Indian Journal of Image Processing and Recognition (IJIPR)
ISSN: 2582-8037 (Online), Volume-4 Issue-3, April 2024

An Overview of Text to Visual Generation Using GAN

Sibi Mathew

Abstract- Text-to-visual generation was once a cumbersome task until the advent of deep learning networks. With the introduction of deep learning, both images and videos can now be generated from textual descriptions. Deep learning networks have revolutionized various fields, including computer vision and natural language processing, with the emergence of Generative Adversarial Networks (GANs). GANs have played a significant role in advancing these domains.A GAN typically comprises multiple deep networks combined with other machine learning techniques. In the context of text-to-visual generation, GANs have enabled the synthesis of images and videos based on textual input. This work aims to explore different variations of GANs for image and video synthesis and propose a general architecture for text-to-visual generation using GANs. Additionally, this study delves into the challenges associated with this task and discusses ongoing research and future prospects.By leveraging the power of deep learning networks and GANs, the process of generating visual content from text has become more accessible and efficient. This work will contribute to the understanding and advancement of text-to-visual generation, paving the way for numerous applications across various industries.

Index Terms—Computer Vision, Image Synthesis, Natural Lan- Guage Processing, Video Synthesis, Filler Images.

I. INTRODUCTION

The concept of text-to-visual generation is founded on the idea that descriptive text can effectively convey information about an image or video frames. This breakthrough has ush- ered in a new era of prompt engineering in the field of visual generation. In the realm of GAN-based visual generation, the primary focus lies in extracting descriptive factors from textual input, which serve as a basis for generating appropriate environmental factors and refining the image to produce the desired output.

Previously, creating an image or designing a video required painstaking manual adjustment of each factor. Designers had to meticulously set up each element of an image, and for video generation, they had to meticulously arrange every frame. However, thanks to technological advancements and deep networks, the process of generating images has become significantly easier. Deep networks, a gift from the progress of artificial intelligence, allow for the generation of images that serve as base frames for video generation.

Manuscript received on 30 March 2024 | Revised Manuscript received on 12 April 2024 | Manuscript Accepted on 15 April 2024 | Manuscript published on 30 April 2024.

**Correspondence Author(s)*
Sibi Mathew*, MTech Scholar, Department of CSE TKM College of Engineering Kollam, Kerala, India, E-mail: 22082@tkmce.ac.in, sibimathew2019@gmail.com, ORCID ID: 0009-0001-7227-8082

This technology has brought about a shift from supervised image generation methods to AI-generated images that leverage auto-encoders and deep networks [2][3][22][23].

A significant breakthrough in text-to-visual generation occurred with the introduction of Generative Adversarial Networks (GANs) in 2014 [1]. GANs utilize two networks, namely a generator and a discriminator. The discriminator's role is to distinguish between real and fake images by assessing how closely an image aligns with the features described in the accompanying text. On the other hand, the generator generates fake images that are refined iteratively until the discriminator perceives them as real images. This basic working principle illustrates the foundation of GAN-based image generation. However, various researchers have proposed improved solu- tions that enhance performance in this field [21].

This review aims to highlight notable advancements in text- to-visual synthesis using GANs. It delves into the methodolo- gies employed in each work and discusses relevant findings that can contribute to further research. The review is structured as follows: the first section provides a brief overview of GANs, emphasizing their significance and functioning. It is then followed by an examination of the research works considered within the scope of this review. Subsequently, the review explores the applications of GANs in text-to-visual synthesis and presents evaluation techniques used for assessing their performance.

By organizing and analyzing the existing research, this review seeks to contribute to the understanding and progress of text-to-visual synthesis using GANs.

II. GENERATIVE ADVERSARIAL NETWORKS

A Generative Adversarial Network (GAN) is a type of machine learning model composed of two neural networks: a generator and a discriminator. GANs were initially proposed by Ian Goodfellow and his colleagues in 2014 as a means of generating realistic and high-quality synthetic data [1]. The fundamental objective of a GAN is to understand and learn the underlying data distribution of a provided training dataset. It accomplishes this by generating new samples that closely resemble the training data. This learning process is facilitated through a competitive framework, where the gen- erator network and discriminator network engage in a game against each other [1]. The generator network is responsible for producing syn- thetic data samples. Initially, its outputs may be random and unconvincing. However, as training progresses, it learns to generate samples that progressively improve in quality and similarity to the real data.

On the other hand, the discriminator network's role is to differentiate between real data samples from the training dataset and fake samples generated by the

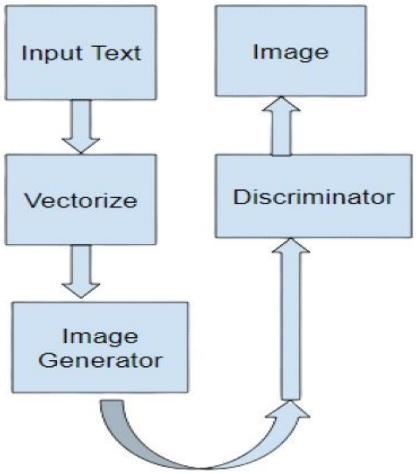

Fig. 1. GAN Based Image Generation General Diagram

generator. The discriminator is trained to distinguish the real from the fake with increasing accuracy over time. As the training progresses, both the generator and dis- criminator networks engage in an adversarial process. The generator aims to generate samples that the discriminator cannot differentiate from real data, while the discriminator strives to accurately identify the real data from the genera- tor's fake samples. Through this adversarial interplay, both networks improve their performance until the generator can produce synthetic data that closely matches the distribution of the real data. By leveraging the competitive nature of the generator and discriminator networks, GANs have become a powerful framework for generating synthetic data that captures the characteristics of the training data. They have found applications in various domains, including computer vision, natural language processing, and more.

