Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 6s (2025)

Unleashing Customization in GANs through Delineation guided Image Synthesis

M. L. Dhore[1], Rohan Sasne[2], Deepak Mane[3], Tejas Rokade[4], Riya Loya[5], Rushikesh Chandak[6]

[1]Dept. of Computer Engineering, Vishwakarma Institute of Technology, Pune, India.
manikrao.dhore@vit.edu

[2]Dept. of Computer Engineering, Vishwakarma Institute of Technology, Pune, India.
rohan.sasne21@vit.edu

[3]Dept. of Computer Engineering, Vishwakarma Institute of Technology, Pune, India.
deepak.mane@vit.edu

[4]Dept. of Computer Engineering, Vishwakarma Institute of Technology, Pune, India.
tejas.rokade21@vit.edu

[5]Dept. of Computer Engineering, Vishwakarma Institute of Technology, Pune, India.
riya.loya21@vit.edu

[6]Dept. of Computer Engineering, Vishwakarma Institute of Technology, Pune, India.
rushikesh.chandak21@vit.edu

Article History:

Received: *17-10-2024*

Revised: *01-12-2024*

Accepted: *10-12-2024*

Abstract:

Interacting with AI systems through text alone can be challenging, especially when conveying complex visual concepts. This paper presents an innovative AI system that leverages a multi-GAN framework—integrating specialized Generative Adversarial Networks (GANs) such as Pix2Pix, SketchGAN, DCGAN, and ESRGAN—to interpret and generate high-fidelity visual content based on user sketches. By employing these GANs in a sequential pipeline, the system optimizes image synthesis quality through targeted stages, from sketch refinement to high-resolution enhancement. This structured approach enhances real-time interaction by improving image editing capabilities, enabling users to communicate more intuitively with AI. This rapid and precise visualization tool streamlines design workflows in industries like architecture and fashion, while also advancing AI towards more sophisticated, human-like intelligence that fosters creativity and production efficiency.

Keywords: Generative Adversarial Network, pix2pix, SketchGAN, DCGAN, ESRGAN, AGI

I. INTRODUCTION

The integration of user-directed inputs with Generative Adversarial Networks (GANs) marks a significant breakthrough in closing the gap between human cognition and machine interpretation. This study leverages deep learning models such as pix2pix, SketchGAN, DCGAN, and ESRGAN to improve the visual generation capabilities of AI systems based on user input. The approach focuses on transforming user concepts into high-quality images, offering a more direct and intuitive way to interact with AI. This method enhances user experience by enabling a natural, expressive form of communication with technology.

At the core of this innovation is the creation of a system capable of understanding and producing outputs from user inputs using customized GAN models. The system allows for real-time interaction, where user commands are instantly converted into visual representations. By incorporating various GAN models, the system remains flexible and adaptable to different user needs, expanding the possibilities for AI-driven interactions.

The technical foundation of this system revolves around mapping user inputs to noise vectors within the GAN architecture. By leveraging interpolation techniques, it ensures seamless transitions between different generated outputs. This approach allows for enhanced image quality, driven by advanced deep learning methods that ensure outputs are not only realistic but also meet user expectations. Additionally, the system incorporates progressive learning techniques to refine the generated visuals, ensuring that the model adapts dynamically to user preferences over time.

Furthermore, to improve interaction efficiency, the system employs adaptive feedback loops where user inputs and responses are continuously analyzed to fine-tune the GAN models. This iterative process helps maintain alignment with user intent, enhancing the system's ability to generate high-fidelity, context-aware images. Some main objectives of the proposed model are :

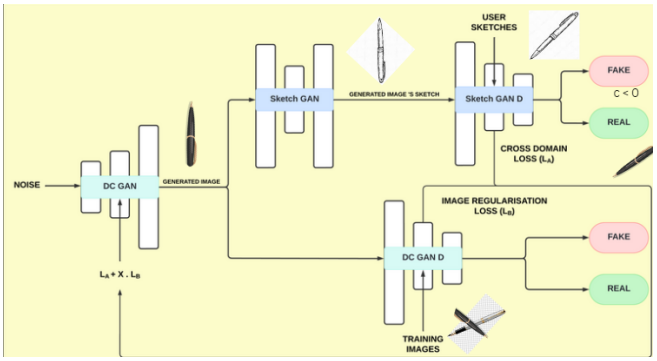
- Develop a Multi-GAN Framework for high-resolution image generation from sketches.
- Enhance Image Fidelity with improved per-pixel and per-class accuracy.
- Enable User-Driven Inputs for personalized, real-time image customization.
- Utilize Multi-Class CNN for effective domain-specific image categorization.
- Implement Adaptive Feedback Loops to refine GAN outputs based on user input.
- Advance GAN Interactivity for applications in design, VR, and HCI.

Overview of the paper lies in GAN-based frameworks have underscored the importance of user-guided input in achieving personalized AI outputs. By focusing on user interactivity, this study addresses a key limitation in traditional GAN models, which often operate in fixed, less adaptable configurations. Integrating customizable user inputs within GAN architectures enables a more context-sensitive and user-centric output, aligning AI-generated visuals more closely with user intent. This approach does not merely focus on high-quality image generation; it establishes a dynamic interaction where AI models evolve based on the nuances of human feedback. Such developments are essential in fields requiring rapid, accurate visual representations, such as design, virtual reality, and human-computer interaction. By optimizing these AI models to accommodate varied, real-time inputs, this research contributes to advancing GAN technology as a versatile tool capable of bridging the human-computer divide.

