Journal of Information Systems Engineering and Management
2025, 10(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Semantic-aware Mapping for Text-to-Image Synthesis

[1]Khushboo Patel, [2]Parth Shah

[1]U and P U. Patel Department of Computer Engineering, Chandubhai S. Patel Intitude of Technology (CSPIT), Faculty of Technology and Engineering (FTE), Charotar University of Science Technology (CHARUSAT), Changa, Gujarat, India E-mail: khushboo30990@gmail.com

[2]Smt. Kundanben Dinsha Patel Department of Information Technology, Chandubhai S. Patel Intitude of Technology (CSPIT), Faculty of Technology and Engineering (FTE), Charotar University of Science Technology (CHARUSAT), Changa, Gujarat, India

E-mail: parthshah.ce@charusat.ac.in

ARTICLE INFO

Received: 15 Oct 2024
Revised: 20 Dec 2024
Accepted: 28 Dec 2024

ABSTRACT

This study explores the fast-progressing domain of Text-to-Image (T2I) synthesis, which aims to bridge the gap between language and visual comprehension. The main emphasis is on the crucial significance of Generative Adversarial Networks (GANs), which have transformed the process of image formation, with a specific emphasis on the impact of conditional GANs. The conditional models enable controlled image generation, and their influence on the production of high-quality images is extensively analyzed. We propose a novel method of generating semantically aware embeddings from the input text description which learns better mapping to generate the output image. Moreover, the paper examines the crucial significance of datasets in T2I research and investigates the development of T2I approaches. Ultimately, the research highlights the persistent difficulties in assessing T2I models, with a particular emphasis on image quality measurements. It emphasizes the necessity for complete evaluation methods that take into account both visual realism and semantic coherence. Experimental results demonstrate that our approach yields considerable performance over existing approaches for text to image generation.

Keywords: text to image synthesis, image generation, semantic aware mapping, GAN, machine translation, embed- dings

1. Introduction

The advancement in deep learning has facilitated substantial advancements in image processing techniques and com- puter vision applications over the past few years. Image synthesis refers to the process of generating new images and modifying existing ones. Image synthesis is a compelling and important task because of its wide range of practical applications in various industries such as virtual reality, computer-aided design, art creation, picture editing, and video games.

Humans possess the ability to promptly generate mental images in their thoughts to depict the essence of things they hear or read. It is uncommon for us to pause and reflect on our innate ability to effortlessly visualize and understand the intricate connection between language and the visual surroundings. Visual mental imaging, sometimes known as "seeing with the mind's eye," plays a crucial role in several cognitive activities such as memory, spatial navigation, and reasoning. Developing a system that can generate graphics that faithfully depict written descriptions and understand the connec- tion between vision and words, drawing inspiration from human perception of scenes, signifies a major advancement in achieving human-level intelligence.

The advent of Generative Adversarial Networks (GANs) enabled the unsupervised training of generative models specifically for images. GANs have generated significant attention and advanced research endeavours in the field of image synthesis [1,2,7,9,13,14,15,18]. The aim of generating an image is treated as a competition between two artificial neural networks. A discriminator network is trained to differentiate between authentic and synthetic images, while a generator network is instructed to generate lifelike samples. The generator's training purpose is to deceive the discriminator. This approach has proven to be effective in various applications, including image super resolution, image in-painting, style transfer, data augmentation, image-to-image translation, high-resolution synthesis of human faces, and representation learning. Although the approaches explored in this study can be used in several image domains, most of the T2I research focuses on methods that generate visually precise, photographic, and authentic images.

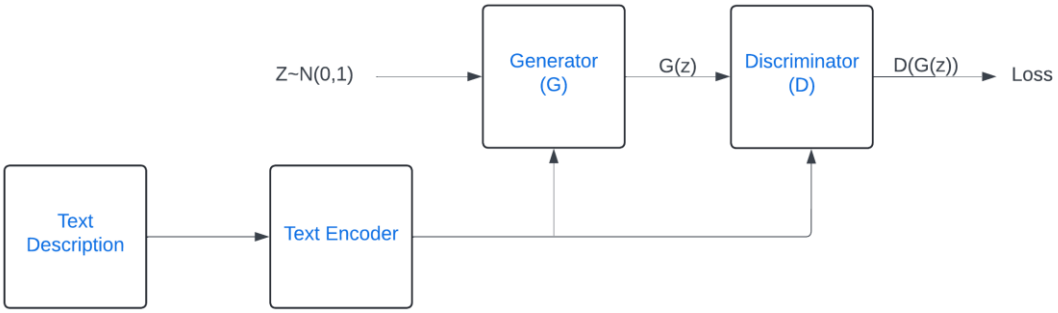

Fig 1. A typical text to image synthesis framework

Fig 2. Typical textual description and corresponding image from the dataset

The most common deep learning models used for this purpose are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and generative models such as Variational Auto-encoders (VAE), Generative Adversarial Networks (GAN) and diffusion models. The key contributions of this paper are:

1. A novel text encoding mechanism is proposed to ensure semantic-aware mapping between input and output.
2. Building a deep learning model for generating images with improved performance.

