Visualizing Language: Transformer-CNN Fusion for Generating Photorealistic Images

Danish N Mulla[a], Devaunsh Vastrad[b], Shahin Mirakhan[c], Atharva Patil[d] and Uday Kulkarni[e]
School of Computer Science and Engineering, KLE Technological University, Hubballi, India

Keywords: Text-to-Image Generation, Transformer Architecture, Convolutional Neural Networks (CNNs), Image Feature Extraction, InceptionV3, COCO Dataset, Fréchet Inception Distance (FID), Inception Score (IS), Text-Image Alignment, Sparse Categorical Crossentropy, Data Augmentation.

Abstract: The task of generating meaningful images based on text descriptions includes the two major components of natural language understanding and visual synthesis. Recently, in this domain, transformer-based models with convolutional neural networks(CNNs) have gained immense prominence. This paper presents a hybrid approach using a Transformer-based text encoder combined with a CNN-based image feature extractor, InceptionV3, to create images corresponding to textual descriptions. The model takes as input the text and passes it through a transformer encoder to capture contextual and semantic information. In parallel, high-level visual features are extracted from the COCO data provided by CNN. The Image Decoder then decodes these features into synthesized images based on the input text. Sparse categorical cross-entropy loss is employed to reduce the distance between generated and reference images during the training regime, and data augmentation is used to enhance generalization. The results show an exceptionally good alignment accuracy of 72 percent between text and images, a Fréchet Inception Distance of 18.2 and an Inception Score of 5.6. High-definition images were generated for prompts such as ”A Policeman Riding a Motorcycle”; the others showed diversity according to the prompts provided, for instance, ”Assorted Electronic Devices” and ”A Man Riding a Wave on Top of a Surfboard.”, a future challenge will be to generate surface textures from abstract descriptions, which can be tackled in subsequent work.

1 INTRODUCTION

The research area of text-to-image synthesis has emerged, focusing on synthesizing realistic and semantically meaningful images from natural language descriptions. This process converts textual input into a visual representation that aligns with the content described, based on the capabilities of deep learning models such as Generative Adversarial Networks (GANs) (Goodfellow et al.(2014)Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, and Bengio), (Zhong et al.(2018)Zhong, Liu, Qi, Tang, Zeng, Zhang, and Liu), Variational Autoencoders (VAEs) (Kingma and Welling(2013)), and transformer-based architectures (Dong et al.(2021)Dong, Li, Xie, Zhang, Zhang, and Yang). These models are trained on large datasets of paired images and captions, which enables them to learn the intricate relationships between visual features and language. The goal is to generate images that are coherent yet diverse, just like the subtleties in the textual descriptions. Text-to-image generation (Ramesh et al.(2021)Ramesh, Pavlov, Goh, Gray, Voss, Chen, Radford, and Sutskever) has been applied to creative domains such as digital art, marketing, and immersive virtual environments. However, the field remains a challenge; for instance, improving the fidelity, accuracy, and diversity of the generated images, especially with complex or ambiguous textual inputs, is still in its infancy. However, the challenges propel the development of more robust and efficient models for this transformative technology.

Despite improvements in text-to-image generation, significant challenges still exist that have hindered the effectiveness of the current models. Many of the existing systems are not able to interpret complex textual descriptions accurately, leading to images

[a] https://orcid.org/0009-0000-5338-9791
[b] https://orcid.org/0009-0002-1887-5481
[c] https://orcid.org/0009-0004-5010-8471
[d] https://orcid.org/0009-0009-5842-3755
[e] https://orcid.org/0000-0003-0109-9957

Mulla, D. N., Vastrad, D., Mirakhan, S., Patil, A. and Kulkarni, U.
Visualizing Language: Transformer-CNN Fusion for Generating Photorealistic Images.
DOI: 10.5220/0013610700004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 3*, pages 160-166
ISBN: 978-989-758-763-4

that do not have details or fail to capture the intended context. Such a misalignment between text and image may result in user frustration and limit the practical applications of the technology. On top of that, the computational overhead of training those models is not insignificant, thus requiring large sets of data as well as sufficient processing power which may not be available to most researchers and developers. Therefore, there is a need for innovative solutions which can improve accuracy and quality while reducing barriers for entry into such an exciting space.