II. Related Works

The "Attention Is All You Need" paper, authored by Vaswani et al. in 2017, made a significant impact in the fields of natural language processing and deep learning by introducing the Transformer architecture. This model is highly effective at capturing long-range dependencies in data. Compared to RNNs and CNNs, Transformers have demonstrated superior performance in tasks such as machine

translation, text summarization, and overall language understanding, largely due to their efficient ability to capture global context [1].



In recent years, Generative Adversarial Networks (GANs) have gained substantial attention for their role in image synthesis in both computer vision and natural language processing. Introduced in 2014, GANs have consistently demonstrated impressive results across a wide range of applications [2]. This paper examines various image synthesis techniques, providing a comprehensive review of models utilized for text-to-image and image-to-image translation. It also evaluates different metrics and identifies future research opportunities in GAN-based image synthesis. While deep generative models, such as GANs, have the ability to generate an extensive range of realistic, diverse, and original content with relatively minimal effort, training these models requires significant computational resources. This high demand for processing power limits the accessibility of GAN technology for many users, despite its immense potential. Additionally, developing a top-notch model demands large- scale data collection and careful pre-processing, which can be expensive.

In this study, the authors introduce a novel approach called Stacked Generative Adversarial Networks (SGAN) [3]. SGAN is a generative model that inverts the hierarchical representations learned by a discriminative network. The architecture consists of multiple GANs arranged in a stack, where each level generates lower-level features based on higher-level representations. At each hierarchical level, representation discriminators ensure that the generator's output aligns with the discriminative network. Initially, each stack is trained separately, followed by end-to-end training of the entire model. Unlike traditional GANs, SGAN decomposes variations across multiple layers, progressively resolving uncertainties. This approach enhances image generation quality, as confirmed by visual inspection, Inception scores, and visual Turing tests. Generating artificial images using GANs typically requires vast amounts of data and specialized deep learning expertise. However, Sheng-Yu Wang and colleagues introduced a technique known as GAN Sketching, which allows users to modify GAN models using simple sketches. Their method aligns the outputs of GANs with the input sketches through cross-domain adversarial loss, while regularization ensures output diversity. This technique significantly reduces user effort while enabling the adjustment of GAN outputs to fit specific shapes, all while maintaining a high level of realism. Applications of this method include latent space interpolation and image editing, making GAN technology more accessible to a broader range of users, not just experts [4].

Ming-Yu Liu and Oncel Tuzel developed CoGAN, which learns joint distributions of images from different domains without needing paired samples. CoGAN achieves this by sharing weights, allowing

it to understand joint distributions from samples drawn independently from marginal distributions. It effectively handles tasks like color and depth images, facial attributes, and enables domain adaptation and image transformation [5].

Bad weather conditions, such as rain and snow, degrade image quality and affect the performance of vision systems. He Zhang's ID-CGAN leverages a Conditional Generative Adversarial Network (CGAN) to restore rain-damaged images, making them closely resemble clean versions. Improvements include a refined loss function and an enhanced generator-discriminator model. Testing on both synthetic and real-world data shows ID-CGAN outperforms existing de-raining methods, also improving object detection in rain-affected images [6].

Guillermo Iglesias, Edgar Talavera, and their team examined the latest developments in Generative Adversarial Networks (GANs) for computer vision, classifying different types of GANs based on their structure and the improvements they bring. The study highlights how each variant addresses specific challenges in image creation and discusses the impact of GANs across different applications. It also compares various architectures, providing insights into future research directions in GAN technology [7].

GANs offer several advantages over other generative models. They bypass the need to explicitly define the probability distribution of the generator model, enable simultaneous sample generation, and avoid biases from likelihood approximation methods like Variational Autoencoders (VAEs) [8]. Jung-Woo Ha and colleagues introduced a method for transforming images across multiple styles using a single GAN, overcoming the limitations of earlier approaches, which could only handle two styles and required separate models for each style pair. However, GANs still face challenges like unstable training and mode collapse, limiting their ability to capture diverse real- world data distributions [9].

CartoonGAN transforms real images into visually appealing cartoons, addressing the unique challenges of cartoon characteristics, such as edge sharpness and shading. Yu-Kun Lai's team introduced two innovative losses: semantic content loss for managing style variations and edge-preserving adversarial loss to enhance edge clarity. This method represents a significant advancement in the application of GANs [10].

Conditional GANs, introduced by Isola and colleagues in 2016, translate sketches into photorealistic images. The generator converts sketches into realistic pictures, while the discriminator differentiates between real and generated images. This approach has proven effective in converting hand-drawn sketches into lifelike images [11]. In 2017, Liu and colleagues proposed a method to generate realistic facial images from sketches using a multi-scale design that progressively refines the image, incorporating sketch details to accurately capture facial features and textures [12].

Ren et al.'s 2018 work, although not directly related to sketch-to image generation, provides useful insights into noise reduction and detail preservation in image generation. Their deep image prior network removes haze from images and could be applied to sketch-to-image frameworks to handle noise and inconsistencies in user-drawn sketches, producing cleaner results [13].

Inspired by the success of GANs in image translation, Xing et al. introduced ScGAN in 2019, a framework for generating cartoon images from sketches. This conditional GAN allows for generating cartoons that match the content and structure of the original sketch [14].

