Text to Image Synthesis using StackGAN

Purva Sandeep Warade
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India
purva.warade23@pccoepune.org

Swati V. Shinde
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India
swati.shinde@pccoepune.org

Bal Virdee
Center of Communication Technology
London Metropolitan University
London, United Kingdom
b.virdee@londonmet.ac.uk

Ashish Khanna
Department of Computer Science
Maharaja Agrasen Institute of Technology
Delhi, India
ashish.khanna@mait.ac.in

Abstract — **Text-to-image synthesis is a difficult task. It involves the translation of descriptive text to visual content, which maps language and image representations. Among several approaches, StackGAN was one of the most notable frameworks because of the pioneer two-stage architecture. The first stage generated low-resolution images based on the global structure. Then, the first stage results were utilized and refined in the second stage into high-resolution, realistic images with improved details. This paper reports the performance of StackGAN on a subset of the Flickr dataset where embeddings of textual descriptions are obtained from the USE. Experimental results will show that the model produces semantically coherent images, which are also visually coherent. A study focuses on the prospect of StackGAN in creative content generation and discusses challenges such as maintaining diversity and mitigating artifacts.**

Keywords— Text-to-Image Synthesis, StackGAN, Generative Adversarial Networks (GANs), Deep Learning, Image Generation.

I. Introduction

The core challenge in AI is generating realistic images based on textual descriptions-that would essentially ask a system to bridge the semantic gap that exists between a given textual input and its corresponding visual representation. Such applications include virtual content creation, e-commerce, and accessibility technologies for the visually impaired, but as it stands, the task is indeed pretty complex, considering that abstraction found in the textual descriptions might be with rather intricate visual interpretations.

GANs emerged as a rather powerful framework in the generation of images. cGANs extended this approach, conditioning the generation of images over some other auxiliary data, such as text. Until now, the existing cGANs are not so well performing in the synthesis of high-resolution images: they suffer from instability during training and lack an ability to capture fine-grained details [3].

In order to address the challenges as stated above, StackGAN introduces a two-stage hierarchical framework. Stage-I GAN produces low resolution images which capture the global structure and basic colors described in the text. Stage-II GAN refines this output into 256×256 images adequately filled with details and corrected defects. Furthermore, Conditioning Augmentation-based training stabilizes training and brings about diversity by smoothing out the manifold of latent conditioning. This paper details the StackGAN architecture, evaluates its performance on the Flickr8k dataset, and compares it with state-of-the-art methods. Experimental results demonstrate StackGAN's ability to generate high-quality, semantically consistent images from text descriptions.

The key innovation of this work was that the Universal Sentence Encoder (USE) was integrated into the StackGAN framework in terms of semantic text embedding. Unlike the majority of the studies done before, USE allows grasping semantic context because it uses word-level or character-length embeddings. Furthermore, to verify the results of the study, both subjective evaluation by human judgement and objective metrics (Inception Score, FID) evaluation will be included. Such a method offers a solid assessment of the semantic integrity and the real rendition of the images within resource-limited settings.

II. Text to Image Synthesis

There have been significant developments in the fields of computer vision and machine learning in recent years related to text-to-image synthesis, which involves creating images based on textual descriptions. This technology allows users to express visual elements through rich and vivid text descriptions, facilitating the automatic generation of images from natural language inputs. Visual materials, like photographs, are generally more engaging and easier to comprehend than written words, making them preferable for sharing and understanding.

Text-to-Image Synthesis refers to the technique of employing computational methods to transform human-written text descriptions of objects (in the form of sentences or keywords) into visually representative images. The best alignment between the visual content and the accompanying text has been achieved through word-to-image correlation analysis combined with supervised synthesis techniques. Recent advances in deep learning have led to the emergence of new unsupervised techniques, particularly deep generative models [1]. These models can generate realistic images using well-trained neural networks.