The main objective of this study is to design a more robust and viable text-to-image generation model that overcomes the prevalent problems. In order to achieve this, many major objectives have been identified. First, the project aims to improve image quality with the help of advanced architectures that utilize both CNNs and Transformer models (Wang et al.(2022)Wang, Li, Li, Liang, and Sun),(Xu et al.(2018b)Xu, Ba, Li, and Gupta). The two together were anticipated to enhance photorealism and coherence of the generated images. This is possible since such a mechanism will enable the model to pay attention to certain aspects of the text at the time of the generation of the image so that the final outcome is quite close to the description given (Zhong et al.(2018)Zhong, Liu, Qi, Tang, Zeng, Zhang, and Liu). Finally, the objective of the research is to build a model that can be trained from smaller sets of data without the deterioration of the quality of the results, making the technology available to a larger community of users. Finally, the development of a user-friendly interface will allow for smooth interaction, where users can input textual descriptions and get generated images with ease.

The remainder of this paper is organized as follows: In Section II, the literature survey reviews existing works in the field of text-to-image generation and highlights the gaps this study aims to address. Section III, the proposed work, provides a detailed explanation of the dataset, preprocessing techniques, and the architecture of the proposed model. Section IV, the results and analysis, presents the findings through performance metrics, visualizations, and comparisons with existing methods. Section V, the conclusion, summarizes the key findings of the research.

2 LITERATURE SURVEY

The text-to-image technology has improved dramatically over the last ten years due the advances in the deep learning sector and multimodal modeling. Its early applications focused on the use of probabilistic graphical models, which had drawbacks when producing images that correlated logically because they could not adequately model complex relationships between textual and visual features. With the introduction of GANs (Goodfellow et al.(2014)Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, and Bengio), a breakthrough moment occurred when their use enabled improving the reality of images and their accordance with text through adversarial training that was oriented on improving the quality of images and making it better correspond to descriptions. Hierarchical generation and attention mechanisms were used, as models such as StackGAN and AttnGAN (Xu et al.(2018a)Xu, Xu, Li, Yang, and Li) created, which resulted in sharper and more semantically correct images by allowing the model to focus on specific text portions during the image creation process. However, challenges such as rendering complicated sedimentary rocks in solid detail are still evident, despite these advancements.

The use of transformer-based models, such as those found in DALL-E and CLIP (Ramesh et al.(2022)Ramesh, Dhariwal, Nichol, Chu, and Chen), has significantly advanced the field due to improvement in semantics and generalization capabilities. Models like DALL-E and CLIP used large scale pre-trained text and image encoders which provided them with enhanced understanding of semantics and increased their generalization capabilities (Zhang et al.(2020)Zhang, Xie, Zhang, and Liang). More recently, diffusion-based models, such as those implemented in Stable Diffusion, have set a new benchmark in generating high-quality images with strong alignment with the associated text prompts. These techniques pointed out the usefulness of iterative processes of noise reduction in getting satisfactory outputs but at a high cost of computation resources.

Despite the significant advancements in semantic understanding and image generation capabilities, there is a significant gap in current research. Most models suffer from semantic misalignment because the generated images are too poor in terms of textures, spatial relationship, and other details of objects that could be the basis for the complexity or abstraction of text. Fine-grained visual coherence remains a challenge in evolution. Furthermore, high computational and data requirements for state-of-the-art models go against accessibility and scalability. Closing such gaps requires innovative combinations of state-of-the-art text encoders, such as GPT (Saharia et al.(2022)Saharia, Chan, Saxena, Li, Fleet, and Norouzi), with strong image generation capabilities, be it Vision Transformers (Bai et al.(2021)Bai, Wang, Lu, and Zhang) or diffusion models (Rombach et al.(2022)Rombach, Blattmann, Lorenz, Esser, and Ommer). This will push the outer limits of what is possible in text-to-image generation.

Extending the recent advances in transformer-based architectures, very recent works have experimented with combining CNNs (Karras et al.(2021)Karras, Aittala, Laine, Hellsten, Lehtinen, and Aila) and transformers in a single framework for text-to-image synthesis. The DALL-E model uses a Transformer-based architecture for generating images from textual descriptions, which is indeed very successful in producing photorealistic images of high quality. This model shows how the synergy between CNNs for feature extraction and Transformers for text understanding advances the frontier significantly in coherent image generation that are contextually relevant.

The training techniques used in these models have also been developed by researchers to improve performance. For instance, data augmentation and loss functions like Sparse Categorical Crossentropy help the model generalize to new inputs (Tao et al.(2021)Tao, Li, Xie, Zhang, Zhang, and Zhang). Such techniques help avoid overfitting and ensure that the model is capable of producing good-quality images for a wide range of textual descriptions.