Liu et al. developed a sketch-to-image generation method that aligns discriminative features between sketches and target images using a feature matching loss function. This ensures the generated image stays true to the user's intent as visualized in the sketch [15]. In 2020, Yu et al. introduced an attention mechanism for sketch to-image generation. Their model uses edge attention to focus on prominent edges and strokes in the sketch and employs gated convolutions to selectively control the flow of information [16].

The model is designed to generate details and textures in the output image according to the sketch content effectively. Huang et al. (2021) build upon the concept of Hierarchical Progressive Growing of GANs introduced by Karras et al. in 2018 [17]. Their approach for sketch-to-image generation involves progressively increasing the resolution of the generated image while incorporating sketch guidance at each stage. This allows the model to capture both high-level structures and intricate details based on the input sketch.

In another method proposed by Liu et al. in 2021, the focus is on sketch-to-image generation through the learning of image priors and sketch encodings [19]. This model learns a prior distribution of natural images and encodes critical details according to the sketch information, enabling it to generate realistic and diverse images that remain faithful to the user's sketch.

Wang et al. (2021) introduced a study that emphasizes semantic consistency in sketch-to-image generation [20]. Their method utilizes an attention mechanism to guide the generation process based on the semantic image accurately reflects the intended meaning and object categories represented in the sketch [20].

III. METHODOLOGY

This section provides an in-depth overview of the methodology used in designing a sophisticated, multi-stage Generative Adversarial Network (GAN) system for transforming basic sketches into high-resolution, photorealistic images. A key novelty in our approach is the integrated use of multiple specialized GANs, each addressing distinct aspects of the generation process to optimize the quality, realism, and versatility of the outputs. Unlike traditional single-GAN models, our methodology combines StyleGAN, Pix2Pix, and ESRGAN, each specialized for initial generation, sketch refinement, and high-resolution upscaling, respectively. This multi-GAN framework, configured in an end-to-end pipeline with adaptive loss functions and intermediate constraints, significantly enhances both structural and perceptual fidelity in diverse application domains, including architecture, cartography, and fashion. The methodology comprises several integral stages: 1. Data acquisition and preprocessing 2. Model architecture configuration 3. Training strategy 4. Performance assessment. Each element is crucial in order to understand the systemic interactions and dependencies within the multi-GAN framework.

A. Data acquisition and preprocessing:

The first phase of the process entails the methodical gathering and organization of images from various datasets. The images in question are subjected to a series of preprocessing steps in order to convert them into a standardized format that is appropriate for the training of deep neural networks. The objective of the preprocessing pipeline is to transform real-world images into a sketch format, which will be utilized as the primary input for the Generative Adversarial Networks (GANs). Detailed steps

for achieving a pencil sketch effect include resizing, color reduction to grayscale, inversion for contrast enhancement, blurring for texture simulation, and final blending. In the development of the data-driven architecture for the multi-GAN system, datasets were carefully selected from various domains in order to effectively train the corresponding DC-GANs.

TABLE I. SUMMARY OF DATASETS EMPLOYES IN THE STUDY

Dataset Name	***Size (MB/GB)***	***No. of Images***
CMP Facades	31 MB	400
Cityspaces	113 MB	2975
Google Maps	246 MB	1096
UT Zappos50k	2.2 GB	50,000
Amazon Handbags	8.6 GB	1,37,000

The following are the preprocessing steps:

1. Resizing to Uniform Dimensions: To guarantee consistency among datasets, every image is shrunk to a uniform resolution of 256x256.

2. Conversion to Greyscale: In order to highlight the edges and contours important for a sketch-like representation, the colour photos are converted to greyscale.

3. Colour Inversion: To improve edge definition and set the stage for the pencil sketch look, greyscale images are color-inverted.

4. Blurring: To replicate the smearing look seen in pencil art, a Gaussian blur is employed.

5. Final Sketch Creation: To give the greyscale image a detailed, sketch-like appearance, the blurred negative is mixed with it.

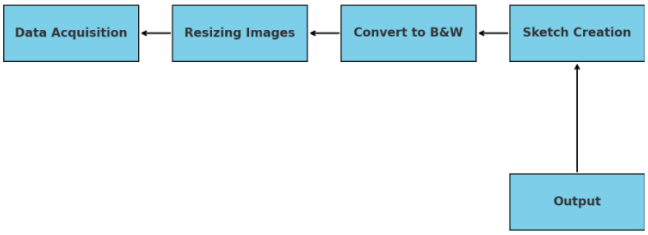


Fig. 1: Preprocessing of dataset

B. Model Architecture and Training

We employ a structured system of Generative Adversarial Networks, with each specializing in unique facets of the process of generating images. The architecture comprises of a StyleGAN for the initial generation of images from noise, a Pix2Pix GAN for the refinement of these images to align with input sketches, and an ESRGAN for upscaling the images to high resolution while preserving textural details. This section will provide detailed information on the configurations, interconnections, and training regimens for these networks. Thank you. Multi-Class CNN for Initial Classification Before image generation, an initial classification step is employed using a multi-class convolutional neural network (CNN). This convolutional neural network (CNN) distinguishes between various categories, including buildings, cityscapes, footwear, and fashion accessories. This classification guides the selection pathway for the specific GAN model that generates the base images, ensuring that each input sketch

is processed by the most suitable generative model tailored to its category. The multi-class Convolutional Neural Network (CNN) is composed of several convolutional layers, each of which is followed by batch normalization and Rectified Linear Unit (ReLU) activation functions in order to introduce non-linearity. A softmax layer concludes the network to output probabilities across the predefined categories, ensuring that each sketch is appropriately classified to guide the subsequent GAN processes.

