Sketch to Image Generation Using GAN

Muneeb Ahmed[1], Jamal Khan[2]
Muneebahmadkhattak@gmail.com, malakjamal2013@gmail.com
Department of Electrical Engineering, COMSATS University Islamabad, Pakistan

***Abstract*—The conversion of hand-drawn sketches into photo-realistic images has become an exciting area of research in computer vision, enabled by advancements in deep learning and generative models. This paper presents a system based on the Pix2Pix model, a type of conditional Generative Adversarial Network (cGAN), designed to learn the mapping from sketches to real images. By training on paired datasets of sketches and corresponding photographs, the model is capable of generating detailed, colorized images from sparse input drawings. The generator, built on a U-Net architecture, captures both global structure and fine details, while the PatchGAN discriminator evaluates the realism of small image patches to ensure texture fidelity. The system was trained on facial sketch datasets, and evaluated using metrics such as SSIM, PSNR, and FID. Experimental results demonstrate that the Pix2Pix model effectively reconstructs realistic images even from incomplete or rough sketches. This research highlights the potential of GAN-based models in creative design, education, and digital media applications.**

I. Introduction

Sketch-to-image generation using Generative Adversarial Networks (GANs) has emerged as a powerful deep learning technique that converts simple sketches into photo-realistic images. This transformation involves learning a mapping between input sketches and corresponding detailed images using paired datasets. The GAN framework comprises two networks: a generator that creates images and a discriminator that evaluates their authenticity. Trained adversarially, the system improves progressively, producing highly realistic images from sparse visual inputs.

This technology has significant implications for fields such as digital art, virtual reality, product design, and medical visualization. Among the various GAN variants, the Pix2Pix model stands out due to its use of conditional GANs (cGANs), a U-Net-based generator, and a PatchGAN discriminator. Unlike traditional GANs that use random noise and unpaired data, Pix2Pix relies on structured inputs and paired datasets, yielding more consistent and semantically accurate results. This study focuses on developing a robust Pix2Pix-based system that effectively generalizes across various sketch styles, manages imperfect inputs, and is accessible to non-expert users, thereby bridging the gap between abstract creativity and realistic image synthesis.

Use the enter key to start a new paragraph. The appropriate spacing and indent are automatically applied.

A. Generative Adversarial Network:

GANs consist of two primary components: the Generator and the Discriminator. The Generator (Red block) is responsible for producing synthetic images from input sketches, while the Discriminator (Blue block) evaluates these images to determine whether they are real or generated. These two networks are trained simultaneously in a zero-sum game, where the Generator (Red block) aims to create increasingly realistic images, and the Discriminator (Blue block) strives to improve its ability to differentiate between real and synthetic data.

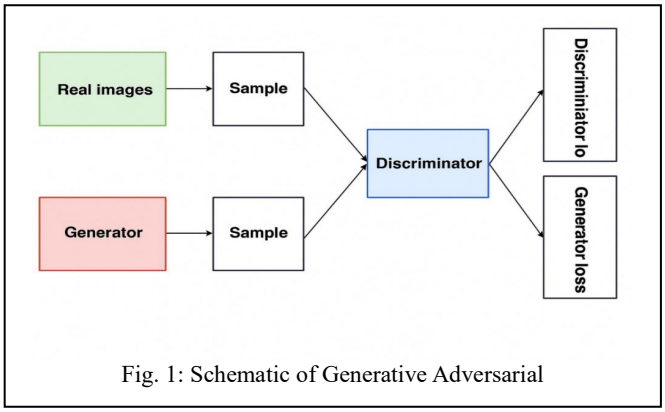

Fig. 1: Schematic of Generative Adversarial

B. Pix2Pix Model

Pix2Pix is a type of cGAN that is designed for image-to-image translation tasks. It employs a U-Net-based Generator and a Patch GAN Discriminator. The model takes an input image and generates a corresponding transformed image as output, enabling tasks such as sketch-to-image conversion, image colorization, and more.