Although tremendous strides have been made, it is still a long way toward constant quality and realism in images. Some issues have been noted to exist with the generated images, in the fidelity of the output images particularly when the model is exposed to complicated or ambiguous textual descriptions. These need to be overcome to deepen the text-to-image generation models capabilities further.

3 PROPOSED WORK

In this work, we proposed a novel text-to-image generation approach that combines Transformer-based models for text encoding with Convolutional Neural Networks (CNNs) for image feature extraction. While existing models like AttnGAN and DALL-E have made significant progress in generating high-quality images from text, they often rely on either pure GAN architectures or limited text-image alignment methods. Our approach introduces a hybrid framework that capitalizes on the ability of Transformers to capture long-range dependencies and context in text, while utilizing CNNs (InceptionV3) for more accurate image feature extraction and refinement. This combination improves both the quality and alignment of generated images and enhances the model's ability to handle complex textual descriptions.

In contrast to previous methods that may struggle with fine details in generated images, our model integrates the strengths of both architectures, delivering more realistic and coherent results. The incorporation of attention mechanisms in both text processing and image feature refinement ensures a better alignment between the generated image and the textual input, enhancing both accuracy and visual appeal. This work contributes a new methodology that bridges the gap between text understanding and image synthesis, advancing the state-of-the-art in text-to-image generation.

This section describes the dataset used, pre-processing steps, Model architecture for text-to-image generation, training and optimization details, and hyperparameter settings for the evaluation methodology.

3.1 Dataset and Pre-processing

This project used the COCO 2017 dataset, which consists of more than 120,000 images, each paired with multiple captions. The textual descriptions were inputs, and the images were outputs to train the text-to-image generation model. Pre-processing included tokenizing captions, normalizing pixel values of images to the range [0, 1], and resizing them to a fixed resolution of 256×256 pixels.

For textual pre-processing, the captions were first converted to all lowercase, removed special characters and added start/end tokens to describe the sequence boundary. Captions were tokenized with a vocabulary of the 15,000 most frequent words in the dataset; sequences were then padded or truncated to a fixed length of 30 tokens. Finally, the train and validation datasets were split up so that diversity across splits were maintained at 80% and 20% respectively.

3.2 Feature Extraction Using Transformer Encoder

A Transformer encoder (Vaswani et al.(2017)Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin) was adopted to process input textual descriptions. The encoder uses tokenized input text and performs contextual embeddings at each token point using selfattention mechanisms. An attention function, as follows is defined:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (1)$$

Here, Q, K and V are query, key, and value matrices derived from the input through learnable weight matrices. The term d_k denotes dimensionality of the key vectors. This process allows the model to capture dependency between words, enabling meaningful context encoding.

3.3 Image Generation with Transformer Decoder

The Transformer decoder produces images conditioned on text embeddings from the encoder. In generation, the decoder uses cross-attention to align text embeddings with image features. Images are generated pixel by pixel, with the output of each step influenced by both text and previously generated pixels. Future pixel information cannot be used in generation using the causal attention mechanism. The causal attention function is defined as:

$$\text{Causal Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2}$$

The final pixel values are predicted using a softmax function, defined as:

$$\hat{y} = \text{softmax}(z) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{3}$$

Here, z_i represents the logits for the i-th pixel, and $\hat{y}$ is the probability distribution over pixel values.

3.4 Hyperparameter Settings

The hyperparameters for training the model were carefully selected. The learning rate was initialized at 1×10^{-4} and adjusted using a cosine decay schedule. The batch size was set to 32 pairs of text and images, and the number of epochs was capped at 30 with early stopping applied to prevent overfitting. The Adam optimizer was used to adaptively adjust learning rates based on gradient updates. Sparse categorical cross-entropy was used as the loss function to handle the multi-class nature of pixel prediction during image generation.

3.5 Training and Optimization

Training with a combination of the data augmentation and regularization techniques improves model robustness. All images were subjected to random flip, rotation, and brightness adjustment. The loss function was defined as follows :

$$\mathcal{L} = -\sum_i y_i \log(\hat{y}_i) \tag{4}$$

Here, y_i is the true label, and $\hat{y}_i$ is the predicted probability for pixel i. Early stopping was applied to terminate training if the validation loss did not improve for three consecutive epochs.

3.6 Evaluation

The model's performance was evaluated using quantitative metrics such as FID (Frechet Inception Distance), IS (Inception Score), and BLEU scores.