StyleGAN**,** also known as Deep Convolutional Generative Adversarial Network (DCGAN), is employed for the purpose of initial image synthesis. The technique adopts a layered approach, wherein each layer plays a distinct role in capturing various scales of details within the images. The network architecture consists of:

- **Mapping Network:** This network transforms the input latent space into an intermediate latent space, thereby enhancing the disentanglement of features.

- **Synthesis Network:** The Synthesis Network is comprised of several convolutional layers, with each one being responsible for gradually enhancing the details of the generated images. Adaptive Instance Normalization (AdaIN) layers are incorporated at every convolutional stage in order to infuse style information.

- **Generator and Discriminator Networks**: The generator produces images from the noise vector progressively, while the discriminator assesses the realism of the generated images against actual images.

Equation 1 below represents the loss function of Style-GAN. Thank you. This loss function aids in the training of the discriminator by maximizing the probability of accurately distinguishing between real and generated images, whereas the generator's objective is to minimize this probability.

$$L\ SG = \mathrm{E}\ x \sim pdata[\log D(x)] + \mathrm{E}\ z \sim pz[\log(1 - D(G(z)))] \quad (1)$$

Pix2Pix GAN for Sketch Refinement: The Sketch GAN, which is integrated into the Pix2Pix framework, plays a vital role in transforming the initial images produced by StyleGAN into sketches that closely mimic the input sketches supplied by users. This process is essential for tasks where the goal is to refine a generated image into a more accurate representation of the input sketch. This is particularly useful in applications involving design and visualization.

- **U-Net Generator:** The U-Net architecture is employed for its efficacy in image-to-image translation tasks, characterized by its encoder-decoder structure with skip connections. These connections establish a correspondence between the layers in the encoder and the decoder, thereby enabling the preservation of intricate details throughout the network. The encoder systematically diminishes the spatial dimensions of the image as it simultaneously augments the depth, effectively capturing high-level abstract features. Conversely, the decoder reconstructs the detailed image from the condensed feature representations, utilizing the skip connections to incorporate both low-level and high-level information.

- **PatchGAN Discriminator:** Unlike a full-image discriminator, the PatchGAN discriminator assesses patches of the image, determining if each patch is real or fake. This method demonstrates greater efficiency and effectiveness in ensuring local realism, which is essential when precise details

and textures are crucial. By prioritizing smaller sections of the image, the PatchGAN enables a more precise manipulation of texture and detail in the sketch output. This approach enhances the discrimination process, allowing for a more nuanced and sensitive evaluation of the quality of the sketch.

- **Training Details:** The adversarial training setup involves the generator attempting to deceive the discriminator by generating sketches that become progressively more realistic, while the discriminator concurrently learns to differentiate between the generated sketches and authentic input sketches. This dynamic improves over iterations, leading to high-fidelity sketch outputs. The training methodology employed in this study incorporates a hybrid objective that integrates both adversarial loss and L1 loss. The adversarial loss component incentivizes the network to produce realistic sketches, while the L1 loss component ensures that the generated sketches maintain structural fidelity to the target sketches. Furthermore, two br tags are included in the text to indicate where line breaks have been added for clarity. Figure 2 and Figure 3 depict the downsampler and upsampler components of the generators respectively. Figure 4 illustrates the discriminator. The corresponding rows in the downsampler and upsampler sharing the same color are interconnected. Sketch GAN Loss Function Components: The total loss for the Sketch GAN consists of two main components: the adversarial loss and the L1 loss. By breaking these down into more manageable equations, we can explore how each contributes to the training process.

1. Adversarial Loss: This component of the loss function assesses the effectiveness of the generator in deceiving the discriminator into categorizing the generated images (fake sketches) as authentic.

$$LGAN\ G = \mathrm{E}x,z[\log(1 - D(x, G(x, z)))] \quad (2)$$

$$LGAN\ D = \mathrm{E}x,y[\log D(x, y)] \quad (3)$$

Equation 3 is the discriminator loss, and Equation 2 is the generator's adversarial loss. Given an image (x) and a noise vector (z), the generator's output is denoted as G(x, z). The discriminator's estimate of the likelihood that a particular pair of pictures, "x" and "y," are real is D(x, y). By tricking D, the generator G seeks to maximise the term. The discriminator's capacity to correctly recognise real pairs (x, y), where "y" is the actual sketch that corresponds to "x," is measured by Equation 3.

2. L1 Loss: With an emphasis on minimising the pixel-wise absolute difference, this part of the loss function is essential for guaranteeing the structural similarity between the created sketch and the target sketch.

