedding to smoothen the conditional manifold from the gaussian distribution replacing the fixed embeddings. Figure 2 shows comparison between GAN-INT-CLS, StackedGAN and StackedGAN++.



Figure 2 shows a comparison between GAN-INT-CLS, Stack GAN and Stack GAN++ [11]

The stackGAN has better color-consistency and show greater details than the other two previous architecture. Diagram 5 shows architecture of stackGAN

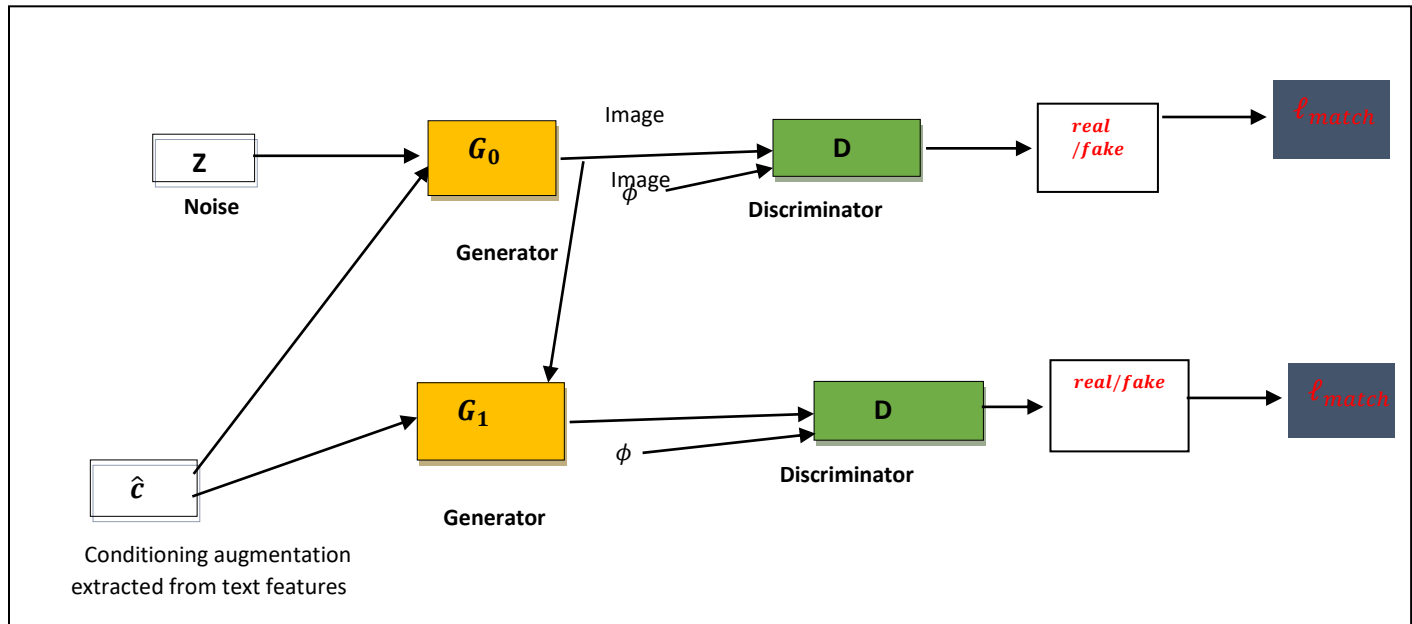


Diagram 4, stack GAN Architecture

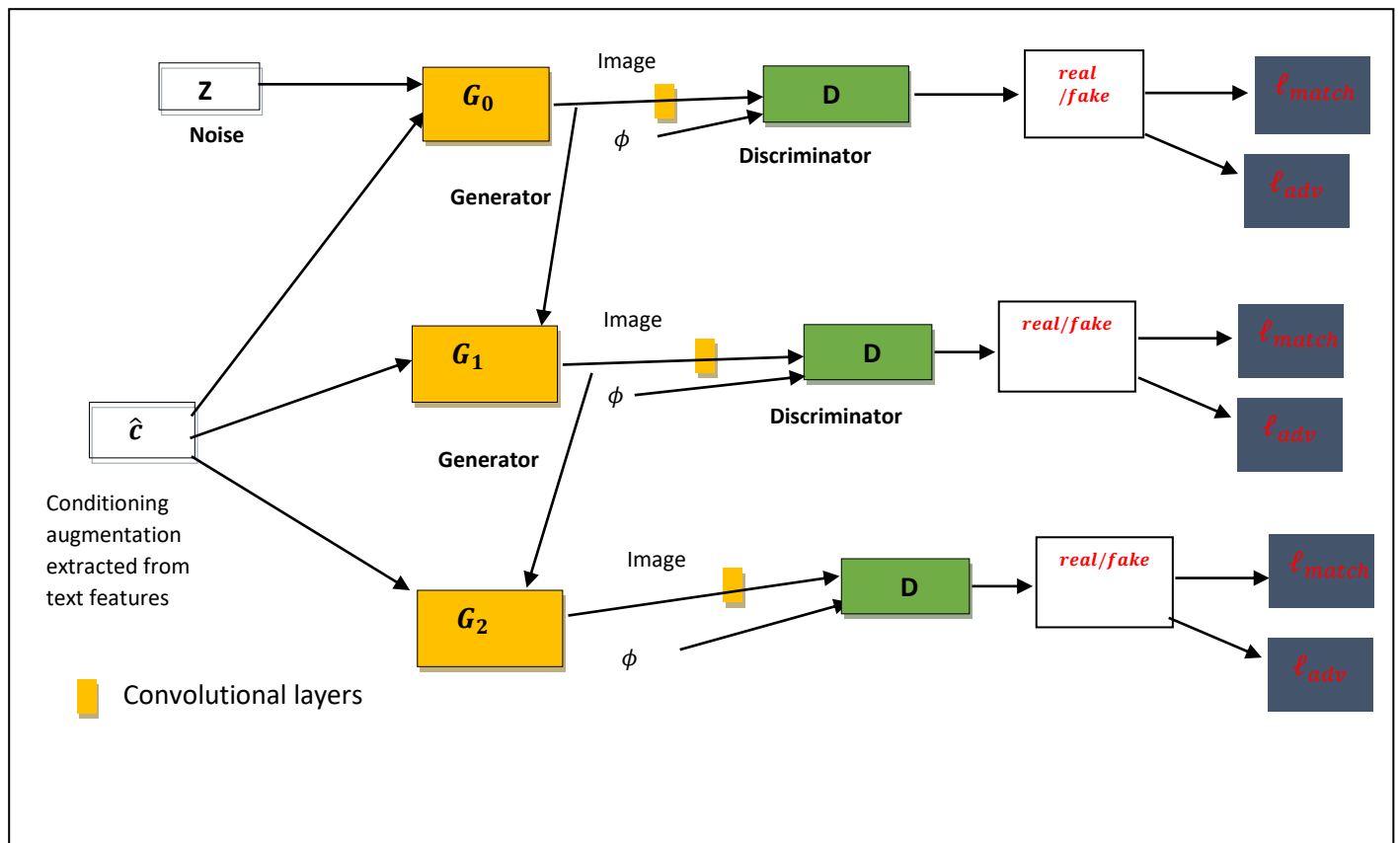


Diagram 5, stackGAN++ Architecture

When the layers jointly trained for simultaneous detecting the image distribution, it enhances the quality of the generated image by regularizing the color consistency. There is challenge introduced in this architecture which is multi-level generators. This challenge led to a new stack-based architecture which works with one generator and three discriminators. The new architecture uses hierarchical-nested network which has adversarial objectives [12]. Besides, having multi-level purpose as it has three discriminators. Figure 3 shows the architecture of the HD-GAN. Figure 3 shows a comparison between the stackGAN and the HD-GAN. From the comparison it can be shown that HD-GAN is finer detailed and has a significantly better scaling.