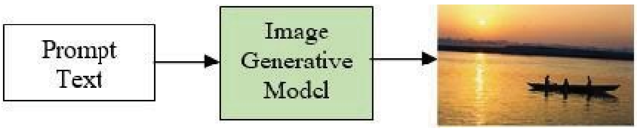

Fig. 1. General architecture of Text to image

Figure 1 illustrates a general architecture of how text-to-image generation may work: feeding into an image generative model is a text prompt, and then the model will use the text description to generate an image.

III. Related Work

Recent advancements in text-to image synthesis, however, took hold with rapid breakthroughs in deep learning as well as generative models. Classic approaches are VAEs and also auto-regressive networks that could not reach rich models in terms of quality and diversity in the output because of the difficulty in the transformation from the textual description to visual content. Hence, there were shortcomings such as minimal fine-grained details and less realistic images.

A. *Effects of GANs:*

Generative adversarial networks completely shifted the game when it is used to build the entire framework to generate images that are firmly grounded in reality, thanks to generator and discriminator. Conditional GANs take this ability even further and introduce additional auxiliary inputs, like text embeddings, to help guide image generation process [4]. Early cGAN-based models, however, were only producing such low-resolution outputs sans fine features [8].

B. *StackGAN Contribution:*

Towards such limitations, StackGAN proposed the two-stage architecture, which primarily targeted the structure and layout of the low-resolution image and complemented that with texture to enhance the resolution in the succeeding stage [2]. This staging establishes the framework for StackGAN to generate high resolutions which are both semantically accurate and visually coherent than the single-stage models. Further improvements were based on the StackGAN framework. For example, AttnGAN utilized attention mechanisms that dynamically attends to relevant parts of the input text during the generation of an image; hence, they better maintained superior text-to-image alignment. Concurrently, transformer-based models such as DALL-E scaled the synthesis task to large datasets and hence leveraged powerful language-image pretraining techniques. However, these models with huge computational overhead pushed the limits of text-to-image synthesis. We are using StackGAN, which really hits a good balance between simplicity and effectiveness. Actually, we are using the Universal Sentence Encoder (USE) for text embedding to try to study its potential toward improving semantic consistency as well as visual quality of generated images and we have chosen StackGAN because of robust performances through the interleaving of two-stage design.

IV. Methodology

This research applies a Text-to-Image Synthesis framework based upon a two-stage Generative Adversarial Network architecture called StackGAN. The research methodology consists of several phases which begin with preparing the dataset to creating the text embeddings and then training and testing the StackGAN model.

A. *Dataset Preparation*

1) *Dataset Description*

The dataset used has been a subset of the Flickr8k, which comprises 2,000 images. Five human-written descriptive captions have been paired with each image. These describe objects, actions, and settings, which makes the dataset highly appropriate for being used in text-to-image synthesis tasks. Paired captions ensure that the captured semantic content within text data is both meaningfully and diversely covered [5].

Although the original Flickr8k dataset has 8,000 images, it is still computationally impractical to use all the images of the substantial dataset, so this study only randomly samples 2,000 images in it. This was a subset that was made to ensure the diversifying of the scenes and captions and to ensure the training was manageable by the available resources. The existing results can be considered a proof-of-concept, and subsequent efforts will be devoted to the confirmation of the model generalization on bigger data sets such as MS-COCO or LAION-400M.

2) *Image Processing*

All images are resized to a fixed resolution of 256×256 pixels to standardize the input to the StackGAN model so that resolutions may be standardized throughout the whole dataset, lowering computational overhead during training. Normalizing pixel values to the range $[-1,1]$ further aids gradient flow during backpropagation, helping to stabilize GANs' training process.

Further key applications of data augmentation techniques are increasing the diversity of the training set without overfitting. They are random crops, rotation, and flipping, bringing variations without altering core semantics in the dataset.

3) *Caption Preprocessing*

Textual captions undergo processing so that they can go well with the embedding model. The preprocessing pipeline would include tokenization of each caption into words and converting all text to lowercase for uniformity. Stop words, special characters, and punctuation marks are removed as these are unnecessary noise.

