Available at https://www.iasj.net/iasj
Iraqi Academic Scientific Journals
Journal homepage:
https://journals.uokerbala.edu.iq/index.php/UOKJ

Research Article

A Survey of GAN-Based Text-to-Face Generation: Models, Forensic Applications, and Open Challenges

[1,]Qamar Haider M.Ali1 [2,] Hiba Jabbar Aleqabie
[1,2] Computer Science Department, College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq.
[2,] Artificial Intelligence Engineering Department, College of Information Technology Engineering,
Al-Zahraa University for Women.

Article Info
Article history:
Received 17 -1-2026
Received in revised form 31-3-2026
Accepted 3-6-2026
Available online 30 -6 -2026
Keywords: Text-to-Face Synthesis ,Multimodal Artificial Intelligence (GANs), Diffusion Models ,AI Ethics.

Abstract

One of the most fascinating aspects of the merger between vision and language Artificial Intelligence is how to create human faces from textual data. This is one of the most challenging tasks in the field. The current state of the art in the field of text to facial generation is reviewed in this paper. From the first GANs to the most recent diffusions and transformers, we describe how and why the models use textual data to create faces, and what attributes and characteristics affect their output including informativeness of text, quality of the data-set, operational inefficiency, and dis/semantic gaps. We describe the bounding industry applications of the technology, including forensic facial reconstruction, security, entertainment, healthcare, and the arts, and we identify both potential and limitations of the technology. The technology is also applied to various fields and several ethics have been mentioned including the lack of human centered methodologies in assessing the data. We describe the most recent progress, the challenges faced in the field and the directions the field might take in the near future, and we present our efforts to ensure the ethical multi-faceted use of the technology. The review also describes in great detail the lack of ethical and evaluative methodologies in examining the potential of such systems. We present the review from this perspective in the hopes of providing the field with more rational and human use of such systems. This review distinguishes itself from prior works that focus on a single model or application. It provides an organized comparative summary of major text-to-face generation models — GANs, diffusion models, transformers — that have been assessed in the literature on dimensions of image fidelity, semantic coherence, efficiency, and ethics. It aims to help researchers pinpoint neglected areas in the field, understand model trade-offs, and formulate new avenues of research in this dynamic field.

Corresponding Author E-mail: qamar.h@s.uokerbala.edu.iq, Hiba.jabbar@uokerbala.edu.iq
Peer review under responsibility of Iraqi Academic Scientific Journal and University of Kerbala.

Introduction

Text-to-image face generation, a branch of text-to-image conversion, has attracted significant interest recently due to its potential applications in several fields, such as security, forensics, entertainment, and healthcare [1], [2], [3]. This review paper provides a structured comparative overview of the present state of text-based face generation using textual descriptions, with particular emphasis on forensic applicability and evaluation methodology. Three primary generative paradigms are examined in this review: Generative Adversarial Networks (GANs), which employ a generator-discriminator framework to produce realistic images through adversarial training [4]; diffusion-based models, which learn to reconstruct images by reversing a noise corruption process [5]; and transformer-based models, which leverage self-attention mechanisms to enhance semantic alignment between text and generated faces [6]. There are many challenges and limitations in generating faces from text descriptions, such as incomplete or ambiguous textual descriptions, limited dataset diversity, and insufficient realism in forensic contexts. For example, Bayoumi et al. [7] showed text clarity influences how accurately facial features are constructed, and Bosheah and Bilicki [8] pointed out that most contemporary models have difficulty with particular descriptive inputs. In criminal investigations [9], [10] , the absence of annotated forensic datasets continues to be a barrier to practical implementation.

This review examines the following principal research questions:

1. How could Generative Adversarial Networks (GANs) be utilized to turn written descriptions of people who are wanted for crimes into genuine human faces?
2. What are the best ways to preprocess and embed text to acquire descriptive information from witness testimony that can help make realistic faces?
3. How can forensic experts and numerical evaluation metrics be used to assess the authenticity of generated faces when compared to real suspects?
4. What are the main challenges that come up when you try to match language features with visible facial traits?
5. How does the suggested GAN model compare to other models that turn texts into faces?

The organization of the paper

- Part 2: A look at the history and theory behind generative models, such as GANs, VAEs, and diffusion models, as well as overviews of datasets [7] ,[11]
- Part 3: A look at different models for text-to-face generation, such as early GAN-based methods, more advanced diffusion and transformer-based methods, and a comparison of performance measures [3], [7], [12], [11], [13].
- Part 4: Discuss the applications in security, entertainment, healthcare, and creative domains [1], [2], [14].
- Part 5: Issues and limitations of current methodologies, emphasizing text quality, semantic accuracy, and computational efficiency [6], [7], [8].
- Part 6: Future study directions, encompassing potential enhancements in text preparation, multimodal alignment, and dataset augmentation [6],[7], [8]
- Part 7: Conclusion that summarizes the findings and highlights significant areas for further investigation.