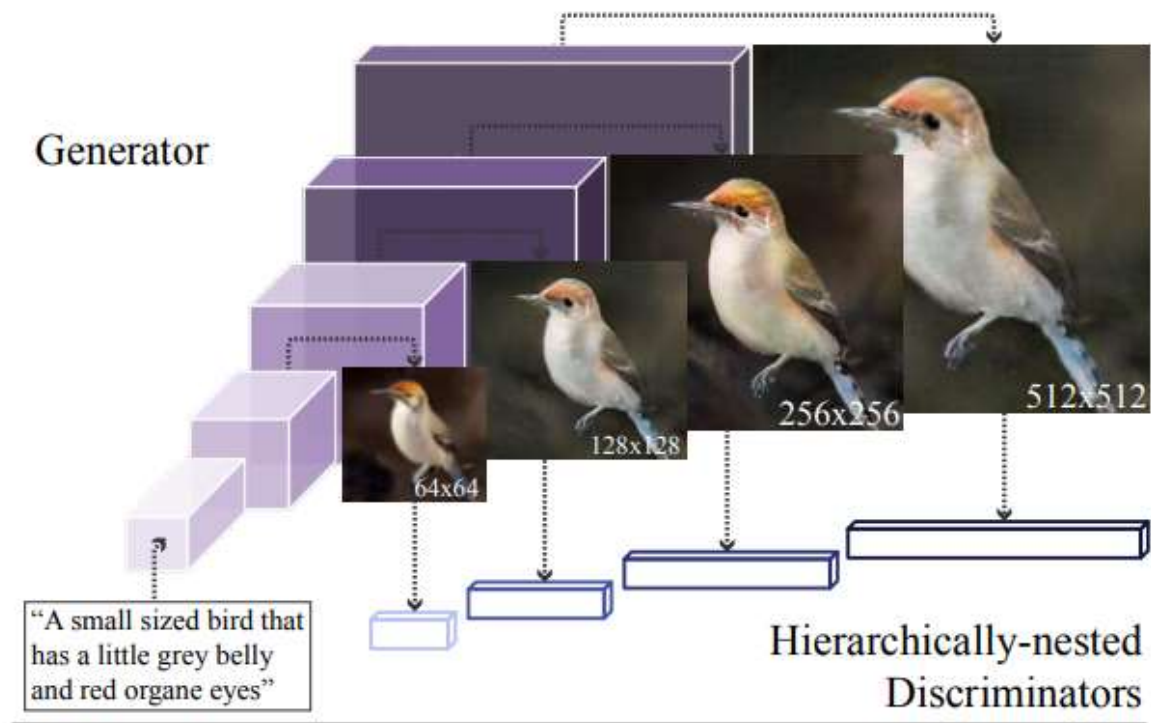


Figure 2 architecture of the HD-GAN [12]



Figure 3 shows a comparison between Stack GAN and HD-GAN [12]

4.3 Attention Architectures

In deep learning context, the attention refers to the focus of specific part of the input by weighting its input correspondingly to its importance. The attention mechanisms have proven to be very effective in applications of visual computing and language processing [13,14,15,16]. Which highly relevant to text-to-image. The attention mechanism in text-to-image architecture is based on Stack-GAN++ [15] they build-in the attention mechanism upon the pipelines of the layers of Stack-GAN++. The AttnGAN achieve that by Deep Attentional Multimodal similarity model (DAMSM) that targets to compute the similarity index between on word-level [17]. Diagram 6 shows the architeture of attention GAN.

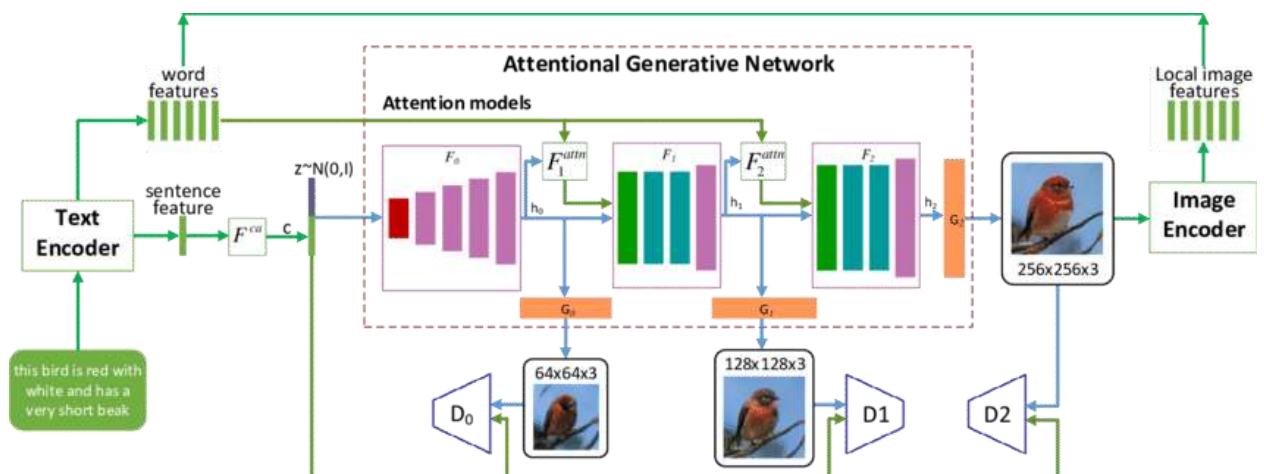


Diagram 6, shows the attnGAN architecture [17]

Tan et al [18] proposed a new approach SEGAN in text-to-image. Refining the idea of AttnGAN by exclusively focusing on visually heavy keywords instead of assigning a weight for each word in the textual description. To know the keywords only the authors used attention regularization term.