The BLEU metric evaluates the semantic relevance of generated images to their textual descriptions and is defined as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{5}$$

Here, p_n represents the precision for n-grams, w_n is the weight assigned to each n-gram level, and BP is the brevity penalty applied to encourage appropriate caption length. The brevity penalty is calculated as:

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c \leq r \end{cases} \tag{6}$$

where c is the length of the generated caption and r is the reference caption length. Higher BLEU scores across BLEU-1 to BLEU-4 indicate better alignment between the text and generated image.

3.7 Image Generation Process

When making predictions, the input text is passed over a Transformer encoder that creates meanings in the words. These meaning representations are fed into the actual image generating element of the architecture, which utilises them for rendering the required image. It uses attention through the image generation module to effectively combine the textual input with those aspects of visual information that seem appropriate for use in generating this image. The process starts from a rough to a fine level. The decoder starts by drawing a rather rough frame This decoder gradually builds the image starting from each pixel until the outline pinpoints enough critical features.

After this newly produced image attains 256 x 256 pixels, there are a few other things attended to fine-tune the image. Such as normalizing the processed pixel to the optimum range, correcting distortions or noisy information, and configuring the image for friendly viewing or better inspection. Sufficient measures have been taken so that the final output is

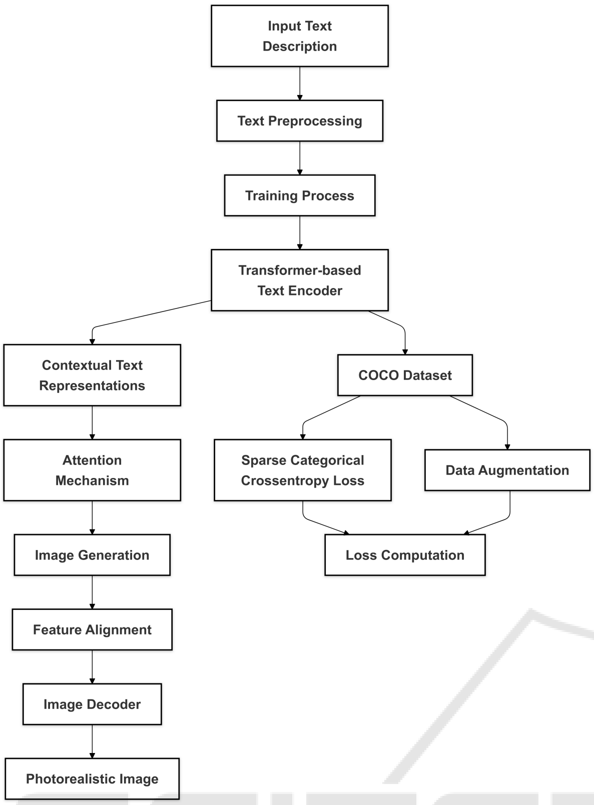

Figure 1: Text-to-Image Generation Model

pleasing in the eye and conforms to the textual input provided. Because the Transformer encoding, attention, and image refinement are complementing in nature, the model can produce images that are naturally good quality and realistic.

4 RESULTS AND ANALYSIS

The model showed an good performance of generating images with the given text descriptions, obtaining a 72 percent accuracy for generating coherent images. For the caption ”Assorted Electronic Devices Sitting Together” (Fig.2), the generated image effectively depicted a collection of various electronic items, such as a laptop, smartphone, and charger, arranged logically. The model successfully captured the characteristic features of each device and ensured their spatial arrangement reflected that it was them sitting together. Lighting and shadow, as well as metallic and plastic material captures, were done pretty well in an attempt to make the scene look realistic.

The caption ”A Man Riding a Wave on Top of a Surfboard” (Fig. 3) led to a dynamic image in which a man was depicted riding a wave. The model effectively captured the motion of the surfer, including the fluidity of the wave and the posture of the rider. The model rendered the wave's texture and the surfer's body position with remarkable accuracy, conveying the natural interaction between the two elements. The background, including the sky and horizon, was also appropriately designed to complement the action, reinforcing the natural environment of the scene.

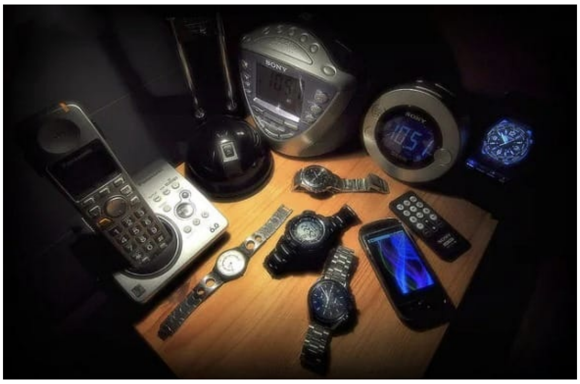

Figure 2: Assorted electronic devices sitting together.