A. Components in GAN

- Generator: The generator network takes random noise or a latent vector as input and transforms it into a synthetic sample. It typically consists of multiple layers, including fully connected layers, convolutional layers, and activation functions. The generator tries to produce samples that are realistic and resemble the training data.
- Discriminator: The discriminator network acts as a binary classifier that aims to distinguish between real and gen- erated samples. It takes a sample, either real or synthetic, as input and predicts the probability of it being a real sample. The discriminator also consists of multiple layers, such as fully connected or convolutional layers, and activation functions.
- Training Process: The GAN training process involves a game-like interaction between the generator and discrim- inator. Initially, both networks are randomly initialized. The training process can be divided into the following steps:

1) Sample Generation: The generator generates a batch of synthetic samples by taking random noise as input and producing synthetic outputs.
2) Real Sample Selection: A batch of real samples is randomly selected from the training dataset.
3) Discriminator Training: The discriminator is trained on both the real samples and the synthetic samples generated by the generator. The discriminator aims to correctly classify the real samples as real (label = 1) and the generated samples as fake (label = 0). The parameters of the discriminator are updated based on the loss calculated from the classification results.
4) Generator Training: The generator is trained to fool the discriminator by producing synthetic samples that are classified as real. The generator aims to generate samples that maximize the probability of the discriminator labeling them as real (label = 1). The parameters of the generator are updated based on the loss calculated from the discriminator's feedback.
5) Iterative Process: Steps 3 and 4 are repeated multi- ple times to iteratively improve both the generator and discriminator. This adversarial process contin- ues until the generator produces synthetic samples that are difficult for the discriminator to distinguish from real samples.

B. Loss Functions

- Generator Loss: The generator loss represents how well the generator is able to deceive the discriminator. It is computed based on the discriminator's output when the generator's samples are fed as input. The generator aims to minimize this loss, pushing it toward producing more realistic samples.
- Discriminator Loss: The discriminator loss measures the ability of the discriminator to correctly classify real and synthetic samples. It is computed based on the discriminator's predictions for both the real and generated samples. The discriminator aims to minimize this loss, improving its ability to distinguish between real and fake samples.

III. TEXT TO IMAGE SYNTHESIS

Text-to-image generation using GANs Figure**??** is an exciting application that aims to generate realistic and coherent images based on textual descriptions. The process involves training a GAN model with a generator and discriminator network, which learn to translate text inputs into corresponding visual representations. These representations can be used to generate images that match the text features [4], [5].

A. General Steps of Text to image synthesis using GAN

- Dataset Preparation: A large dataset is required for train- ing the GAN model. This dataset consists of pairs of textual descriptions and corresponding real images. For example, the dataset could include captions describing images or paired text-image datasets.
- Generator Network: The generator network takes textual descriptions as input and generates synthetic images as output. It typically consists of recurrent neural networks (RNNs), such as long short-term memory (LSTM) or transformers, which are capable of processing sequential input data. The generator aims to produce images that match the given textual description.

Retrieval Number: 100.1/ijipr.A804113010524
DOI: 10.54105/ijipr.A8041.04030424
Journal Website: www.ijipr.latticescipub.com

Published By:
Lattice Science Publication (LSP)

- Discriminator Network: The discriminator network acts as a binary classifier, distinguishing between real and generated images. It takes an image, either real or synthetic, as input and predicts the probability of it being a real image. The discriminator helps the generator to improve by providing feedback on the realism of the generated images.
- Text-to-Image Mapping: To facilitate the mapping between text and image, additional techniques can be used. One common approach is to employ an embedding network, which converts the textual descriptions into a fixed-length feature vector representation. This embedding vector is then fed into the generator network to guide the image generation process based on the given text.
- Adversarial Training: The training process consists of an adversarial game between the generator and discriminator. The steps involved are as follows:

1) Real Image Discrimination: The discriminator is trained using pairs of real images and their cor- responding textual descriptions. The discriminator aims to correctly classify real images as real (label = 1) and generated images as fake (label = 0). The parameters of the discriminator are updated based on the loss calculated from the classification results.
2) Text-to-Image Generation: The generator takes tex- tual descriptions and generates synthetic images. The generated images are then fed into the discrim- inator, which provides feedback on their realism. The parameters of the generator are updated based on the loss calculated from the discriminator's feed- back, aiming to generate images that can fool the discriminator into classifying them as real.
3) Iterative Training: Steps a and b are repeated itera- tively to improve both the generator and discrimina- tor networks. The adversarial process continues until the generator produces high-quality images that are difficult for the discriminator to distinguish from real images.

- Evaluation and Fine-tuning: After training the GAN model, evaluation techniques are employed to assess the quality of the generated images. Objective metrics, such as Inception Score or Frechet Inception Distance, can be used to measure the realism and diversity of the generated images. Fine-tuning and optimization steps can be applied to further enhance the quality of the generated images.

