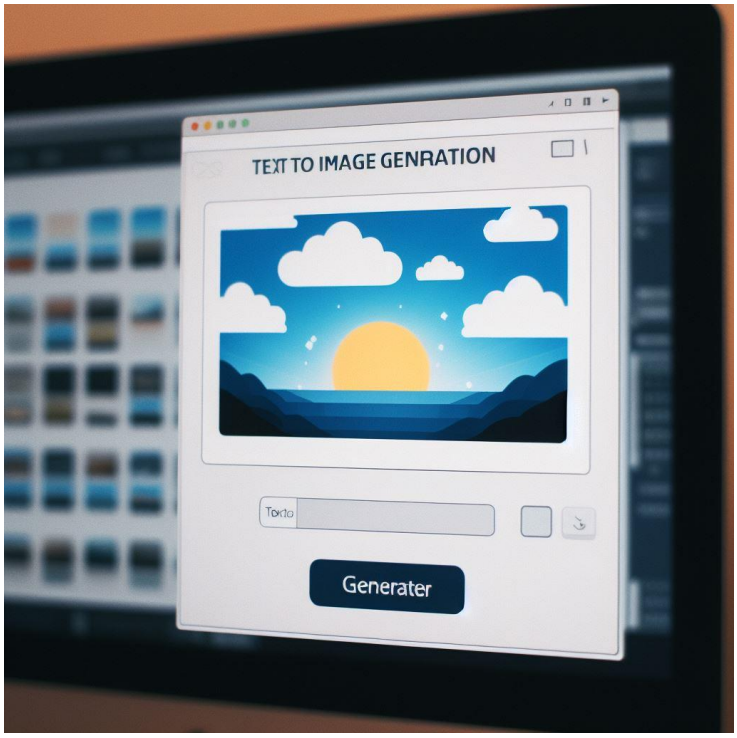
Advances in Text-to-Image Generation: From Basic Techniques to Deep Learning Innovations

Godavarthi Sai Nikhil and Anish Borkar
(CSE, B-TECH 3rd Year, BML MUNJAL UNIVERSITY)



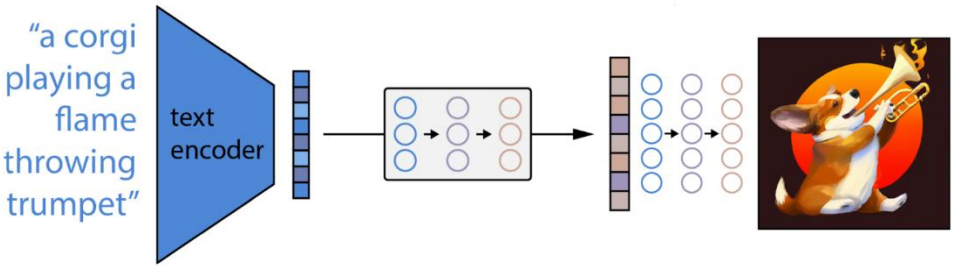
Abstract

This research focuses on the imperative process of text-to-image generation in today's digital landscape. Despite the proposal of numerous deep learning methods for this task, their efficiency in swiftly creating meaningful images from text remains a challenge. Our study comprehensively evaluates various deep learning approaches, assessing their proficiency in image generation while addressing challenges related to image quality, feature extraction, and the overall effectiveness in creating coherent visual representations. We explore the intricacies of accurately capturing details, discuss issues such as low image quality, and evaluate these methods across diverse datasets. Drawing on existing research, we propose strategies to assess and enhance image quality. The study highlights practical applications derived from literature findings and discusses commonly used tools in these experiments. Notably, we reference specific models like GANs, StackGAN, BERT, DALL-E, and others, providing a nuanced understanding of our investigation. Finally, we outline future research directions aimed at advancing text-to-image generation using deep learning.

Keywords: Text-to-Image Generation · Deep Learning · Visual Representation · Feature Extraction · Evaluation Strategies.

1. Introduction:

Text-to-image generation represents a compelling convergence of natural language processing (NLP) and computer vision, driven by advancements in deep learning. This innovative field aims to translate textual descriptions into coherent and meaningful visual content. The potential applications span diverse domains, from creative content creation to aiding the visually impaired, highlighting its significance in modern technology.



The challenges inherent in text-to-image generation are multifaceted, ranging from capturing nuanced semantics in text to producing high-fidelity, contextually relevant images. Deep neural networks, especially Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have emerged as pivotal tools in addressing these challenges in [5] and [7]. Through training on diverse datasets of paired text-image examples, these models learn to map textual inputs to corresponding visual outputs, capturing intricate relationships between words and their visual representations.

The training process involves exposing the model to a diverse dataset, enabling it to learn the mapping between textual prompts and visual outputs. This equips the model to generate images aligning with given textual descriptions. The outcome is a model capable of creative expression and content synthesis, offering potential applications in art, design, e-commerce, virtual environments, and accessibility.

Text-to-image generation finds applications across various domains, empowering content creators to automate visual content generation for storytelling, marketing, and design. In e-commerce, it enhances the shopping experience by creating product images based on textual descriptions. Notably, the technology can assist individuals with visual impairments by providing detailed visual descriptions, showcasing its potential to drive positive societal impact.

As the field continues to evolve, researchers are exploring ways to refine generated images for improved realism and diversity. Ethical considerations, such as responsible technology use and mitigating biases in content generation, are crucial aspects of ongoing research. The dynamic landscape of text-to-image generation contributes to the broader goal of seamlessly integrating language and vision, offering novel applications and enriching our digital interaction.

In this journey, various models like LeicaGAN, AttaGAN, StackGAN, DALL-E, and others have played crucial roles, shaping the landscape and pushing the boundaries of what text can inspire visually in [1], [2] and [10].