The paper is organized as follows. Section 2 covers the background of text to image synthesis wherein we describe the foundational framework of generative adversarial networks which are majorly used to accomplish this task. We also discuss in this section different components of T2I systems including text encoder, and variants of T2I models available in literature. The proposed approach is explained in Section 3 and experimental setup alongwith dataset, performance metrics and discussion of results is carried out in Section 4. Finally, Section 5 concludes the paper.

2. Background

2.1. *Generative Adversarial Networks (GAN)*

GANs [26] are comprised of two components: the generator and the discriminator. Generator generates new data instances, typically images, from random noise. The goal of the generator is to

produce data that is indistinguishable from real data. Discriminator, on the other hand, evaluates the generated data, determining whether it's real or fake. The discriminator's objective is to correctly classify real and fake data.

These two networks are trained simultaneously in a competitive setting. The generator improves its ability to produce realistic data by receiving feedback from the discriminator, which in turn becomes better at distinguishing real from fake data. Over time, this adversarial process leads to the generator producing increasingly convincing data.

Figure 3 illustrates the fundamental concept of GANs, which involves a dynamic adversarial game between a generator and a discriminator during the training phase.

A typical GAN comprises a discriminator network D(x), where x represents images generated from a distribution pg, and a generator network G(z) that generates images using noise z drawn from a prior noise distribution pz. The training procedure employs a two-player game to instruct the generator on generating deceptive images that can fool the discriminator. This is achieved by capturing the real data distribution and training the discriminator to distinguish between artificially made images and authentic ones. This implies that, in a formal context, the training can be conceptualized as a value function V (D, G) in a two-player minimax game. In this game, the generator G(z) is taught to minimize the likelihood of being classified as fake by the discriminator, while the discriminator D(x) is trained to maximize the log- likelihood it allocates to the proper class.

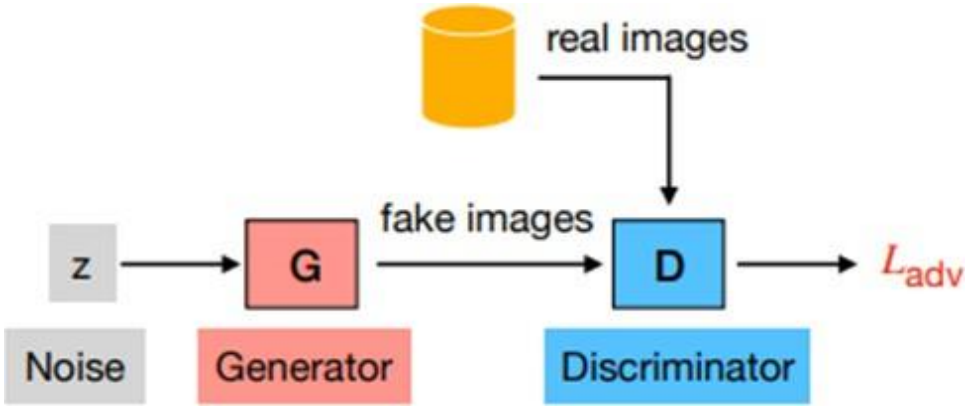

Fig 3. GAN Architecture typically used in T2I scenario

2.2. *Conditional GANs*

While creating novel, lifelike samples is intriguing, mastering the process of creating images has great practical significance. The conditional GAN (cGAN) was introduced by Mirza et al. to determine the digit to be generated with the help of an imposed condition on the model. Thus, the model was trained to generate images of specific class label only. To improve conditional GAN training and compute word features corresponding to image subregions (image-text similarity at the word level), the cGAN objective function was extended in several ways. During training, the BiLSTM matches the intermediate features of an image classifier that has already been learned. The majority of the works that followed adopted the use of BiLSTM in AttnGAN [10] for caption encoding since it was first introduced. Nonetheless, some recent works obtain text embeddings by using transformer-based models that have already been trained, like BERT.