Lastly on this section, Li et al introduced the controlGAN architecture which is capable of achieving the text-to-image synthesis and editing the textures, and visual attributes without affecting the background of the image and other content [19]. This is achieved by creating a word-level spatial; besides, it is supplied with a channel-wise attention to have more accurate region-based synthesis from the attention mechanism. Furthermore, a ‘bird’ is very likely to be the central region also specifying head, chest and wing of the bird is subregions. Moreover, this attention approach can differentiate between regions and their colours which adds up to a better-quality image. Moreover, the spatial domain attention (Word-level) focuses mainly on colours while the channel-wise attention focuses on the semantic parts of the images. To visually illustrate the word-level discriminator, Li et al. has provided the controlGAN model trained without the word-level discriminator as shown in Figure 4. It can be shown that the model failed to correctly correlate the colours with the regions given, such as, the last image that states the head must be white, but it output a brown-headed bird. In figure 5, the model is trained without the channel-wise attention the small changes of the regions of the image such as belly and the head made the generator create a very different looking image in terms of background and the pose. While the channel-wise attention allows the model to have better control over the image and easily create a semantic change to the image while keeping other details correctly aligned to the text. Figure 6 shows an output comparison between StackGAN++, AttnGAN and ControlGAN.









Input	This yellow bird has grey and white wings and a red head .	This yellow bird has grey and white wings and a red belly .	The bird is small and round with white belly and blue wings.	The bird is small and round with white head and blue wings.
Ours without channel-wise attention				
Ours				

Figure 4 shows the importance of the channel-wise attention to the model (ControlGAN) by showing the same model with it and without it. [19]

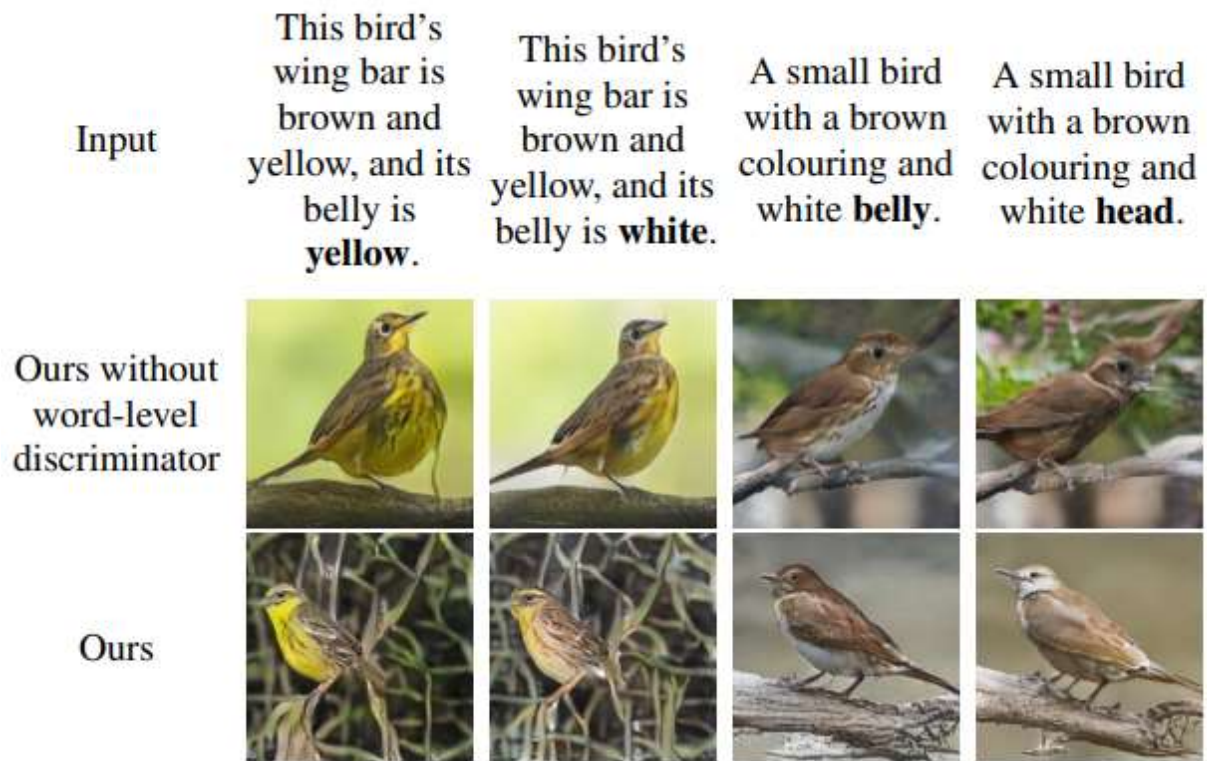


Figure 5 shows the importance of the channel-wise attention to the model (ControlGAN) by showing the same model with it and without it. [19]