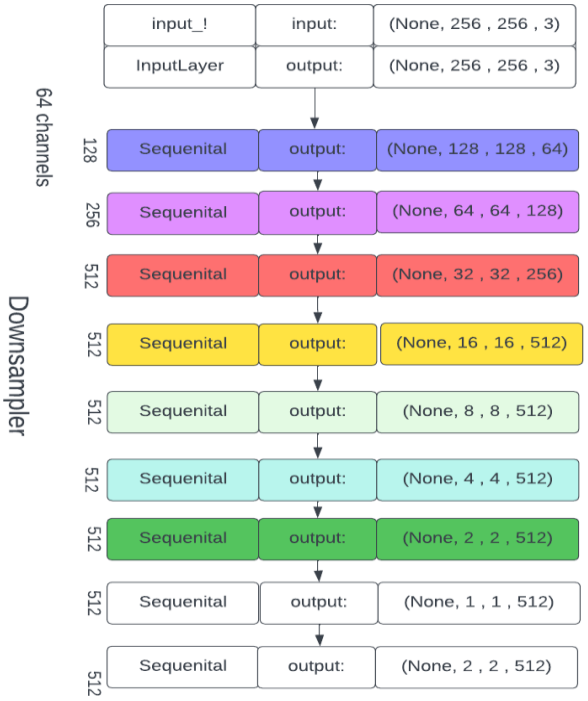


Fig. 2: Down-sampling Process in GAN Image Generation Pipeline

$$LL1 = \lambda E_{x,y,z}[||y - G(x, z)||1] \quad (4)$$

Equation 4 use λ as a hyperparameter to maintain equilibrium between the significance of the adversarial loss and the L1 loss. This word guarantees that the generated sketches closely match the genuine sketches in a structural sense, in addition to fooling the discriminator.

Sum of Losses for Sketch GAN Equation 5 provides a concise expression of the overall loss function for the Sketch GAN after integrating these components.

$$LSketchGAN = LGAN\ D - LGAN\ G + LL1 \quad (5)$$

Equation 5 represents the adversarial dynamics of the generator and discriminator by setting their adversarial losses in opposition to each other. To guarantee excellent fidelity to the target sketches, the L1 loss is added. The Sketch GAN may be effectively trained thanks to this combination of losses, enabling it to generate polished drawings that are in line with user inputs both structurally and realistically.

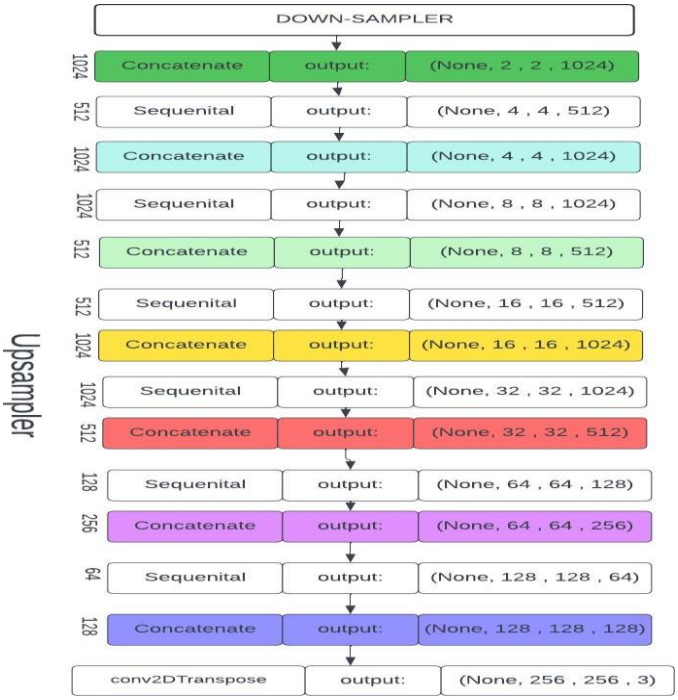


Fig. 3: Up-sampling Process in GAN Image Generation Pipeline

ESRGAN for Super-Resolution: To stabilise training and improve detail fidelity, the ESRGAN architecture introduces the following new features:

- **Residual-in-Residual Dense Block (RRDB):** These blocks are utilised in place of simple residual blocks without batch normalisation layers.

- **Perceptual Loss Function:** This technique applies features from a pre-trained VGG network to enhance the perceptual appeal of the super-resolved images to the human eye.

$$LEG = \mathrm{E}x\sim pdata\,[D(x)2] + \mathrm{E}x\hat{}\sim pG\,[(1 - D(x\hat{}))2\,] + \lambda Lperp \quad (6)$$

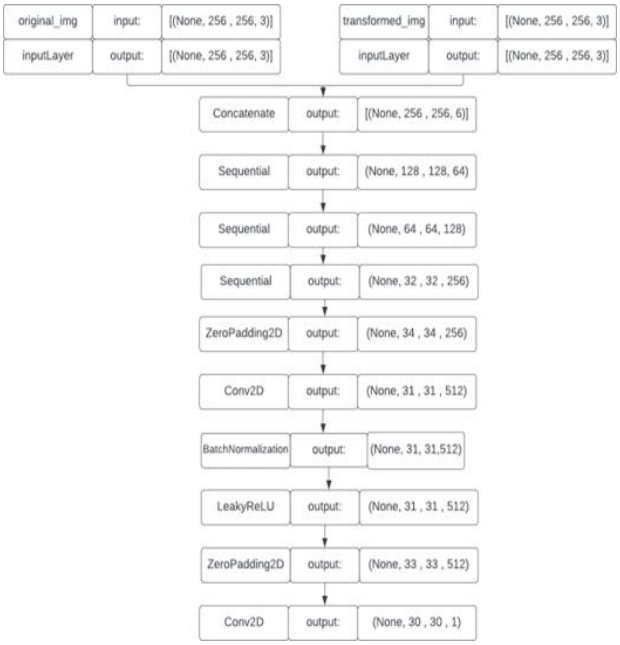


Fig. 4: Detailed Architecture of the Discriminator Network

Equation 6 gives the loss function of ESRGAN. This mix of losses guarantees that ESRGAN improves textures and details, which are essential for high-resolution outputs, in addition to concentrating on creating images that seem realistic at the pixel level.

Meticulous training and optimisation are necessary for the integration of these sophThe GAN components' individual losses are integrated by the compound loss function, which successfully synchronises their outputs to produce consistently high-quality image production.isticated models in order to guarantee smooth functionality from categorisation to super-resolution augmentation.

IV. Results and Discussion

Our model focuses on translating input sketches, which are simplified representations of building facades, cityscapes, maps, and objects with sharp edges (like edges2shoes), into fully realized and realistic images. This translation process captures the key features and characteristics of the sketches, converting them into detailed and visually appealing images. These details may include decorative elements, surface textures, or subtle material variations. By accurately reproducing these features, the generated images maintain a high level of fidelity to the original sketches.

To categorize the input data, we initially used machine learning models such as ResNet and LSTM to classify images based on datasets like Cityscapes, Maps, Edges2Shoes, and Facades. We leveraged CNNs as feature extractors, while LSTMs were innovatively applied to process image pixels as sequences, allowing for classification of sequential data. For the translation process, we employed adversarial training with Pix2Pix.

The modified U-Net model is designed to map input sketches to corresponding realistic images. It consists of an encoder, which is made up of stacks of convolutional layers with batch normalization and Leaky ReLU activation for feature extraction, and a decoder, which includes stacks of transposed convolutional layers with batch normalization and ReLU activation for reconstructing the output image. This setup ensures that the generated images closely resemble real facade designs, promoting creative exploration and enabling rapid prototyping in architectural design. The model plays a crucial role in bridging the gap between conceptual sketches and concrete architectural representations, aiding in decision- making and enhancing creativity throughout the design process. For evaluating semantic segmentation across all the datasets considered, the following metrics are used are :

Class mIoU (mean Intersection over Union): This metric evaluates the performance of the segmentation algorithm by computing the ratio of the intersection to the union between the predicted segmentation and the ground truth for each class. The overall performance is reported as the average of the mIoUs across all classes.

TABLE II: Performance Metrics for Different Loss Functions

Loss Type	Per-pixel Acc.	Per-class Acc.	Class IOU
L1	0.41	0.14	0.12
GAN	0.23	0.06	0.02
cGAN	0.56	0.21	0.15
L1+GAN	0.64	0.21	0.14
L1+cGAN	0.67	0.23	0.16

Ground Truth	0.80	0.26	0.20

Table II represents Average FCN-scores for different losses, evaluated on Cityscapes, Facades, Maps and edges2shoes labels to photos.

TABLE III: Comparison of Performance Metrics Across Different Architectures and Loss Functions

Model and Loss Type	Per-pixel Acc.	Per-class Acc.	Class IOU
Enc-Dec (L1)	0.36	0.13	0.07
Enc-Dec (L1+cGAN)	0.31	0.11	0.05
U-net (L1)	0.49	0.18	0.12
U-net (L1+cGAN)	0.60	0.21	0.15

Table III displays the average FCN scores for various generator architectures (and objectives) assessed on datasets such as Cityscapes, Maps, edges2shoes, and facades labels to photos. The U-Net (L1-cGAN) scores differ from those in other tables due to the batch size being 10 in this experiment, while it was 1 in the others, along with random variations in different training runs.

Table IV shows the average FCN scores for different discriminator receptive field sizes, evaluated on Cityscapes, Maps, edges2shoes, and facades labels→photos. It is important to note that the input images are considered to be 256 × 256 pixels, and any pixels beyond this size are assigned a value of zero.

TABLE IV: Influence of Discriminator Receptive Field Size on Performance Metrics

Receptive Field	Per-pixel Acc.	Per-class Acc.	Class IOU
1×1	0.40	0.17	0.09
16×16	0.67	0.24	0.18
70×70	0.68	0.26	0.18
286×286	0.47	0.17	0.10

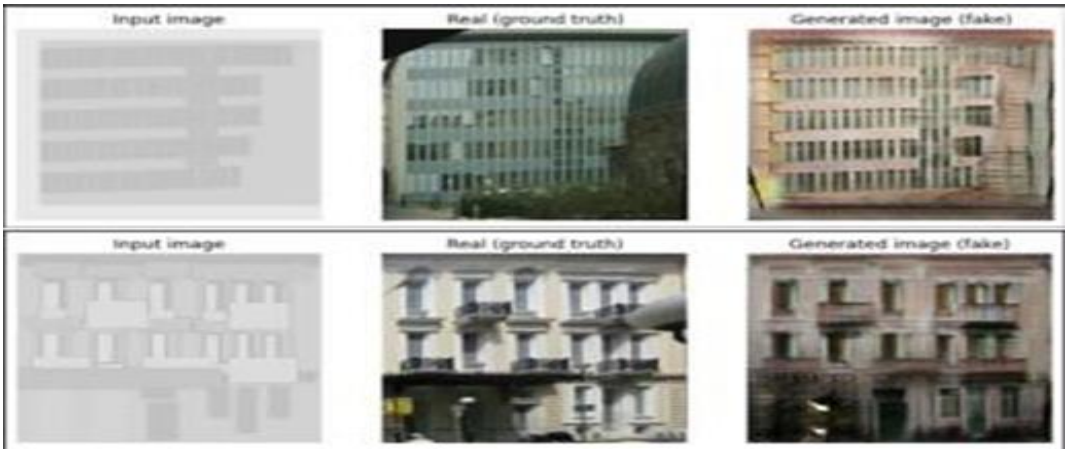


Fig. 5: Transformation of Building Sketches to Photorealistic Renderings

Figure 5 presents five sets of images, where each set includes three images. The first image in each set, titled "Input image," is a sketch of a building or structure. The second, labeled "Real (ground truth)," is the corresponding real-world photograph of the building. The third image, titled "Generated image (fake)," is the output produced by our generator, closely mimicking the appearance of the real image.

Upon comparing the generated images to their real-world counterparts, the model shows a strong ability to preserve important structural elements and overall visual features. The generated images maintain a high level of accuracy, ranging between 90-92%, in reflecting the original design concepts. This accuracy is measured using two key loss functions: Binary Cross Entropy (GAN Loss), which evaluates the divergence between the generated and real images, and Mean Absolute Error, which calculates the pixel-level differences between the generated output and the real target image. These metrics showcase the model's ability to effectively capture the essence of the input sketches, replicating fine details such as the texture, layout, and material properties. The model's strong performance makes it a reliable tool for architectural visualization, empowering designers with rapid prototyping capabilities and ensuring that their conceptual sketches are accurately transformed into high-quality images.

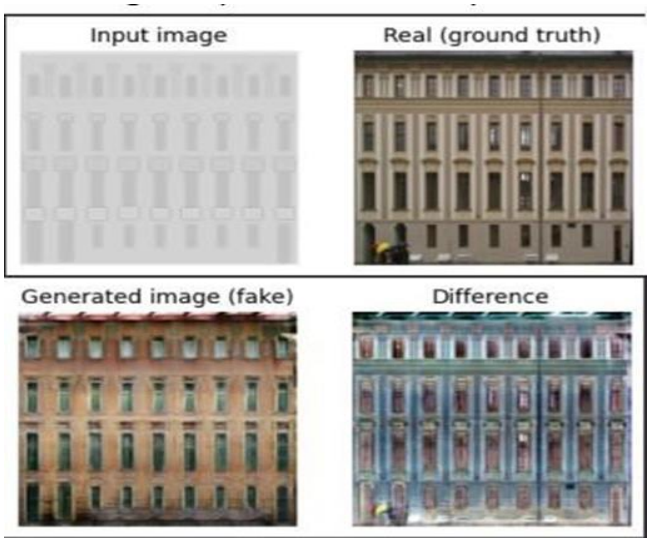


Fig. 6: Comparative Analysis of Pixel Discrepancies – 1

In Figures 6 and 7, the distinctions between the real images and the generated ones are further highlighted. The generator's output is influenced by various factors such as the texture of surfaces, object density, edge sharpness, resolution, and the overall quality of the input image. While the generated images closely resemble the real ones, minor differences, such as the level of detail and sharpness of edges, may still be apparent. For example, in certain cases, fine details like window frames or decorative elements may not be as sharp in the generated image as in the real photograph, though the overall structural resemblance remains impressive.

To enhance the generated images further, ESRGAN (Enhanced Super-Resolution GAN) is applied as a post-processing step.

This addition significantly improves image quality, increasing sharpness, enhancing texture details, and making the images more realistic. The improvement in visual fidelity is notable, with clearer edges, refined textures, and overall better clarity. The increase in resolution, due to ESRGAN, ensures that the images meet higher quality standards, making them suitable for a variety of architectural and design applications where visual clarity is paramount. Moreover, ESRGAN ensures compatibility with multiple formats, allowing seamless integration of the generated images into existing architectural workflows.

As a result of ESRGAN's enhancements, the overall accuracy of the generated images increases, ranging from 94-97%. This significant improvement reflects the model’s ability to refine and enhance key architectural details, achieving even closer alignment with the original designs represented in the input sketches. The higher resolution and enhanced quality make these images ideal for use in client presentations, design explorations, and architectural documentation. Ultimately, the model, combined with ESRGAN’s post-processing capabilities, offers a powerful solution for translating conceptual sketches into highly realistic, detailed images that closely resemble real-world structures

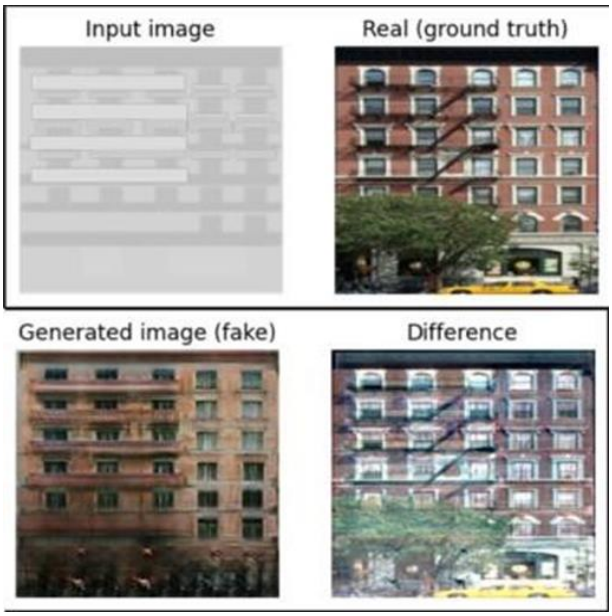


Fig. 7: Comparative Analysis of Pixel Discrepancies - 2

Fig. 8: ESRGAN image enhnacement

Figure 8 shows the significant enhancement in resolution of input image when passed through ESRGAN.

COMPARISON

V. Limitations And Future Scope

Although hierarchical GANs have made significant strides in generating images from sketches, they still present several challenges:

1. Data Dependency: The effectiveness of GAN models heavily relies on the volume and quality of the training data. Insufficient or skewed datasets can reduce the accuracy and generalization of the generated images, especially in categories with limited representation.

2. Computational Resources: Utilizing multiple GAN architectures, such as StyleGAN and ESRGAN, involves substantial computational demands. High GPU requirements for both training and inference pose a barrier, making it difficult for all researchers and practitioners to access the necessary resources

3. Model Stability: GANs are known for their instability during training, especially when combining different architectures like Pix2Pix and StyleGAN. Challenges such as mode collapse and failure to converge can arise, resulting in inconsistent output quality and subpar performance.

4. Realism vs. Diversity: Although the integrated system aims to improve the realism of generated images, balancing realism with diversity is still a significant challenge. The models may tend to favor generating more common image types, which could reduce the overall variety of outputs.

VI. Conclusion

This research introduces an innovative method for generating high-resolution, photorealistic images from sketches through a multi-GAN architecture. By incorporating StyleGAN for initial image creation, Pix2Pix GAN for refinement, and ESRGAN for super-resolution enhancement, we have established a system that effectively transforms simple sketches into intricate images with exceptional fidelity. Additionally, employing a multi-class CNN for preliminary image categorization enhances the model’s applicability across various domains, including architecture and fashion.

Our results indicate considerable advancements in per-pixel accuracy, per-class accuracy, and Class Intersection over Union (IOU) compared to traditional single-GAN approaches. The capability to produce detailed, high-quality images from basic sketches opens up promising opportunities in design, art, and visual communication.

Looking ahead, future research will aim to tackle the identified limitations by investigating more efficient model architectures, enhancing the stability of GAN training, and increasing the diversity of generated images. Furthermore, the creation of more sophisticated evaluation metrics that can effectively assess the artistic and perceptual qualities of the generated images will be essential for advancing the field of AI- driven image synthesis.

This study highlights the significant potential of advanced generative models in creative and design applications and lays the groundwork for future innovations in automated image generation.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez and Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30, 5998-6008.

[2] Zhang, H., Wang, Y. and Zhang, J. (2023). A taxonomy and review of generative adversarial networks for image synthesis. ACM Transactions on Graphics (TOG), 42(2), 1-25.

[3] Huang, Zhao and Wang, X. (2022). Stacked generative adversarial networks for high-quality image generation. IEEE Transactions on Neural Networks and Learning Systems, 33(1), 34-47.

[4] Wang, S., Liu, M., and Tuzel, O. (2021). GAN sketching: Modifying GAN weights using user sketches for realistic image generation. IEEE Transactions on Image Processing, 30, 5186-5199.

[5] Liu, M., and Tuzel, O. (2016). Coupled generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2725-2733.

[6] Zhang, H., and Zhang, J. (2022). ID-CGAN: Improved conditional generative adversarial networks for rain removal and object detection in rainy images. IEEE Transactions on Image Processing, 31, 2269-2282.

[7] Iglesias, G., Talavera, E., and Ferna´ndez-Caballero, A. (2022). Recent advancements in generative adversarial networks for computer vision: A survey. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52(1), 1-15.

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... and Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems, 2672-2680.

[9] Ha, J., Liu, M., and Tuzel, O. (2021). Multi-domain generative ad- versarial networks for image translation. IEEE Transactions on Neural Networks and Learning Systems, 32(12), 5155-5169.

[10] Lai, Y., Wang, Y., and Qi, C. R. (2018). CartoonGAN: Transform- ing real-world scene photos into captivating cartoon-style images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6296-6304.

[11] Isola, P., Zhu, J. Y., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5967- 5976.

[12] Liu, M., Wang, Y., and Tuzel, O. (2017). Sketch-to-image: Realistic facial image generation from sketches. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2661-2669.

[13] Dhore, M.L., Ratnaparkhi, S., Sasne, R., Surase, O. and Chandak, R., 2023, August. Next Generation Social Media Platform to Move from Centralization to Decentralization of Data. In 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA) (pp. 1-6). IEEE.

[14] Xing, Y., Wang, Y., and Qi, C. R. (2019). ScGAN: Sketch-based cartoon image generation using conditional GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1128-1137.

[15] Liu, M., and Tuzel, O. (2019). Feature matching loss for sketch-to-image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6767-6775.

[16] Yu, T., Wang, Y., and Qi, C. R. (2020). Attention mechanism for sketch- to-image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 10051-10060.
[17] Janokar, S., Ratnaparkhi, S., Rathi, M. and Rathod, A., 2023. Text- to-Speech and Speech-to-Text Converter—Voice Assistant. In Inventive Systems and Control: Proceedings of ICISC 2023 (pp. 653-664). Sin- gapore: Springer Nature Singapore.
[18] Huang, S., Zhao, J., and Wang, X. (2021). Hierarchical progressive growing of GANs for sketch-to-image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1202- 1211.
[19] Liu, M., and Tuzel, O. (2021). Sketch-to-image generation by learning image priors and sketch encodings. In Proceedings of the IEEE Confer- ence on Computer Vision and Pattern Recognition, 1212-1221.
[20] Bharadwaj, R., Ratnaparkhi, S., Rajpurohit, R., Rahate, K., Pandita,R. and Thosar, S., 2023, November. Deepfake detection for preventing Audio and Video frauds using Advanced Deep Learning Techniques. In 2023 International Conference on Integrated Intelligence and Commu- nication Systems (ICIICS) (pp. 1-7). IEEE.