While text-to-image generation holds promise, challenges persist in accurately translating textual nuances into diverse and realistic visuals. Models like LeicaGAN and AttaGAN in [1], [6] contribute to ongoing innovations, addressing complexities in this dynamic field. Researchers strive for more sophisticated algorithms and ethical content generation, emphasizing realism and inclusivity.

The future of text-to-image generation looks promising, with ongoing research focusing on human-centric impacts. Models like StackGAN and DALL-E in [4], [11] pave the way for inclusive and ethical technology use. As these capabilities expand, the integration of these models into daily life has the potential to reshape digital interactions, offering enriched experiences across diverse domains.

2. Dataset Description

2.1 Introduction to Datasets:

Datasets are crucial for the development of text-to-image (T2I) models, serving as a foundational resource for learning complex relationships between textual descriptions and visual content.

2.2 Characteristics of Datasets:

1. **Diversity of Visual Concepts:**

   Datasets encompass a wide range of visual concepts, object categories, and scenes, exposing T2I models to diverse contexts for accurate image generation.
2. **Contextual Understanding:**
   Datasets with contextual scenes contribute to T2I models understanding of object relationships in real-world scenarios, enhancing the coherence of generated images.
3. **Training Robust Models:**
   Large-scale datasets, like ImageNet or MS COCO, provide ample training samples crucial for building robust T2I models that generalize well and generate diverse, high-quality images.
4. **Benchmarking and Evaluation:**
   Datasets serve as benchmarks for evaluating T2I model performance. Researchers can compare outputs on standardized datasets, facilitating quantitative assessment of capabilities and limitations.

2.3 Significance in Text-to-Image Generation Model Training & Evaluation:

Datasets used in text-to-image generation research serve dual roles crucial to model development:

Training:

For text-to-image generation models, especially those leveraging deep learning, robust training is essential. This process teaches the model how to proficiently convert textual descriptions into meaningful visual content. The diversity and intricacy of the textual prompts in the dataset directly influence the model's capacity to handle real-world text-to-image synthesis tasks effectively.

Evaluation and Testing:

Datasets play a pivotal role in evaluating the performance of text-to-image generation models. When testing models on a set of textual descriptions they haven't encountered during training, researchers can gauge their accuracy, efficiency, and generalization capabilities. This evaluation process is crucial for assessing the model's real-world applicability and overall effectiveness in generating visual content from diverse textual inputs.

Dataset	Used in Papers	Category	Description	Why Use in Text to Image
CUB 200	[1], [2], [3], [5], [8], [12], [13]	Birds	Contains 200 bird species with 11,788 images.	It offers a rich variety of bird poses and backgrounds.
LSRVC2012	[4], [11]	Objects	Contains over 1.2 million images covering 1,000 object categories.	It serves as a comprehensive resource for training models on various objects.
MS COCO	[2], [6], [9], [14], [15]	Objects	Contains 3,28,000 images covering 80 objects categories.	Comprises over images covering 80 object categories, making it suitable for diverse visual recognition tasks.
Oxford-102 flower	[2], [5], [7], [10], [14]	Flowers	Contains 8,189 images spanning 102 flower categories.	It provides a specialized dataset for models focused on floral imagery in text-to-image generation.

Table: Main Datasets Used Description

3. History of Text-to-Image Generation:

3.1 Early Approaches:

Text-to-image generation has undergone significant evolution. In its early stages, the focus was on rule-based systems and basic templates to create rudimentary visual representations of textual input.

1. **Rule-Based Systems (Pre-2000s):** Before the advent of deep learning, early attempts at text-to-image generation involved rule-based systems and handcrafted algorithms. These systems often struggled to produce realistic and diverse images.

2. **Early Generative Models (2000s):** In the early 2000s, generative models like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) began to gain popularity. These models showed promise in generating images but not initially optimized for text-to-image synthesis.

Advantages of Early Approaches: Initial exploration of generative models for image generation. And their drawbacks are Limited realism and diversity in generated images.

3.2 Deep Learning Era:

1. **Introduction of GANs (2014):** The breakthrough moment for text-to-image generation came with the introduction of GANs by Ian Goodfellow and his colleagues in 2014. GANs, with their adversarial training framework, demonstrated remarkable capabilities in generating high-quality images from random noise.
2. **Conditional GANs (2015):** Conditioning GANs on textual descriptions opened the door to text-to-image synthesis, allowing guided image generation.
3. **StackGAN (2017):** Addressing the hierarchical nature of image generation, StackGAN significantly improved the quality of generated images in [8] and [15].
4. **LeicaGAN and AttaGAN**: These models contributed to the advancement of text-to-image generation, each bringing unique approaches to enhance the quality and diversity of generated images in [1].
5. **ControlGAN:** ControlGAN added to the growing repertoire of techniques for fine-tuning and controlling the output of GANs in text-to-image synthesis.
6. **SAGAN and AC-GAN:** These contributed to improving the spatial attention and class conditioning aspects of GANs.

Advantages of Deep learning Era: Improved spatial attention and class conditioning. And drawbacks are Increased computational complexity.

3.3 Introduction of Neural Networks:

Neural networks, particularly CNNs, RNNs and LSTMs, played a pivotal role in capturing sequential dependencies within text used in [4] and [10].

Advantages of Neural Networks: Enhanced sequential modeling for better text understanding. and drawbacks are Limited ability to capture long-term dependencies in certain cases.

3.4 Progress in Natural Language Processing (NLP):

1. **BERT and Transformers (2018):** The introduction of BERT and transformer models marked a significant advancement in natural language understanding [9], providing better context-aware representations for text.
2. **CLIP (2021):** OpenAI's CLIP demonstrated breakthroughs in aligning vision and language, showing remarkable performance in cross-modal tasks, including text-to-image generation used in [11].

Advantages of NLP: Improved alignment between vision and language. and drawbacks are Increased computational requirements during training.