B. *Text Embedding*

1) *Purpose of Text Embedding*

Text embeddings form an intermediary between the input text-that is, captions-and the generation of images. Embeddings are used here, which describe the captions in terms of numerical representations that correlate with a dense vector format so that semantic meanings can be captured. In this way, encoding captions to fixed-length embeddings helps the model achieve knowledge of the relationships between text and visual features.

2) *Embedding Process*

The Universal Sentence Encoder utilizes caption embeddings into 512-dimensional dense vector representations. Each caption is fed into the USE model, which will then produce an embedding vector ztext that captures the semantic meaning of the caption. This can be expressed mathematically as:

ztext = USE(captioni)

Such embeddings are normalized to unit length to have their presentation standard . It also provides consistency and reduces the probability of coming across numerical instability at the time of training.

3) *Conditioning Augmentation*

To introduce variabilities and to improve the robustness of text embeddings, a condition augmentation technique is provided. It involves the generation of Gaussian noise from the embeddings for better generalization across the unseen data. The formula to calculate the augmented embedding z_{aug} is given by:

$$z_{aug} = N(\mu,\sigma 2)$$

where μ and σ are derived from the original embedding z_{text}. This augmented representation is then used as input to the StackGAN model.

C. *StackGAN Architecture*

The main architecture of StackGAN is the two-stage Generative Adversarial Network, mainly utilized for text-to-image synthesis. The coarse-to-fine approach of StackGAN works; Stage-I generates input text into a low-resolution image based on embeddings created from the text, and Stage-II refines it into a high-resolution image.

The StackGAN involves two stacked stages:

- Stage-I GAN: This stage generates a low-resolution image (64×64) from the input text embeddings and captures the overall structure and basic layout as described in the text.
- Stage-II GAN: This stage refines the image from Stage I, adding high-frequency details and improving resolution to 256×256, which produces a photo-realistic and semantically consistent image.

Each phase contains two primary modules:

A Generator (G): The generator incorporates images conditioned on an input noise vector and text embedding.

A Discriminator (D): Measures the quality of the generated images and their semantic consistency with the text descriptions.

1) *Stage-I GAN: Coarse-to-fine generation:*

In Stage I, GAN generates a coarse image at the resolution of 64 × 64 pixels capturing the global structure and layout of the input caption.

- Inputs to Stage-I

i. Text Embedding (z_{text}): A semantic meaning of the input caption that is represented using USE with dimension 512.

ii. Noise Vector (z noise): It is a random vector sampled from the N(0,1) Gaussian distribution. This means, the added randomness to the generative process of images.

iii. Input to Generator: First, it passes the concatenation of text embedding and noise vector input to the generator

$$z_{input} = [\, z_{text}, z_{noise}].$$

- Stage-I Generative Architecture

The concatenated input is transformed into an intermediate representation of the features with dense layers (fully connected layers). Transposed convolutional layers also known as fractionally strided convolutions progressively upsampling the feature maps with spatial resolution 64 × 64.Batch normalization and ReLU activations are employed everywhere to stabilize the training procedure and improve the feature representations.

The final output is a pixel image of size 64 × 64 × 3 with values scaled using a Tanh activation function.

- Stage-I Discriminator Architecture

The discriminator takes as input either a generated image at 64×64 or a real image and feeds it through some convolutional layers where the output are features from the image and then combines it with the text embedding, which is passed through fully connected layers.The dense layers sum up to some sigmoid output that lets it know whether the input image is real or fake and whether it justifies the text description.

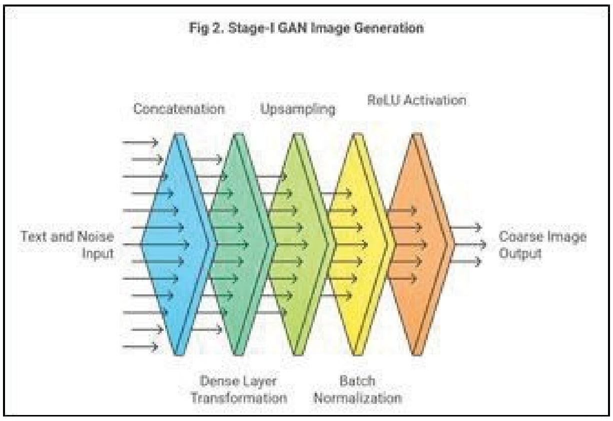

Fig. 2. Stage 1 GAN Image Generation

2) *Stage-II GAN: Fine-to-Coarse Refinement*

The Stage-II GAN refines the coarse images coming from Stage-I, and the generated images are of size 256x256 with realistic textures and details.

- Inputs to Stage-II:

i. Low-Resolution Image: The output image of size 64×64 from Stage-I.

ii. Text Embedding (z_{text}): This is reused from Stage-I for semantic coherence.

- Stage-II Generator Architecture:

64×64 image is upsampled and then concatenated with text embedding. Residual blocks are employed to refine the features, introducing high-frequency details while preserving low-resolution features[7]. Transposed convolutional layers further upsample the image, progressively increasing the resolution to 256×256.

The output is a refined, high-resolution image with enhanced realism, generated using a Tanh activation function.

- Stage-II Discriminator Architecture:

Similar to Stage-I, the discriminator evaluates the generated 256×256 image or a real image through convolutional layers. Text embedding (z_{text}) is passed through parallel processing and added to the image features. The dense layers take the combined representation, and with the output, it determines if the image is real or fake and whether the image aligns with the text.

StackGAN's two-stage coarse-to-fine hierarchical architecture enables high-resolution image synthesis with semantic accuracy directly from text descriptions. Stage-I lays down the structural foundation, and Stage-II aims at pushing the visual fidelity and realism while ensuring a strong text-to-image synthesis pipeline.

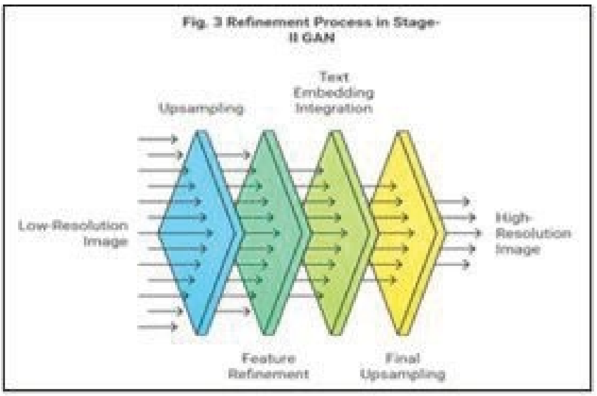

Fig. 3. Refinement Process in Stahe II GAN

V. Experiments and results

This section shows results of experiments using a smaller subset of the Flickr8k dataset, 2,000 images with their captions for evaluating the ability of the StackGAN framework in carrying out the text-to-image synthesis task.

A. *Model Configuration and Training*

The StackGAN framework was trained on a two-stage architecture. In Stage-I, there was a low resolution of 64×64 produced from the text embedding and random noise vector. The output of the Stage-I fed into Stage-II in which the image was refined to come up with a high resolution of 256×256. The generators and discriminators are trained using the Adam optimizer at a learning rate of 0.0002. It was trained on 600 epochs with a batch size of 64.

B. *Evaluation Metrics*

To evaluate the model's performance, we used the following evaluation metrics:

1) *Inception Score (IS):*

The assessment evaluates the quality and diversity of the generated images. Good performance in terms of IS correlates well with high-quality images being generated and diverseness.

2) *Fréchet Inception Distance (FID):*

The FID checks the similarity of distributions of real images and generated images. This means that a lower FID score means the generated images have a closer distribution to the real images.

3) *Human Evaluation:*

Human judges do subjective evaluation with respect to the realism and coherence of the images generated by relating it to their input text captions.

This research paper has done an evaluation of StackGAN based on the standard generative metrics that include the Inception Score (IS), and the Fréchet Inception Distance (FID). The Stage-I and Stage-II models have generated a limited number of images and the measurement of the metrics was made via the available tools since the resources used were limited. The size of the sample was not too big, but the metrics gave comparative evidence of the realism of the images produced and their variety raised by the model. This will be followed up with large- scale assessment to make more statistically valuable comparison.

C. *Experimental Results*

1) *Quantitative Results:*

TABLE I.

Metric	**Stage-I**	**Stage-II**
Inception Score (IS)	3.12 ± 0.07	4.68 ± 0.11
FID Score	85.43	36.47

Stage-II improved significantly over Stage-I, with higher Inception Score (4.68 vs 3.12) and a lower FID score (36.47 vs 85.43). This shows that the process of refinement in Stage-II considerably enhances the quality and realism of generated images.

2) *Human Evaluation:*

Human evaluators rated the images on a scale of 1 to 5 based on their **realism** and **relevance** to the captions. The average ratings for each stage were:

TABLE II.

Stage	**Realism**	**Relevance**
Stage-I	2.9	3.0
Stage-II	4.3	4.6

3) *Qualitative Results:*

Example:

Text: “A sunset and water.”

Stage-I Output:

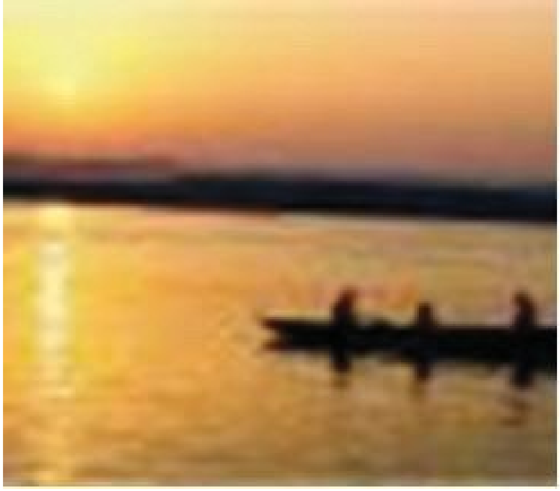

Fig. 4. low resolution image

Stage-II Output:

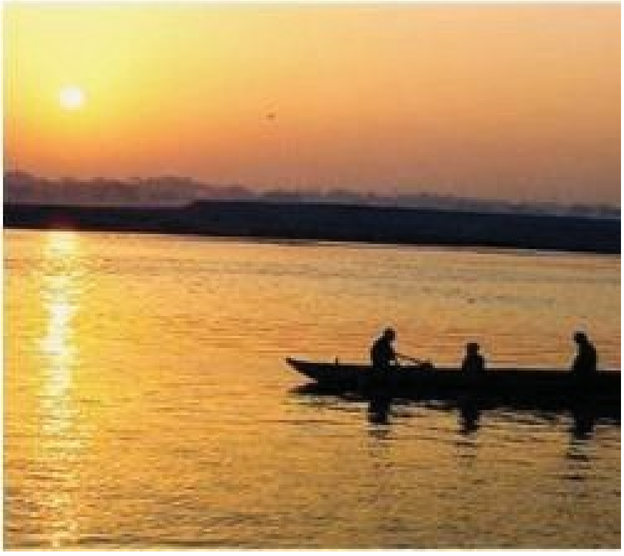

Fig. 5. high resolution image

The results have shown that the StackGAN framework is capable of producing very high-quality images from text descriptions. Furthermore, a very distinct improvement in both Inception Score and FID Score from Stage-I to Stage-II clearly addresses the significance of refinement in improving the realism of generated images along with their textual alignment. Further, the quality of the generated images is ensured by these human evaluation results, thus proving that Stage-II produces not only more realistic but also better aligned images concerning the input captions. These confirm the proposed architecture of two stages to be effective in synthesizing visually appealing images and semantically accurate from the text description.

