Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
JCRES

VOLUME:7, ISSUE:2, PAPER NUMBER:56

Performance Analysis Of Text Based Generative Adversarial Network Models

Chandana Cheera[1], Surya Teja Burra[2], Manideep Nampelli[3], Dr. Sreedhar Potla[4]

[1,2,3,] Student, and [4] Associate Professor
Department of Information Technology,
Sreenidhi Institute of Science and Technology, Hyderabad, India
cheerachandana@gmail.com [1],suryateja01092002@gmail.com [2],itsmemani77@gmail.com [3],
Sreedhar.p@sreenidhi.edu.in [4]

Abstract: *Generative Adversarial Networks (GAN) are essentially applied to the task of image generation from textual descriptions. We want to do a performance analysis on a few of the text-based GAN models, such as Stack-GAN, DF-GAN (Deep Fusion GAN), and the baseline Attn-GAN model. Attn-GAN, known for its attention mechanism, offers a unique approach to text-to-image generation. Through qualitative and quantitative assessments using the CUB dataset, Attn-GAN demonstrated promising results, showcasing a balance between image realism and diversity. DF-GAN showed competitive results in terms of realism and diversity, reflected in its high Inception Score of 4.86, compared to Stack-GAN's Inception Score of 4.04. Stack-GAN demonstrated quality consistency with low Fréchet Inception Distance (FID) values (18.45), compared to DF-GAN (18.49), indicating a distribution closer to real images.*

Keywords— *Generative Adversarial Networks (GAN), Image generation, Text-based GAN models, Stack-GAN, DF-GAN (Deep Fusion GAN), Attn-GAN, Performance analysis, Qualitative measures, Quantitative measures, CUB dataset, Inception Score (IS), Fréchet Inception Distance (FID).*

1. INTRODUCTION:

In the human body, Neural Networks are essentially the connections between neurons. In the field of Machine Learning, Neural Networks have been built using nodes, taking inspiration from this biological configuration [1]. Neural Networks are utilized for data processing, pattern recognition, and decision-making [2]. Numerous neural network types exist, including convolution, deconvolutional, feed-forward, recurrent, and modular networks [3]. Each of these networks has a specific function and can be used for tasks like image processing, sequence prediction, modularization, and generative tasks [4,5].

The generator and discriminator neural networks, which feature in the Generative Adversarial Neural Network, compete with one another to produce false images [6]. Numerous GAN models exist, such as Stack-GAN, which requires repeated attempts with the generator and discriminator in order to produce high-quality images [7]. Deep Convolutional GAN, on the other hand, is also known as DC-GAN, indicating its use of convolution and convolution transpose in both the generator and the discriminator process [8]. Ming Tao [9] et al. stated that "Deep Fusion GAN (DF-GAN) features deep text to image fusion blocks that are

connected to a pair of generator and discriminator for image generation". Dynamic Memory GAN (DM-GAN) introduces the Memory Network to refine fuzzy image contents by incorporating contextual information and refining details through iterative refinement steps [11].

We explored various datasets for our study as preferred by the domain research such as MNIST, CUB, and COCO, which have their own unique features and complexities. Among them were MNIST for recognizing handwritten digits and COCO for image descriptions [15,16]. Ultimately, we focused on the CUB dataset, known for its diverse collection of 200 bird species [14]. Its complexity challenges text-to-image synthesis models to capture intricate visual details. We believe CUB is an ideal benchmark for assessing our methods in text-to-image synthesis.

Our goal is to do an analysis for text-based GANs, "We chose GANs for text-to-image tasks to improve realism in content creation. Many existing methods struggle with semantic consistency and low-detail images. We're focusing on Deep Fusion GAN and Stack GAN for their innovative architectures, aiming to assess their effectiveness and potential for enhancing text-to-image synthesis [22].

2. RELATED WORK:

Over the past few years, producing a visual image with a textual description has attracted a lot of attention, leading to a rise in the development of different GAN-based image generating methodologies. Given that the majority of works use changed training objectives, Reinforcement Learning, or continuous-based outputs such Gumbel-SoftMax or Soft-Argmax distributions in an attempt to override the model's optimal result. The most important contribution of a study is to critically evaluate and offer a distinct source of current GAN-based text generation research, much of which spans the years 2016 to 2020[10,17,9,11].