Figure 6 showing a comparison between StackGAN++, AttnGAN and Control GAN [19]

4.3 Cycle Consistency Architecture

In the context of unsupervised learning the cycle consistency is the process which allows the model to reconstruct an image x from the latent variable z . The cycle consistency has been integrated with various GAN architectures such as cycleGAN [21], DiscoGAN [22] and DualGAN [23]. Qiao et al [24] proposed a new architecture that takes the advantage of encoders and decoders networks that are used for image captioning to check the loss entropy of the text reconstruction of the image to see how much the synthetic image correlate with the given text [25,26] to guide the generators from the given text embeddings.

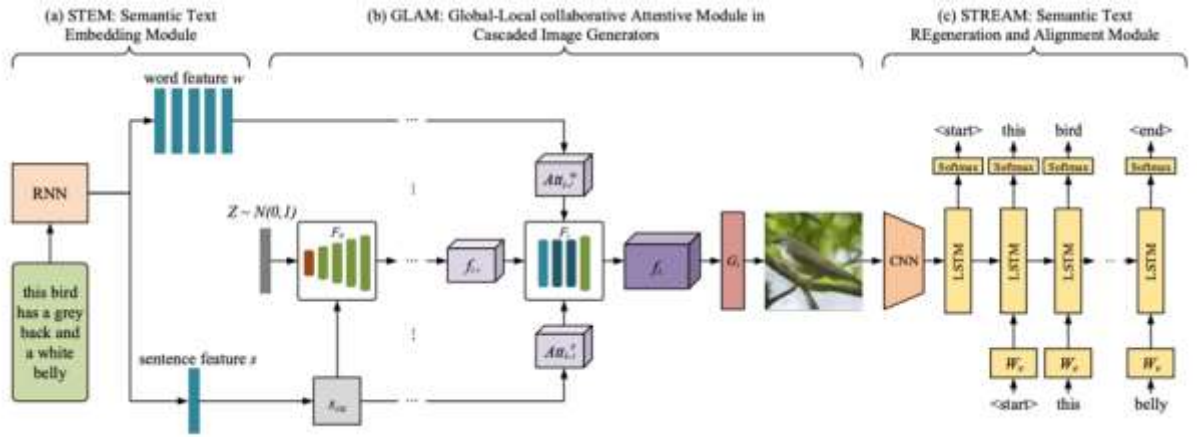


Diagram 7, shows the architecture of the MirrorGAN [24]

Figure 7 shows a comparison for MirrorGAN and AttnGAN. Qualitatively, the MirrorGAN outperforms the older AttnGAN.



Figure 6 showing a comparison between mirrorGAN and its older counterpart attnGAN [24]

Lao et al [27] proposed a new approach which infer two latent variables (style and content) from a real image by the cycle consistency approach. These latent variables are correspondingly inserted into the generator to synthesis image too close to the inferred variables.

4.4 Brief Cover of Interactive Text-to-image Approach

The approaches and architecture discussed was mainly unsupervised approaches and relies on only one given caption or description. Sequential Attention GAN sqnAttnGAN was proposed by Cheng et al [28] a dialogue-based text-to-image approach which generate image based on user interactivity as shown in Figure 8. The approach is based on interactively gaining information by asking questions. This approach enhances image editing by automating the editing by given a base image.

