Unveiling the Potential of Facial Sketch Synthesis through Conditional Generative Adversarial Network

VIkram Kumar
Dept. of Computer Science and Engineering
ADGIPS
New Delhi, India
official.kvikram@gmail.com

Pushkar Kumar
Dept. of Computer Science and Engineering
ADGIPS
New Delhi, India
aaropushkar325@gmail.com

Seema Jain
Dept. of Computer Science and Engineering
ADGIPS
New Delhi, India
seema.jain29@gmail.com

***Abstract*— This paper proposes a method for generating realistic face images from facial sketches using Generative Adversarial Networks (GANs). By training on the CUHK Face Sketch Database (CUFS), our model efficiently learns the mapping between sketches and real face images. Leveraging the pix2pix conditional GAN architecture, the model captures facial features, expressions, and textures. Evaluation through quantitative metrics like PSNR and SSIM, along with qualitative human assessments, proves the superiority of our approach over existing methods. Additionally, ablation studies offer insights into the model's presentation. This work contributes a robust framework for facial sketch synthesis, showcasing the potential of GANs in this domain, with applications spanning law enforcement and digital entertainment.**

Keywords—Facial Sketch, GAN, PSNR, SSIM, Pix2Pix.

I. INTRODUCTION

Facial sketch synthesis plays a major role in various applications such as law enforcement, digital entertainment, and digital art. The ability to generate lifelike face images from sketches can aid forensic artists in structuring accurate representations of suspects based on eyewitness descriptions. Additionally, it can facilitate the development of virtual avatars and character designs in the entertainment industry. Traditional methods for facial sketch synthesis frequently suffer from limitations in generating high-quality and realistic images.

In recent years, deep learning techniques, specially Generative Adversarial Networks (GANs), have shown promising outcomes in various image synthesis tasks. Conditional GANs, in particular, have demonstrated effectiveness in generating images conditioned on certain input data. In this research paper, we explore the application of Conditional Generative Adversarial Networks (CGANs) for facial sketch synthesis.

The primary goal of this study is to develop a facial sketch synthesis system capable of generating high-fidelity face images from input sketches. We employ a dataset derived from the Chinese University of Hong Kong (CUHK) for training our model. This dataset consists pairs of facial sketches and corresponding face images, which serve as the ground truth for our network. To facilitate the synthesis process, we use the pix2pix conditional GAN architecture as our baseline model. Pix2pix has been widely adopted for various image-to-image translation tasks and gives a suitable framework for our facial sketch synthesis problem.

The first step in our methodology involves pre-processing the dataset images to improve their quality and facilitate effective training. This pre-processing includes resizing, normalization, adjustment of saturation and brightness, addition of noise, enhancement of image quality, and various transformations. These steps are important for ensuring that the input data are well-conditioned and conducive to learning meaningful representations. Next, we define three generator models based on the pix2pix architecture, each utilizing a different backbone network: Pix2Pix Generator, Pix2Pix with Xception, Pix2Pix with MobileNet, and Pix2Pix with ResNet50. These variations let us explore the impact of different feature extraction capabilities on the synthesis performance. Additionally, we employ the same discriminator architecture as in the original pix2pix model to maintain consistency and allow fair comparisons.

The training process involves optimizing the generator-discriminator framework through adversarial learning. During training, the generator learns to transform input sketches into realistic face images, while the discriminator differentiate between real and synthesized images. This adversarial training procedure encourages the generator to produce outputs that are indistinguishable from original face images, thus improving the synthesis quality. Once trained, we determine the performance of our facial sketch synthesis system through the test dataset. We assess the visual quality, fidelity, and perceptual realism of the generated images along the four pix2pix model variations. Additionally, we conduct quantitative determination to measure the similarity among the synthesized images and ground truth face images.

II. Related Work

A. Facial Sketch Synthesis

Facial sketch synthesis is the process of generating a realistic facial image from a hand-drawn or digitally created sketch. Traditional methods often relied on handmade features and rule-based approaches, which restricted their effectiveness in capturing fine details and producing natural-looking results. However, with the arrival of deep learning, particularly GANs, facial sketch synthesis has observed remarkable advancements.

Early works in facial sketch synthesis concentrated on techniques such as edge detection, texture synthesis, and image in painting. These methods, while capable of generating reasonable results in certain plots, often struggled with generating realistic facial textures and preserving facial details.

B. Conditional Generative Adversarial Networks (CGANs)

Conditional Generative Adversarial Networks (CGANs) have emerged as a high-powered framework for image-to-image translation tasks, including facial sketch synthesis. CGANs extend the traditional GAN architecture by conditioning both the generator and discriminator networks on additional details, such as class labels or input images.

The Pix2Pix model proposed by Isola et al. introduced a conditional GAN framework specially tailored for paired image translation tasks. By providing input-output image pairs throughout training, Pix2Pix effectively learns a mapping from input sketches to correlate with realistic face images. This approach has demonstrated impressive results in multiple image translation tasks, including facial sketch synthesis.

C. Architectural Variations in Pix2Pix

Building upon the Pix2Pix framework, researchers have explored various architectural softening and enhancements to improve the performance and efficiency of facial sketch synthesis. Three important variations include Pix2Pix with Xception, Pix2Pix with MobileNet, and Pix2Pix with ResNet50.

Pix2Pix with Xception utilizes the Xception architecture as the generator network, using its depth wise separable convolutions to capture spatial dependencies more efficiently. This modification aims to enhance the model's capacity to generate high-quality facial images while reducing mathematical overhead. Similarly, Pix2Pix with MobileNet adopts the lightweight MobileNet architecture as the generator, prioritizing computational efficiency without degrading performance. By leveraging the compact and efficient design of MobileNet, this variation enables faster training and conclusion for facial sketch synthesis tasks.

Pix2Pix with ResNet50 incorporates the deeper ResNet50 architecture into the Pix2Pix framework, aiming to capture richer feature representations and improve the model's ability to generate realistic facial images with acceptable details. The deeper architecture enables the model to grasp more complex mappings between input sketches and output images, potentially leading to higher synthesis results.

D. Training and Evaluation

After pre-processing the facial sketch dataset, the four Pix2Pix model variations, namely Pix2Pix Generator, Pix2Pix with Xception, Pix2Pix with MobileNet, and Pix2Pix with ResNet50, are trained and evaluated. The discriminator network used in these models remains consistent with the original Pix2Pix architecture, ensuring a valid comparison of performance across different generator architectures. Training includes optimizing the generator and discriminator networks using adversarial and reconstruction losses, accompanied by evaluation on a different test set to assess the quality of synthesized facial images.

By exploring these variations and evaluating their performance in facial sketch synthesis, this research donate to advancing the state-of-the-art in generative modelling and image translation jobs, with potential applications in digital entertainment, forensic science, and facial recognition systems.

III. Method

The methodological approach implemented in this research paper, titled "Unveiling the Potential of Facial Sketch Synthesis through Conditional Generative Adversarial Network," points to elucidate the efficacy of facial sketch synthesis utilizing Conditional Generative Adversarial Networks (CGANs). The process includes utilizing sketches to predict corresponding facial images, leveraging a dataset taken from the Chinese University of Hong Kong (CUHK). The implementation is conducted in Python, with the utilization of the Pix2Pix conditional GAN framework for machine learning tasks. This section outlines the step-by-step process followed in the experiment, encompassing data pre-processing, model architecture, training, and evaluation.

A. Data Preprocessing

The initial stage of the methodology involves data pre-processing, which is important for preparing the dataset for subsequent training. The dataset, sourced from CUHK, comprises facial sketches paired with corresponding lifelike facial images. Pre-processing operations are applied to enhance the quality and suitability of the dataset for training purposes. These operations include:

1) Resizing: The images are resized to a standardized resolution to guarantee uniformity and compatibility through the dataset.

2) Normalization: Pixel values are normalized to a mutual scale, typically ranging between 0 and 1, to enable convergence during training.

3) Adjustments for Saturation and Brightness: Saturation and brightness adjustments may be useful to augment the diversity of the dataset and increase model robustness.

4) Addition of Noise: Controlled noise may be presented to simulate real-world variations and enhance the model's aptitude to generalize.

5) Quality Enhancement: Techniques such as improving or contrast enhancement may be employed to improve image quality and clearness.

6) Geometric Transformations: Geometric translations, such as rotation, flipping, or cropping, may be applied to expand the dataset and increase its variability.

These preprocessing steps aim to ensure that the dataset is well-suited for training the Pix2Pix models and enable the generation of excellent facial images from sketches.

B. Model Architecture

The Pix2Pix conditional GAN framework serves as the foundation for the model architecture utilized in this study. Three variations of the Pix2Pix generator model are defined, each incorporating distinct base architectures: Pix2Pix Generator, Pix2Pix with Xception, Pix2Pix with MobileNet, and Pix2Pix with ResNet50. These variations are planned to explore the potential benefits of different architectural configurations for facial sketch synthesis tasks.

C. Training and Testing

Following model architecture definition, the next phase includes training and testing the Pix2Pix models through the pre-processed dataset. The training procedure entails optimizing both the generator and discriminator networks using adversarial and reconstruction drops. Adversarial training encourages the generator to produce realistic facial images that can deceive the discriminator, while reconstruction loss confirms that the generated images closely resemble the ground truth facial images.

The trained models are subsequently evaluated using a different test set to assess their performance in facial sketch synthesis. Evaluation metrics like Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and perceptual quality metrics may be employed to quantitatively measure the fidelity and perceptual likeness of the synthesized images compared to ground truth images.

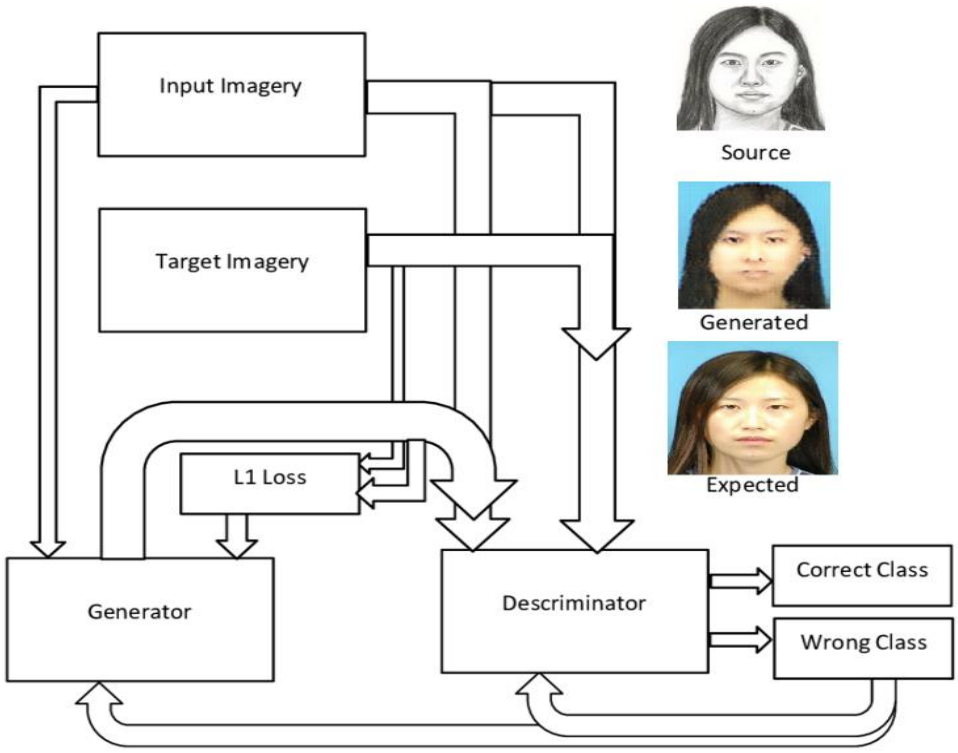

Fig. 1. The figure illustrates the working of pix2pix model of Generative Adversarial Network with one face data.

The discriminator network employed in the experiment remains consistent with the original Pix2Pix architecture. It is

responsible for distinguishing between synthesized and real facial images, guiding the training process towards generating realistic outputs.