Generative Adversarial Networks (GAN) were first introduced by Goodfellow, Ion [6]. They proposed a new framework for estimating generative models via an adversarial process, in which they simultaneously train two models: a generative model G and discriminator model D [6]. By referencing Generative Adversarial Networks [6] Conditional-GAN was introduced which is the conditional version of generative adversarial nets, which can be constructed by simply feeding the data, y, conditioned on to both the generator and discriminator [12]. A class of CNNs called deep convolutional generative adversarial networks (DC-GANs) was introduced, that have certain architectural constraints, and a strong candidate for unsupervised learning [8]. Stack-GAN model was introduced which consists of a top-down stack of GANs, each learned to generate lower-level representations conditioned on higher-level representations [7]. The Attentional Generative Adversarial Network (Attn-GAN) allows attention-driven, multi-stage refinement for fine-grained text-to-image generation [10]. The Style-GAN has architecture that leads to an automatically learned, unsupervised separation of high-level attributes (e.g., pose and identity when trained on human faces) and stochastic variation in the generated images (e.g., freckles, hair), and it enables intuitive, scale-specific control of the synthesis [11]. A simpler but more effective Deep Fusion Generative Adversarial Networks (DF-GAN), that has a novel one-stage text-to-image backbone that directly synthesizes high-resolution images without entanglements between different generators and a novel Target-Aware Discriminator composed of Matching Aware Gradient Penalty and One-Way Output [9].

3. METHODOLOGY:

Evaluation of text-based GANs involves extensive training of the models. The focus is on assessing the performance of DF-GAN and Stack-GAN, alongside the baseline Attn-GAN model, which serves as the foundation for comparison. These evaluations aim to gauge the effectiveness of each model in generating high-quality images from textual descriptions. Attention is also given to the attentional mechanisms employed within these models, as described previously, to understand their impact on the generation process [10]. Through both qualitative and quantitative analyses, the capabilities and limitations of each GAN variant are thoroughly examined to provide insights into their efficacy in text-to-image synthesis.

3.1 Deep Fusion Gan (DF-GAN)

The proposed Deep Fusion GAN (DF-GAN) introduces a novel one-stage text-to-image backbone that directly synthesizes high-resolution images without entanglements between different generators [9]. This backbone utilizes a single generator and discriminator pair, requiring more layers than previous stacked architectures. To effectively train these layers, residual networks are incorporated to stabilize the training of deeper networks [9].