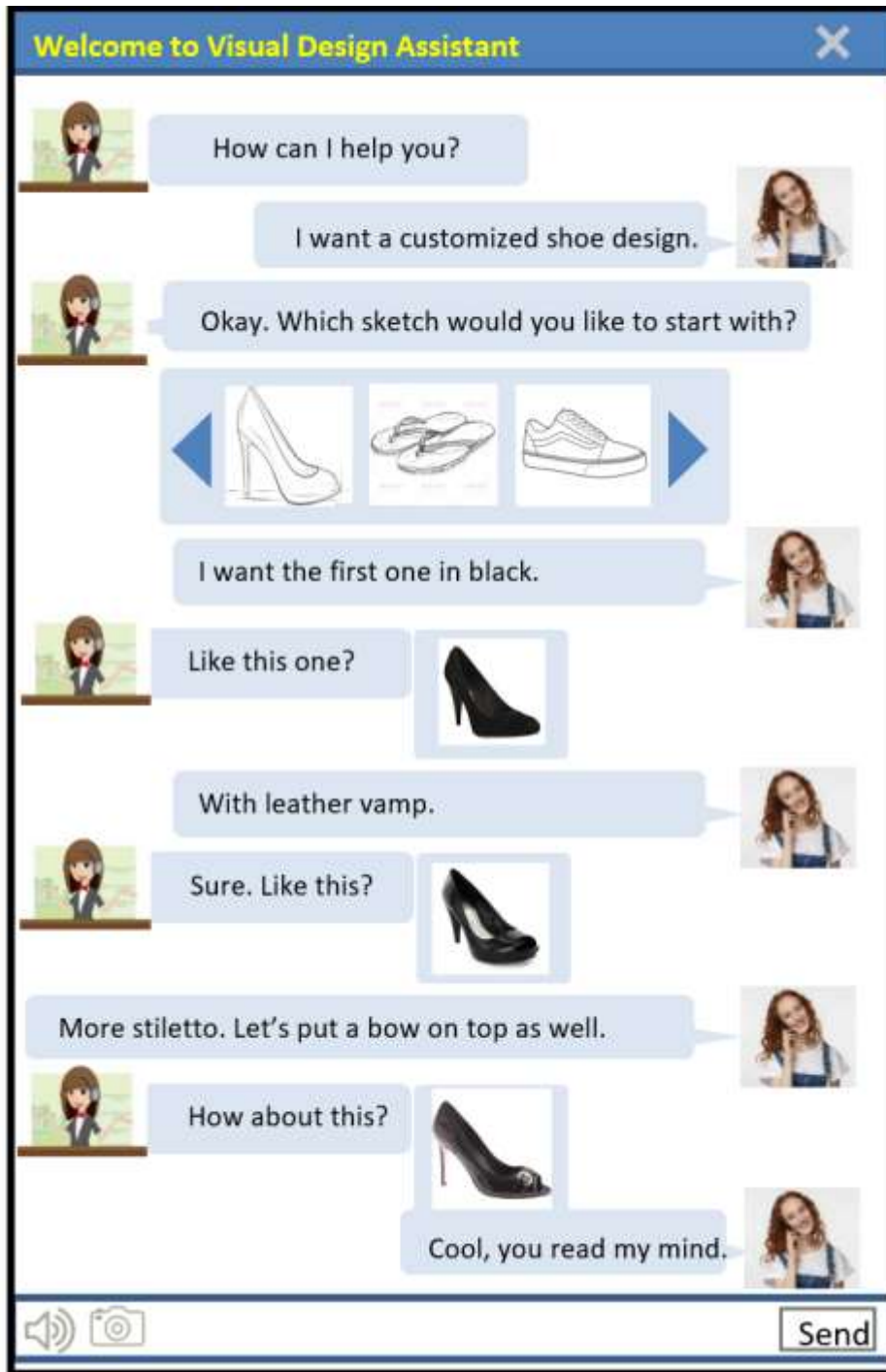


Figure 7: example of dialogue-based text-to-image [28]

4.5 The Evaluation Challenge

The aforementioned approaches enhanced the output of the text-to-image all the way from the first approach down the line to the most recent architectures. However, there is an

underlying problem of measuring the performance. The current measures, at some cases, give a higher result value for a clearly looking synthetic image over the real image which is very deceiving. Hence, if the Generator training was objectively trying to optimize the scores of the metrics it can greedily optimize the benchmark score with sacrificing the result quality of the image. It can be inferred that the architecture mentioned before tries to constrain the generator to keep the semantic alignment and reflect the image to the text otherwise the generator tries to form any image that can deceive the generator. There is no single metric can evaluate the two aspects of product which are the generated image and the relevance of text. Thus, there are two types of generating the image. The first category are the image evaluators which are the Inception Score (IS) [29] and Fréchet Inception Distance (FID) [30]. The IS takes two factors into consideration the first is distinctively of the image and the diversity of the generated images. The IS takes the conditional probability distribution $p(y|x)$ the meaningfulness of the image is inversely proportional with the entropy of the distribution. Furthermore, a more meaningful image will result in a low entropy. The diversity of the image is measured with integral margin of the probability $\int (p(y|x = G(z)))dz$ a more diverse images results in high entropy values. The two objective requirements are measured by the KL-Divergence of $p(y|x)||p(y)$. Formally, the Inception score is defined by