1. Background and Foundations

The integration of textual description and visual data is important and one of the modern tasks in the field of artificial intelligence and computer vision [1]. In addition, generating

images using textual description has the potential to change the ways of perceiving and interacting with visual data. Competitive generative networks, which were introduced in 2014, consist of two main sub-models: the generator and the discriminator [4]. The generator's role is to create fake images, while the discriminator's role is to distinguish between images as to whether they are real or fake. Many models based on GANs have been developed to address the previous gaps and have begun to be used in many different fields, including creating more realistic images and cartoon characters, increasing accuracy, and converting images to text and vice versa[4]. Recently, diffusion models have been considered a subset of deep generative models, due to the important results they have shown that surpass competitive generative network models [15]. Diffusion models have proven successful in generating images with high accuracy, as well as in coloring, editing, and translation. The diffusion model generates a Markov chain of diffusion steps. Stable diffusion utilizes a latent diffusion model architecture where VAE, U-Net, and the text encoder operate in a low-dimensional latent space, significantly reducing processing requirements compared to pixel-based diffusion models [15]. Transformer models such as BERT, GPT, and XLNet provide new opportunities for generating images from text. XLNet uses a transformer architecture and an attention mechanism to encode input data by predicting the symbol based on all other symbols in the sequence, allowing complex data dependencies to be taken into account [1]. The attention mechanisms in stable diffusion models enable the model to adaptively focus on different parts of the input data and enhance the expression of important features. Cross-media attention allows each word in the text to interact with a region in the image, thus ensuring that the resulting content matches the input text description, while self-attention captures the dependencies between image regions, thus ensuring structural coherence and image content consistency [1].

2.1 Generative Models Overview: GANs ،VAEs ،Diffusion Models.

Almost all text-to-face synthesis relies on the generative model. Among the classic neural networks, Generative Adversarial Networks (GANs) consist of a generator and a discriminator that compete to create incredibly realistic images [7], [3], [11]. On the other hand, images created from the Variational Autoencoder (VAE) models tend to be less realistic than products of a GAN. Recently, the diffusion-based algorithms have gained popularity due to their proficiency in transforming noise to create a variety of complex images, especially images of faces, as used in the detailed work required [12], [13], [15].

2.2 Text-to-Image vs. Text-to-Face

Generating faces from text is a specialized form of image generation from textual description. Face modeling differs from creating general objects or scenarios because it requires a deep understanding of human anatomy, emotional expression, and the characteristics that constitute an individual's identity. If textual descriptions lack clarity, the resulting facial expressions may appear less authentic. This highlights the importance of having both meaningful and well-crafted text [7], [8]. Models should not simply convert text into images. Faces generated by models must be coherent, consistent, and compatible with human perception [2], [3], [16].

2.3 Datasets and Benchmarks

Multiple datasets have been employed to train and assess text-to-face models. CelebA has a lot of famous face photographs with

descriptions for different features. This lets you design artwork based on those traits [16]. Multi-modal datasets combine text descriptions with facial images, making it easier to train models to work with different types of data [1],[3],[12]. Even though they are small, specialized forensic datasets are very important for criminal investigations because they help models make faces that look like what witnesses say [10],[9].The Frechet Inception Distance (FID) and the Inception Score (IS) are two common ways to evaluate the quality of the generated image. Research depending on human preference examines the relationship between textual inputs and produced facial representations [2],[3],[8].

Recently, the research has introduced multi-modal datasets, including Deep Fashion-MultiModal and Face2Text, designed to integrate textual and visual modalities by providing corresponding text-image samples. However, as demonstrated in Text-to-Image Synthesis with Generative Models [1], these datasets often lack sufficient diversity in text types or adequate detail in descriptions, complicating the attainment of appropriate semantic alignment. There is difficulty in judging the models correctly due to the lack of standard tests for the degree of correspondence between the texts and the generated faces.

3 . Models of Text – to- Face Generation

3.1 Early Approaches: GAN-based frameworks (StackGAN, AttnGAN)

The last few years have been very positive with the development of Generative Adversarial Networks (GANs) for cross-modal applications. This is particularly the case with text-to-image generation which enables the automation of design tasks by making the generation of realistic images from text descriptions quick and easy. GANs have been and continue to be successful in various applications. Despite all of that, text-to-image generation through the use of GANs is still very difficult. First is the quality of the images created from the text. Second is the uncertainty that comes from the training of the GAN which leads to images that lack detail and diversity. Because of the interest in this field and the use of GANs in cross-modal applications, there has been a lot of new work in the field. One of the notable contributions in this area is the Stacked Generative Adversarial Networks (StackGAN) which introduced a multi-stage process to generate high resolution images over multiple iterations [1],[7]. Another very notable advancement is the Attentional Generative Adversarial Networks (AttnGAN) which introduced a multi-level attention mechanism to improve the correlation between text and image features to generate more detailed images [2],[16]. As we stated, poor stability of the model and poor quality of images still present challenges in text to image generation. Additionally, successfully managing intricate text descriptions continues to be a pivotal research emphasis at present. This research study focuses on the evaluation and analysis of the fundamentals and advancements of technology centered on GAN text to image generation. First, an introduction is given to the basic concepts of the, and then, there is an overall analysis of the experimented efficiency of the dominant architectures in the field. The study also evaluates the advantages and disadvantages, and the potential of these models in future development.

