10.48047/jocaaa.2024.33.08.342

Semantic Face Generation from Natural Language using GANs

Mrs. G. Lavanya
Assistant Professor
Artificial Intelligence And Data Science
Vignan Institute of Technology And Sciences
Hyderabad,India

M.Kushal
UG Student, Artificial Intelligence And Data Science
Vignan Institute of Technology And Sciences
Hyderabad,India
kushalvirendrachowdary.mandu ri @gmail.com

G. Gayathri
UG Student, Artificial Intelligence And Data Science
Vignan Institute of Technology And Sciences
Hyderabad,India
ggayathri1809@gmail.com

M. Revanth
UG Student, Artificial Intelligence And Data Science
Vignan Institute of Technology And Sciences
Hyderabad,India
mrevanthreddy17@gmail.com

Abstract— Generating semantic faces from natural language text is an emerging area and is largely generating momentum for visually creating a photo-real human face from an input text description. This project focuses on using Generative Adversarial Networks with Natural Language Processing pretrained models, which applies the process of encoding the semantic meaning of the text as the mapping constraints and the resulting image features. Candidate GAN paradigms were StackGAN, AttnGAN, and StyleGAN2 with and without attention-based ideas and multi-modal loss functions to train faces that exhibit high realism and semantic constraints. The system is intended to work with datasets of image-caption pairs for input. The implementation uses a modular three-part pipeline, with plaintext description to a generated face on the first part, generating the face image as the second part, and then the third part takes any user critique/input based on the generated image. Applications and discussion suitable for system designs include creating virtual avatars and generating content like forensic artist sketches and a more inclusive digital identity system. How this method facilitates user experiences in conjunction with system design features allows for controllable face generation based on a text description.

Keywords— Text-to-Face Generation, Generative Adversarial Networks, Deep Learning, CLIP, StyleGAN2.

I. INTRODUCTION

Semantic face generation from natural language is a meaningful effort that demonstrates that text descriptions can be turned into realistic human faces through deep learning methods. This approach serves as a bridge from natural language interpretation to visual creation with broad applications, including digital avatars, entertainment, forensics, and virtual reality. Historically, representing human-like faces or faces based on user input involved manual design or rigid template-based design systems. These systems were impractical due to their time-consuming approaches, limited scalability, and lack of semantic accuracy. They do not effectively support a realistic approach to complex descriptions and face diversity, thereby limiting their use in dynamic real-world applications. However, with the availability of generative adversarial networks (GANs), including StackGAN, AttnGAN, and StyleGAN2, in conjunction with integrated language models like BERT and CLIP, a method for generating high-quality faces with semantic alignment exists. GAN systems can be generalized and scaled meaningfully, do not require design templates to learn fine-grained visual-to-text mappings, and provide interactive refinement to align images to user descriptions in real-time. While not universally applicable, advances in AI for multimodal generation can drive new end-user adoption across identity modeling, personalized content, and human-computer interaction.

Recently, advances in computer science and technology have made it plausible for technology to produce realistic human facial images based on natural language. Human faces possess many attributes—like age, emotional state, hairstyle, and facial expression—which may be semantically interpreted and textually described. Advances in deep learning and generative models have the means to convert descriptive language to visual outputs with reasonable accuracy. Prior attempts at text-to-face generation used very simple templates or small scale models that were not able to represent some of the finer features of human faces and lost the ability to represent various descriptions of real-world positions.

In this work, we will seek to overcome those challenges through using large pre-trained language and vision models, under a GAN framework. In particular, we will combine text encoders (e.g. BERT, CLIP) with GAN architectures (e.g. StackGAN, AttnGAN, StyleGAN2) that allow high-resolution generated outputs with semantic alignment. Our method would be trained on datasets such as CelebA and CelebA-HQ, supplemented by detailed textual descriptions, to allow greater generalization over a larger quantity of facial features and identity.

10.48047/jocaaa.2024.33.08.342

This work represents a major advance in text-to-face synthesis by integrating extremely powerful multimodal embeddings with strong generative architectures. This integrated solution offers dramatic improvement over previous approaches by both increasing image realism and ensuring that the outputs accurately reflect the perceptions of the input text. This concept is revolutionary for personalized avatars, media for virtual media creation, forensic visualization, inclusive digital identity tools, and ultimately, a natural, yet scalable, form of human-computer interaction.

A. Contirbution

In order to further the goal of realistic human facial image synthesis from natural language, this paper provides the following contributions:

We put forth a new GAN-based framework which connects and unifies natural language processing with image synthesis to semantically map textual descriptions into high-fidelity facial images. We leverage state-of-the-art pre-trained models such as BERT and CLIP to extract semantic embeddings, and subsequently use those embeddings to condition state-of-the-art generative models including StackGAN, AttnGAN, and StyleGAN2.

We use two publicly accessible facial datasets to train and evaluate our system: CelebA and CelebA-HQ, which contain many descriptive captions to allow for supervised learning and text-to-image generation. We show substantial improvements over visual quality, semantic alignment, and diversity in generated outputs.

Section II gives a brief overview of published work in the field of text-to-image synthesis and face generation. Section III describes our proposed approach, including architecture and training strategy. Section IV describes the experimental setup, datasets, and evaluation metrics we adopted to evaluate our approach. Section V outlines the experimental findings, as well as the main findings. Lastly, Section VI summarizes this paper, providing perspectives and recommendations for future work.

II. LITERATURE SURVEY

Recently, text-to-image synthesis, more specifically the generation of realistic human faces from natural language descriptions, increased in prominence due to rapid advances in deep learning and in particular Generative Adversarial Networks (GANs). GANs were first presented by Goodfellow et al. [1] and laid the foundation for adversarial training. Since then, conditional models like cGANs [22] which were conditioned on external data like text or attributes, have been developed.

The first research to apply a GAN to the text-to-image synthesis task was done by Reed et al. [9][13]. They showed that while textual input could be used as a prompt to generate images with moderate levels of semantic alignment. StackGAN [12] developed the notion of producing images in a multi-stage fashion; sequentially producing image outputs which became progressively more complete and higher resolution than previous images. Moreover, the StackGAN architecture to StackGAN++ [4] applied similar techniques to build on the performance of the StackGAN. AttnGAN [15] introduced a mechanism for the model to "attend" to different parts of the input text in its image generation so that different areas of an image could align semantically with different parts of the input text while also allowing higher detail of the generated image.

Substantial improvements included the introduction of "MirrorGAN" [16] which included a redescription module to recreate text from the generated image, ensuring the description and output maintain semantic alignment. "Hierarchical-nested adversarial networks" [17] also establish visual realism through multi-scale discriminators and nested learning objectives.

Several works considered additional conditioning methods. For example, "FTGAN" [31] included facial-specific features for the text-to-face case, while "Di and Patel" [27] developed a method for synthesizing faces from textual attributes through intermediate sketches using the VAE and GAN methods. "ChatPainter" [14] showed the merits of dialogue- based input, which reinforces the flexibility and interaction of text-to-image systems.

In early work regarding datasets and evaluations, datasets like CUB [5], Oxford flowers [6], and MS-COCO [7] were utilized with rich image-caption pairings. However, these datasets are not as useful for facial synthesis. Later work, such as Face2Text [18] and MS-Celeb-1M [19], addressed these issues by using text using faces, enabling training specific to semantic face generation tasks.

Modern work also attempts to use powerful pre-trained models. For example, OpenAI's CLIP was trained on millions of text-image pairs and was used in e.g., StyleCLIP [not cited here, but important], for text-based manipulation of StyleGAN-generated faces. Furthermore, VQGAN + CLIP or StyleGAN2 [21] can also generate high-quality facial images, and researchers must also use care to condition on text to achieve semantic control.

In spite of these advancements, there are still limitations to consider. Some models lack sufficient fine-grained control of

10.48047/jocaaa.2024.33.08.342

facial attributes, while other models suffer from the quality of training datasets or bias due to the training datasets. One of the salient challenges has been to consistently maintain identity, realism, and semantic relevance all at once. This work seeks to address these gaps through a modular architecture combining multimodal encoders(e.g., CLIP/BERT) as input to GANs, with improved semantic alignment and visual realism for text-to-face synthesis.

III. METHODOLOGY

A. Preprocessing Data

Our project relied on face image datasets, like CelebA and Face2Text, with associated textual descriptions, to train and evaluate our semantic face generation system. The datasets contain variations in terms of pose and lighting, image quality, and diversity of attributes. In order to provide consistent inputs for our model, we used a preprocessing pipeline.

The first task was to detect and align the facial regions using pre-trained face detectors such as MTCNN, ensuring all images were centered and cropped properly. The facial crops were then resized to a fixed size (e.g., 64×64 or 128×128) based on the model constraints. In addition to normalization, augmentations of the data, such as horizontal flipping, random rotations, and image brightness were added to improve generalization capabilities and robustness to variations due to different facial expressions or orientations.

B. Text Embedding and Conditional Input

To systematically convert text descriptions into semantically useful facial features, we will use pre-trained language encoders. We experimented with two popular ones.

BERT: A transformer-based language model which output contextual embeddings with terms seated in an embedded representation that has deep semantic meaning. We use BERT to encode the descriptive phrases such as "a young man with glasses and curly hair" into fixed-length embeddings.

CLIP: CLIP embeds text and images in the same semantic space where the embedding for a given text is close to its associated embedding for an image. We capitalize on its strong cross-modal alignment in order to improve feature correlation between the input description and the generated face images.

Their contributions, as embeddings, will be the conditional input specification for the generator, that will constrict the output features of the generator via textual constraints.

C. GAN Architecture and Training Process

Our semantic face generation system is developed based on a conditional Generative Adversarial Network (cGAN) framework. The essential components are the following:

Generator: The generator receives the concatenated vector of random noise and text embedding and generates a synthetic face image. In early experiments, we used a lightweight fully-connected architecture and for the more advanced models like StackGAN or AttnGAN, we added CNN layers or additional refinement steps.

Discriminator: The discriminator determines if an image is real or generated and simultaneously determines whether it matches the text condition. It processes the image and the text embedding separately before merging them to produce a matching score.

To enhance performance and stability during training, we implement Global Average Pooling, which reduces the dimensionality of intermediate representations. Additionally, we have added a dropout layer with a 50% rate to limit overfitting, especially in the dense layers. We will apply GELU (Gaussian Error Linear Unit) as our activation function because it provides an good ability to generalize with some noise, as well as a good ability to model non-linear patterns in data.

We select Adam with differential learning rates. A lower rate for pre-trained layers in the fine-tuning phase, and a higher learning rate for the new layers. We accomplish this using MultiOptimizer in TensorFlow Addons, which ensures stable convergence. Further learning rate scheduling and early stopping were also implemented to prevent overfitting and ultimately improve training efficiency.

A. Fine-Tuning and Feature Alignment

In instances where a pretrained backbone such as CLIP or StyleGAN2 is used, we only selectively fine-tune the deeper layers. For example, we train the final dense layers for the specific task of text-to-image fundamentals from scratch, leaving the previous layers to retain knowledge from pretraining on a massive scale. This approach enables the model to quickly composite our face generation task while not forgetting the generalized visual-textual relationships underlying pretraining.

Moreover, utilizing strong text embeddings in combination with a well-featured GAN pipeline ensures there is both pixel realism and semantic fidelity with respect to the input text representation when generating face images.

10.48047/jocaaa.2024.33.08.342

IV. EXPERIMENTS

A .Datasets Used

We used both of the well-known public datasets to evaluate our facial semantic generation approach outlined above:

1. CelebA Dataset

The CelebA dataset has 200,000+ celebrity images of faces annotated with 40 possible facial attributes (e.g., smiling, glasses, beard) and identification labels. For this project we additionally expanded the set of labels available with natural language descriptions using attribute-to-text conversion and human annotators. This dataset provides an extensive and representative set of face images which allow us to provide training on visual fidelity and semantic matching.

Fig. 1. CelebA Dataset

2. Face2Text Dataset

The Face2Text dataset contains images of the human face along with rich and free-form textual descriptions. The textual annotations contain many valuable attributes such as facial expressions, age, hair color, gender, and accessories. The relationship between the images and the text describes the semantic relevance in text-to-image generation, allowing us to test the extent to which generated images accurately depict their corresponding text-based inputs.