Text-to-image generation using GANs has applications in various fields, including computer vision, multimedia generation, and creative design. It enables the synthesis of visual content based on textual input, expanding the possibilities for generating realistic and contextually relevant images.

B. Text to Video Generation

This review couldn't define a general structure for a text-to-video generation model as most works follow different strategies, but in general, videos are generated as frames of images. They form a visual illusion in which the human eye cannot distinguish the frames of images. Here in the case of GAN images generated by GAN become base images to the previous frame in the case of frames of a video.

IV. RELATED WORKS

The first GAN model proposed for text-to-image generation is called Generative Adversarial Text-to-Image Synthesis (GAN-INT-CLS) [17]. Introduced by Reed et al. in 2016, this model combines a deep convolutional GAN (DCGAN) with an auxiliary text classifier. The main idea behind GAN-INT-CLS is to generate realistic images based on textual descriptions. The generator network takes a random noise vector as input and is conditioned on a text description. It generates images that aim to align with the given text. The discriminator network is responsible for distinguishing between real images and the images generated by the generator. It is trained to improve its ability to classify whether an image is real or fake

In addition to the generator and discriminator, GAN-INT-CLS incorporates an auxiliary text classifier. This classifier is trained to predict the class label of the given text description. The presence of this classifier helps guide the generator during the training process. By conditioning the generator on both the text description and the class label, it is encouraged to generate images that not only resemble the text but also correspond to the specific class. Through an adversarial training process, where the generator tries to fool the discriminator while the discriminator learns to distinguish real from generated images, GAN-INT-CLS learns to generate images that align with textual descriptions.

In another work [7] proposed machine Stacked Generative Adversarial Network (StackGAN) is used to create 256x256 photos primarily based totally on textual content descriptions. The textual content-to-photographic technology is split into two ways in terms of usage of StackGAN: Stage-I GAN and Stage-II GAN. Stage-I GAN creates a low-decision pho- tograph and Stage-II GAN corrects defects within the low- decision photograph created via way of means of Stage-I and creates a high-decision photograph. But this method requires improvement in the way that StackGAN calls for a big quantity of GPU.

StackGAN++ was proposed [13] as an enhanced version of StackGAN, introducing a tree-like structure with multiple gen- erators and discriminators. Unlike StackGAN's paired setup, StackGAN++ utilizes a hierarchical approach for both condi- tional and unconditional generative tasks.

The framework of StackGAN++ involves multiple stages, where each stage generates progressively higher-resolution images. In the initial stage, a noise vector and a condition vector are fed into the generator to produce lower-resolution images, following a similar approach as StackGAN. However, in subsequent stages, the previous output and the conditional variable are used to generate higher-resolution images. The need for a noise vector is eliminated in these stages since randomness is already preserved in the output of the first stage. The multi-stage training in StackGAN++ aims to ap- proximate multiple distributions, including multi-scale image distributions, and joint conditional and unconditional image distributions. This approach enhances the quality of image generation and improves training stability. Additionally, Stack- GAN++ introduces a color-consistency regularization term to ensure coherence among samples generated by different generators at various scales.

Both quantitative and qualitative evaluations demonstrate the significant improvements achieved by StackGAN++ in both conditional and unconditional image generation tasks. The framework produces high-resolution (256x256) images while maintaining color consistency and generating coherent samples. StackGAN++ has demonstrated its efficacy in advancing the quality and diversity of generated images compared to its predecessor, StackGAN.

In 2018 another improved version was introduced that uses attention for analysing the text further and enhanced model was also introduced [14], [15]AttnGAN and E-ATTN GAN are both variants of Generative Adversarial Networks (GANs) that incorporate attention mechanisms to improve the quality and details of generated images. AttnGAN, proposed by Xu et al. in 2018, focuses on generating images conditioned on text descriptions. It employs an attention mechanism to enhance the generation process based on the textual information provided. The attention mechanism in AttnGAN enables the generator to focus on specific regions of the image that correspond to the given textual description. By attending to relevant image regions, AttnGAN generates more contextually consistent and visually plausible images.E-ATTN GAN (Enhanced Attention Generative Adversarial Network) is an improved version of GAN that incorporates an enhanced attention mechanism. It was proposed to capture fine-grained details in generated images. E-ATTN GAN uses attention not only to focus on relevant regions but also to allocate resources for capturing intricate details. It consists of multiple generators and discriminators arranged in a tree-like structure. Additionally, E-ATTN GAN jointly approximates multiple distributions, including multi-scale and conditional/unconditional distributions, to further enhance image generation quality.both AttnGAN and E-ATTN GAN employ attention mechanisms to improve image generation. AttnGAN specifically focuses on generating images conditioned on text descriptions, using attention to align the generated image with the given text. On the other hand, E-ATTN GAN extends the attention mechanism to enhance the generation of fine-grained details and employs a multi-generator, multi-discriminator framework to capture multi-scale and conditional/unconditional distributions.