3.5 Recent Advances:

1. **DALL-E (2021):** Developed by OpenAI, it showcased the potential of generative models in creating diverse and high-quality images from textual prompts as shown in [11] and [14].
2. **GAN-INT-CLS, StackGAN++, HDGAN, and DualAttn-GAN:** These models were built upon the earlier successes, bringing improvements in various aspects of text-to-image synthesis such as image resolution in [10], attention mechanisms, and integration with language understanding.
3. **MirrorGAN:** It contributed to the ongoing exploration of novel architectures and techniques for generating realistic images from textual descriptions [13].

Method	Advantages	Limitations	Drawbacks	Common Applications
Rule-Based Systems	Simplicity and early exploration	Struggles to produce realistic and diverse images	Lack of adaptability, limited complexity	Early visual representation, basic image synthesis
Early Generative Models	Introduction to generative modeling	Not initially optimized for text-to-image synthesis	Limited capability to handle complex textual input	Basic image generation

GANs (2014)	Breakthrough in high-quality image generation	Requires careful training, mode collapse issues	Sensitive to hyperparameters, training instability	Diverse image synthesis, realistic content creation
Conditional GANs (2015)	Enables guided image generation based on textual input	May require large datasets for effective training	Mode collapse remains a challenge	Image synthesis guided by textual descriptions
StackGAN (2017)	Addresses hierarchical nature of image generation	Complex architecture, increased training complexity	Sensitivity to hyperparameters, potential for mode collapse	Improved quality in two-stage text-to-image synthesis
LeicaGAN and AttaGAN	Enhances quality and diversity of generated images	Model complexity may lead to increased training times	Limited scalability to large datasets	Image enhancement and diversity
ControlGAN	Provides fine-tuning and control over GANs output	Increased complexity and potential challenges in implementation	May require extensive fine-tuning for specific tasks	Controlled generation of specific image features
SAGAN and AC-GAN	Improves spatial attention and class conditioning	Increased computational complexity	Potential for increased training time	Improved spatial attention and class conditioning
Neural Networks (RNNs, LSTMs)	Improved capture of sequential dependencies within text	May struggle with long-range dependencies	Training can be computationally intensive	Sequential understanding in text-to-image synthesis

BERT and Transformers (2018)	Significant advancement in natural language understanding	Limited by context window, may struggle with certain tasks	Requires large pre training datasets	Context-aware representations for text
CLIP (2021)	Breakthrough in aligning vision and language	Computational complexity, resource-intensive training	Limited fine-tuning examples for specific tasks	Cross-modal tasks, text-to-image alignment
DALL-E (2021)	Demonstrates potential for diverse and high-quality image synthesis	Resource-intensive training, potential for mode collapse	Limited control over specific image features	Creative image generation based on textual prompts
GAN-INT-CLS, StackGAN++, HDGAN, and DualAttn-GAN	Improved image resolution, attention mechanisms, and language understanding integration	Increased model complexity, potential for overfitting	Requires careful tuning of hyperparameters	Enhanced text-to-image synthesis with improved resolution, attention, and language understanding integration
MirrorGAN	Contributes to ongoing exploration of novel architectures and techniques	Limited scalability to large datasets	Requires careful tuning of hyperparameters	Exploration of novel architectures for realistic image generation from textual descriptions

Table: Different Methods with Advantages and Drawbacks

4. Previous Contributions:

[1]. Generative Adversarial Text to Image Synthesis:

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016, June) proposed a method called LeicaGAN, consisting of a textual-visual co-embedding network (TVE), a multiple priors aggregation network (MPA) and a cascaded attentive generator (CAG). The text encoder was a pre-trained Bi-directional LSTM and the visual encoder was built upon the Inception-v3 model. The MPA network fused the sentence-level embeddings. This acted as an input in the CAG, where an attention block, two residual blocks, an up-sampling block, and a convolution layer make up the generator. Word and Sentence-context features were produced. Two adversarial losses were employed: a visual realism adversarial loss to ensure that the generators generate visually realistic images and a text-image pair-aware adversarial loss to guarantee the semantic consistency between the input text and the generated image. For effectiveness, LeicaGAN was compared with AttnGAN. CUB and Oxford-102 datasets were used and evaluation was done based on the Inception Score. LeicaGAN outperformed AttnGAN, on both the datasets.

[2]. Learn, Imagine, and Create: Text-to-Image Generation from Prior Knowledge:

Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019) proposed ControlGAN. For this, they introduced a word-level spatial and channel-wise attention-driven generator that could disentangle different visual attributes. Also, they proposed a word-level discriminator. The backbone architecture they used was AttnGAN and the text encoder was a pre-trained bi-directional RNN. Conditioning Augmentation was applied. The generator exploited the attention mechanism by incorporating a spatial attention module and a channel-wise attention module. The spatial attention module dealt with words with individual spatial locations. The model was experimented on CUB and MS COCO datasets. The model proposed was compared with AttnGAN and StackGAN++ and the performance metrics were Inception Score, R-precision, and L2 Error. ControlGAN gave the best results among the three for the CUB dataset. (Inception Score = 4.58 ± 0.09, Top-1 Acc(%) = 69.33 ± 3.23, L2 error = 0.18). For the COCO dataset, AttnGAN was the best in Inception Score and Top-1 Acc(%), but ControlGAN had the lowest L2 error, i.e., 0.17.

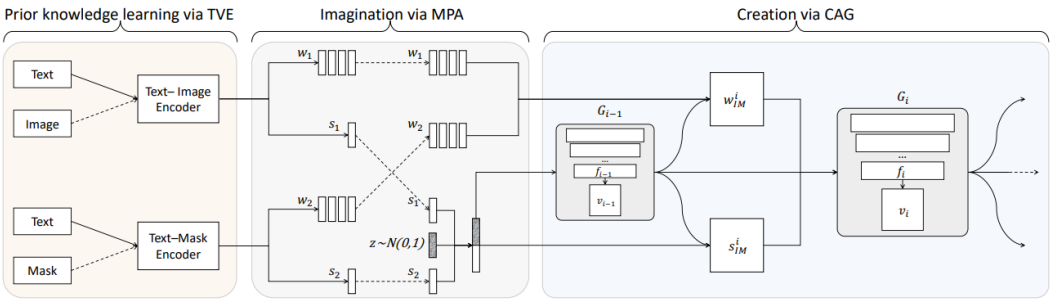


Figure 1: The LeicaGAN framework, which tackles the T2I problem by decomposing it into three phases, which are 1. multiple priors learning via text-visual co-embedding (TVE), 2. imagination via multiple priors aggregation (MPA) and 3. creation via a cascaded attentive generator (CAG).