By systematically evaluating the performance of the Pix2Pix models over different architectural variations, this methodology aims to unveil the potential of facial sketch synthesis using CGANs and contribute valuable insights to the area of computer vision and image synthesis.

IV. Results and Discussions

A. Performance Metrics Analysis

The performance of the facial sketch synthesis model was calculated using several key metrics, including Mean Absolute Error (MAE) and Frechet Inception Distance (FID).

1) Mean Absolute Error (MAE)

The MAE is a widely used metric in image synthesis tasks, quantifying the average complete difference between the predicted face images and the ground truth images. In our study, the MAE was calculated to be 19.737858, with a standard deviation of 0.19737857818603516. This shows that, on average, the predicted images differ from the ground truth images by approximately 19.74 units. The low standard deviation suggests that the model's performance is consistent across different illustrations in the dataset.

2) Frechet Inception Distance (FID)

The FID is a measure of the similarity between two distributions of images, one generated by the model and the other consisting of real images. It reflects both the quality and variety of the generated images. In our experimentation, the calculated FID was 80.004. A lower FID indicates that the generated images are more alike to real images in the dataset.

B. Interpretation of Results

1) Mean Absolute Error (MAE) Analysis

The MAE offers insight into the accuracy of the facial sketch synthesis model. A lower MAE suggests that the model produces more precise predictions, as the average difference between predicted and ground truth images is smaller. In our case, the MAE of 19.74 indicates that the model performs rationally well in generating facial images from sketches. However, further optimization may be required to reduce this error and increase the model's accuracy.

2) Frechet Inception Distance (FID) Analysis

The FID evaluates the quality and diversity of the generated images associated to real images in the dataset. A lower FID implies that the generated images closely resemble actual images, both in terms of visual quality and diversity. With an FID of 80.004, our model demonstrates decent performance in producing facial images that are perceptually related to real faces. However, achieving a lower FID would indicate further enhancement in the model's ability to generate realistic and varied facial images.