3.2 Advanced Approaches

3.2.1 Diffusion Models (Stable Diffusion extensions)

Recent developments around face generation have provided modifications to GANs, which now use diffusion techniques, and transformers, which have seen improvements

in photo-realism, semantics, and control. Several models have made diffusion techniques more effective; among them, the most successful is Stable Diffusion. These models employ a learned reverse-diffusion technique to incrementally remove noise from a random noise vector, thereby generating a realistic image. This approach enables efficient and scalable production while maintaining outstanding perceptual quality [5].In creating images of faces based on textual descriptions, multiple studies have further developed the diffusion framework and proposed methods for the integration of textual and sketch-based conditioning, improving the modeling of the intricacies of faces [4], [5]. The study by [17] proposed a multi-modal diffusion technique that integrates text, sketches, and facial attributes to enhance the coherence and control of the synthesis process. This also enhances Stable Diffusion's customization of face generation. With textual input, users can modify the image to reflect changes in emotions, age, and hair. When it comes to visual fidelity and semantic alignment, these methods are much better than regular GANs. However, these methods remain computationally intensive and require significant resources, which is a big problem, especially when you need to do them quickly or with few resources, such as forensic investigations.

3.2.2 Transformer-based Models (CLIP-guided generation)

The use of Transformer models like Contrastive Language-Image Pre-training (CLIP) has opened up an entirely new world of enhanced face generation using text-to-image models. The combination of text-to-image models, or, as it is commonly known, multimodal models, creates more nuanced meaning than models restricted to one format. OpenAI's CLIP is a large model that is able to assign both an image and a textual context to an embedding in the same semantic space, thanks to its training on almost everything available on the internet. CLIP contains two encoders. The text encoder runs a sort of a Transformer, while the image encoder is a Vision Transformer or a CNN, both of which produce image and text embedding vectors that are equivalent. In applications for face generation, CLIP is able to contextualize the relationship and alignment of a description or other text input to an image or a set of images of a simulated generated face. For illustration, imagine the works of Xiang et al. [15], who provided a set of simulacra of age, expression, and hairstyles in the final image of face generation, the Stable Diffusion-based framework relied on CLIP's text encoder and the input description served as the semantics to manage the denoising via cross-attention. Models employing CLIP aren't faultless either, as is the case for the CLIP-guided applications of Wu et al. [2], who found that their preference classifier at least in terms of error, did a better job than CLIP, who got 32.9% of the results correct and the classifier got 43.5% correct, demonstrating that while CLIP can almost always potentially guide semantic representation, it still lacks an ability for capturing and discerning certain qualities of the face e.g. matched realism that a human face would possess.

3.3 Comparative Analysis

Evaluating text-to-image models shows that achieving a balance between text alignment, image quality and computational efficiency is a complex, interrelated process. Two widespread metrics used to assess an image's credibility are Inception Score (IS) and Fréchet Inception Distance (FID). While IS evaluates whether recognizable and/or multiple different objects are present in an image, FID analyses the distributions of real and/or fake images in a certain deep feature space. However, IS and FID perceive the world differently to how people do. IS scores may be invalid for images that are not akin to

ImageNet, while FID may perceive noise and compression artefacts as low-quality images. Based on FID, people perceived CogView2 to be the best model of the three: GLIDE, CogView2 and Stable Diffusion. In contrast, Stable Diffusion had real and accurate annotations, which were preferred by the human annotators. This difference highlights the inadequacy of automated metrics in assessing perceptual realism. Computational efficiency remains a key concern. Diffusion models are more effective than GAN-based models at sustaining coherency and enhancing the image. However, they are more resource-intensive than GAN-based models, primarily due to the numerous denoising steps they must perform. Transformer-based models, on the other hand, are transformer-based models.

4. Applications

4.1 Security and Forensics

Police can create visual representations based on witness accounts using this technology, allowing for more precise and efficient visual impressions, and thereby the improvement of criminal investigations.

4.2 Entertainment and Gaming

In the entertainment industry, for gaming and other entertainment purposes, designers can create brand-new, original characters from text descriptions in a matter of moments, thanks to text-to-face generation. This system can create avatars, digital actors, and non-playable characters (NPCs) without the requirement for 3D modeling. Furthermore, the innovative applications [1],[3], [14] which use diffusion-based and transformer-guided models can produce different avatars and diverse characters at a lower cost while maintaining a higher realism.

4.3 Healthcare and Assistive Tools

Assisting people in especial needs or memory loss can be coupled with healthcare and assistance applications. In healthcare, these models can assist in tracking a patient or aid in the location of a missing person by reconstructing a face from a description to assist from partial accounts [1],[9] Sensitive health care applications need to be reliable and create images of the health care applications, which needs text to be interpreted and grow the images based on the description [1], [8].

4.4 Creative Industries

The creative industries can leverage this text-to-image technology in numerous ways, including marketing, education, and product design. Users can easily incorporate digital images, illustrations, infographics, and various visual assets thanks to the ability to merge text with images.

5. Challenges and limitations

5.1 Text Quality and Semantic Ambiguity

One of the most important factors for the generation of text-to-image systems is the quality and the clarity of the text descriptions entered. If the input text is ambiguous, incomplete, or inconsistent, it is likely that the facial outputs generated will be unrealistic or inaccurate, and will be of poor quality. Output models must be able to identify small nuances, especially in forensic applications, where small details like the shape of the eyes, hair style, and the presence of facial scars can be crucial. Output models tend to struggle with small, subtle, or specific details which can be the reason why some outputs do not match or belong to the desired faces or overall identities.