3.1.1 Architecture

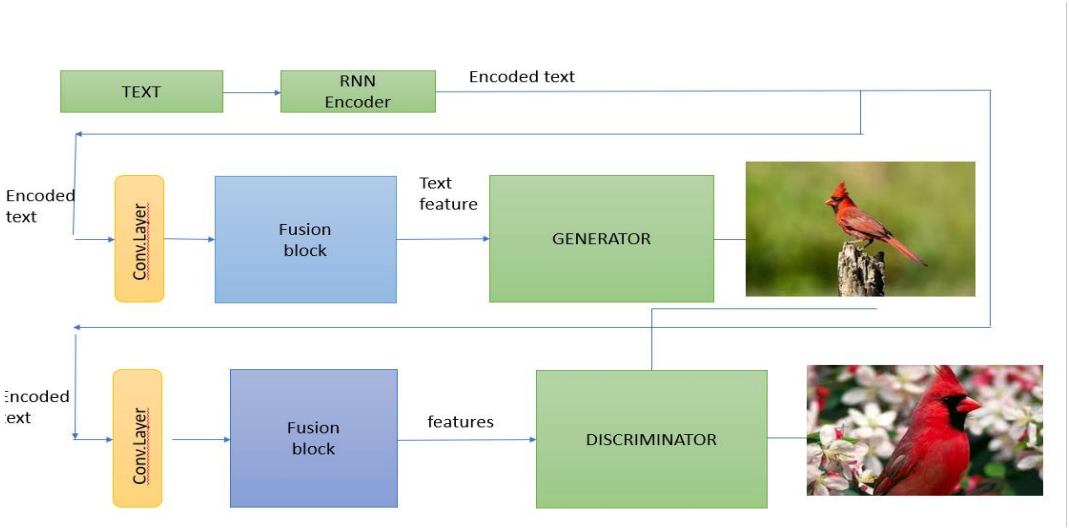

Fig-1: DFGAN architecture

Deep Fusion GAN has a fusion block that works on the textual description by computing text features that map to picture features. These are then connected to the generator and discriminator, which compete to create high resolution images (256x256).

The formulation of the one-stage method with hinge loss is as follows:

$$LD = -Ea \sim fr\,[min\,(0,-1 + D\,(a,sv))] - (1/2)\,EG(z) \sim fg\,[min\,(0,-1 - D(G(z),sv))] - (1/2)\,Ea \sim fmis\,[min\,(0,-1 - D\,(a,sv))] \quad (1)$$

$$LG = -EG(z) \sim fg\,[D(G(z),sv)] \quad (2)$$

Here, $\boldsymbol{z}$ represents the noise vector sampled from a Gaussian distribution, while $\boldsymbol{sv}$ denotes the sentence vector. $\boldsymbol{f_g}$, $\boldsymbol{f_r}$, and $\boldsymbol{f_{mis}}$ correspond to the synthetic data distribution, real data distribution, and mismatching data distribution, respectively.

The Target-Aware Discriminator, a key component, facilitates the generation of more realistic and text-image semantic-consistent images. It consists of Matching-Aware Gradient Penalty (MA-GP) and One-Way Output.MA-GP, known as Matching-Aware zero-centered Gradient Penalty, enhances text-image semantic consistency by applying gradient penalty on real images with matching text [20]. This strategy promotes text-visual semantic consistency, crucial for accurate text-to-image synthesis.

Traditional text-to-image GANs employ image features extracted in two ways, leading to slower convergence of the generator. To address this, DF-GAN adopts One-Way Output, concatenating the sentence vector and image feature, and outputs only one adversarial loss through two convolution layers.Through the fusion of text and image data throughout the image generation phase, DF-GAN facilitates easier interpretation of textual context. The fusion method ensures that image features are transformed into images by convolution layers.

3.2 Stack-GAN

Stack-GAN introduces a novel hierarchical approach to text-to-image synthesis, utilizing two stages of image generation which enables the creation of high-resolution images with fine-grained features, driven by textual input [17]. By conditioning the image generation process on textual descriptions at multiple levels, Stack-GAN achieves a multi-scale transformation that enhances visual realism [19].

As previous model operates through a two-step processive, In Stage I, low-resolution images capturing rough shapes and colours are generated. These images serve as input to Stage II, where a high-resolution image with additional features is synthesized [17]. This hierarchical architecture enables the model to progressively enhance image resolution and detail representation.