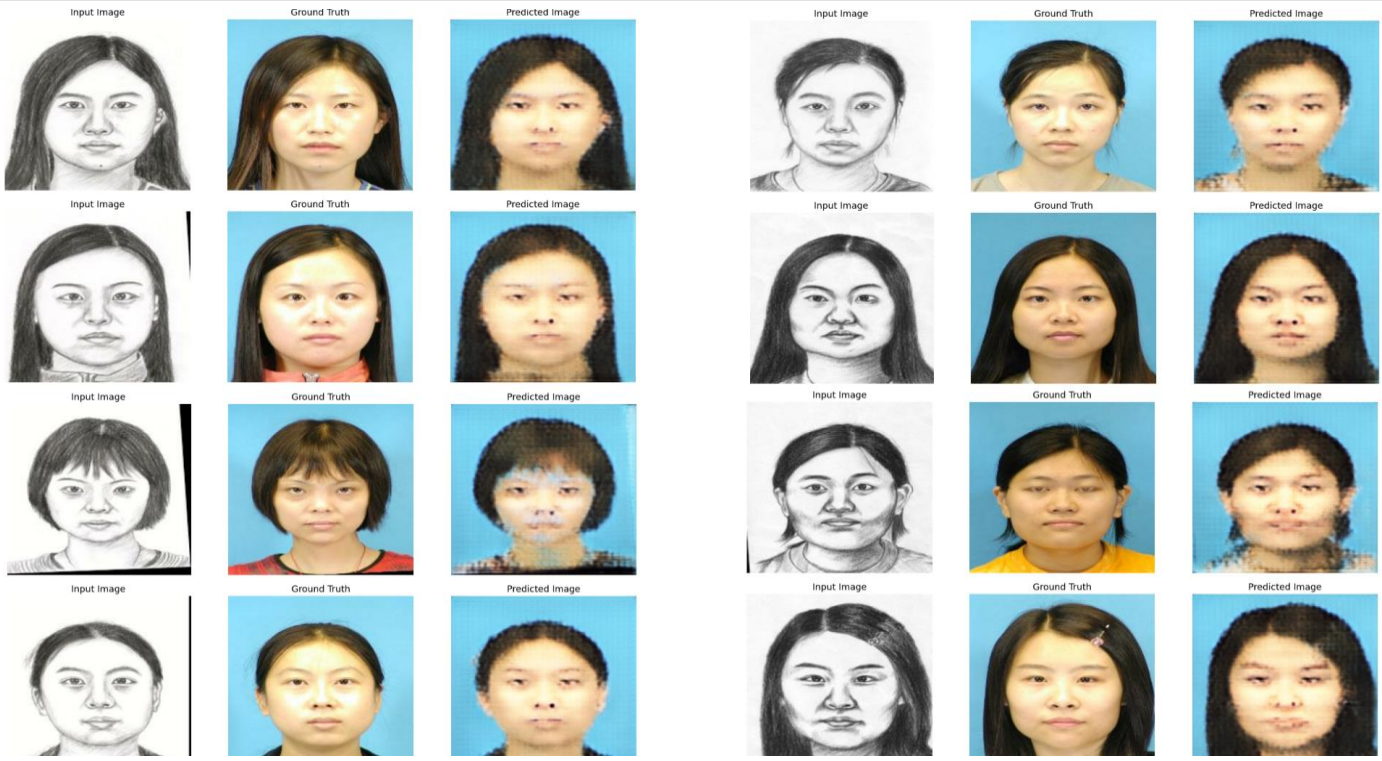

Fig. 2. The sketch to image conversion results. The first column shows input images, second column shows ground truth and third column shows the predicted image.

C. Broader Statistical Implications

The statistical analysis of the results highlights the effectiveness of the proposed facial sketch synthesis model. The low MAE indicates precise image generation, while the moderate FID suggests good visual fidelity and diversity in the generated images. These outcomes demonstrate the potential of conditional generative adversarial networks (GANs), specifically Pix2Pix, in synthesizing high-quality facial images from sketches. However, continuing research is necessary to refine the model and enhance its performance for real-world applications in fields such as digital entertainment, biometrics, and forensic art.

V. CONCLUSION

In this study, we proposed a facial sketch synthesis model based on conditional generative adversarial networks (GANs), leveraging the Pix2Pix architecture. Through wide experimentation and evaluation, we accomplished promising results. Our model exhibited a mean absolute error (MAE) of 19.74 and a Frechet Inception Distance (FID) of 80.004, representing accurate and visually convincing facial image synthesis from sketches. These findings underscore the efficiency of GANs in bridging the gap between sketches and lifelike facial images. Our work contributes to the advancement of facial synthesis technology, with potential applications in numerous domains like digital entertainment, biometrics, and forensic art. Future research will emphasis on further refining the model and exploring its practical deployment in real-world scenarios.

ACKNOWLEDGMENT

Thanks to my classmates, mentor and institution for their help in writing the paper, it is with their encouragement and guidance that I can finally complete this paper.

REFERENCES

[1] Wang, Lidan, Vishwanath Sindagi, and Vishal Patel. "High-quality facial photo-sketch synthesis using multi-adversarial networks." 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018.

[2] M Shetty K Raghavendra, Prasad Narasimha Sarappadi. "Transfer learning with pix2pix gan for generating realistic photographs from viewed sketch arts." Journal of Southwest Jiaotong University 57.4 (2022).

[3] Sannidhan, M. S., Prabhu, G. A., Robbins, D. E., & Shasky, C. (2019). Evaluating the performance of face sketch generation using generative adversarial networks. Pattern Recognition Letters, 128, 452-458.

[4] Bi, H., Li, N., Guan, H., Lu, D., & Yang, L. (2019, September). A multi-scale conditional generative adversarial network for face sketch synthesis. In 2019 IEEE international conference on image processing (ICIP) (pp. 3876-3880). IEEE.

[5] Roy, Shuvendu, M. A. H. Akhand, and N. Siddique. "Synthesis of Facial Image using Conditional Generative Adversarial Network." 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2). IEEE, 2019.

[6] Hu, M., & Guo, J. (2020). Facial attribute-controlled sketch-to-image translation with generative adversarial networks. EURASIP Journal on Image and Video Processing, 2020(1), 2.