MirrorGAN is a text-to-image synthesis model 2019 [16][25]. The key idea behind MirrorGAN is to leverage a multi- level mirror mechanism to generate high-quality images from text descriptions. The model consists of a mirror generator and a mirror discriminator. The generator aims to synthesize images that match the given textual description, while the discriminator's role is to distinguish between real and generated images. The mirror mechanism is designed to capture both global and local details in the generated images. The mirror mechanism is a key innovation in MirrorGAN. It enables the generator to capture both global and local details by leveraging the generated images at higher levels as references for the lower-level generation process. This allows the generator to refine and align the generated images at different scales, resulting in more visually coherent and contextually relevant images. MirrorGAN follows the standard GAN training procedure, where the generator and discriminator play an adversarial game. The generator aims to produce images that can fool the discriminator, while the discriminator learns to distinguish between real and generated images. Through iterative training, the generator improves its ability to generate realistic images that align with the given textual descriptions.

DM-GAN (Dynamic Memory Generative Adversarial Network) is a text-to-image synthesis model proposed by Tran et al. [18]. It introduces a dual-memory mechanism, consisting of global and local memory modules, to capture long-range and short-range dependencies in the textual descriptions. The generator network utilizes these memory modules to generate coherent and semantically accurate images that align with the given text. The discriminator network provides feedback to the generator during adversarial training, enabling the model to improve the quality of the generated images. DM-GAN has demonstrated promising results in generating high-quality images that correspond to textual descriptions by effectively leveraging the dual- memory mechanism. The paper title "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis" provides detailed insights into the model and its experimental results.

VQGAN+CLIP [20] is a powerful combination of two deep learning models, VQGAN and CLIP, used for image generation and manipulation based on textual prompts. VQGAN (Vector-Quantized Generative Adversarial Network) is a generative model that generates images from random noise vectors. It employs a codebook to quantize the continuous image space into discrete codes, enabling better control over the generated images. CLIP (Contrastive Language-Image

Pretraining) is a model that learns the relationship between images and their textual descriptions. It can understand and rank the similarity between images and text. By combining VQGAN and CLIP, the VQGAN+CLIP model enables image generation conditioned on textual prompts. The process involves optimizing the latent code of the VQGAN generator using gradient ascent to maximize the similarity between the generated image and the desired textual description, as measured by the CLIP model. This allows users to input specific textual prompts or descriptions and generate corresponding images that align with those prompts. According to another research [6], T2V(Text-To-Vision) is a innovative method that creates TV-program-like Computer Graphics animation is generated automatically from a script and can be considered one of the first of its kind. T2V Player is created using this technology. The software allows users to create animated videos by typing text. The T2V Player employs a framework that converts text to animation. Using this technology, we can map text to animated images. T2V player provides a limited number of characters and commands for creating an animated video. The adaption of artificial intelligence and modern technology can let the application expand its functionalities.

A more advanced [8] model uses Cyclic GAN, for creating pictures of size 128 × 128 that go along with the substance of the information content.

This model is created using the Oxford-102 blossoms dataset, which has, for each picture, a class name and in any event five content depictions. For actualizing the TAC-GAN system utilizes the Tensor stream execution of a Deep Convolutional Generative Adversarial Network (DCGAN), in which G is displayed as a Deconvolutional Neural Network, and D is demonstrated as DCGAN-tensor stream a Convolution Neural Network (CNN). This approach uses a version of Cyclic GAN, which is named to synthesize text images. The main disadvantage of this system is that the dataset contains only 102 classification datasets, consisting of 102 blossom classes.

MoCoGAN [9] framework creates a video clip by utilizing an image generator to sequentially generate images. Each random vector comprises a content component and a motion component. The content component remains constant, while the motion component is treated as a stochastic process. To achieve unsupervised learning of motion and content decomposition, we introduce a novel adversarial learning scheme that incorporates both image and video discriminators. Through extensive experiments on challenging datasets, this work compares approaches with state-of-the-art methods, analyzing qualitative and quantitative results to validate the effectiveness of this framework. To improve the quality of created videos, the framework can make use of developments in picture production in the GAN framework. MoCoGAN's key drawback is that it has a higher FID, which indicates that the generated images will be less similar to the training data. Temporal Generative Adversarial Net (TGAN) [10][24] can create an entirely new video by learning representation from an unlabeled video data set. The generator in this model is formed by two sub-networks: a temporal generator and an image generator. The temporal generator, in particular, initially produces a series of latent variables, each of which correlates to a latent variable for the image generator. The image generator then converts these latent variables into a video with the same frame count as the variables. MoCoGAN and TGAN have higher Frechet Inception Distance (FID), which indicates that the generated images will be less similar to the training data. However, notwithstanding these examples, the quantity of research on textual-to-video stays small.MoCoGAN creates a video clip by generating video frames in a specified order. An image-generative network maps a random vector into an image at each time step. The random vector is composed of two components, the first of which is drawn from a content subspace and the second from a motion subspace. We model the content space using a Gaussian distribution and use the same realization to generate each frame in a video clip because the content in a short video clip is usually the same. A recurrent neural network is used to sample from the motion space, and the network parameters are learned during training.

VGAN is a generative adversarial network for video with a Spatio-temporal convolutional architecture untangles the scene's subject from the background, according to Carl [11]. This model can produce small films at a full-frame rate for up to a second. This model uses data to train an autoencoder, and the decoder uses a two-stream generator network. After that, it feeds instances through the encoder and fits a 256-component Gaussian Mixture Model (GMM) over the 100-dimensional hidden space. This model takes a sample from the GMM and feeds it through the decoder to create a new video. The created scenarios in films generated with this model are mostly very crisp, and the movement patterns are often realistic for the scenario. The lack of object resolution is a key flaw in this approach.