3.2.1 Architecture

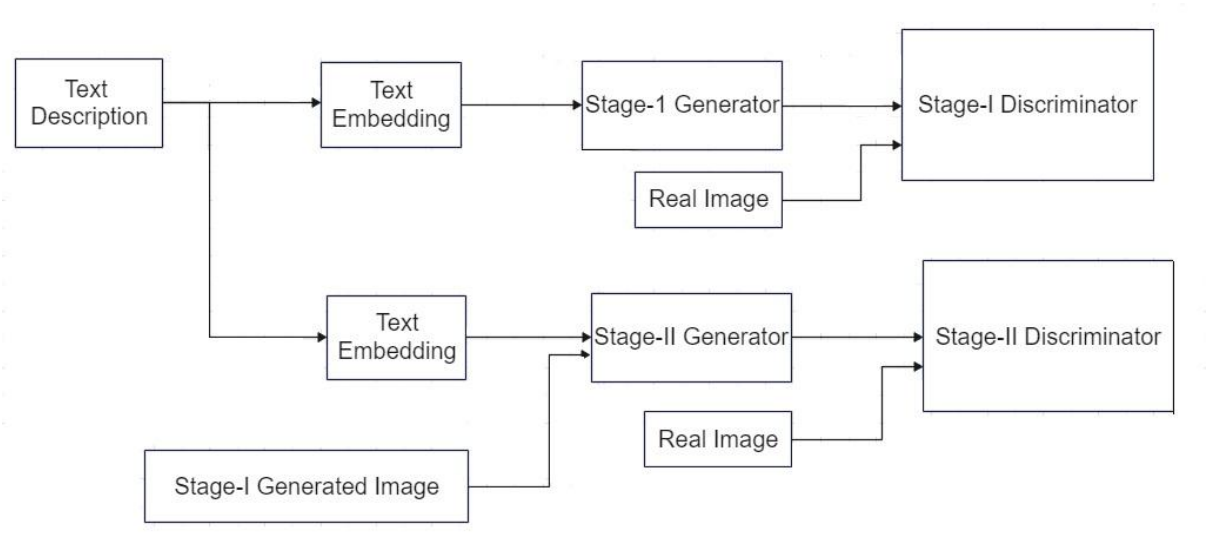

Fig-2: Stack GAN architecture

In Stage I of Stack-GAN, a text description is first converted into a vector, serving as a noise input for the generator. This vector is also connected to the discriminator, where it represents the textual information as if it were an image from the dataset. The generator produces low-

resolution images, which are then refined in Stage II to yield high-resolution images (256x256).

To elaborate on the Stage-I process, the text embedding ϕt of the description is obtained using a pre-trained encoder. Gaussian conditioning variables $(\mu 0(\phi t), \Sigma 0(\phi t))$ are sampled from a distribution to capture variations in the meaning of ϕt[17]. The Stage-I GAN, conditioned on $\hat{c}o$ and a random noise vector **z**, trains the discriminator $\mathbf{D_0}$ and generator $\mathbf{G_0}$ to maximize $L\{D_0\}$ and minimize $L\{G_0\}$, respectively. $L\{D_0\}$ involves maximizing the likelihood of real images and minimizing the likelihood of generated images, with a regularization term balancing the two objectives, as mentioned in eq (3) and (4).

$$\begin{aligned} LD_0 &= E(I_0, t) \sim fdata\,[\log D_0(I_0, \phi_t)] + Ez \sim fz, t \\ &\sim fdata\,[\log(1 - D_0(G_0(z, \hat{c}_0), \phi_t))] \end{aligned} \quad (3)$$

$$\begin{aligned} LG_0 &= Ez \sim fz, t \\ &\sim fdata\,[log(1 - D_0(G_0(z, \hat{c}_0), \phi_t))] \\ &+ \lambda DKL(N\,(\mu_0(\phi_t), \Sigma_0(\phi_t)) || N\,(0, I) \end{aligned} \quad (4)$$

Despite the benefits of Stage-I generation, low-resolution images may lack vivid object details and could contain shape distortions. To address these limitations, Stage-II GAN is introduced. Building upon the results of Stage-I, Stage-II GAN generates high-resolution images conditioned on the low-resolution results and text embedding [17,19]. This allows for the correction of defects in Stage-I results and the incorporation of previously ignored text information to produce more photo-realistic details.

In Stage II, conditioning on the low-resolution result $\boldsymbol{s0} = \boldsymbol{G0}(\boldsymbol{z}, \hat{c}\boldsymbol{o})$ and Gaussian latent variables $\hat{\mathbf{c}}$, the Discriminator **D** and Generator **G** are trained to maximize **LD** and minimize **LG** [17], similar to the Stage-I training process which is mentioned in eq (5) and (6).

$$\begin{aligned} LD &= E(I, t) \sim fdata\,[log\,D(I, \phi_t)] + Es^0 \sim fG^0, t \sim \\ &fdata\,[log(1 - D(G(s^0, \hat{c}), \phi_t))] \end{aligned} \quad (5)$$

$$\begin{aligned} LG &= Es^0 \sim fG^0, t \sim \\ &fdata\,[log(1 - D(G(s^0, \hat{c}), \phi_t))] + \\ &\lambda DKL\left(N\,(\mu(\phi_t), \Sigma(\phi_t))| \quad |N\,(0, I)\right) \end{aligned} \quad (6)$$

Through this iterative refinement process, Stack-GAN achieves enhanced image quality and fidelity, driven by textual descriptions.