Fig 4.1: Model Architecture

[3]. Controllable Text-to-Image Generation:

Li, B., Qi, X., Lukasiewicz, T., & Torr, P. (2019) proposed the Self-Attention Generative Adversarial Network (SAGAN). This compared to convolutional GANs, helps with modeling long-range, multi-level dependencies across image regions. Due to self-attention, the generator can draw images in which fine details at every location are carefully coordinated with fine details in distant portions of the image. Moreover, the discriminator can also more accurately enforce complicated geometric constraints on the global image structure. Spectral Normalisation was used in the generator and discriminator. Spectral normalization in the generator can prevent the escalation of parameter magnitudes and avoid unusual gradients. The experiment was done on the ILSVRC 2012 dataset. The evaluation metrics chosen were Inception Score and Fréchet Inception distance. The model was compared with AC-GAN and SNGAN-projection, in which SAGAN performed best, having Inception score of 52.52 and Fréchet Inception distance of 18.65.

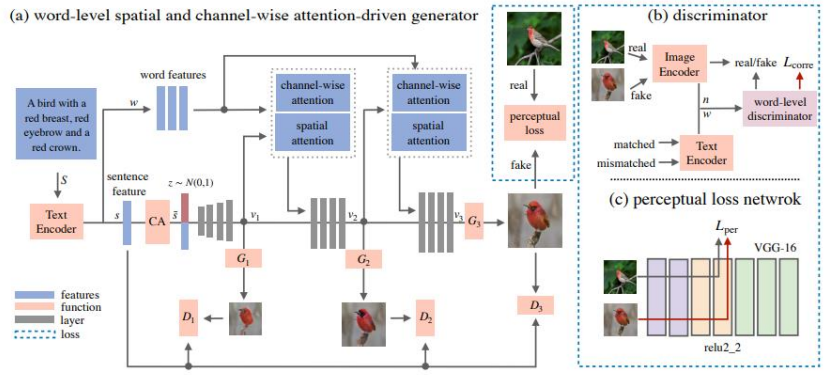


Figure 2: The architecture of our proposed ControlGAN. In (b), $\mathcal{L}_{\text{corre}}$ is the correlation loss discussed in Sec. 3.3. In (c), $\mathcal{L}_{\text{per}}$ is the perceptual loss discussed in Sec. 3.4.

Fig 4.2: Model Architecture

[4]. Self-Attention Generative Adversarial Networks:

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019, May) proposed a StackGAN model. The model is built in two stages: Stage 1 GAN giving Low Resolution images and Stage 2 GAN giving High resolution images. The model first processes the text input and generates corresponding text embeddings to feed into the generative adversarial networks. The model included a text encoder and decoder implemented with a word-level bidirectional recurrent neural network (RNN) consisting of two long short-term memory (LSTM). The generator and discriminator receive a conditioning variable. Dataset used was COCO. The pre-trained StackGAN model has decent performance on generating images from a text input that is similar to its training set, although, when the input contains multiple objects, StackGan fails to generate the correct number of instances with clear boundaries and spatial relationships.

[5]. Text-to-image generation using multi-instance stackgan:

Fu, A., & Hou, Y. (2017) implemented DC-GAN conditioned on text features encoded by a hybrid character-level convolutional recurrent neural network. In the generator, a text query

was encoded. The description embedding was compressed using a fully connected layer followed by LeakyReLU as the activation function. This was then concatenated with the noise. The discriminator consisted of several layers of strides-2 convolution with spatial batch normalization followed by LeakyReLU. The experiment was done on the CUB and Oxford-102 datasets. The GAN baseline was compared with GAN-CLS with image-text matching discriminator, GAN-INT learned with text manifold interpolation and GAN-INT-CLS which combined both.

[6]. Zero-Shot Text-to-Image Generation:

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July) described a transformer trained to autoregressive model the text and image tokens as a single stream of data. A two-stage training procedure was used: Training a discrete Variational Autoencoder to compress an 256 x 256 image into a 32 x 32 grid of image tokens and concatenating up to 256 BPE-encoded text tokens with the 32 × 32 = 1024 image tokens and train an autoregressive transformer to model the joint distribution over the text and image tokens. For a text-image pair, the lowercase caption is BPE-encoded using at most 256 tokens with vocabulary size 16,384. The image is encoded using 32 × 32 = 1024 tokens with vocabulary size 8192. The image tokens are obtained using argmax sampling from the dVAE encoder logits. The text and image tokens are concatenated and modeled autoregressive as a single stream of data. The experiment was carried out of Conceptual Captions, an extension of MS COCO. The model is compared with AttnGAN, DM-GAN and DF-GAN. The evaluation metrics were Inception Score and Fréchet Inception Distance. The zero-shot model obtained an FID score on MS-COCO within 2 points of the best prior approach, despite having never been trained on the captions. But, the model fares significantly worse on the CUB dataset, for which there is a nearly 40-point gap in FID between it and the leading prior approach, i.e., DM-GAN.

[7]. Text to Image using Deep Learning:

Singh, Akanksha & Anekar, Sonam & Shenoy, Ritika & Patil, Sainath. (2022) implemented a GAN-CLS including a generator and discriminator. The inputs were batches of images and the matching text. Both the matching and mismatching text description are encoded, noise is added. Three inputs are passed to the Discriminator: Correct Text with actual image, Incorrect Text with actual image and Correct Text with fake image. These help in the better training of Discriminator. The dataset used was the Oxford-102 flower dataset.