Text-Filter conditioning Generative Adversarial Network (TFGAN) [12] is a conditional Generative Adversarial Network (GAN) model designed specifically for text-to-video gener- ation. It incorporates a unique multi-scale text-conditioning scheme to enhance the association between text descriptions and video content. The goal of TFGAN is to generate high- quality videos that accurately represent complex real-world scenarios based on textual input.

The distinguishing feature of TFGAN is its conditioning strategy, which leverages multi-scale information from the text descriptions. This approach allows the model to capture and utilize fine-grained details present in the text, resulting in improved associations between the generated videos and the provided textual descriptions.

To achieve this, TFGAN integrates the conditioning scheme with a deep GAN architecture. The GAN framework consists of a generator network and a discriminator network. The generator takes the text input as a condition and generates video frames that align with the provided description. The discriminator network is trained to distinguish between real and generated video frames.

TFGAN's conditioning strategy enables the generation of videos depicting concepts that were not encountered during the training phase. This suggests that the model has the ability to generalize and create videos based on novel concepts, showcasing its capacity for creative output.

By combining the conditioning scheme with a deep GAN architecture, TFGAN aims to produce high-quality films that accurately represent the content described in the input text. This model holds promise for text-to-video generation tasks, particularly in complex real-world scenarios, where capturing intricate details and associations is crucial.

TiVGAN [19] is a text-to-video generation model introduced in the referenced paper [19][26]. This model utilizes the Kinetics dataset as a source of training data. The primary objective of TiVGAN is to generate a video based on a given textual description.

The process starts by using a GAN network to generate a single image that corresponds to the provided text description. This image serves as the initial frame of the video. Then, TiVGAN employs an iterative approach to produce subsequent frames of the video. The model continuously refines the gen- erated frames, aiming to minimize the loss function associated with each iteration.

During the iterative process, TiVGAN evaluates the realism of the generated video by comparing it to real videos. This in- volves employing a discriminator network to discern between real and fake videos. The model seeks to generate a video that can pass the discriminator's scrutiny, indicating that it closely resembles real videos.

By iteratively generating and refining frames while minimizing the loss function, TiVGAN gradually constructs a full- length video based on the provided textual input.

The use of the Kinetics dataset provides a diverse range of training examples, enabling the model to learn the dynamics and characteristics of different actions and scenes, resulting in more realistic and contextually consistent video generation. TiVGAN represents an approach to text-to-video generation that combines GAN networks, iterative refinement, and adver- sarial training to generate videos that capture the essence of the provided textual descriptions.

V. APPLICATIONS

Text-to-visual generation, also known as text-to-image syn- thesis, has numerous applications across various domains. Here are some of the notable applications:

- Content Creation and Storytelling: Text-to-visual generation can be used to enhance content creation and storytelling. It allows writers, authors, and content creators to bring their written descriptions to life by generating visual representations of their ideas, enabling them to create visually engaging content for books, articles, advertisements, and more.
- Design and Advertising: Text-to-visual generation can assist in the design process by automatically generating visual elements based on textual descriptions. Designers and advertisers can use this technology to quickly generate visual assets such as logos, illustrations, and product designs, saving time and resources.
- Virtual Environments and Gaming: Text-to-visual genera- tion can be utilized to generate visuals for virtual environ- ments and gaming applications. It enables the automatic generation of realistic scenes, characters, and objects based on textual descriptions, enhancing the immersive experience of virtual reality (VR) and augmented reality (AR) applications.
- Personalization and Customization: Text-to-visual gener- ation can facilitate personalized content creation. It can generate custom visual representations based on user preferences, allowing for personalized avatars, profile pictures, and visual content tailored to individual users.
- E-commerce and Product Visualization: Text-to-visual generation can be leveraged in e-commerce to provide visual representations of products based on textual descriptions. This enables customers to visualize and interact with products before making purchasing decisions, enhancing the overall shopping experience.

VI. EVALUATION OF GAN

Evaluating a GAN-based network involves assessing the generated samples' quality, diversity, and fidelity. Here are some common evaluation metrics and techniques used to evaluate GAN models:

- Visual Inspection: Visual inspection involves manually examining the generated samples. Evaluators assess the generated images' overall quality, realism, and coherence. They look for artifacts, blurry regions, inconsistencies, and any deviations from the desired output. While sub- jective, visual inspection provides valuable insights into the qualitative aspects of the generated samples.
- Inception Score (IS): The Inception Score measures the quality and diversity of generated images. It utilizes a pre- trained Inception model to compute the class probabilities of the generated samples. A higher Inception Score indicates higher image quality and diversity. However, it has limitations and may not capture all aspects of image quality.
- Frechet Inception Distance (FID): FID is another popular metric that assesses the similarity between the distribution of real images and generated images. It calculates the distance between feature representations extracted by an Inception model. Lower FID scores indicate higher simi- larity and better quality in terms of distribution matching.
- Precision and Recall: Precision and recall metrics evaluate the performance of GANs in generating specific objects or classes. For conditional GANs, precision mea- sures the percentage of correctly generated samples for a specific class, while recall measures the percentage of real samples belonging to that class that were successfully generated.
- Perceptual Metrics: Perceptual metrics, such as SSIM (Structural Similarity Index) and PSNR (Peak Signal-to-Noise Ratio), compare the generated images against

TABLE I SUMMERY TABLE

S. No	Network Name	Specifics	Limits
1	GAN	The main idea behind GAN-INT-CLS is to generate realistic images based on textual descriptions. The discriminator network is responsible for distinguish- ing between real images and the images generated by the generator. It is trained to improve its ability to classify whether an image is real or fake.	First, the proposed GAN framework may struggle to generate high-resolution images with fine-grained details. Secondly, the evaluation is primarily qual- itative, lacking extensive quantitative metrics for objective comparison. Lastly, the generalizability of the approach beyond the specific datasets used in the paper is not thoroughly explored.