Fig. 2. Face2Text Dataset

The datasets provide a wide distribution of facial features and semantic content useful to train and validate the proposed text-to-face GAN framework comprehensively.

A. Experimental Setup Used

We ran the experiments within the Kaggle platform, using the cloud computing capabilities of the Kaggle platform. We utilized a GPU (Graphics Processing Unit) enabled jupyter notebook environment with the following specifications:

RAM: 16 GB

Disk: 100 GB

GPU: NVIDIA T4 (dual accelerators)

To stave off overfitting, we utilized Early Stopping, which cancels the training process if validation loss doesn't improve in 5 consecutive epochs. We also used data augmentation (flipping, rotation, etc.) and dropout layers to regularize training and improve generalization.

All models implemented used either TensorFlow or PyTorch, depending upon the architecture implementation and compatibility with pretrained modules (e.g., CLIP).

B. Evaluation Metrics Used

In the event that a set of face images were generated, we used both quantitative and qualitative evaluation metrics to evaluate the quality of the images generated:

Inception Score (IS):

A common evaluation metric when assessing the quality of generated images to measure realism and diversity using a pre-trained classifier. Higher IS scores indicates better quality.

Frechet Inception Distance (FID):

A frequently used metric to measure the distance between features of distributions of real images and generated images. Lower FID indicates more realism and images of high fidelity.

CLIP Similarity Score:

Assesses the degree to which the generated image is semantically aligned to the text input. The metric determines the cosine similarity between CLIP's text and image embeddings. This is a metric that reflects how well the image captures the textual description directly.

These metrics provide a comprehensive evaluation framework, balancing both visual realism and semantic consistency, which are crucial in text-to-face synthesis tasks.

10.48047/jocaaa.2024.33.08.342

V. RESULT

This section presents the evaluation process of different models for semantic face generation from natural language descriptions, specifically focusing on quantitative and qualitative evaluation. Use standard metrics such as FID (Frechet Inception Distance), which describes the realism of the generated images, and the CLIP Similarity Score, which measures the semantic similarity between generated and input descriptions.

A. Quantitative Evaluation

We evaluated the effectiveness of three of the main architectures—StackGAN, AttnGAN, and StyleGAN2-CLIP—alongside the CelebA and Face2Text datasets. The main metrics of interest were:

FID Score: A lower FID indicates better visual fidelity and similarity of distributions to real face images.

CLIP Similarity: A higher score indicated better semantic similarity between the generated image and input text description.

TABLE- 1
EVALUATION RESULTS ON CELEBA DATASET

Model	FID Score	CLIP Similarity
StackGAN	54.2	0.61
AttnGAN	46.8	0.67
StyleGAN2-CLIP	32.4	0.74

Table 1 shows that StyleGAN2-CLIP significantly outperforms other models in both visual quality and semantic alignment, achieving the lowest FID and highest CLIP similarity on the CelebA dataset.

TABLE 2
EVALUATION RESULTS ON FACE2TEXT DATASET

Model	FID Score	CLIP Similarity
StackGAN	57.9	0.58
AttnGAN	48.6	0.64
StyleGAN2-CLIP	35.7	0.71

In Table 2, we observe a similar trend on the Face2Text dataset. The StyleGAN2-CLIP model consistently achieves superior results across all evaluation metrics.

B. Qualitative Observations

The sample images tell us that StackGAN produces very low-resolution faces that do have some details, while AttnGAN produces faces that have a little bit better attribute realism and better semantic alignment. However, StyleGAN2-CLIP produces the most realistic face and semantically aligned face that can capture detailed hair color, age, expression, and even the accessibility.

C. Model Insights

StackGAN performs poorly with complex textual input and struggles with blurring of the image at higher resolution.

AttnGAN improves on semantic correspondence by utilizing attention modules to improve consistency, but the output was still inconsistent with strange artifacts

StyleGAN2-CLIP works well due to its joint embedding space, and its capability to capture new subtle semantic cues from the text descriptions.