[8]. Stylized Text-to-Image Generation:

Vincent, E., Chandran, D. implemented a StackGAN to generate a stylised output image directly from the model. The Stage 1 GAN generated low resolution images and a new conditioning is added to the Stage 2 GAN to generate higher resolution images in a given style. The discriminator is trained on stylised 256 x 256 images. The datasets used were COCO and CUB. No meaningful results were obtained from this method. The reason given, too much time to reasonably generate enough stylized training data and to perform a hyper-parameter search with enough iterations each time to find the best settings.

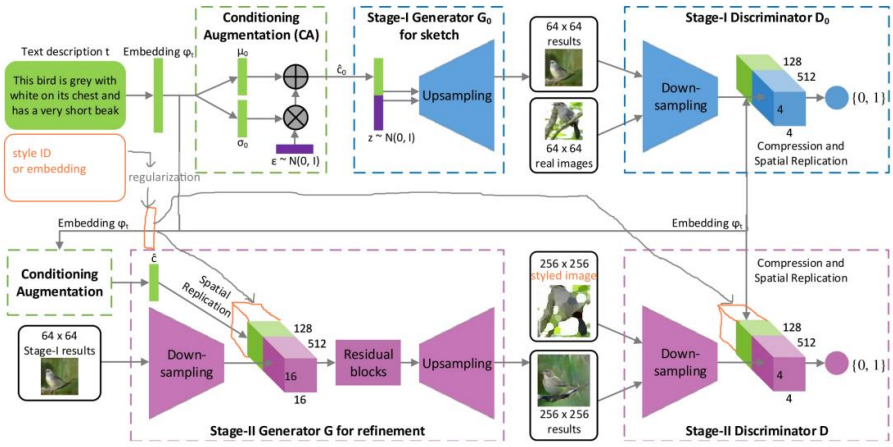


Figure 1: Diagram of the new architecture modified from StackGAN.

Fig 4.3: Model Architecture

[9]. Text-to-Image Generation Grounded by Fine-Grained User Attention:

Koh, J. Y., Baldridge, J., Lee, H., & Yang, Y. (2021) proposed TReCS (Tag - Retrieve - Compose - Synthesize) that uses descriptions to retrieve segmentation masks and predict object labels aligned with mouse traces. These alignments are used to select and position masks to generate a fully covered segmentation canvas; the final image is produced by a segmentation-to-image generator using this canvas. The dataset used Localised Narratives, has detailed natural language descriptions of images paired with mouse traces that provide a sparse, fine-grained visual grounding for phrases. In TReCS, a tagger predicts object labels for every word. A BERT model is trained for this. • A text-to-image dual encoder retrieves images with semantically relevant masks. This helps select contextually appropriate masks for each object. Selected masks are composed corresponding to trace order, with separate canvases for background and foreground objects. Finally, a realistic image is synthesized by inputting the complete segmentation to mask-to-image translation models. The evaluation metrics considered were Inception Score and Fréchet Inception Distance. The model is compared with AttnGAN. Both the models are evaluated on the COCO validation set of Localized Narratives (LN-COCO) and on a held out test set of Open Images data that is covered by Localized Narratives (LN-OpenImages). In the case of LN-COCO, TReCS has a better IS and FID, i.e., 21.3 and 48.7, respectively. On the other hand, for LN-OpenImages, AttnGAN has a better IS and FID, i.e., 15.3 and 56.6, respectively.

[10]. Text-to-Image Generation Using Deep Learning †:

Ramzan, S., Iqbal, M. M., & Kalsum, T. (2022) proposed Recurrent Convolutional Generative Adversarial Network (RC-GAN). Conditional GANs were used with recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for generating meaningful images from a textual description. RNN was used for capturing the contextual information of text sequences by defining the relationship between words at altered time stamps. Text-to-image mapping was performed using an RNN and a CNN. The CNN recognized useful characteristics from the images without the need for human intervention. An input sequence was given to the RNN, which converted the textual descriptions into word

embeddings with a size of 256. These word embeddings were concatenated with a 512-dimensional noise vector. Semantic information from the textual description was passed into the generator. This generated image was used as input in the discriminator along with real/wrong textual descriptions and real sample images from the dataset. The dataset used was Oxford-102 flowers. The evaluation metrics were Inception Score and PSNR. The PSNR value of the model obtained was 30.12 dB. The model was compared with GAN-INT-CLS, StackGAN, StackGAN++, HDGAN, and DualAttn-GAN for the Inception Score. RC-GAN gave the best score, i.e., 4.15 ± 0.03.

[11]. Hierarchical Text-Conditional Image Generation with CLIP Latents:

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022) proposed a two-stage model: a prior that generates a CLIP image embedding given a text caption, and a decoder that generates an image conditioned on the image embedding. The decoder used diffusion models to produce images conditioned on CLIP image embeddings. To generate high-resolution images, two diffusions up sampler models were trained: one to up sample images from 64×64 to 256×256 resolution, and another to further up sample those to 1024×1024 resolution. For the first up-sampling stage, gaussian blur was implemented, and for the second, a more diverse BSR degradation was to make the decoder more robust. For the prior, two model classes were explored: Autoregressive and Diffusion. In the Autoregressive (AR)Model, the dimensionality of the CLIP image embeddings is reduced by applying Principal Component Analysis. After applying PCA, the principal components were ordered by decreasing eigenvalue magnitude, quantizing the dimensions into discrete buckets, and predicting the resulting sequence using a Transformer model with a causal attention mask. The AR prior is conditioned on the text caption and the CLIP text embedding by encoding them as a prefix to the sequence. For the diffusion prior, a decoder-only Transformer was trained with a causal attention mask on a sequence consisting of, in order: the encoded text, the CLIP text embedding, an embedding for the diffusion timestep, the noised CLIP image embedding, and a final embedding whose output from the Transformer is used to predict the un-noised CLIP image embedding. No conditioning was done. The model is not directly trained on the MS-COCO training set, but can still generalize to the validation set zero-shot. With the diffusion prior, the zero-shot FID score obtained is 10.39.