Even though this paper is focused on StackGAN, it takes note of the progress of the other architectures like AttnGAN and DALL·E. Although AttnGAN includes attention modules and DALL·E scales up on transformer-to-pretrained generation, both models consume much more computational resources. In comparison, StackGAN provides a sensible balance between model intricacy and output excellence, particularly when there are limited resources. This will be used in the future to make quantitative comparisons with these models under standardized data to emphasize relative advantages and disadvantages [9].

VI. Conclusion

This work proves the capability of the StackGAN framework in text-to-image synthesis, utilizing a two-stage architecture to generate high-quality and semantically accurate images from textual descriptions. The coarse-to-fine design of StackGAN enables a model to first produce a more elementary image structure in Stage-I, and then to refine it into a high-resolution and very realistic image in Stage-II. Results on a subset of Flickr8k dataset showed that the two-stage approach leads to significant gains both in terms of visual fidelity and semantic relevance of the generated images. Quantitative evaluations done against metrics such as Inception Score and Fréchet Inception Distance reported Stage-II outputs superior to those from Stage-I. Human evaluation also showed that the images produced by Stage-II were scored higher on realism and caption alignment.

Results validate that StackGAN could generate images that, not only are visually aesthetic but also semantically aligned with the text descriptions given, making it a significant tool for text-to-image synthesis tasks. Future work could be on further optimization of the model or adaptation to more complex datasets and diverse applications, such as video generation or real-time synthesis [6].

REFERENCES

[1] S. K. Alhabeeb and A. A. Al-Shargabi, "Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction," in IEEE Access, vol. 12, pp. 24412-24427, 2024,

doi: 10.1109/ACCESS.2024.3365043.

[2] Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas D. “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks.” arXiv [Internet]. 2017 [cited 2024 Nov 18]; Available from: https://arxiv.org/abs/1612.03242

[3] X. Wu, K. Xu and P. Hall, "A survey of image synthesis and editing with generative adversarial networks," in Tsinghua Science and Technology, vol. 22, no. 6, pp. 660-674, December 2017,

doi: 10.23919/TST.2017.8195348.

[4] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. &amp; Lee, H.. (2016). Generative Adversarial Text to Image Synthesis. <i>Proceedings of The 33rd International Conference on Machine Learning</i>, in <i>Proceedings of Machine Learning Research</i> 48:1060-1069 Available from https://proceedings.mlr.press/v48/reed16.html

[5] Y. L. Sahithi, N. Sunny, M. M. L. Deepak and S. Amrutha, "Text-to-Image Synthesis using stackGAN," 2023 Global Conference on Information Technologies and Communications (GCITC), Bangalore, India, 2023, pp. 1-6,

doi: 10.1109/GCITC60406.2023.10426184.

[6] Shinde, S., Kulkarni, U., Mane, D., Sapkal, A. (2021). Deep Learning-Based Medical Image Analysis Using Transfer Learning. In: Patgiri, R., Biswas, A., Roy, P. (eds) Health Informatics: A Computational Perspective in Healthcare. Studies in Computational Intelligence, vol 932. Springer, Singapore.

https://doi.org/10.1007/978-981-15-9735-0_2

[7] S. B. Chavan and S. V. Shinde, "An Experimental Investigation of U-Net-Based Deep Learning Segmentation for Histopathology Images," 2023 1st DMIHER International Conference on Artificial Intelligence in Education and Industry 4.0 (IDICAIEI), Wardha, India, 2023, pp. 1-6,

doi: 10.1109/IDICAIEI58380.2023.10406634.

[8] Waghmare, P., Shinde, S. (2022). Image Caption Generation Using Neural Network Models and LSTM Hierarchical Structure. In: Das, A.K., Nayak, J., Naik, B., Dutta, S., Pelusi, D. (eds) Computational Intelligence in Pattern Recognition Advances in Intelligent Systems and Computing, vol 1349. Springer, Singapore. https://doi.org/10.1007/978-981-16-2543-5_10

[9] A. Deo, S. Shinde, T. Borde, S. Dhamak and S. Dungarwal, "A Comprehensive Review of Image Colorization Methods," 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 2023, pp. 1-6, doi: 10.1109/I2CT57861.2023.10126250.