5.2 Dataset Limitations

Unfortunately, well organized, annotated, and easy to access multimodal data is one of the main challenges for facial data synthesis systems. Although dataset collections with facial images, like CelebA, offer valuable

resources for training, there is a lack of sufficient textual descriptions that can be used to train reliable models for text-to-face synthesis. In forensic and security applications, there is a lack of sufficient specialized training data in the form of eyewitness descriptions to train and validate models in real-world scenarios of the system. In addition, not having enough datasets with varying facial images hinders the ability to improve the overall accuracy and effectiveness of a multimodal system. Some of the systems lack the ability to generalize across varying ethnicities, age groups, and different facial expressions.

5.3 Model Performance and Computational Constraints

New diffusion systems and advanced transformer system frameworks are able to generate very high quality systems, but they require much more computing power and hardware resources in order to perform at peak efficiency. These systems require many high quality datasets for them to be able to perform at their best. This in turn creates a very risky situation in the real world, since a lot of training is needed for inefficient and inaccessible systems. Building systems which can perform at high efficiency and which are able to generate good quality outputs is very time consuming, and therefore training the models will be hinder, especially for smaller research facilities and for applications that require instantaneous performance. To meet the text Diffusion Models computational requirements within time-sensitive scenarios, future works must seek model simplifications paired with aggressive deployment optimizations, including fewer denoising iterations, then sacrificing some image quality, as a near-term solution for achieving demonstrable, real-time face generation from witness descriptions [5], [15]. Furthermore, it is difficult to calibrate image realism and semantic alignment since models pursuing visual realism may end up losing textual fidelity, and the other way around [2],[3].

5.4 Assessment Standards

Assessing the text-to-face models should be considered one of the most difficult challenges. It is true that the standard metrics, e.g. FID and IS, do metric image quality, however, it is not true that image quality captures the alignment with the text [2], [3], [8]. It is true that human evaluation is the most reliable metrics concerning the alignment of the description and the generated face, but it is subjective and not easy to standardize [1], [2]. Hence, the field remains short of an objective and reproducible assessment standard that incorporates quality and alignment.

5.5 Ethical and Practical Issues

Text-to-face generation, and particularly text-to-face generation in forensics, is not at all free of ethical issues, such as privacy, misidentification, and the potential for abuse [10],[9] . When faces based on textual descriptions are generated, models may unintentionally transfer biases when faces are generated based on non-representative datasets[8], [12] Thus, the lack of fairness, accountability, and transparency in the design and implementation of the models is one of the most important issues to be solved in the future [12].

6. Future Research Directions

6.1 Improving Text Quality

Further research should continue to target improving the quality and depth of the text inputs. Face attributes should be better captured with the use of sophisticated text preprocessing and natural language enhancements, as well as embedding refinement [7], [8]. The text-to-image fidelity, especially for forensic applications, can be augmented by standardizing and

enriching eyewitness accounts with large language models [1],[8].

6.2 Alignment and Representation

Across Modalities the inclusion of multi-modality academic frameworks offers highly useful opportunities for synergetic research. Some improvements to semantic alignment have been shown through the use of transformer models and CLIP-like embeddings [3],[18] The realism and coherence of the images generated from these models should be improved by exploring better mechanisms of alignment to capture the fickle textual nuances.

6.3 Models Generalization through Data Diversity and Expansion

The generalization of the models can be improved through the creation of large-scale multi-concept datasets that are highly annotated. More attention to datasets that concentrate on specific forensics use with detailed descriptions and datasets that capture diverse ethnicities, ages, and expressions is needed. The use of GANDiffFace and Collaborative Diffusion [9], [10], [12] should also be considered for realistic datasets to be complemented by synthetic datasets.

6.4 Assessment Techniques and Human-Centered

Evaluation In future research, we ought to design thorough assessment mechanisms that couple quantitative components (e.g., FID, IS) with qualitative, human-centered ones [2],[3],[8]. Creating metrics that ascertain how semantically faithful generated images are to their paired text would yield fairer comparisons across various models and would push advancements in systems that generate text-to-face.

6.5 Ethical and Responsible AI Practices

The ethical, equitable, and responsible use of text-to-face generation suffers from a lack of research, especially in relation to forensic and security contexts [10], [9], [8]. Future research should include bias mitigation, transparency in model instantiation, and privacy-preserving systems to reduce the potential for misuse of generated facial imagery.

7. Related Work

In Table 1, we summarize 20 studies on text-to-image synthesis and face generation, published from 2022 to 2025. The studies address work with a variety of methods and architectures in deep learning such as Generative Adversarial Networks (GANs), Latent Diffusion Models (LDMs), and hybrid models. The first generation of models focused on GAN-based architectures such as DCGAN, StackGAN, and AttnGAN. While these models produced relatively realistic generator outputs, these models tended to be low resolution and not diverse, and they lacked fine-grained realistic details. Diffusion-based models, especially Stable Diffusion, have been recently used increasingly within this domain. These models produce better text-and-image alignment and overall a better quality image.
One of the trends in the literature is the use of CLIP as a bridge that is semantic between text and image which offers better control for generating images. Utilizing multiple modalities, a number of studies have focused on the generation of text, sketches, and 3D morphable models and masks all of which improve the degree of control of the generator. While these models have improved, there are a considerable number of studies which focus on the high cost of computation, the need for large transformer

architectures for pretraining, and the generation of images that are structurally complex and high quality. The literature does a great job of summarizing the rapid improvement in the domain of generative modeling for face and image synthesis, and it outlines the need for models that are more interpretable, efficient, and create a synthesis in a more ethical manner.