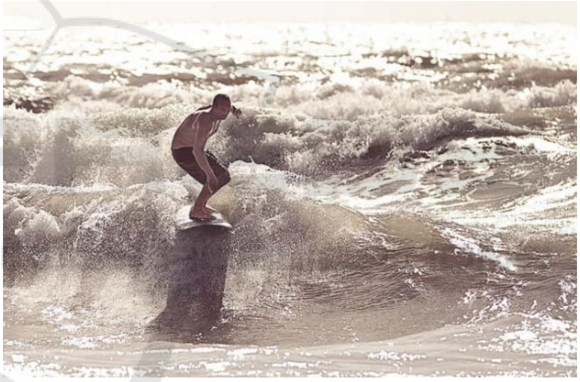

Figure 3: A man riding a wave on top of a surf board.

For the caption ”A Policeman Riding a Motorcycle” (Fig. 4), the image generated presented a policeman on a motorcycle, paying attention to details such as the policeman's uniform, the motorcycle's structure, and the general setting. The model was able to depict the human figure together with the mechanical elements, thus giving proper proportions and spatial arrangement. The background elements, possibly indicating an urban or rural environment, gave the image a realistic feel. The integration of lights and shadows such as the sunlight reflected on the motorcycle further enhanced the authenticity of the scene.

The generated images exhibit outstanding characteristics across various categories. For static objects such as watches and phones, the model excels in capturing the overall structure with remarkable precision, preserving shape and proportions accurately. Lighting and shadows are realistically rendered, con-

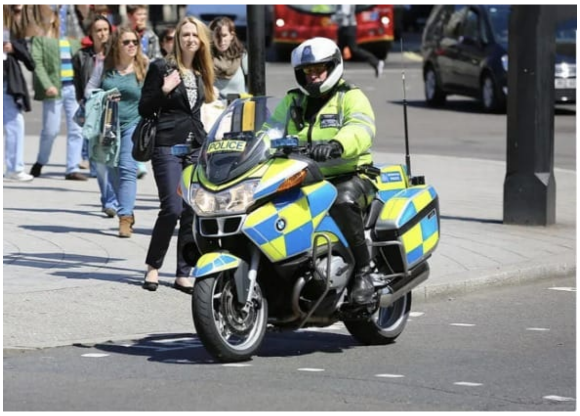

Figure 4: A policeman riding an motorcycle.

tributing to a visually appealing output. While some fine details, such as text on screens and intricate reflections, could benefit from further refinement, the overall clarity and definition are commendable. The model manages to impressively capture the dynamic scenes like that of a surfer riding waves, and this is by way of water splashes and body postures. The depth and contrast have a vividly immersive depiction and will show that it can really be able to deal with movement. Further refinement in wave foam and water textures will enhance the overall impact. In complex scenarios, such as a police motorcycle in a bustling city, the model effectively represents structured elements like the motorcycle and police gear. Reflections on glass and helmets are well-executed, demonstrating the model's capacity to simulate real-world lighting interactions. Background crowd elements appear natural, though minor blending artifacts could be improved to achieve even greater realism.

Table 1: Comparison of Text-to-Image Generation Models.

Model	FID ↓	IS ↑	Text-Image Alignment (%)
Proposed Model	18.2	5.6	72
AttnGAN	23.3	4.4	70
DALL-E	10.4	6.2	85
BigGAN	12.8	6.8	65
StackGAN++	27.1	4.1	62

The above table compares the performance of the proposed model with several existing text-to-image generation models using three key evaluation metrics: Fréchet Inception Distance (FID), Inception Score (IS), and Text-Image Alignment Accuracy. These metrics are essential for assessing the quality, realism, and relevance of the generated images with respect to the given textual descriptions.

In terms of FID, the proposed model achieves a score of 18.2, which reflects a reasonable level of image quality. A lower FID score indicates that the generated images are closer to real images in terms of feature distribution. However, the proposed model's FID is higher than that of DALL-E (10.4), suggesting that there is still room for improvement in making the generated images more realistic.