$$IS = \exp(E_x(KL(p|x) || p(y))) \quad (4)$$

The FID metric is more stable and consistent than the inception score metric it is based on the Euclidian distance of the probability distribution of the real image and the synthetic image [31]. However, this metric assumes that the features follow a gaussian distribution. This assumption is not always true; thus, a Kernel Inception Distance (KID) introduced as a better metric and unbiased as in FID [32].

As an example, for text relevance, Hinz et al [33] introduced a semantic Object Accuracy, the idea of this metric is based on image captioning. Furthermore, if the caption was “A man standing on a chair in an office” then there must be a man, a chair, and an office as recognizable objects. This metric detects only the explicit mentions of the objects. The inferred details of images and other objects that was not mentioned in the text are not metered; thus, the meaningfulness of the image is not assessed in this metric. There are other approaches that measures the text relevance such as R-precision such in [17] and image captioning.

The different metrics are represented in table 1

Metric	Image Quality	Image Diversity	Text relevance	Explainable
IS	Yes	No	No	No
FID	Yes	Yes	No	No
R-Precision	No	No	Yes	No
SOA	No	No	Yes	No
Captioning	No	No	Yes	No

Table 1 compare different metrics [34]

5 Gantt Chart

Diagram 8 shows the working plan for the project.

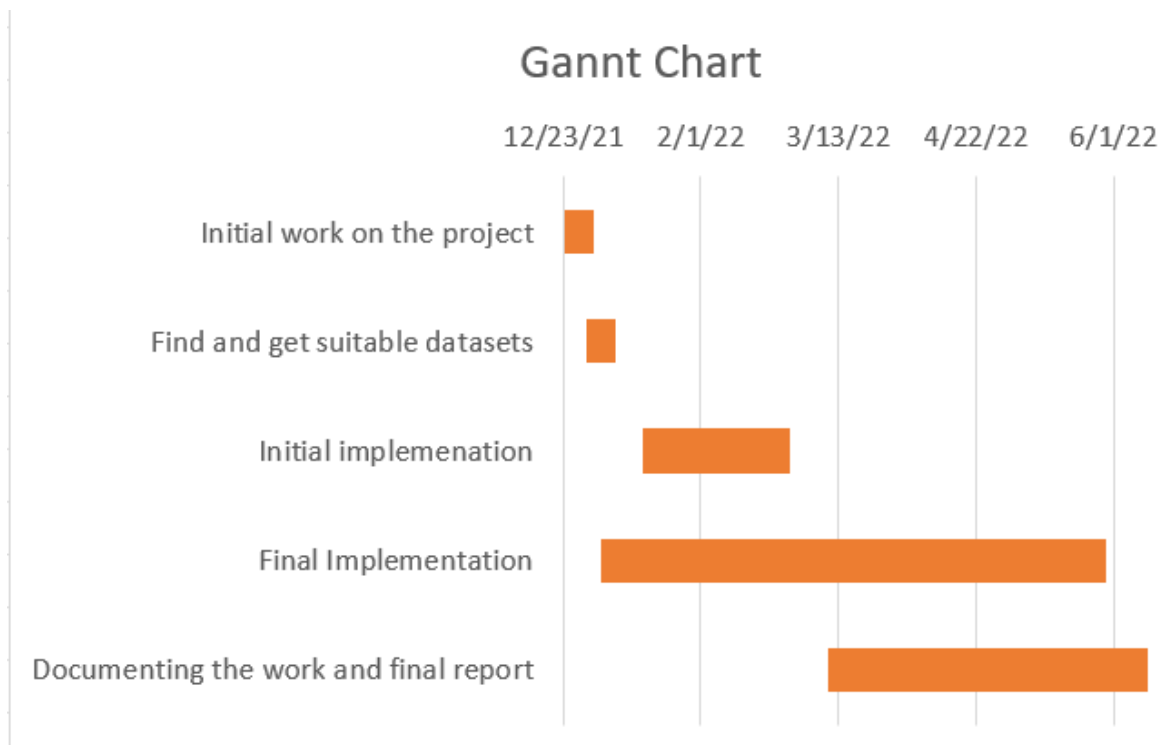


Diagram 8: Gantt Chart

Reference

- [1] Kosslyn, Stephen M., Giorgio Ganis, and William L. Thompson. "Neural foundations of imagery." *Nature reviews neuroscience* 2.9 (2001): 635-642.
- [2] Barua, Sukarna, et al. "Quality evaluation of gans using cross local intrinsic dimensionality." *arXiv preprint arXiv:1905.00643* (2019).
- [3] Reed, Scott, et al. "Generative adversarial text to image synthesis." *International Conference on Machine Learning*. PMLR, 2016.
- [4] Zhou, Rui, Cong Jiang, and Qingyang Xu. "A survey on generative adversarial network-based text-to-image synthesis." *Neurocomputing* 451 (2021): 316-336.
- [5] Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." *arXiv preprint arXiv:1411.1784* (2014).
- [6] Odena, Augustus, Christopher Olah, and Jonathon Shlens. "Conditional image synthesis with auxiliary classifier gans." *International conference on machine learning*. PMLR, 2017.
- [7] Zhang, Han, et al. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.
- [8] Souza, Douglas M., Jônatas Wehrmann, and Duncan D. Ruiz. "Efficient Neural Architecture for Text-to-Image Synthesis." *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020.
- [9] Wang, Tianren, Teng Zhang, and Brian Lovell. "Faces à la Carte: Text-to-Face Generation via Attribute Disentanglement." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021.
- [10] Pavllo, Dario, Aurelien Lucchi, and Thomas Hofmann. "Controlling style and semantics in weakly-supervised image generation." *European Conference on Computer Vision*. Springer, Cham, 2020.
- [11] Zhang, Han, et al. "Stackgan++: Realistic image synthesis with stacked generative adversarial networks." *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018): 1947-1962.