Table 1: Summary of Related Works on Text-to-Image and Face Generation Using Deep Learning Techniques

#	Author(s), Year	Title of the Study	Method	Dataset Used	Objectives	Key Techniques	Results	Limitations	Contribution
1.	Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer-2022	A Sketch Is Worth a Thousand Words: Image Retrieval with Text and Sketch [19]	TASK-former (dual-encoder with CLIP backbone)	COCO 5k, Flower 102, synthetic sketch generation	Improve image retrieval using combined sketch + text queries	Late-fusion dual-encoder, multi-label classification, caption generation, synthetic sketch augmentation	Higher Recall@1 (0.609), Recall@5 (0.847), Recall@10 (0.917) compared to SOTA	Performance influenced by sketch quality; limited real-sketch datasets	Introduced TASK-former and new sketch dataset; proved complementarity of text+sketch
2.	Wenlong Xiang, Shuzhen Xu, Cuicuil v, AND Shuo Wang -2024	A Customizable Face Generation Method Based on Stable Diffusion	Modified Stable Diffusion with LoRA and customized VQ-VAE	CelebA (low-resolution), FFHQ references	Create a controllable facial generation model with improved naturalness	CLIP-guided diffusion, cross-attention, LoRA fine-tuning, customized VQ-VAE	Improved facial realism, stronger text–image alignment, TQDA evaluation shows quality gains	High-res datasets may cause overfitting; customization limited by training data	Provides a stable diffusion variant optimized for face customization and reduces overfitting

		Model [15]							
3.	Priya Yadav, Ramandeep Kaur, Dr. U. Hariharan-2023	A Study on Customized Face Generator using GANs [20]	GAN-based Customized Face Generator (CFG) using GANs and VAEs	Not explicitly stated (likely generic facial datasets)	Develop a customizable and ethical face-generation system	GANs, VAEs, progressive growth, conditional image generation	High realism, adjustable facial attributes, improved user-controlled generation	Potential ethical issues (bias, misuse); dataset not detailed	Introduces CFG system combining realism, diversity, and customization
4.	Sarita d D. Deshpande, Pallavi N. Shejwal, Vandana G. Dixit, Sampada A. Kulkarni, Swapna S. Bhavsar-2023	Automatic Synthesis of Realistic Human Images From Text Using GANs [21]	VQGAN + CLIP + GAN-based T2F	CelebA	Generate realistic faces from text	GAN, VQGAN, CLIP	Good realism but limited detail	Small dataset, artifacts	Extends T2F using GANs

5.	Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, Hongsheng Li-2023	Human Preference Score: Better Aligning Text-to-Image Models with Human Preference [2]	CLIP Fine-tuning + HPS + LoRA SD	98k images + 25k human choices	Align T2I with human preference	CLIP fine-tuning, human feedback	Improved human-preferred outputs	Metrics fail to match human prefs	First large-scale human preference dataset + HPS
6.	Zenab Bosheah, Vilmos Bilicki-2025	Challenges in Generating Accurate Text in Images [8]	Evaluation of T2I models (DALL·E3, Ideogram, SD3.5)	Custom benchmark datasets	Evaluate accuracy of text rendering	GPT-4 evaluation, structured prompts	Models fail at structured/multilingual text	OCR issues, layout/text errors	First benchmark for text accuracy in T2I
7.	Ziqi Huang, Kelvin C. K. Chan, Yuming Jiang, and Ziwei Liu-2023	Collaborative Diffusion for Multi-Modal Face Generation and Editing [17]	Collaborative Diffusion framework combining pre-trained uni-modal diffusion models using a Dynamic Diffuser	CelebA Mask-HQ, CelebA-Dialog	Enable multi-modal (text + mask) face generation and editing without retraining models	Dynamic Diffuser, spatial-temporal influence functions, multi-modal denoising, pixel-wise softmax blending	Higher consistency and image quality than TediGAN and Composable Diffusion; strong identity preservation	Depends on availability of pre-trained uni-modal models; inference time is slower due to multi-stage	First framework to merge separate diffusion models for multi-modal control without retraining