2	StackGAN	Stacked Generative Adversarial Network (Stack-GAN) is used to create 256x256 photos primarily based totally on textual content descriptions. The textual content-to-photographic technology is split into two ways in terms of usage of StackGAN: Stage- I GAN and Stage-II GAN. Stage-I GAN creates a low-decision photograph and Stage-II GAN corrects defects within the low-decision photograph created via way of means of Stage-I and creates a high- decision photograph. The approach involves pretrain- ing a text-conditional GAN and then training an image-conditional GAN, leading to improved image synthesis capabilities.	First, the approach heavily relies on accurate and informative textual descriptions, which may not al- ways be available or reliable. Secondly, while the model shows promising results on specific datasets, its generalizability to diverse domains and real- world scenarios is not extensively explored. Lastly, the training process involves a two-step procedure, which can be computationally intensive and time-consuming compared to single-stage GAN models.
3	StackGAN++	StackGAN++, an improved version of the original StackGAN model. StackGAN++ utilizes a multi- stage refinement process with multiple generators and discriminators, enabling the generation of high- quality and diverse images from text descriptions. Experimental results demonstrate that StackGAN++ achieves superior performance compared to previ- ous models, producing more realistic and visually appealing images.	First, the multi-stage refinement process in Stack-GAN++ increases the complexity of the model and requires longer training times. Secondly, while the generated images exhibit improved realism, there may still be instances where the correspondence between the textual descriptions and the gener- ated images is not accurate. Lastly, the evaluation of StackGAN++ primarily focuses on qualitative assessments, and more comprehensive quantitative metrics for assessing image quality and diversity could be beneficial.
4	AttnGAN	AttnGAN, a model that generates detailed and re-alistic images from textual descriptions. AttnGAN incorporates an attention mechanism to selectively focus on different regions of the image while con- ditioning on the text, resulting in more accurate and fine-grained synthesis. Experimental results on benchmark datasets demonstrate that AttnGAN out- performs previous methods in terms of visual quality and fidelity.	Firstly, the attention mechanism in AttnGAN relies heavily on the quality and specificity of the textual descriptions, which can be challenging to obtain in real-world scenarios. Secondly, the model's perfor- mance may vary when applied to complex scenes or when dealing with ambiguous or abstract textual descriptions. Lastly, while AttnGAN achieves im- pressive results, there is still room for improvement in terms of generating images with even higher levels of detail and realism.
5	E-ATTN GAN	It presents an improved version of the AttnGAN model for generating fashion images from textual descriptions. The enhanced attentional generative adversarial network (GAN) incorporates semantic consistency constraints to ensure that the generated images align closely with the provided textual de- scriptions. Experimental results demonstrate that the proposed approach achieves better semantic consis- tency and generates fashion images that are visually coherent and consistent with the given textual inputs.	First, the model's performance and generalizability may be influenced by the quality and specificity of the textual descriptions used as input. Ambiguous or vague descriptions could lead to less accurate or less coherent image synthesis. Secondly, the eval- uation of the proposed approach primarily focuses on qualitative assessments, and more comprehen- sive quantitative metrics for assessing the quality and fidelity of the generated fashion images could provide a more objective evaluation. Lastly, while the enhanced attentional GAN improves semantic consistency, there may still be instances where the generated fashion images do not fully capture the desired fashion style or details, indicating further room for improvement.
6	MirrorGAN	MirrorGAN, a text-to-image generation model that utilizes redescription to enhance the generation pro- cess. The proposed approach aims to improve the coherence and diversity of the generated images by iteratively refining the textual descriptions and syn- thesizing corresponding images. Experimental results demonstrate that MirrorGAN produces more diverse and visually appealing images compared to baseline models.	Firstly, the redescription process in MirrorGAN re-lies on the initial textual descriptions, which may in- troduce biases or limitations if the initial descriptions are inaccurate or incomplete. Secondly, the iterative refinement process may result in increased com- putational complexity and training time compared to other text-to-image generation models. Lastly, while MirrorGAN shows improvements in generat- ing diverse images, there may still be challenges in producing highly detailed or high-resolution images that match the complexity of real-world scenes.
7	DM-GAN	DM-GAN, a model that incorporates a dynamic memory module to improve the text-to-image syn- thesis process. The dynamic memory module allows the model to store and retrieve relevant information from past training samples, enhancing the generation of diverse and high-quality images. Experimental results demonstrate that DM-GAN outperforms pre- vious methods in terms of image quality, diversity, and relevance to the given text descriptions.	First, the dynamic memory module introduces ad-ditional complexity to the model, which can in- crease training time and computational requirements. Secondly, while DM-GAN demonstrates improve- ments in generating diverse images, there may still be limitations in capturing fine-grained details and achieving photorealistic results. Lastly, the evalua- tion of DM-GAN primarily focuses on qualitative assessments, and more comprehensive quantitative metrics for assessing the quality and diversity of the generated images could provide a more objective evaluation.