VI. CONCLUSION

Our study addressed the problem of generating human face images that are both semantically and photorealistically accurate from natural language descriptions. Understanding the need for intelligent multimodal generation systems, we developed a deep learning-based approach that combined natural language understanding through the semantic approach and generative image synthesis through gan-style models.

Both quantitative (FID, CLIP similarity) and qualitative evaluations confirmed that StyleGAN2-CLIP consistently results in the best overall performance in terms of semantic fidelity and image quality. Our study has made a meaningful contribution to demonstrating the compelling link between descriptive language and facial representation, and how modern GAN frameworks can leverage descriptive language correspondence to achieve an accurate synthesis.

For future work, we will expand the diversity and richness of training datasets to implement multilingual descriptions and different demographic attributes (i.e., ethnicity, age, emotion). We will also expand our study of training models on decentralized devices and using Federated Learning as a privacy-preservation mechanism when training on user-generated or sensitive data.

We think our method has promising applications in personalized avatars, virtual assistants, digital content creation, and law enforcement. It may also enable more organic human-computer interactions, allowing computers to generate human-like faces from the way we describe each other—using words.

10.48047/jocaaa.2024.33.08.342

VII. REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ``Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014,pp. 2672_2680.

[2] S. Hong, D. Yang, J. Choi, and H. Lee, ``Inferring semantic layout for hierarchical text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput Vis. Pattern Recognit.*, Jun. 2018, pp. 7986_7994.

[3] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, and A. Graves,``Conditional image generation with pixelcnn decoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4790_4798.

[4] H. Zhang, T. Xu, H. Li, S. Zhang, X.Wang, X. Huang, and D. N. Metaxas,``StackGANCC: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947_1962, Aug. 2019.

[5] C.Wah, S. Branson, P.Welinder, P. Perona, and S. Belongie, ``The caltechucsd birds-200-2011 dataset," California Inst. Technol., Pasadena, CA,USA, Tech. Rep., 2011.

[6] M.-E. Nilsback and A. Zisserman, ``Automated _ower classi_cation over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722_729.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dolláar, and C. L. Zitnick, ``Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis*. Cham, Switzerland: Springer, 2014, pp. 740_755.

[8] D. P. Kingma and M. Welling, ``Auto-encoding variational Bayes,"2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114

[9] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee,``Generative adversarial text to image synthesis," 2016, *arXiv:1605.05396*. [Online]. Available: http://arxiv.org/abs/1605.05396

[10] A. Radford, L. Metz, and S. Chintala, ``Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[11] H. Dong, S. Yu, C.Wu, and Y. Guo, ``Semantic image synthesis via adversarial learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5706_5714.

[12] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, ``StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907_5915.

[13] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, ``Learning what and where to draw," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 217_225.

[14] S. Sharma, D. Suhubdy, V. Michalski, S. Ebrahimi Kahou, and Y. Bengio,``ChatPainter: Improving text to image generation using dialogue," 2018, *arXiv:1802.08216*. [Online]. Available: http://arxiv.org/abs/1802.08216

[15] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, ``AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316_1324.

[16] T. Qiao, J. Zhang, D. Xu, and D. Tao, ``MirrorGAN: Learning text-toimage generation by redescription," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1505_1514.

[17] Z. Zhang, Y. Xie, and L. Yang, ``Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6199_6208.

[18] A. Gatt, M. Tanti, A. Muscat, P. Paggio, R. A. Farrugia, C. Borg, K. P. Camilleri, M. Rosner,

and L. van der Plas, ``Face2Text: Collecting an annotated image description corpus for the generation of rich face descriptions," 2018, *arXiv:1803.03827*. [Online]. Available: http://arxiv.org/abs/1803.03827

[19] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, ``Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 87_102.

[20] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, ``Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," Tech. Rep., 2008.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang, ``Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730_3738.

[22] M. Mirza and S. Osindero, ``Conditional

10.48047/jocaaa.2024.33.08.342

generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/ 1411.1784

[23] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, ``High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798_8807.

[24] [24] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, ``Towards open-set identity preserving face synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6713_6722.

[25] R. Huang, S. Zhang, T. Li, and R. He, ``Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2439_2448.