								processing	
8	Faezeh Gholamrezaie, Mohammad Manthouri — 2022	Cycle Text2Face: Cycle Text-to-Face GAN via Transformers [18]	Encoder-Decoder CycleGAN with Sentence Transformer + GAN-based face generator	CelebA (with auto-generated captions)	Convert text descriptions into a realistic face and reconstruct the text to ensure semantic consistency	Sentence-BERT text embeddings, GAN generator/ discriminator, cyclic text-to-face-to-text reconstruction	Achieved FID = 3.458; improved quality and faster processing compared to prior GAN-based models	Limited fine-grained facial detail; relies on automatically generated text; lower realism than diffusion-based models	First cycle-based text-to-face model integrating Transformers and GAN to enforce text-image consistency
9	Minchul Kim, Feng Liu, Anil Jain, Xiaoming Liu— 2023	DCFace: Synthetic Face Generation with Dual Condition Diffusion Model [12]	Dual-Condition Diffusion Model (Identity + Style) with Patch-wise Style Extractor and Time-step Depende	FFHQ (ID), Style Bank (real face images), CASIA-WebFace (evaluation)	Generate high-quality synthetic faces for training face recognition while maintaining identity consistency and	Dual-condition DDPM, patch-wise style extractor, time-dependent ID loss, two-stage generation pipeline	+6.11% improvement over previous synthetic datasets in 5 FR benchmarks; high subject uniqueness and diversity	Requires large compute; not designed for text-to-face; quality depends on style bank diversity	First method to jointly control identity and style to create realistic, identity-consistent synthetic datasets for face recognition

			nt ID Loss		high diversity				
1(	Ximing Xing, Chuang Wang, Haitao Zhou, Jing Zhang, Qian Yu, Dong Xu-2023	DiffSketcher: Text-Guided Vector Sketch Synthesis through Latent Diffusion Models [13]	Latent diffusion model + Bézier curve optimization + extended SDS Loss	No dataset required; uses pretrained diffusion models	Generate diverse vectorized free-hand sketches directly from natural language	Score Distillation Sampling (SDS), attention-based stroke initialization, differentiable rasterizer	Outperforms prior sketching methods; produces abstract and detailed sketches	High computational cost; dependent on pretrained diffusion quality	First method enabling text-to-vector-sketch generation without supervised text-sketch datasets
1:	Songwei Ge, Taesung Park, Jun-Yan Zhu, Jia-Bin Hu=] [ang-2023	Expressive Text-to-Image Generation with Rich Text [22]	Region-based diffusion guided by rich-text formatting attributes	Uses pretrained text-to-image diffusion models (no new dataset)	Enable precise, interpretable, and controllable text-to-image generation using rich text	Attention-map-based region segmentation, region-specific diffusion guidance, rich-text parsing	Produces more accurate colors, details, and styles than plain-text baselines	Processing pipeline is complex; relies on attention-map accuracy	First framework leveraging rich-text attributes (color, size, style, footnotes) for image generation

1	Rungu o Wang-2024	A Compa rative Analysi s of StackG AN and AttnG AN in Text-to-Image Gener ation [23]	StackGA N (2-stage GAN), AttnGAN (attentio n-based GAN)	CUB-200-2011 birds dataset	Compa re StackG AN vs AttnGA N in realism , detail, and text–image consist ency	Multi-stage generation, attention mechanism , text embedding	IS: 3.7 → 4.36, FID: ~50 → 23.98; AttnGAN superior	Limited to birds dataset; difficult y with comple x scenes; limited diversit y	Shows advantages of attention mechanisms; identifies weaknesses in early GAN-based T2I
1	Debin Meng, Christ os Tzelep is, Ioanni s Patras , Georgi os Tzimir opoul os-2024	MM2L atent: Text-to-Facial Image Gener ation and Editing with Multi modal Assista nce [24]	StyleGA N2-based multimo dal GAN (text + mask/sk etch/3D MM)	FFHQ, CelebA-HQ, CelebA-Mask-HQ, CelebA-Dialog	Enable controll able multim odal facial generat ion + real-image editing	StyleGAN2, FaRL text encoder, autoencode rs, pseudo-text embedding s, mapping network	CLIP score 24.59%, mask accuracy 85.61%, CMMD 1.43, fastest inference	Focused only on faces; requires multipl e modalit y encoder s	Unified multimodal generation + editing; fast robust inference; eliminates hyperparameter tuning
1	Chaoyi Tan et al., 2025	Gener ating Multi modal Images with GAN: Integra ting Text, Image, and Style [25]	Conditio nal GAN + text encoder + image feature extractio n + style encoder	COCO Caption , Oxford-102 Flowers	Genera te images combin ing text semant ics, refere nce image details, and style	Transforme r text encoder, CNN extractor, style encoder, multimodal losses	Lower FID & higher IS vs prior models; high semantic & style consisten cy	High comput ational cost; requires style images; comple x architec ture	Integrates text + image + style into unified GAN; fills gap in multimodal artistic generation

15	Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Dominik Lawatsch, Florian Domin, Maxim Schaubert-2023	GANDiffFace: Controllable Generation of Synthetic Datasets for Face Recognition with Realistic Variations [26]	Hybrid GAN + Diffusion pipeline (StyleGAN3 + DreamBooth)	Synthetic identitities (StyleGAN3), VGG2, IJB-C	Generate synthetic, demographically controllable datasets with realistic intra-class variation	Latent-space manipulation, StyleGAN3, DreamBooth personalization	Improved intra-class variation, realistic identities, balanced demographics	Reliance on pretrained models, illumination less reliable	First hybrid GAN–Diffusion synthetic dataset addressing privacy & demographic imbalance
16	Hao Zhu, Linjia Huang, Yiyu Zhuang, Yuanxun Lu, Xun Cao-2023	High-Fidelity 3D Face Generation from Natural Language Descriptions [16]	Two-stage text-to-3D pipeline (CLIP, descriptive code, 3DMM, StyleGAN2 textures)	DESCRIBE3D dataset (1,627 annotated 3D faces)	Generate high-fidelity 3D face models from fine-grained text descriptions	CLIP parsing, descriptive code, 3DMM, StyleGAN2 texture generator, abstract refinement	High-fidelity 3D faces matching fine-grained text	Small dataset, limited ear details, complex text-to-3D mapping	First fine-grained text-annotated 3D face dataset + strong baseline