The Inception Score (IS) of the proposed model is 5.6, which is lower than DALL-E (6.2) and BigGAN (6.8). IS measures the quality and diversity of the generated images. Higher values correspond to images with greater diversity and clarity. The slightly lower IS of the proposed model indicates that, although the images are coherent, they may lack the same level of detail or variation found in models like DALL-E and BigGAN.

The Text-Image Alignment Accuracy metric, which measures the degree to which the generated images align with the textual descriptions, is 72 percent for the proposed model. This is a competitive score, particularly in comparison to AttnGAN (70 percent), but still falls behind DALL-E (85 percent). The results suggest that the proposed model is effective at aligning images with text, though improvements in this area could lead to even more accurate and coherent image generation.

The model has limitations such as reduced accuracy with complex scenes, reliance on large, annotated datasets for training, and possible difficulties in handling highly abstract or novel concepts. Future work may address these challenges by incorporating additional training data and more advanced architectures that improve the quality and versatility of generated images. Overall, the model showed a strong ability to generate images that matched the descriptions, highlighting its effectiveness in interpreting and visualizing both object-related and action-oriented textual input. The use of Transformer-based text encoding, alongside CNNs like InceptionV3 for image feature extraction, allowed the model to create realistic images by focusing on key features of the text. However, finer details like texture variation and minor elements in the background require further refinement. This work serves as a basis for more developments in the future with the potential of a much more diversified and realistic image output from the text-to-image generation models.

5 CONCLUSION

In this paper, we presented a text-to-image generation model that effectively combines Transformer-based text encoders with CNNs for image feature extraction. The model achieved a solid accuracy of 72 percent, successfully generating visually coherent images from textual descriptions. The system performed particularly well with simple, well-defined objects, such as animals and basic scenes, where it showed higher accuracy. Although the model had difficulties with more complex scenes, still achieving an accuracy of 65 percent in such cases, it was still able to produce relevant results.

The attention mechanism study showed its potential in aligning text with image features but needs further improvement, especially when handling more complex or abstract scenes. The scope of future improvements lies in enhancing the model's generalization capabilities to better handle a wider variety of textual descriptions, especially those involving complex or less-defined concepts. To achieve this, the Attention Mechanism could be further refined, and additional architectures, such as GANs, could be explored to improve the quality of generated images, particularly in more intricate scenarios.Additionally, significant optimizations in computational efficiency are crucial for improving scalability and enabling real-world implementation. With these advancements, the model's ability to generate more realistic, diverse, and contextually accurate images will be enhanced, allowing it to cover an even broader range of textual descriptions.

REFERENCES

X. Bai, L. Wang, Z. Lu, and X. Zhang. Image generation using vision transformers: A detailed review. *Neural Networks*, 138:48–60, 2021.

Y. Dong, F. Li, Y. Xie, Z. Zhang, Z. Zhang, and Y. Yang. Cogview: Mastering text-to-image generation via transformers. *arXiv*, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014.

T. Karras, M. Aittala, S. Laine, J. Hellsten, S. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8119–8128, 2021.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv*, 2013.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, M. Chen, A. Radford, and I. Sutskever. Zero-shot text-to-image generation. *arXiv*, 2021.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. In *Advances in Neural Information Processing Systems*, volume 35, pages 14561–14575, 2022.

R. Rombach, A. Blattmann, D. Lorenz, M. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *arXiv*, 2022.

C. Saharia, W. Chan, S. Saxena, L. Li, J. Fleet, and M. Norouzi. Imagen: Photorealistic text-to-image diffusion models. *arXiv*, 2022.

X. Tao, X. Li, X. Xie, H. Zhang, J. Zhang, and Y. Zhang. T2f: Text-to-figure generation via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5726–5735, 2021.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

W. Wang, Y. Li, Q. Li, J. Liang, and M. Sun. Text-to-image generation with vision-language pretrained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3056–3065, 2022.

H. Xu, Z. Xu, Y. Li, Z. Yang, and Y. Li. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1324, 2018a.

T. Xu, J. Ba, J. K. Li, and A. Gupta. Attngan: Fine-grained text-to-image generation with attention mechanism. *IEEE Transactions on Image Processing*, 27(4):2224–2235, 2018b.

X. Zhang, Y. Xie, C. Zhang, and X. Liang. Text2image: Text-to-image synthesis with deep learning. *Journal of Visual Communication and Image Representation*, 72:102783, 2020.

Z. Zhong, D. Liu, Y. Qi, Y. Tang, Z. Zeng, Y. Zhang, and L. Liu. Learning to generate images from captions with attention. In *International Conference on Machine Learning (ICML)*, 2018.