3.3 Attentional Generative Network(ATTN-GAN):

There is currently a shortage of fine-grained word-level information in the current GAN-based text generation models [10, 17, 9, 11]. This attentional generative network (Figure 3) comprises m generators that accept hidden states and to create pictures [23].

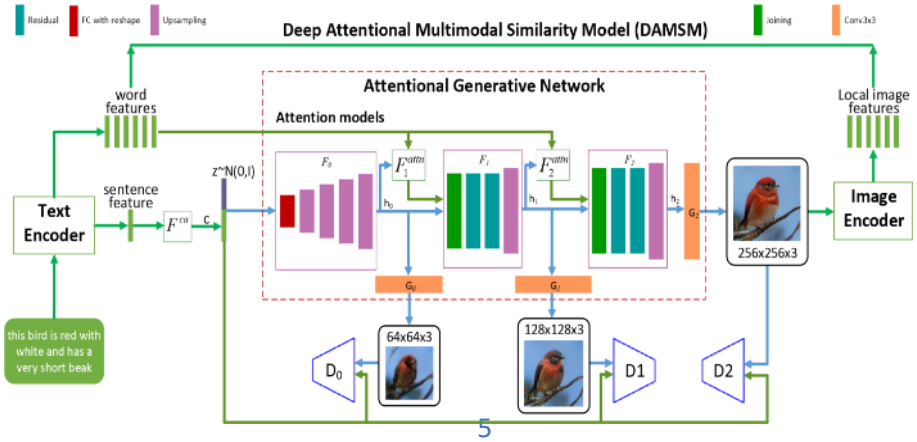

Fig. 3. This shows how the GAN is supposed to work. The DAMSM gives the generative network a loss for fine-grained image-text matching. Each attention model automatically picks out the criteria (word vectors) for making different parts of the picture. [23]

Specifically,

$$h_0 = F_0(z, F^{ca}(\bar{e}))$$
$$h_i = F_i\left(h_{i-1}, F_i^{attn}(e, h_{i-1})\right) \; for \; i = 1,2,\dots,m-1;$$
$$\hat{x}_i = G_i(h_i)$$

The noise-based vector **z** is drawn from an example of a normalized distribution in this case. Each word in a sentence is placed in a row called an e vector. The Conditioning Augmentation [17] moves the sentence vector e to the conditioning vector, which is shown by F^{ca}. At the i^{th} level of the GAN, F^{attn} is the proposed attention model. The neural networks F^{ca}, F^{attn}, F_i, and G_i are modelled.

The attention model $F^{attn}(e,h)$ uses the word features $e \in R^{D\times T}$ and the picture features $h \in R^{\hat{D}\times N}$ from the preceding hidden layer (e,h). Adding a new perceptron layer, $e^0 = Ue$, converts word characteristics into shared semantic space.

Then, depending on the image's hidden features, a word-context vector is computed for each sub-region (query). Each column of h represents a subregion of the picture... The word-context matrix for the j^{th} sub-region is a dynamic depiction of word vectors pertinent to h_j, derived prior text to image algorithms [17,9,10], our solution uses the CUB and COCO datasets. The CUB dataset is first pre-processed. Table 1 summarises the dataset. Evaluation.

Dataset	CUB [28]		COCO [14]	
	train	test	train	test
#samples	8,855	2,933	80k	40k
caption/image	10	10	5	5

Table-1 Statistics of datasets

4. PERFORMANCE ANALYSIS

Evaluation of text-based GAN's is carried out by extensive training of the models. For both models the CUB dataset is used. In the context of evaluating Generative Adversarial Network (GAN) models, both Fréchet Inception Distance (FID) and Inception Score (IS) are commonly used metrics, so we used them both to evaluate.

4.1 Qualitative Evaluation

In our qualitative analysis, we delve into the visual quality and coherence of the images produced by both models. By leveraging the high-quality CUB birds dataset, we maintain consistency and reliability in our evaluations. Additionally, using the same text encoder for both models ensure a fair and comparable assessment process. Through qualitative examination, we closely inspect elements like image sharpness, detail accuracy, and overall realism. This scrutiny provides valuable insights into how well each model performs in faithfully translating textual descriptions into visually captivating images.

For training Stack-GAN and DF-GAN, we customized the CUB Bird's dataset to suit the specific requirements of each model. This involved preprocessing the dataset to ensure compatibility with the model architectures and training objectives. Despite the customization, the essence of the original CUB dataset remained intact, providing a robust foundation for model learning and image synthesis. Additionally, both models utilized the customized dataset in tandem with the same text encoder, ensuring consistency in textual representation across training iterations.

Outcome of DF-GAN- Deep Fusion GAN integrates a fusion block that operates on textual descriptions, extracting text features that correspond to image features. These features are subsequently linked to both the generator and discriminator, engaging in competition to generate high-resolution images, typically sized at 256x256 pixels.

Fig. 4. Output of DF-GAN with caption: A red bird with black eyes and short beak

Outcome of Stack-GAN- For Stack-GAN there are two stages where stage-1 produces 64x64 images which are basically blur some of them were low resolution images. These images were used as input source for stage-2. In stage-2 the model is trained to produce high resolution images of size 256x256.

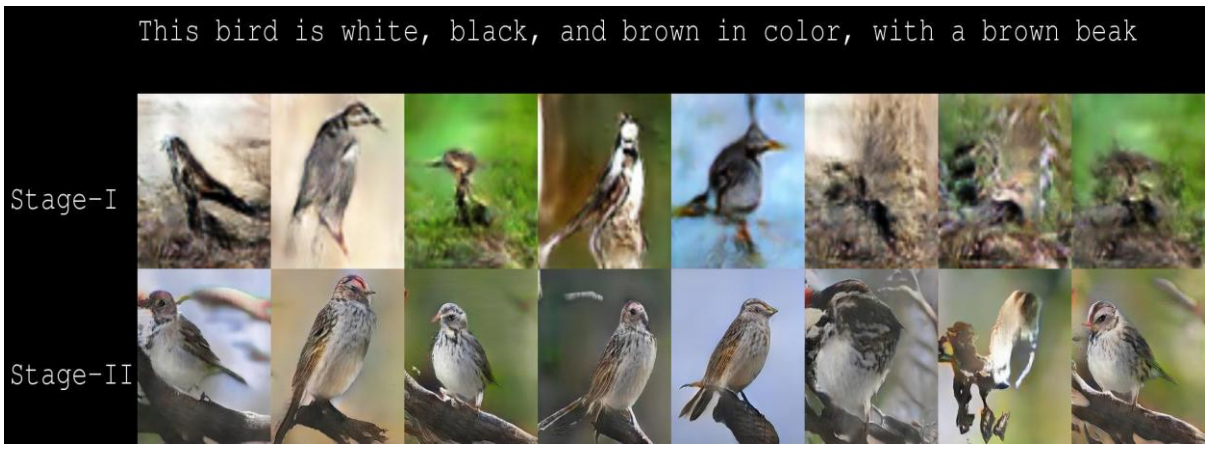

Fig 5: Output of Stack-GAN

Outcome of Attn-GAN- For enable attention-driven, multi-stage filtering for fine-grained picture production, with a novel attentional generative network that considers the essential words, our application can synthesise fine-grained features in various picture subregions.

Fig. 6. Output of Attn-GAN

4.2 . Qualitative Evaluation

Quantitative Evolution is a method of examining, interpreting, and drawing conclusions from numerical data. It involves the use of statistical techniques and mathematical models to analyse data and identify patterns, trends, and relationships [21].

For qualitative analysis Inception Score (IS) and Fréchet Inception Distance (FID) are used. Inception Score is used to determine quality of image generated by GAN. It measures how realistic and diverse the output images are. Higher IS values indicate better performance, as they suggest that the generated images are both realistic and diverse [18].

Table-2: outcome of Inception score of the models

Methods	**Inception Score**
Attn-GAN	4.36 ± 0.03
Stack-GAN	4.04 ± 0.06
DF-GAN	4.86 ± 0.04

FID resembles the distance between distribution of generated images and the real images in the feature space of a pre-trained inception v3 network [3]. It gives a measure on the quality of the images generated. It correlates with human judgement on quality of images. Lower FID values indicate better performance, as they suggest that the generated images are closer in distribution to the real images.

Table-3: FID score of outcome of the models

Methods	**CUB-FID**
Attn-GAN	23.98
Stack-GAN	18.45
DF-GAN	18.49

5. CONCLUSION

In our analysis of three GAN models—DF-GAN, Stack-GAN, and Attn-GAN—on the CUB dataset, DF-GAN impressed us with its fast training and ability to create high-quality images, scoring lower on FID and achieving a top Inception Score. Stack-GAN improved its

Inception Score but took longer to build models due to its complex design. Attn-GAN stood out for its skill in generating excellent images from textual descriptions, significantly boosting its Inception Score. Overall, DF-GAN had the highest Inception Score, while Stack-GAN had the lowest FID Score. Throughout our evaluation, all three consistently produced impressive images.

6. REFERENCES

[1] Shao, Feng, Shen, Zheng (9 January 2022). "How can artificial neural networks approximate the brain?". Front Psychol. 13: 970214.

[2] Thorat, S. B.; Nayak, S. K.; Jyoti P Dandale (2010). "Facial Recognition Technology: An analysis with scope in India". arXiv:1005.4263 [cs.MA].

[3] Somnath Mukherjee, Bikash Sadhukhan, Nairita Sarkar, Debajyoti Roy, Soumil De, "Stock market prediction using deep learning algorithms". 2019

[4] Nebauer, C. (1998) “Evaluation of convolutional neural networks for visual recognition.” IEEE Transactions on Neural Networks 9 (4): 685- 696.

[5] Jurgen Schmidhuber, "Deep Learning in Neural Networks: An Overview", Technical Report IDSIA-03-14 / arXiv:1404.7828 v4 [cs.NE] 2014

[6]. Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

[7]Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[9] Ming Tao, Hao Tang, Fei Wu, Xiaoyuan Jing, Bing-Kun Bao, Changsheng Xu. Deep Fusion GAN(DFGAN): A Simple and effectiveness for text-image synthesis. In arxiv.org, cs, 2020.

[10] Xu, Tao, et al. "Attn-GAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks." IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

[11] Karras, Tero, et al. "A Style-Based Generator Architecture for Generative Adversarial Networks." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.

[12] Mehdi Mirza, Simon Osindero, “Conditional Generative Adversarial Nets” 2014

[13] Minfeng Zhu, et al. Dm-gan: Dynamic memory generative adversarial networks for textto-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5802– 5810, 2019.

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-T

[15] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-Based Learning Applied to Document Recognition" in Proceedings of the IEEE, 1998

[17] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris Metaxas. “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks” arXiv:1612.03242v2 [cs.CV] 5 Aug 2017.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In CVPR, 2016.

[19] Tominaga, Rihito, and Masataka Seo. 2023. "Image Generation from Text Using StackGAN with Improved Conditional Consistency Regularization" Sensors 23, no. 1: 249. https://doi.org/10.3390/s23010249

[20] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In International Conference on Machine Learning, pages 3481–3490, 2018.

[21] prahlad gaur, "what is quantitative data analysis?", in geeksforgeeks,2024.

[22] Yin, Guojun & Liu, Bin & Sheng, Lu & Yu, Nenghai & Wang, Xiaogang & Shao, Jing. (2019). Semantics Disentangling for Text-To-Image Generation. 2322-2331. 10.1109/CVPR.2019.00243.