TABLE II. Summery Table

S. No	Network Name	Specifics	Limits
8	VQGAN+CLIP	VQGAN+CLIP is a powerful combination of two deep learning models, VQGAN and CLIP, that are utilized for generating and manipulating im- ages based on textual prompts. VQGAN, a Vector- Quantized Generative Adversarial Network, gener- ates images from random noise vectors by quantizing the continuous image space into discrete codes, al- lowing for better control over the generated images. CLIP, on the other hand, is a model that learns the connection between images and their corresponding textual descriptions, enabling effective understanding and manipulation of images based on textual guid- ance.	Multiple models introduce additional complexity to the model, which can increase training time and computational requirements. Further evaluation of results is not applicable.

9	Cycle GAN	Cycle GAN presents a method for translating textual descriptions into corresponding images using Cycle GAN. The Cycle GAN framework is employed to learn the mapping between the text and image do- mains and generate realistic images based on textual inputs. The proposed approach aims to bridge the gap between text and image modalities, facilitating text-to-image translation tasks.	First, the proposed method relies on the availability of paired text-image data for training, which can be challenging to obtain in certain domains or may require extensive manual labeling. Secondly, the quality and fidelity of the generated images may vary depending on the complexity of the textual descriptions and the dataset used for training. Lastly, the evaluation of the proposed approach may pri- marily rely on qualitative assessments, and a more comprehensive quantitative analysis could provide a better understanding of the model's performance and limitations.
10	TFGAN	TFGAN, a conditional GAN model with a novel multi-scale text-conditioning scheme that improves text-to-video associations. TFGAN creates high- quality films from the text on complex real-world video data sets by integrating this conditioning strat- egy with a deep GAN architecture.	First, the video quality is very low, secondly, the videos have chances to deviation from the prompt.
11	TiVGAN	TivGAN, a method for generating videos from tex- tual descriptions. The proposed approach utilizes a step-by-step evolutionary generator that combines text-to-image synthesis and image-to-video genera- tion. Experimental results demonstrate the effective- ness of TivGAN in generating coherent and visually appealing videos from text inputs.	First, the step-by-step evolutionary generator ap- proach employed by TivGAN may increase com- putational complexity and training time compared to other text-to-video generation methods. Secondly, the quality and realism of the generated videos may be influenced by the accuracy and specificity of the textual descriptions, as well as the limitations of the underlying image and video generation models. Lastly, while TivGAN shows promising results, there may still be challenges in generating videos with complex dynamics or handling diverse visual scenar- ios. Further improvements are required to enhance the realism and fidelity of the generated videos.
12	TGAN	The Temporal Generative Adversarial Net (TGAN) is a model capable of generating new videos by learn- ing representations from unlabeled video datasets. The generator in TGAN consists of two sub- networks: a temporal generator and an image gener- ator. The temporal generator generates a sequence of latent variables that correspond to latent variables in the image generator. These latent variables are then transformed by the image generator into a video with a matching number of frames.	Firstly, the quality and realism of the generated videos heavily rely on the training data and the complexity of the underlying video dataset. If the training dataset is limited or lacks diversity, the generated videos may exhibit limitations in terms of variation and visual fidelity. Secondly, TGAN may face challenges in capturing long-term temporal dependencies or complex dynamics in videos, as the model's architecture and training process may struggle to effectively model such intricate patterns. Lastly, like other GAN-based approaches, TGAN can suffer from issues such as mode collapse or lack of convergence, requiring careful hyperparame- ter tuning and training strategies to overcome these challenges.

the ground truth images. These metrics evaluate image similarity in terms of structure, texture, and pixel-level fidelity. However, they may not capture the higher-level semantic aspects of image quality.

- User Studies: User studies involve gathering human feed- back by conducting surveys or preference tests. Partici- pants rate or rank the generated images based on quality, realism, and other desired attributes. User studies provide valuable insights into the subjective perception of the generated samples.

It's important to note that no single metric can fully capture the quality and performance of a GAN-based network. It is recommended to use a combination of evaluation metrics and techniques to obtain a more comprehensive assessment. Additionally, evaluation should consider the specific objectives and requirements of the application to ensure the chosen metrics align with the desired outcomes.

VII. CONCLUSION

The review focused on studying the applications of GAN and different visual generation models. The study pointed out some of the work done in the field of image synthesis and video synthesis from text input. This work also discussed some evaluation matrices that can be used to evaluate GAN.

DECLARATION STATEMENT

Funding	No, I did not receive
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Not relevant.
Authors Contributions	I am only the sole author of the article

REFERENCES

1. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley,
2. S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in Proc. 27th Int. Conf. Neural Information Processing Systems, Montreal, Canada, 2014, pp. 2672–2680.
3. Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
4. Gregor, K., Danihelka, I., Graves, A., Rezende, D., & Wierstra, D. (2015, June). Draw: A recurrent neural network for image generation. In International conference on machine learning (pp. 1462-1471). PMLR.
5. Huang, H., Yu, P. S., & Wang, C. (2018). An introduction to image synthesis with generative adversarial nets. arXiv preprint arXiv:1803.04469.