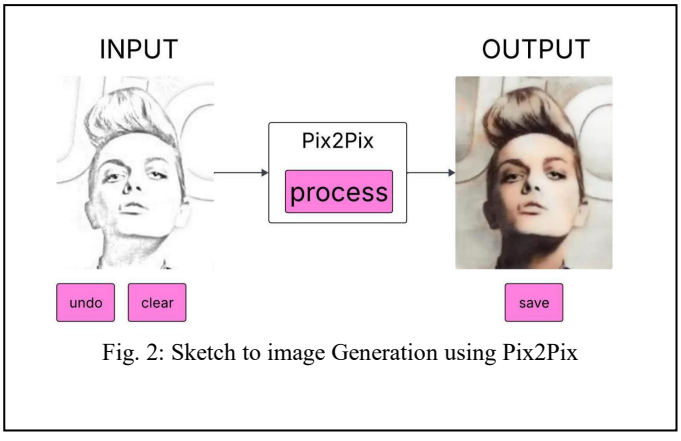

Fig. 2: Sketch to image Generation using Pix2Pix

II. LITERATURE REVIEW

A. Deep Learning Foundations in Image Synthesis:

The intersection of deep learning and computer vision has revolutionized how machines interpret and generate visual data. One particularly exciting application is sketch-to-image generation, where hand-drawn sketches are transformed into realistic images. As discussed in [1], this task is crucial for domains like creative design, forensics, illustration, and human-computer interaction. These advancements have largely been fueled by the evolution of deep neural architectures, particularly Generative Adversarial Networks (GANs).

The rise of deep learning in image synthesis can be traced back to improvements in convolutional neural networks (CNNs) for recognition tasks [2]. However, creating images—not merely classifying them depends on novel methods. A major milestone was the introduction of GANs by Goodfellow et al. [3], where two neural networks (generator and discriminator) compete in a minimax game. This adversarial training allows the generator to progressively produce more realistic outputs. Early GANs generated images from random noise in [4], but advancements like conditional GANs (cGANs) allowed models to generate images based on specific input data, including semantic labels or sketches.

B. Sketch-to-Image Generation:

Foundational Work: Sketch-to-image translation, a subset of image-to-image translation, is uniquely challenging due to the abstract and sparse nature of sketches. Initial approaches [5] employed basic GANs to map edge-like representations to photos. Although promising, these models struggled with maintaining detail and producing diverse outputs.

A breakthrough came with Pix2Pix, a conditional GAN architecture proposed by Isola et al. [6], which addressed the general image-to-image translation problem. While not initially designed for sketches, Pix2Pix provided an effective framework for tasks requiring paired training data. In [7], researchers introduced a context-aware generation model that utilized deep inference and semantic features to deal with incomplete sketches—highlighting the importance of handling real-world, ambiguous drawings. **Architectural Advancements** As research matured, limitations of standard GANs in handling noisy or abstract sketches led to the development of enhanced architectures and techniques. **Neural Style Transfer and Feature Learning** to better match sketches with photos, style transfer methods like those in [8] used multi-scale feature pyramids, enabling both global and local feature extraction. Meanwhile, [9] applied CNNs and transfer learning to align synthetic sketches with real photos in facial recognition tasks, leveraging pretrained models and data augmentation. **Unsupervised and Semi-Supervised Approaches** Due to the difficulty in acquiring paired sketch-photo datasets, unsupervised and semi-supervised methods gained popularity. Kazemi et al. [10] proposed an approach that combined autoencoders with adversarial training to learn facial geometry without paired samples. Other works introduced sketch-photo translation models trained solely on unpaired data, improving flexibility in resource-constrained environments. **Vaes and Probabilistic Models** in [11], Conditional Variational Autoencoders (CVAEs) were combined with GANs to enable diverse and controllable image synthesis. These probabilistic models sampled from latent spaces to improve generalization and diversity. further extended this with gender-preserving facial generation models focused on attribute retention—relevant to forensic and privacy-sensitive domains. **Transformers and Attention Mechanisms** Recent advances have introduced transformer-based architectures to sketch-to-image generation. These models leverage self-attention to model long-range dependencies, improving global consistency. Hybrid approaches, combining CNN backbones with transformers, have shown promise in improving both diversity and image quality. **Datasets for Sketch-to-Image Generation** High-quality datasets are essential for training robust sketch-to-image models. Notable examples include:

CUFS and CUFSF [7]: Paired photo-sketch datasets focused on facial synthesis.