[12]. Text to Image Generation with Semantic-Spatial Aware GAN:

Liao, W., Hu, K., Yang, M. Y., & Rosenhahn, B. (2022) introduced a framework called Semantic-Spatial Aware GAN for synthesizing images from input text which consists a simple and effective Semantic-Spatial Aware block, which learns semantic-adaptive transformation conditioned on text to effectively fuse text features and image features and learns a semantic mask in a weakly-supervised way that depends on the current text-image fusion process to guide the transformation spatially. The model had a text encoder that learns text representations, a generator that had 7 SSA blocks for deepening text-image fusion and improving resolution, and a discriminator that is used to judge whether the generated image is semantically consistent to the given text. The text encoder was a bidirectional LSTM and pre-trained using real image-text pairs by minimizing the Deep Attentional Multimodal

Similarity Model (DAMSM) loss. Each SSA block consisted of an up-sample block, a semantic mask predictor, and a Semantic-Spatial Condition Batch Normalization block with a residual connection. The up-sample block was used to double the resolution of image feature maps by bilinear interpolation operation. The residual connection was used to maintain the main contents of the image features to prevent text-irrelevant parts from being changed and the image information being overwhelmed by the text information. The discriminator concatenated the features extracted from the generated image and the text vector for computing the adversarial loss through two convolution layers. Associated with the Matching-Aware zero-centered Gradient Penalty (MA-GP), it guided the generator to synthesize more realistic images with better text-image semantic consistency. The datasets used were COCO and CUB and the evaluation metrics were Inception Score (IS), Fréchet Inception Distance (FID), and R-precision. The model was compared with StackGAN++, AttnGAN, ControlGAN, SD-GAN, DM-GAN, DF-GAN and DAE-GAN. For the CUB dataset, the model had the best IS, i.e., 5.17±0.08 and for the COCO dataset, the model had the least FID, i.e., 19.37. DAE-GAN had best R-precision in CUB and COCO, 85.4±0.57 and 92.6±0.50, respectively.

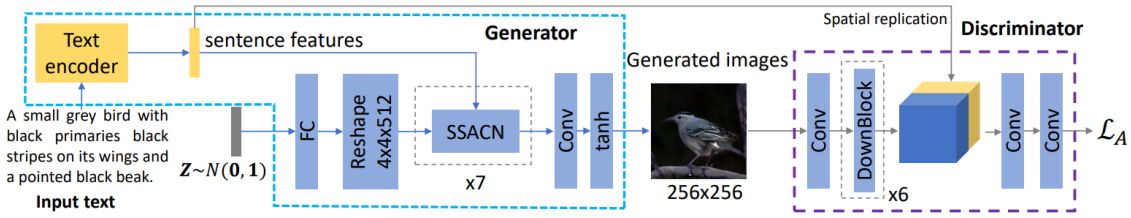


Figure 2: A schematic of our framework SSA-GAN. It has one generator-discriminator pair. The generator mainly consists of 7 proposed SSA blocks which fuse text and image features through the image generation process and guarantee the semantic text-image consistency. The gray lines indicate the data streams only for training.

Fig 4.4: Model Architecture

[13]. MirrorGAN: Learning Text-to-image Generation by Redescription:

Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019) introduced MirrorGAN. It consisted of three modules: a semantic text embedding module (STEM), a global-local collaborative attentive module for cascaded image generation (GLAM), and a semantic text regeneration and alignment module (STREAM). In STEM, a RNN is used to extract semantic embeddings from the given text description, which include a word embedding and a sentence embedding. Conditioning Augmentation was done. In GLAM, three image generation networks were stacked sequentially. It took a word embedding and a visual feature as the input. The word embedding was first converted into an underlying common semantic space of visual features by a perception layer. It was multiplied with the visual feature to obtain the attention score. For sentence-level model, the same actions were performed, just like with the word-level model. In STREAM, the text description is regenerated from the generated image, which was semantically aligned with the given text description. The image encoder was a CNN pre-trained on ImageNet and the decoder was a RNN. The datasets used were COCO and CUB and the evaluation metrics were Inception Score (IS) and R-precision. The model was compared with AttnGAN. The IS was best in MirrorGAN for both CUB and COCO, i.e., 4.56 ± 0.05 and 26.47 ± 0.41, respectively.

[14]. ChatPainter: Improving Text to Image Generation using Dialogue:

Sharma, S., Suhubdy, D., Michalski, V., Kahou, S. E., & Bengio, Y. (2018) proposed a way to improve the quality of the images using dialogue. VisDial dialogues were used along with the MS COCO dataset. The model is built upon the StackGAN model. It generates an image in two stages where Stage-I generates a course 64 × 64 image and StageII generates a refined 256 × 256 image. Caption embedding was done using a pre-trained encoder. The dialogue embeddings were done by two methods: Non-recurrent encoder and Recurrent encoder. In a Non-recurrent encoder, the entire dialogue is collapsed into a single string and encoded with a pre-trained Skip-Thought encoder. In the Recurrent encoder, Skip-Thought vectors are generated for each turn of the dialogue and encoded with a bidirectional LSTM-RNN. Conditioning Augmentation is done on the concatenation of the caption and dialogue embeddings, which is passed as the input. In Stage 1, the conditioned variables are concatenated with noise. The generator up-samples this input. In the discriminator, the conditioned variable is concatenated with the down-sampled image. This was further downsampled to a scalar value between 0 and 1. In Stage 2, the generator, the conditioned variable is concatenated with the downsampled generated image from Stage 1. For Stage-II training, in the case of the recurrent dialogue encoder, the RNN weights are copied from Stage-I and kept fixed. The concatenated input is passed through a series of residual blocks and is then up-sampled. In the Stage-II discriminator, the conditioned variable was concatenated with the down-sampled image, which was further downsampled to a scalar value between 0 and 1. The evaluation metric was the Inception Score (IS). The model was compared with the original StackGAN model. The IS for the recurrent model was 9.74 ± 0.02 and for the non-recurrent model was 9.43 ± 0.04, both the scores higher than the IS of the StackGAN, i.e, 8.45 ± 0.03.