2.3. *Encoding Text*

Producing an embedding from textual representations that functions effectively as a network conditioning variable is challenging. Reed et al. use a pre-trained character-level convolutional recurrent neural network (char-CNN-RNN) to retrieve the text encoding of a textual description. With the help of the class labels, the char-CNN-RNN is pretrained to identify a correspondence function between text and image. Text encodings with visual discrimination result from this.

More text embeddings were generated during training by simply extending the embeddings of two training captions. TAC-GAN [3] employed vectors referred to as Skip-Thought vectors. Rather than

using the fixed text embedding produced by a pre-trained text encoder, Conditioning Augmentation (CA) was suggested in the StackGAN architecture [15]. This can be accomplished by randomly sampling the latent variable from a Gaussian distribution whose covariance matrix and mean are functions of the text embedding. The regularization term employed in training was the Kullback-Leibler (KL) divergence between the conditioned Gaussian distribution and the standard Gaussian distribution. By doing this, the training manifold becomes smoother and more training pairs are formed. This method was used by several T2I techniques that followed. In the authors' view, Sentence Interpolation (SI) is comparable to CA since it is deterministic and offers a continuous and smooth embedding space throughout training. The authors replaced the char-CNN RNN with the bi-directional LSTM (BiLSTM) in order to extract feature vectors in AttnGAN. Next, a feature matrix for every word was created by concatenating the BiLSTM's hidden states. Concatenating the most recent hidden states yields the global sentence vector. Pretraining a Deep Attentional Multimodal Similarity Model yields the text encoder.

2.4. Text to Image Synthesis Methods

The first T2I method, proposed by Reed et al., relies on the whole sentence embedding from a text encoder that has already been trained for the generating process. Through training, the discriminator learns to distinguish between genuine and fake image-text pairs. One may see the first T2I model as a logical extension of a cGAN, since a text embedding ϕ eliminates the requirement for conditioning on a class label y. Three separate pairings are given to the discriminator in GAN-INT-CLS [16]: an artificially generated image and its corresponding text, an actual image and text that do not match, and an actual image and text that match. This method forces the discriminator and the generator to focus on timing the realistic images with the input text in addition to creating them. Using one-hot encoded class labels, TAC-GAN adds an additional auxiliary classification loss—inspired by AC-GAN. It differs from GAN-INT-CLS in this way.

2.5. Stacked Architectures

GAN-INT-CLS generated low-resolution 64 × 64 pixel images, whereas TAC-GAN generated 128 × 128 pixel images. Numerous scholarly articles that followed recommended using multiple stacked generators to allow T2I models to synthe- size better-quality images. A textual conditioning vector and a random noise vector are used in StackGAN's first stage to produce a coarse 64x64 pixel image. This first image and the text embedding are then passed to a second generator, which creates an image of 256 × 256 pixels. During both stages, a discriminator is trained to distinguish between image-text pairs that match and those that don't. By using an end-to-end framework that trains three generators and discriminators jointly, StackGAN++ improved the design. The simultaneous approximation of the multi-scale, conditional, and uncondi- tional image distributions is the goal of this framework. The authors suggested sampling text embeddings from a Gaussian distribution in place of fixed text embeddings in order to produce a smooth conditioning manifold. An additional color- consistency regularization term was added to reduce the disparities in pixel average and dispersion between sizes. This motivates the network to produce images with a consistent color scheme and structure at every scale.

Two generators are utilized by FusedGAN [20]. One for conditional image synthesis and another for unconditional picture synthesis. Because of their partially overlapping latent space, these generators can be used for both unconditional and conditional creation. This method entails training unconditional and conditional distributions simultaneously.

To generate 512x512-resolution images, HDGAN [11] used hierarchically-nested discriminators at several interme- diary levels. This method does away with the requirement for multiple generator networks. Stated differently, each resolution level in the adversarial game uses different discriminators at different points in the generator's evolution. The discriminators are trained to maximize the matching aware pair loss and to discriminate between generated and actual pic- ture patches. Regularization of the hidden layers of the generator is achieved by using discriminators at greater resolutions to improve output consistency across scales.

In the same way, PPAN [21] makes use of one generator and three different discriminators. In order to combine high- resolution, semantically less significant features with low-resolution, semantically more significant attributes through lateral links, the PPAN generator uses a hierarchical structure. Using characteristics taken from a pre-trained VGG network, the researchers also added an auxiliary classification loss and a perceptual loss to the training process. HfGAN, on the other hand, makes use

of a hierarchically fused design and a single discriminator. The procedure entails adaptively combining multi-scale global parameters that are extracted from many phases. This makes it possible for lower-resolution, spatially coarse feature maps that incorporate and regulate the overall semantic structure of the final image to serve as a guide for the generation of fine details. The authors took their cues from ResNet and employed identity addition, weighted addition, and shortcut connections.

2.6. Attention Mechanisms

Researchers have worked to embed attention mechanism into this task also in the past [9,13,19]. By giving relevant input components a higher weight than unimportant ones, attention techniques enable the network to focus on specific input properties. The remarkable capacity of attention as a method has tremendously benefited in the development of language and vision applications. StackGAN++ [4] is the foundation for the iterative process AttnGAN, which adds attention mechanisms. In addition to the global sentence vector, the attention mechanism enables the network to provide detailed characteristics depending on pertinent words. It is recommended that, when building the image, the network concentrate on the most pertinent phrases for each aspect of the image. Sentence and word level information is integrated into the Deep Attentional Multimodal Similarity Model (DAMSM) [5] to determine the similarity between the input text and the output picture. Training involves performing this calculation, which yields a loss. Huang et al. added a mechanism to the grid-based attention model that links word phrases and object-grid regions. Auxiliary bounding boxes determine the object-grid zones. In addition to sentence and word data, phrase features are extracted using part-of-speech tagging.

The authors of SEGAN [6] proposed an attention competition module that would simply focus on essential words, as opposed to giving each word in the sentence an attention weight (as is done in AttnGAN). They achieved this by using an attention regularization term that only keeps the attention weights for visually meaningful words. ControlGAN

[8] is capable of both visual property modification and Text-to-Image (T2I) production. It can update the description without changing the backdrop or posture, but it won't change the category, texture, or color. Word-level spatial and channel-wise attention are used by the authors to create a generator that can create visual regions that match the most relevant words. Channel-wise attention creates associations between words (e.g., "head" and "wings" for CUB-200 birds) and semantically relevant elements. However, color information is the primary focus of spatial attention. A word-level discriminator uses the correlation between words and areas of an image to distinguish between various visual attributes and give the generator comprehensive training signals.

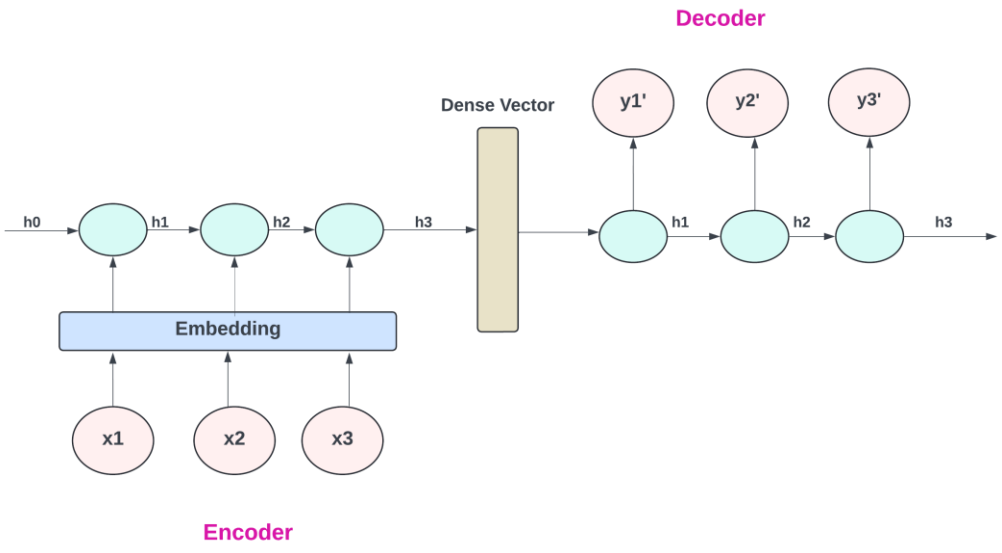

Fig 4. Proposed semantic aware text-encoder approach

3. Method

Initializing a machine learning model with appropriate probability distribution yields more relevant and quick con- vergence. Qiao et al. in [16] utilized available information to conduct the generation of images using prior knowledge. Motivated from their work, we propose a novel NMT-guided text encoder mechanism for text-to-image synthesis which helps to derive semantic aware embeddings from the given input text description. This embedding helps the model to learn the mapping between words of a text description and the image to be generated. Using an LSTM based semantic aware text embeddings to encode text input, we train a deep convolutional generative adversarial network (DC-GAN). Feed-forward inference is carried out by both the discriminator network D and the generator network G, depending on the text characteristic.

A neural machine translation (NMT) is an application of sequence-to-sequence learning. Such a system is comprised of encoder and decoder module. The encoder is responsible for encoding the source sentence into a dense vector which contains the entire source sentence information into itself. This vector is used as input to decoder which gets initialized to generate target language sentence. We do not use this encoded representation for decoding purpose, instead we feed it as text embedding to our T2I framework.

Fig. 4 illustrates the architecture of the proposed method. As can be seen, for example, a source text of length 3 (number of words) is given as input to consecutive time stamps to LSTM units at respective time stamps. The vector "h3" contains information for the entire source sentence, and is used as dense vector. This vector is not used for decoding purpose, but is extracted from the network and used as input to the text encoder module of our system.

4. Experiments

In this section, we describe the experimental methodology used in this work.

4.1. Dataset

The datasets Oxford120 Flowers, Caltech UCSD Birds (CUB-200) [10], and COCO [23] are frequently utilised in T2I studies. We evaluate the suggested model using the CUB bird dataset. There are 11,788 images in the CUB collection that faithfully represent 200 different species of birds. For every bird image, there are ten textual descriptions. The CUB bird dataset has 8,855 training images (150 species) and 2,933 test images (50 species).

4.2. Performance Evaluation

Measuring progress and ensuring fair comparisons require access to automated assessment tools that reliably evaluate performance. Since numerous features (such visual reality and variety) may resemble an outstanding image, evaluating created photographs is especially challenging. However, creating lifelike images is only one aspect of a robust T2I model. Another important stage is to assess how well the prepared images and text descriptions match semantically. The following sections cover the automated metrics currently used by the T2I community and the approach used in user studies. We next evaluate the necessary conditions for valid measurements, highlight the flaws in current assessment protocols, and provide suggestions for assessing T2I techniques with the available metrics.

Fig 5. Generated Bird images

Table 1. Performance comparison of our approach on CUB dataset with existing systems

Architecture	FID(↓)	IS(↑)
StackGAN [14]	51.89	3.7
AttnGAN [9]	23.98	4.36
DF-GAN [12]	12.14	4.86
Text2Image [22]	-	5.0178
ManiGAN [17]	-	8.47
Our approach	14.85	9.22

4.2.1 Inception score

It is employed to provide a numerical summary of the generated images' quality [24]. The model is used to classify a large number of generated images in order to determine the inception score. More specifically, each class's probability of the image being in it is predicted. To determine how an image fits into each class and how diverse the collection of photographs is, these probabilities are summed together. Higher value of inception score indicates high quality images.

4.2.2 Fre´chet Inception Distance (FID)

This approach is a recently developed measure that extracts features from an intermediary layer of images using an inception check [25]. Then, a multi-variate Gaussian distribution with the mean and covariance is used to extract features from this, modeling it as a data distribution. The score is determined by calculating the distance between the synthetic and actual images. The synthetic and actual data distributions are closer together the lower the FID score. Better picture quality and variety are therefore indicated by lower FID values.

4.3. Results and Discussion

Fig 5 and 6 show some sample images generated by the model for a given text description. Our approach gener- ates images semantically consistent images with appropriate backgrounds. The model is essentially able to capture the information from all the attributes.

Results in Table 1 show a comparison of our proposed approach with existing architectures in terms of FID and Incep- tion score. Higher the IS value, better the performance; and a smaller value of FID is preferable for good performance. It can be seen that our approach is able to outperform the models in Inception Score. Further, FID score is also showing promising performance.

6a

5. Conclusion and Future Work

In this paper, we bring forth a novel technique that learns superior mapping to produce the output image by creating semantically aware embeddings from the input text description. The outcomes of

our experiments show that our method outperforms other methods for text to image generation by a significant margin. It is evident that our method can beat the models in Inception Score. Additionally, the FID score demonstrates encouraging performance.

Although it is recognized that the scores can vary based on the implementation used, the quality of the photos, and the number of samples, there are many anomalies that are hard to explain. The absence of evaluation code in open-sourced programs and the lack of detailed explanations of the assessment techniques are common problems. Furthermore, scores that are different from those reported in published studies can be obtained by updating the baseline method codes. While most of the discrepancies are insignificant and do not impact the final score, a few are substantial enough to cast doubt on the fairness and dependability of the comparisons. To improve repeatability, we advise researchers to provide thorough justifications for their assessment procedure, elucidate any disparities that may occur, and make their evaluation code publicly accessible.