- [12] Zhang, Zizhao, Yuanpu Xie, and Lin Yang. "Photographic text-to-image synthesis with a hierarchically-nested adversarial network." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
- [13] Sood, Ekta, et al. "Improving natural language processing tasks with human gaze-guided neural attention." *arXiv preprint arXiv:2010.07891* (2020).
- [10] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning*. PMLR, 2015.
- [14] You, Quanzeng, et al. "Image captioning with semantic attention." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [15] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144* (2016).
- [16] Zhao, Bo, et al. "Diversified visual attention networks for fine-grained object classification." *IEEE Transactions on Multimedia* 19.6 (2017): 1245-1256.

- [17] Xu, Tao, et al. "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
- [18] Tan, Hongchen, et al. "Semantics-enhanced adversarial nets for text-to-image synthesis." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
- [19] Li, Bowen, et al. "Controllable text-to-image generation." *arXiv preprint arXiv:1909.07083* (2019).
- [20] Qiao, Tingting, et al. "Mirrorgan: Learning text-to-image generation by redescription." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
- [21] Almahairi, Amjad, et al. "Augmented cycleGAN: Learning many-to-many mappings from unpaired data." *International Conference on Machine Learning*. PMLR, 2018.
- [22] Kim, Taeksoo, et al. "Learning to discover cross-domain relations with generative adversarial networks." *International Conference on Machine Learning*. PMLR, 2017.
- [23] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. DualGAN: Unsupervised dual learning for image-to-image translation. In ICCV, 2017.
- [24] Qiao, Tingting, et al. "Mirrorgan: Learning text-to-image generation by redescription." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
- [25] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
- [26] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
- [27] Lao, Qicheng, et al. "Dual adversarial inference for text-to-image synthesis." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
- [28] Cheng, Yu, et al. "Sequential attention GAN for interactive image editing." *Proceedings of the 28th ACM International Conference on Multimedia*. 2020.
- [29] Barratt, Shane, and Rishi Sharma. "A note on the inception score." *arXiv preprint arXiv:1801.01973* (2018).
- [30] Heusel, Martin, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." *Advances in neural information processing systems* 30 (2017).
- [31] Bynagari, Naresh Babu. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium." *Asian Journal of Applied Science and Engineering* 8 (2019): 25-34.
- [32] Bińkowski, Mikołaj, et al. "Demystifying mmd gans." *arXiv preprint arXiv:1801.01401* (2018).
- [33] Hinz, Tobias, Stefan Heinrich, and Stefan Wermter. "Semantic object accuracy for generative text-to-image synthesis." *arXiv preprint arXiv:1910.13321* (2019).
- [34] Frolov, Stanislav, et al. "Adversarial text-to-image synthesis: A review." *arXiv preprint arXiv:2101.09983* (2021)