Sketchy Dataset [12]: Contains over 75,000 hand-drawn sketches across 125 object categories.

Edge2Photo [13]: Used in Pix2Pix and other frameworks, includes edge maps paired with real-world photos.

Due to scarcity of paired data, augmentation techniques such as geometric distortion, sketch simulation, and noise injection are widely adopted to enhance generalization.

C. Challenges in Existing Literature

Despite progress, several core challenges persist in sketch-to-image synthesis.

1) Limited Data and Annotations:

Creating large, high-quality paired datasets is time intensive. While semi-supervised models offer partial relief, lack of diversity still hampers performance on complex translation tasks.

2) Image Quality and Realism:

Early models produced images with artifacts and lacked realism. Techniques like perceptual loss and multi-scale training have improved fidelity, but sparse sketch inputs remain difficult to reconstruct accurately.

3) Training Stability:

Training GANs remains inherently unstable [3]. Issues like mode collapse, gradient vanishing, and sensitivity to hyperparameters demand more reliable and adaptive training strategies.

4) Computational Requirements

Training deep generative models is computationally expensive. This limits experimentation, especially in regions with restricted access to high-end GPUs or cloud computing resources.

III. METHODOLOGY

The diagram outlines a typical pipeline for a deep learning project using GAN-based architecture, likely for image-to-image translation or segmentation tasks. It starts with a **statement of problem**, followed by acquiring a suitable **dataset** (e.g., from Kaggle). The **model architecture** is defined, comprising a **Generator** (U-Net) and a **Discriminator**, forming a conditional GAN structure. The generator creates outputs that are evaluated by the discriminator to distinguish between real and generated data. These outputs feed into the **loss functions**, which guide model learning through adversarial and reconstruction losses. The model is then **trained** using these losses, and once optimized, it proceeds to the **inference** stage where it is used to make predictions on new data. This workflow ensures that the generator learns to produce realistic and accurate outputs that satisfy the given problem.

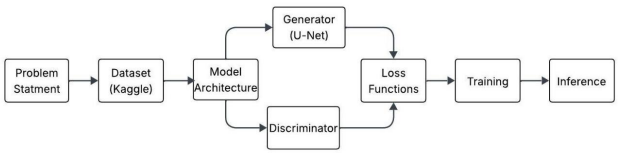

Fig. 3: Project Implementation of Sketch to Image Generation using GAN

A. Dataset Preparation

For our sketch-to-image generation task using the Pix2Pix model, we did not manually prepare a custom dataset; instead, we utilized a pre-existing paired dataset available on Kaggle. This dataset contains aligned pairs of sketches and their corresponding real images, which is essential for training the Pix2Pix model, as it relies on supervised learning with image-to-image translation. By using a readily available dataset, we ensured that the input-output pairs were properly structured and formatted, allowing us to focus on model training and evaluation without the additional overhead of data collection and preprocessing.

Fig. 4: Sketches in Dataset

B. Model Architecture

The diagram illustrates the architecture of a sketch-to-image generation model based on an encoder-decoder framework combined with a discriminator. The process starts with an input sketch, which is passed through the encoder to extract high-level features and compress them into an embedding image. This embedding is then fed into a decoder, which reconstructs a colored face image known as the transferred style map. This generated image is then compared with the target styled map (the real-colored image) by the discriminator, which determines whether the generated image is fake or real. This structure reflects the typical working of GANs, particularly suitable for tasks like image-to-image translation (e.g., Pix2Pix).

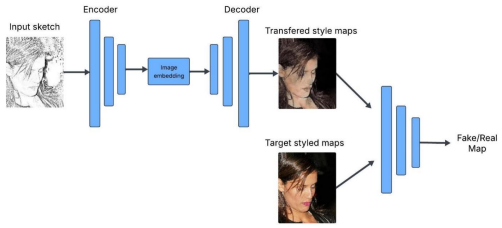

Fig. 6: Sketches in Dataset

1) Loss Functions

Multiple loss functions are employed to balance fidelity, realism, and stability:

- **Adversarial Loss:** Encourages the generator to produce images indistinguishable from real data.
- **L1 Loss (Reconstruction Loss):** Penalizes pixel-wise differences between generated and ground truth images, promoting structural similarity.
- **Perceptual Loss:** Utilizes pretrained feature extractors (e.g., VGG-19) to enforce similarity at higher semantic levels.
- **Cycle-Consistency Loss (if using unpaired data):** Ensures that translating from sketch to image and back reconstructs the original sketch.

The total loss is a weighted combination of these terms, with hyperparameters tuned empirically.

1. **X-axis**: Number of **epochs** (training iterations), from 0 to 40.
2. **Y-axis**: **Loss** value: a measure of how well the model is performing. Lower loss = better performance.
3. **Blue line with circles**: **Training loss**
4. **Orange line with crosses**: **Validation loss**

In the **initial phase** (epochs 0–5), both the training and validation losses decrease sharply, indicating that the model

Fig. 5:Real images in Dataset for Figure 4

is rapidly learning and adjusting its parameters effectively. During the **middle phase** (epochs 5–20), the rate of loss reduction slows down, and the gap between training and validation loss begins to stabilize, suggesting consistent performance across both datasets. In the **later phase** (epochs 20–40), the training loss continues to decline gradually, reaching approximately 0.0061 by the final epoch. At the same time, the validation loss plateaus around 0.01, which implies that the model maintains good generalization and is not over-fitting the training data.

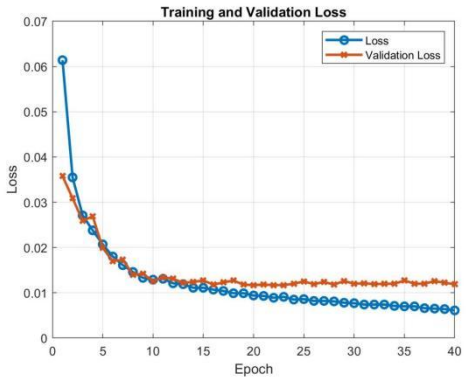

Fig. 7: Plotting of loss and val_loss w.r.t Epoch

2) FlowChart

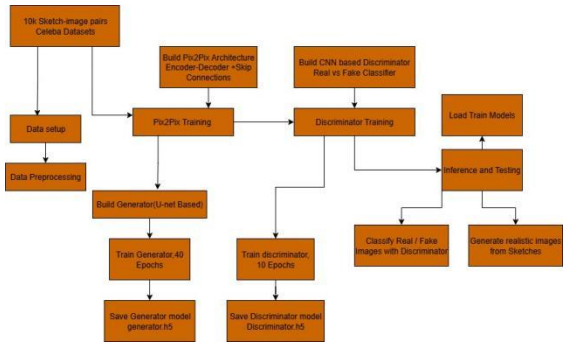

Fig. 8: Flow Chart

The diagram illustrates the complete workflow of the Pix2Pix-based sketch-to-image generation system. It begins with a dataset of 10,000 paired sketch and image samples (from CelebA), which undergoes setup and preprocessing to prepare the data for training. The system architecture is then built, consisting of a U-Net-based generator and a CNN-based discriminator. The Pix2Pix model is trained in two parts: the generator is trained for 40 epochs to learn sketch-to-image translation, while the discriminator is trained for 10 epochs to classify real versus generated images. Both trained models are saved as `.h5` files for reuse. In the testing phase, these models are loaded to perform inference, where the generator creates realistic images from new sketches, and the discriminator evaluates their authenticity. This end-to-end pipeline enables the system to convert abstract sketches into detailed, photo-realistic outputs efficiently.

3) FrontEnd:

The "Sketch to Face Generator" interface provides an interactive platform for users to convert sketches into photo-realistic face images. On the left side, users can upload their sketch via drag-and-drop or by clicking the upload box. After submitting the image using the **"Submit"** button, the application processes the sketch through a trained GAN (Pix2Pix) model. The output, i.e., the generated realistic face, is displayed in the right-hand panel under **"Generated Face"**. Additional options like **"Clear"** reset the input, and **"Flag"** can be used for reporting issues. The interface is served locally (127.0.0.1:7860), indicating it's running in a local development environment. Built with Gradio, it supports both web interaction and API integration for broader accessibility.

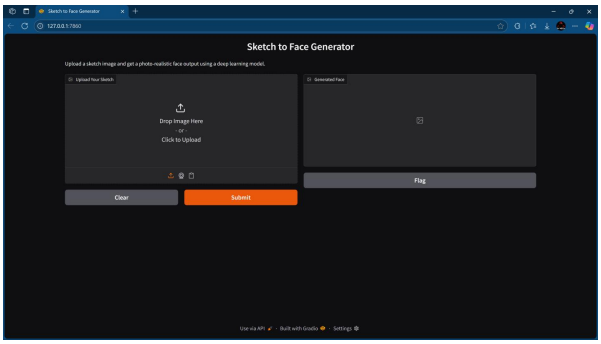

Fig. 9:The Basic User Interface for sketch to image Generation

IV. RESULT AND DISCUSSION

The primary objective of this research was to develop and evaluate a deep learning system capable of generating realistic images from simple or incomplete sketches using advanced GAN techniques. This chapter presents the results of extensive experiments conducted to assess the model's performance from both quantitative and qualitative perspectives. Additionally, the chapter offers a thorough discussion of findings, placing them within the broader context of the literature and highlighting the project's contributions, practical implications, and limitations.

A. Qualitative Results:

The qualitative results of sketch-to-image generation using GANs demonstrate the model's ability to produce visually realistic and coherent images from input sketches. By leveraging the adversarial training between the generator and discriminator, the GAN effectively learns the intricate mapping between structural outlines and detailed, photorealistic outputs. The generated images exhibit accurate texture, colorization, and object shapes that align closely with the intended content of the sketches. Visual inspection reveals that the model preserves the overall structure and style of the input while enriching it with lifelike features, highlighting the capability of GANs to

translate abstract representations into high-quality images

Fig. 10: Successfully generated a real image from a sketch

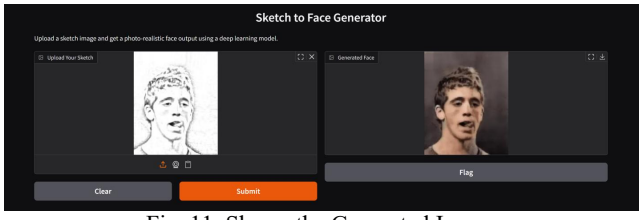

Fig. 11: Shows the Generated Image

B. Quantitative Results

Quantitative evaluation in image-to-image translation tasks, such as sketch-to-image generation using GANs like Pix2Pix, involves measuring how closely the generated images resemble real images using objective metrics. Commonly used metrics include PSNR (Peak Signal-to-Noise Ratio), which quantifies pixel-level similarity and image clarity; SSIM (Structural Similarity Index Measure), which evaluates perceived structural and textural similarity; and FID (Fréchet Inception Distance), which assesses the distribution-level similarity between real and generated image sets. These metrics collectively provide insights into the fidelity, realism, and perceptual quality of generated images. Lower FID and higher PSNR/SSIM values generally indicate better performance. Quantitative evaluation helps compare models objectively and is essential for validating improvements over baseline architectures or during training.

My Project	Scores
PSNR	20.02
SSIM	0.770
FID	203.89

C. Future Work

The future work are given below

- **High-Resolution Image Generation**

Upgrade the model to **Pix2PixHD** or **StyleGAN** to generate higher-resolution and more photo-realistic images.These advanced architectures improve detail, texture, and overall image quality compared to standard Pix2Pix.

- **Unpaired Data Support**

Integrate **CycleGAN** or **CUT (Contrastive Unpaired Translation)** to enable training on **unpaired datasets**.This reduces the need for paired sketch-photo data, making the system more flexible and easier to train with diverse datasets.Useful in real-world scenarios where paired data is limited or unavailable.

V. CONCLUSION

To enhance the quality and flexibility of the sketch-to-image generation system, future improvements can include upgrading the architecture to **Pix2PixHD** or **StyleGAN**, which are capable of producing high-resolution, highly detailed, and photo-realistic images. These models outperform the standard Pix2Pix in terms of visual fidelity and output resolution. Additionally, to overcome the limitation of needing paired training data, the system can be extended to support **unpaired datasets** by integrating models like **CycleGAN** or **CUT (Contrastive Unpaired Translation)**. This would allow the model to learn sketch-to-image translation even when exact sketch-photo pairs are not available, increasing training flexibility and making the system more practical for real-world applications.

ACKNOWLEDGMENTS

"We begin by expressing our gratitude to the almighty ALLAH (S.W.T) for granting us the strength and opportunity to fulfill our responsibilities as electrical engineering students and complete this project. We extend heartfelt thanks to our supervisor, Dr. M Mohsin Riaz whose inspiration, guidance, recommendations, feedback, and companionship were invaluable throughout the duration of the endeavor. Our appreciation also goes to all the lecturers in the electrical engineering department for their valuable guidance, suggestions, and input during the project. We are especially grateful to the department staff and technicians for their assistance and both direct and indirect contributions to the project's completion. Lastly, we extend special thanks to our families for their unwavering support and encouragement. Without their support. "

REFERENCES

[1] J.-Y. Z. T. Z. a. A. A. E. P. Isola, "Image-to-image translation with conditional adversarial networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017.

[2] C. Proceedings, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst., Unknown, 2012.

[3] J. P.-A. M. M. B. X. D. W.-F. S. O. A. C. a. Y. B. I. Goodfellow, " Generative adversarial nets," in Advances in Neural Information Processing Systems, Unknown, 2014.

[4] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv preprint, vol. arXiv:1411.1784, 2014.

[5] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 11, p. 1955–1967, 2009.

[6] W. L. J. Y. J. S. B. Z. a. S. B. Z. Yu, "SketchMate: Deep Hashing for Million-Scale Human Sketch Retrieval," in Proc. IEEE/CVF Conf.

Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019.

[7] Y. Lu, Y.-W. Tai and C.-K. Tang, "Attribute-Guided Face Generation Using Conditional CycleGAN," arXiv preprint, vol. arXiv:1705.09966, 2017.

[8] W. C. a. J. Hays, "SketchyGAN: Towards diverse and realistic sketch to image synthesis," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, 2018.

[9] C. Galea and R. A. Farrugia, "Forensic face photo-sketch recognition using a deep learning-based architecture," IEEE Signal Process. Lett., vol. 24, no. 11, p. 1586–1590, 2017.

[10] Z. Q. Z. L. a. H. W. Y. Liu, "Auto-painter: Cartoon Image Generation from Sketch by Using Conditional ," in Proc. Int. Conf. Neural Inf. Process., Guangzhou, China, 2017.

[11] X. W. a. A. N. Q. Huang, "Sketch2Face: Deep face generation from sketches," in Proc. IEEE Int. Conf. Image Process. (ICIP), Beijing, China, 2017.

[12] P. Sangkloy, N. Burnell, C. Ham and J. Hays, "The Sketchy Database: Learning to retrieve badly drawn bunnies," ACM Trans. Graph. (TOG), vol. 35, no. 4, p. 119:1–119:12, 2016.

[13] S. M. K. a. D. Poole, "SimplE Embedding for Link Prediction in Knowledge Graphs," in Adv. Neural Inf. Process. Syst. 31, Montréal, Canada, 2018.