17	Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Bjorn Ommer -2022	High-Resolution Image Synthesis with Latent Diffusion Models [5]	Latent Diffusion Models (LDM) + pretrained autoencoder with cross-attention	ImageNet, LAION, LSUN, CelebA-HQ, FFHQ	Efficient high-resolution image synthesis with reduced compute cost	Autoencoder latent compression, cross-attention, UNet diffusion	State-of-the-art results with far less compute	Reconstruction depends on AE quality; extreme compression reduces fidelity	Foundation of modern text-to-image models (e.g., Stable Diffusion)
18	Rishitha, Chandana Reddy, Ashritha - 2023	Face Generation Using General Adversarial Networks [4]	GAN, DCGAN, Conditional GAN variations	General image datasets (not explicitly specified)	Investigate generating faces conditioned on identity using GAN modifications	DCGAN architecture, Keras/Tensorflow, identity-conditioned GANs	Generator/Discriminator ~50% accuracy; output resembles real images but lacks detail	Difficulty generating high-resolution images; needs improved discriminator strength	Outlines GAN variants; proposes conditional GAN-based framework for improved synthesis
19	Kushal Kumar Jain (2025)	Face Sketch Generation and Recognition [14]	StyleGAN, PS-StyleGAN, Stable Diffusion, CLIP4Sketch	FS2K, CUHK, APDrawing, PRIP-Composites, synthetic sketch datasets	Study artistic portrait generation, sketch synthesis, and forensic sketch-to-mugsh	Attentive Affine Transforms, latent W+ space manipulation, diffusion-based identity preservation	State-of-the-art performance in sketch synthesis; improved recognition accuracy with synthetic datasets	Training complexity; reliance on high-quality pretrained models; limited real paired	Introduces PS-StyleGAN, CLIP4Sketch, and novel stylization methods advancing forensic and artistic sketching

					ot matching			datasets	
2(	Xin Wang, Hui Guo, Shu Hu, Ming-Ching Chang, Siwei Lyu (2023)	GAN-generated Faces Detection: A Survey and New Perspectives [11]	Survey of DL-based, physical-based, physiological-based detection methods	Multiple GAN-face datasets (FFHQ, CelebA-HQ, StyleGAN, ProGAN, etc.)	Provide a comprehensive review of GAN-face detection methods and challenges	Analysis of corneal highlights, pupil shapes, CNN-based detectors, color component methods	Summarizes performance across methods; identifies detection weaknesses	GAN-faces becoming hard to detect; human accuracy low; detection lacks interpretability	First comprehensive survey organizing detection into four categories and outlining future directions

8. Conclusion

Text-to-face generation is a complicated and intricate subject and the current review explains the issues and growing pains of the topic in great detail. Face generation from the text is an issue that started with the rudimentary use of GANs. However, the newest transformer-based technologies not only surpass GANs, and diffusion in their ability to exercise real control over the images, but also in the maintenance of a unified piece of text. Nevertheless, the generation of images from the text is still constrained by the clarity and accuracy of the prompt, and the lack of large-scale, annotated databases, especially in the forensic domain, and the immense computational resources required to accomplish face generation from text. The Fréchet Inception Distance (FID) and Inception Score (IS), which serve as the primary means of evaluation of such frameworks, lack fundamental measures of alignment with connotative meaning within a corpus, as well as the means to compare that corpus with respect to other corpuses. Until key issues surrounding equity, bias and privacy are resolved, the use of text-to-face generation systems will yet remain conceptual. More focus on alignment of text systems, and the ability to encode and further increase the volume of the datasets will allow these systems to be useful in a greater variety of real life applications, and a much broader audience. And last but not least, a new field of study will arise in relation to how to deal with outcomes of these processes in order for

the field to be able to evolve in a variety of ways to understand these features. This is the last point. Text-to-face generation technologies, for example, in case likenesses of human beings can be produced, animated, and drawn, it would be a good fit for the fields of security, assistance, and creativity. It might seem to theoretically have potential for implementation in some tasks. However, maximizing its ethical, accurate, and reliable performance in execution, in an entirely automated text-to-face paradigm, is an area of study that still needs considerable effort and constitutes a challenge**.**

9. Conflict of Interest

The authors declare no conflict of interest

References:

[1] S

. K. Alhabeeb and A. A. Al-Shargabi, "Text-to-Image Synthesis with Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction," *IEEE Access*, vol. 12, pp. 24412–24427, 2024, doi: 10.1109/ACCESS.2024.3365043.

[2] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, "Human Preference Score: Better Aligning Text-to-Image Models with Human Preference," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2303.14420

[3] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "TediGAN: Text-Guided Diverse Face Image Generation and Manipulation," 2021. [Online]. Available: https://github.com/weihaox/TediGAN.

[4] C. Reddy and R. Ashritha, "Face Generation Using General Adversial Networks," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 07, no. 01, Jan. 2023, doi: 10.55041/ijsrem17409.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," 2022. [Online]. Available: https://github.com/CompVis/latent-diffusion

[6] D. Ivezić and M. B. Babac, "Trends and Challenges of Text-to-Image Generation: Sustainability Perspective," *Croatian Regional Development Journal*, vol. 4, no. 1, pp. 56–77, Jun. 2023, doi: 10.2478/crdj-2023-0004.

[7] R. Bayoumi, M. Alfonse, M. Roushdy, and A. B. M. Salem, "Text-to-image generation based on AttnDM-GAN and DMAttn-GAN: applications and challenges," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 1180–1188, Apr. 2023, doi: 10.11591/eei.v12i2.4199.

[8] Z. Bosheah and V. Bilicki, "Challenges in Generating Accurate Text in Images: A Benchmark for Text-to-Image Models on Specialized Content," *Applied Sciences (Switzerland)*, vol. 15, no. 5, Mar. 2025, doi: 10.3390/app15052274.

[9] Tim. Newburn, Tom. Williamson, and Alan. Wright, *Handbook of Criminal Investigation*. Taylor and Francis, 2012.

[10] PH. D. DEVERE D. WOODS JR., "O'HARA'S FUNDAMENTALS OF CRIMINAL INVESTIGATION," 2013.

[11] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, "GAN-generated Faces Detection: A Survey and New Perspectives," Nov. 2023, [Online]. Available: http://arxiv.org/abs/2202.07145

[12] M. Kim, F. Liu, A. Jain, and X. Liu, "DCFace: Synthetic Face Generation with Dual Condition Diffusion Model," 2023.

[13] X. Xing, C. Wang, H. Zhou, J. Zhang, Q. Yu, and D. Xu, "DiffSketcher: Text Guided Vector Sketch Synthesis through Latent Diffusion Models," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2306.14685

[14] K. K. Jain and A. M. Namboodiri, "Face Sketch Generation and Recognition," 2025.

[15] W. Xiang, S. Xu, C. Lv, and S. Wang, "A Customizable Face Generation Method Based on Stable Diffusion Model," *IEEE Access*, vol. 12, pp. 195307–195318, 2024, doi: 10.1109/ACCESS.2024.3520719.

[16] M. Wu, H. Zhu, L. Huang, Y. Zhuang, Y. Lu, and X. Cao, "High-fidelity 3D Face Generation from Natural Language Descriptions." [Online]. Available: https://github.

[17] Z. Huang, K. C. K. Chan, Y. Jiang, and Z. Liu, "Collaborative Diffusion for Multi-Modal Face Generation and Editing," 2023. [Online]. Available: https://github.com/ziqihuangg/Collaborative-Diffusion

[18] F. Gholamrezaie and M. Manthouri, "cycle text2face: cycle text-to-face gan via transformers," Jun. 2022, [Online]. Available: http://arxiv.org/abs/2206.04503

[19] P. Sangkloy, W. Jitkrittum, D. Yang, and J. Hays, "A Sketch Is Worth a Thousand Words: Image Retrieval with Text and Sketch," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2208.03354

[20] P. Yadav, R. Kaur, and U. Hariharan, "A Study on Customized Face generator using Generative Adversarial Networks (GANs)," in *2023 7th International Conference on Electronics, Materials Engineering and Nano-Technology, IEMENTech 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/IEMENTech60402.2023.10423429.

[21] A. Sampada and P. E. S. Kulkarni, "AUTOMATIC SYNTHESIS OF REALISTIC HUMAN IMAGES FROM TEXT USING GANS शोध प्रभा AUTOMATIC SYNTHESIS OF REALISTIC HUMAN IMAGES FROM TEXT USING GANS," *Shodha Prabha (UGC CARE Journal*, vol. 48, p. 2023, 2023, doi: 10.13140/RG.2.2.34479.55207.

[22] S. Ge, T. Park, J.-Y. Zhu, and J.-B. Huang, "Expressive Text-to-Image Generation with Rich Text," Jan. 2025, [Online]. Available: http://arxiv.org/abs/2304.06720

[23] R. Wang, "A Comparative Analysis of StackGAN and AttnGAN in Text-to-Image Generation," *Applied and Computational Engineering*, vol. 105, no. 1, pp. 9–15, Nov. 2024, doi: 10.54254/2755-2721/105/2024tj0055.

[24] D. Meng, C. Tzelepis, I. Patras, and G. Tzimiropoulos, "MM2Latent: Text-to-facial image generation and editing in GANs with multimodal assistance," Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.11010

[25] C. Tan, W. Zhang, Z. Qi, K. Shih, X. Li, and A. Xiang, "Generating Multimodal Images with GAN: Integrating Text, Image, and Style," in *Proceedings of the 2025 2nd International Conference on Computer and Multimedia Technology, ICCMT 2025*, Association for Computing Machinery, Inc, Sep. 2025, pp. 16–21. doi: 10.1145/3757749.3757753.

[26] P. Melzi *et al.*, "GANDiffFace: Controllable Generation of Synthetic Datasets for Face Recognition with Realistic Variations," May 2023, [Online]. Available: http://arxiv.org/abs/2305.19962