[15]. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding:

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022) introduced Imagen. It consisted of a text encoder that maps text to a sequence of embeddings and a cascade of conditional diffusion models that map these embeddings to images of increasing resolutions. The model explored three pre-trained text encoders: BERT, T5 and CLIP. The weights were frozen of these text encoders. The model utilized a pipeline of a base 64×64 model and two text-conditional super-resolution diffusion models to upsample a 64 × 64 generated image into a 256 × 256 image and then to 1024 × 1024 image. Noise conditioning augmentation is used for both the super-resolution models. For the base model, the U-Net architecture is adapted. The network is conditioned on text embeddings via a pooled embedding vector, added to the diffusion timestep embedding. Further conditioning on the entire sequence of text embeddings was done by adding cross attention over the text embeddings at multiple resolutions. For the Super-resolution model, modifications were made on the U-net model like removing the self-attention layers. Cross attention is used in the super-resolution models. The input for this was the 64x64 low-resolution images and output was the upsampled 1024 × 1024 images as outputs. The dataset used was COCO and the evaluation metric was Zero-shot FID-30K, in which the model had the distance of 7.27.

4.1 Future Trends and Potential Uses:

As the field of text-to-image generation continues to evolve, several future trends and potential applications emerge, paving the way for exciting advancements. The exploration of these trends can guide researchers and practitioners in shaping the trajectory of text-to-image research.

Future Trends:

1. **Semantic Understanding and Rich Context:**

   Advances in models like GAN-INT-CLS, StackGAN++, HDGAN, and DualAttn-GAN indicate a move toward deeper semantic understanding in text-to-image synthesis, enriching generated images with nuanced context [2].
2. **Multimodal Fusion for Enhanced Creativity:**

   Future trends involve multimodal fusion, combining diverse information types to unlock new dimensions of creativity. Models like MirrorGAN explore novel architectures for imaginative image synthesis.
3. **Fine-Grained Control and User Interaction:**

   The future emphasizes giving users precise control over the generation process. Models like ControlGAN contribute to the theoretical foundation of user-centric text-to-image synthesis.
4. **Ethical and Responsible AI:**

   Theoretical discussions highlight the importance of ethical considerations, transparency, and accountability in text-to-image models, ensuring fair and unbiased image generation.
5. **Transfer Learning Across Modalities:**

   Transfer learning advancements aim to enhance adaptability across modalities in text-to-image models. Models like Vision Transformers (ViT) showcase unified frameworks for handling both NLP and computer vision.

Potential Uses:

Text-to-image models revolutionize advertising and content creation by transforming textual descriptions into visually appealing content, E-Commerce and Product Design, Education and Training, Virtual and Augmented Reality, Healthcare and Medical Imaging and Entertainment and Gaming like MirrorGANs [13].

5. Methodology:

5.1 Challenges and Future Directions:

Challenges:

Text-to-image generation faces challenges in achieving a perfect alignment between textual input and generated visual output. Ensuring that the generated images faithfully represent the nuanced details described in the text remains a formidable challenge. Furthermore, striking the right balance between model complexity and interpretability is crucial, as overly complex models may risk becoming inscrutable.

Future Directions:

To overcome these challenges, future research should focus on refining attention mechanisms and semantic understanding within models. Integrating explainability into advanced architectures, such as GAN-INT-CLS and DualAttn-GAN, can pave the way for more interpretable text-to-image synthesis. Additionally, exploring novel techniques for capturing abstract concepts and emotions from text can push the boundaries of generative models in representing diverse textual prompts.

5.2 Current Limitations:

Limitations:

The current landscape of text-to-image generation grapples with limitations such as the occasional generation of artifacts and the struggle to capture fine-grained details. Models like ControlGAN and MirrorGAN, while advancing the field, still face challenges in producing consistently realistic and high-fidelity images. Moreover, there is a tendency for certain models to overemphasize prominent features, leading to imbalances in the generated visual content.

Addressing Limitations:

Overcoming these limitations involves refining the training methodologies and loss functions used in models. Leveraging insights from human perceptual studies can guide the development of more robust evaluation metrics. Integrating advanced filtering techniques within the training process, as demonstrated by HDGAN, can contribute to reducing artifact generation and enhancing the overall quality of the synthesized images.

5.3 Handling Complex Scenes and Substantial Missing Data:

Challenges:

Textual descriptions often involve complex scenes and intricate details, posing a challenge for models to accurately translate them into coherent images. Additionally, dealing with substantial missing data in the textual input, where some details are left unspecified, remains a challenge.

Models like StackGAN++ and DualAttn-GAN, while adept, may struggle in the absence of complete information.

Future Strategies:

Future research directions should explore methods to handle incomplete textual descriptions and enhance the ability of models to fill in missing details intelligently. The integration of reinforcement learning approaches, as seen in StackGAN++, can guide models in making informed decisions about scene completion. Advancements in leveraging external knowledge bases to supplement missing information can contribute to more comprehensive text-to-image synthesis.

5.4 Dependency on Extensive and Diverse Training Data:

Challenge:

Many text-to-image models heavily rely on extensive and diverse training datasets to generalize well across various inputs. However, this dependency poses challenges in scenarios where obtaining such datasets may be impractical or where the data distribution is highly skewed. Models like SAGAN and AC-GAN, while powerful, may struggle to generalize effectively in data-scarce domains.