6. Agnese, J., Herrera, J., Tao, H., & Zhu, X. (2020). A survey and taxonomy of adversarial neural networks for text-to-image synthesis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(4), e1345. https://doi.org/10.1002/widm.1345
7. Hayashi, M., Inoue, S., Douke, M., Hamaguchi, N., Kaneko, H., Bachelder, S., & Nakajima, M. (2014). T2v: New technology of converting text to cg animation. ITE Transactions on Media Technology and Applications, 2(1), 74-81. https://doi.org/10.3169/mta.2.74
8. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas,
9. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 5907-5915).
10. Kambhampati. Monica, Duvvada Rajeswara Rao," Text to Image Translation using Cycle GAN", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April 2020 https://doi.org/10.35940/ijeat.D8703.049420
11. Tulyakov, S., Liu, M. Y., Yang, X., & Kautz, J. (2018). Mocogan: Decomposing motion and content for video generation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1526-1535). https://doi.org/10.1109/CVPR.2018.00165
12. Saito, M., Matsumoto, E., & Saito, S. (2017). Temporal generative adversarial nets with singular value clipping. In Proceedings of the IEEE international conference on computer vision (pp. 2830-2839). https://doi.org/10.1109/ICCV.2017.308
13. Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating videos with scene dynamics. Advances in neural information processing systems, 29.
14. Balaji, Y., Min, M. R., Bai, B., Chellappa, R., & Graf, H. P. (2019, August). Conditional GAN with Discriminative Filter Generation for Text-to-Video Synthesis. In IJCAI (Vol. 1, No. 2019, p. 2). https://doi.org/10.24963/ijcai.2019/276
15. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas,
16. N. (2018). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE transactions on pattern analysis and machine intelligence, 41(8), 1947-1962. https://doi.org/10.1109/TPAMI.2018.2856256
17. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1316-1324). 1316–1324. https://doi.org/10.1109/CVPR.2018.00143
18. Ak, K. E., Lim, J. H., Tham, J. Y., & Kassim, A. A. (2020). Semantically consistent text to fashion image synthesis with an enhanced attentional generative adversarial network. Pattern Recognition Letters, 135, 22-29. https://doi.org/10.1016/j.patrec.2020.02.030
19. Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). Mirrorgan: Learning text- to-image generation by redescription. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1505- 1514). https://doi.org/10.1109/CVPR.2019.00160
20. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee,
21. H. (2016, June). Generative adversarial text to image synthesis. In International conference on machine learning (pp. 1060-1069). PMLR.
22. Zhu, M., Pan, P., Chen, W., & Yang, Y. (2019). Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5802-5810). https://doi.org/10.1109/CVPR.2019.00595
23. Kim, D., Joo, D., & Kim, J. (2020). Tivgan: Text to image to video generation with step-by-step evolutionary generator. IEEE Access, 8, 153113-153122. https://doi.org/10.1109/ACCESS.2020.3017881
24. Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., & Raff, E. (2022, October). Vqgan-clip: Open domain image generation and editing with natural language guidance. In European Conference on Computer Vision (pp. 88-105). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19836-6_6
25. 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). https://doi.org/10.1109/CVPR42600.2020.00790
26. Karthika*, N., Janet, B., & Shukla, H. (2019). A Novel Deep Neural Network Model for Image Classification. In International Journal of Engineering and Advanced Technology (Vol. 8, Issue 6, pp. 3241–3249). https://doi.org/10.35940/ijeat.f8832.088619
27. Sumanth, A. G., R. Hema, Sumanth, R. H., Chowdary, A. C. V., Shashank, A., & Sravan, T. (2020). Real Time Image Captaioning. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 6, pp. 1707–1709). https://doi.org/10.35940/ijitee.f4566.049620
28. Bai, D. M. R., Sreedevi, Mrs. J., & Pragna, Ms. B. (2020). Enhanced Unsupervised Image Generation using GAN based Convolutional Nets. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 6, pp. 5312–5316). https://doi.org/10.35940/ijrte.f9856.038620
29. Radhamani, V., & Dalin, G. (2019). Significance of Artificial Intelligence and Machine Learning Techniques in Smart Cloud Computing: A Review. In International Journal of Soft Computing and Engineering (Vol. 9, Issue 3, pp. 1–7). https://doi.org/10.35940/ijsce.c3265.099319
30. Nair, V. K., Jose, R. R., Anil, P. B., Tom, M., & P.L., L. (2020). Automation of Cricket Scoreboard by Recognizing Umpire Gestures. In International Journal of Innovative Science and Modern Engineering (Vol. 6, Issue 7, pp. 1–7). https://doi.org/10.35940/ijisme.g1235.056720

AUTHOR PROFILE

Sibi Mathew, Education: Pursuing MTech in Computer Science and Engineer- ing at TKM College of Engineering, Kol- lam, Kerala, India. **Interests:** Specializing in Image Processing, Image-based Fault Diag- nosis, Deep Learning, and Machine Learn- ing. **Background:** Holds a Bachelor's degree in Computer Science and Engineering from Amal Jyothi College of Engineering. Pas- sionate about leveraging advanced technolo- gies to solve real-world problems in diverse domains.