In future, we aim to investigate the impact of the proposed approach on other datasets for text to image generation.

Evaluation using other performance metrics will also be our target to assess the efficacy of the

6. References

[1] Ming Tao, Hao Tang, Fei Wu, Xiaoyuan Jing, Bingkun Bao and Changsheng Xu: A Simple and Effective Baseline for Text-to-Image Synthesis.

[2] Z. Wang, Z. Quan, Z. Wang, X. Hu, Y. Chen, Text to image synthesis with bidirectional generative adversarial network, in: IEEE International Conference on Multimedia and Expo, 2020, pp. 1–6.

[3] Dash, A., Gamboa, J. C. B., Ahmed, S., Liwicki, M., Afzal, M. Z. (2017). Tac-gan-text conditioned auxiliary classifier generative adversarial network. arXiv preprint arXiv:1703.06412.

[4] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D. N. (2018). Stackgan++: Realistic image syn- thesis with stacked generative adversarial networks. IEEE transactions on pattern analysis and machine intelligence, 41(8), 1947-1962.

[5] Karras, T., Laine, S., Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

[6] Cha, M., Gwon, Y. L., Kung, H. T. (2019, July). Adversarial learning of semantic relevance in text to image synthesis. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 3272-3279).

[7] H. Tan, X. Liu, X. Li, Y. Zhang, B.-C. Yin, Semantics enhanced adversarial nets for text-to-image synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10501–10510.

[8] Li, B., Qi, X., Lukasiewicz, T., Torr, P. (2019). Controllable text-to-image generation. Advances in neural informa- tion processing systems, 32.

[9] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, Attngan: Fine-grained text to image generation with attentional generative adversarial networks, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2017, pp. 1316–1324.

[10] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P. (2010). Caltech-UCSD birds 200.

[11] Z. Zhang, Y. Xie and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial net- work", Proc. CVPR, pp. 6199-6208, Jun. 2018.

[12] Tao, M., Tang, H., Wu, F., Jing, X. Y., Bao, B. K., Xu, C. (2022). Df-gan: A simple and effective baseline for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16515-16525).

[13] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In International conference on machine learning, pages 7354– 7363. PMLR, 2019.

[14] 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.

[15] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis.

[16] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Learn, imagine and create: Text-to-image generation from prior knowledge.

[17] Li, B., Qi, X., Lukasiewicz, T., Torr, P. H. (2020). Manigan: Text-guided image manipulation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7880-7889).

[18] Bingchen Liu, Kunpeng Song, Yizhe Zhu, Gerard de Melo, and Ahmed Elgammal. Time: text and image mutual- translation adversarial networks.

[19] Yuchuan Gou, Qiancheng Wu, Minghao Li, Bo Gong, and Mei Han. Segattngan: Text to image generation with segmentation attention. arXiv preprint arXiv:2005.12444, 2020

[20] Bodla, N., Hua, G., Chellappa, R. (2018). Semi-supervised FusedGAN for conditional image generation. In Pro- ceedings of the European conference on computer vision (ECCV) (pp. 669-683).

[21] G. Lianli, C. Daiyuan, S. Jingkuan, X. Xing, Z. Dongxiang, S. Hengtao, Perceptual pyramid adversarial networks for text-to-image synthesis, Proc. AAAI Conf. (2019) 8312–8319.

[22] Liao, W., Hu, K., Yang, M. Y., Rosenhahn, B. (2022). Text to image generation with semantic-spatial aware gan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 18187-18196).

[23] Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S. (2016). Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140.

[24] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In Advances in Neural Information Processing Systems, pp. 2234–2242, 2016.

[25] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in Neural Information Processing Systems, pp. 6629–6640, 2017.

[26] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

Authors' Profiles

Khushboo Patel, Khushboo Patel is a research scholar at Charusat university, Changa, Gujarat. Her research area includes machine learning and deep learning including generative models. She has published book chapter on international level.

Parth Shah Dr Parth Shah received his PhD in cloud computing at the Charusat University, Changa, Gujarat. He is the head and professor at the Department of Information Technology, Charusat University, Changa, Gujarat.

His research interests include issues related to machine learning and deep learning, cloud computing , Internet of things and cyber security. He is author of many journal papers, conference papers and book chapters which are published at international and national level.