Addressing Dependency:

Future directions involve investigating techniques for mitigating the dependency on vast datasets. This includes exploring methods for transfer learning, domain adaptation, and data augmentation to enhance the adaptability of models like ViT. Additionally, incorporating techniques for unsupervised learning can contribute to better generalization in the absence of extensive training data.

5.5 Model Generalizability and Overfitting:

Challenge:

Ensuring the generalizability of text-to-image models beyond the training distribution is a persistent challenge. Overfitting to specific textual patterns can result in models producing unrealistic outputs when faced with novel inputs. Models like DALL-E, while innovative, may encounter challenges in maintaining coherence and relevance when confronted with diverse and unconventional textual prompts.

Strategies to Mitigate Overfitting:

Future research directions should focus on strategies to enhance model generalizability. Incorporating regularization techniques, diverse training scenarios, and adversarial training against overfitting can contribute to models that perform consistently across a broader spectrum of textual descriptions. Moreover, exploring methods for controllable diversity in generated outputs, as in DALL-E, can mitigate overfitting by promoting varied synthesis.

6. Conclusion and Future Research:

6.1 Summary of Findings:

In the dynamic landscape of text-to-image generation, our exploration of various models has uncovered a trajectory marked by significant milestones. Early rule-based systems and generative models set the stage, but it was the advent of GANs in 2014 that truly revolutionized the field. Subsequent models like Conditional GANs, StackGAN, and the more recent GAN-INT-CLS, StackGAN++, HDGAN, and DualAttn-GAN used in [1], [2], [4], [6] and [10] have built upon these foundations, showcasing advancements in image resolution, attention mechanisms, and language understanding integration. MirrorGAN stands out for its role in pushing the boundaries of creativity and novel architectural exploration.

6.2 Theoretical Insights:

Looking ahead, the future of text-to-image generation holds exciting theoretical prospects. Models such as GAN-INT-CLS and StackGAN++ suggest a trajectory toward semantic richness, emphasizing a deeper understanding of the textual context for image synthesis. The exploration of multimodal fusion, as seen in MirrorGAN, opens new avenues for enhanced creativity by combining diverse information types. The pursuit of fine-grained control, exemplified by ControlGAN, foresees a future where users have precise influence over the generated content. Ethical considerations, transparency, and accountability are gaining prominence in the theoretical landscape, ensuring responsible AI practices. Moreover, the theoretical shift towards transfer learning across modalities, demonstrated by Vision Transformers (ViT), signifies increased adaptability in text-to-image models.

6.3 Final Thoughts:

In conclusion, the journey through text-to-image generation reveals not only the historical evolution of models but also promising directions for the future. Theoretical advancements underscore the importance of not only achieving visual realism but also understanding and incorporating richer contextual meanings from textual inputs. As models evolve to grant users more control and incorporate ethical considerations, the applications in advertising, e-commerce, education, virtual reality, healthcare, and entertainment promise transformative impact. The collaborative synergy between theoretical insights and model development is poised to redefine how we translate language into compelling visual narratives, opening new frontiers in the intersection of artificial intelligence and creative expression.

6.4Future Research Areas:

Exploration of Multimodal Fusion:

Future research should delve into advanced multimodal fusion techniques that seamlessly combine textual, visual, and potentially other modalities. Models like MirrorGAN provide a

starting point for exploring how different information types can be effectively integrated to enhance creativity and context in text-to-image synthesis.

Enhanced User Interaction:

Theoretical research areas should include the development of models that facilitate enhanced user interaction, allowing users precise control over the image generation process. Exploring techniques similar to those used in ControlGAN can contribute to theoretical frameworks that empower users in the creative process.

6.5The Road Ahead:

Integration of Ethical Considerations:

As text-to-image generation becomes more pervasive, the road ahead involves integrating ethical considerations, transparency, and accountability into model development. Research in line with responsible AI practices, as hinted by models like CLIP, can guide the development of models that prioritize fairness and unbiased image generation.

Advancements in Transfer Learning:

The future trajectory includes continued advancements in transfer learning across modalities. Unified frameworks, as demonstrated by ViT, showcase the potential for models to seamlessly adapt knowledge gained from one domain to enhance performance in another, laying the groundwork for more versatile text-to-image synthesis models.

7. References:

[1]. http://proceedings.mlr.press/v48/reed16.pdf

[2]. https://proceedings.neurips.cc/paper/2019/file/d18f655c3fce66ca401d5f38b48c89af-Paper.pdf

[3]. https://proceedings.neurips.cc/paper_files/paper/2019/file/1d72310edc006dadf2190caad5802983-Paper.pdf

[4]. http://proceedings.mlr.press/v97/zhang19d/zhang19d.pdf

[5]. http://cs231n.stanford.edu/reports/2017/pdfs/324.pdf

[6]. https://proceedings.mlr.press/v139/ramesh21a/ramesh21a.pdf

[7]. https://www.researchgate.net/publication/359441889_Text_to_Image_using_Deep_Learning

[8]. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/custom/15722067.pdf

[9]. https://openaccess.thecvf.com/content/WACV2021/papers/Koh_Text-to-Image_Generation_Grounded_by_Fine-Grained_User_Attention_WACV_2021_paper.pdf

[10]. https://www.mdpi.com/2673-4591/20/1/16

[11]. https://3dvar.com/Ramesh2022Hierarchical.pdf

[12]. https://openaccess.thecvf.com/content/CVPR2022/papers/Liao_Text_to_Image_Generation_With_Semantic-Spatial_Aware_GAN_CVPR_2022_paper.pdf

[13]. https://openaccess.thecvf.com/content_CVPR_2019/papers/Qiao_MirrorGAN_Learning_Text-To-Image_Generation_by_Redescription_CVPR_2019_paper.pdf

[14]. https://arxiv.org/pdf/1802.08216.pdf

[15]. https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf