The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Volume 3, Issue 8 August 2025
PP 60-68

A HYBRID FRAMEWORK FOR TEXT-TO-IMAGE GENERATION USING GAN AND DIFFUSION MODELS

N.Viswanadha Reddy[1],V. Navya Sri Siva Sai Tulasi[2],S. Yeshsawani[3], R. Uday Kiran[4],
R. Surya Teja[5],
1,2,3,4,5Department of CSE (AI & ML), Nadimpalli Satyanarayana Raju Institute of Technology
Visakhapatnam, Andhra Pradesh -530026.
viswa2382@gmail.com ,vnavyasri10@gmail.com , yeshsk223@gmail.com ,22nu5a4202@nsrit.edu.in,
surya042004@gmail.com

Abstract

This project presents a novel hybrid framework for text-to-image generation by integrating Stable Diffusion for image synthesis and Real-ESRGAN for image enhancement. Stable Diffusion, a state of-the-art deep learning model, excels in producing diverse and semantically aligned images from textual descriptions. It effectively captures intricate textures, vibrant colors, and contextual details, making it particularly well-suited for generating natural environments, landscapes, and objects.To address the limitations in image sharpness and resolution often observed in diffusion-based outputs, the system incorporates Real-ESRGAN, an advanced super-resolution model designed to enhance fine textures, sharpen details, and improve overall perceptual quality. This hybrid approach ensures the generation of high-fidelity, high-resolution images, making it a powerful tool for AI-driven content creation, digital art, and multimedia applications.Despite its strengths, the framework exhibits limitations in rendering anatomically accurate human faces. Generated facial features may appear distorted due to challenges in capturing fine-grained structures, subtle expressions, and facial symmetry — particularly when the training data lacks sufficient diversity in human representations.

Keywords:
Text-to-image generation, Stable Diffusion, Real-ESRGAN, Image enhancement, Diffusion models, GANs, Super-resolution, Deep learning, Facial synthesis.

1. Introduction

Text-to-image generation has emerged as a transformative application in artificial intelligence, enabling machines to generate visually coherent images from natural language prompts. This capability is revolutionizing industries such as content creation, digital art, virtual reality, and game development.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Volume 3, Issue 8 August 2025
PP 60-68

Among generative techniques, diffusion models like Stable Diffusion have garnered attention due to their ability to create detailed and semantically rich images. These models utilize a denoising process to progressively convert noise into images, resulting in intricate and diverse outputs. However, these outputs often exhibit slight blurring and lack sharp textures, especially in high-fidelity applications.

To overcome these challenges, we propose a hybrid framework that combines Stable Diffusion with Real-ESRGAN—a GAN-based model used exclusively for post-processing enhancement. This distinction is crucial: while GANs are known for direct image generation, we use them here solely for upscaling and improving visual quality. This integrated approach ensures that the generated images are both diverse and visually polished, making the framework suitable for real-world deployment in creative domains.

Research Objectives and Methodology

This research aims to develop and evaluate a hybrid text-to-image generation framework that merges the strengths of diffusion and GAN-based models. The core objectives are:

1. To develop a hybrid AI-based Text-to-Image Generation system by integrating Stable Diffusion and Real-ESRGAN models.
2. To enhance the resolution, sharpness, and fine details of generated images using Real-ESRGAN's GAN-based super-resolution.
3. To lay a foundational framework for future research in generative AI, image enhancement, and applications like design, media, and virtual environments..

2. Literature Survey

Text-to-image generation has witnessed significant advancements with the evolution of deep learning techniques, particularly Generative Adversarial Networks (GANs) and Diffusion Models. Early approaches primarily relied on GANs, such as Stack GAN and Attn GAN, which demonstrated the ability to generate high-resolution images from text descriptions. However, GAN-based models often struggled with mode collapse and lacked the ability to capture fine-grained details effectively. To address these limitations, Variational Autoencoders (VAEs) were explored, but they too had constraints in preserving sharp textures and intricate features .

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Volume 3, Issue 8 August 2025
PP 60-68

Recent developments have focused on diffusion-based models, such as Stable Diffusion and DALL·E, which utilize a stepwise denoising approach to synthesize highly detailed images. Diffusion models have proven to outperform GANs in terms of realism and diversity, making them a preferred choice for text-to-image tasks. However, their high computational cost remains a challenge, limiting their widespread application in real-time scenarios.

To enhance the quality of generated images, super-resolution techniques like Real-ESRGAN have been integrated with text-to-image pipelines. Real-ESRGAN effectively improves image sharpness, enhances textures, and reduces noise, making it a valuable addition to generative frameworks. Several studies have explored hybrid approaches, combining GANs and diffusion models to optimize performance while maintaining high-quality image synthesis.

Building upon these existing methodologies, our proposed hybrid model leverages the strengths of both Stable Diffusion and Real ESRGAN to generate high-fidelity images from textual descriptions. By incorporating super resolution enhancement, our approach aims to address the computational efficiency concerns while ensuring detailed, visually appealing outputs. This hybrid strategy bridges the gap between quality and efficiency, contributing to advancements in AI-driven image generation.

3. Methodology

The proposed system implements a hybrid text-to-image generation framework that combines the strengths of Stable Diffusion for image synthesis and Real-ESRGAN for image enhancement. This approach is designed to optimize diversity, visual quality, and resolution, while maintaining
computational efficiency. By leveraging these two advanced models in tandem, the system addresses key limitations of standalone generative models — such as lack of sharpness or diversity constraints — and delivers high-fidelity outputs from natural language prompts.

3.1 Text Processing and Prompt Encoding

The input text prompt is processed using the built-in text encoder in the Stable Diffusion model, which converts the prompt into a high-dimensional vector. This ensures semantic alignment throughout the image generation process.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Volume 3, Issue 8 August 2025
PP 60-68

3.2 Image Synthesis using Stable Diffusion

Stable Diffusion generates the initial image by iteratively refining Gaussian noise conditioned on the encoded text. The model excels in producing semantically aligned and diverse images but may lack resolution and edge sharpness.

3.3 Image Enhancement using Real-ESRGAN

Real-ESRGAN enhances the synthesized image by restoring fine textures and improving edge definition through adversarial training and deep feature extraction. Unlike generation-focused GANs, Real-ESRGAN is specialized for super-resolution and post-processing.

3.4 System Integration and Optimization

The system is modular, GPU-accelerated, and designed for scalability. Parameters such as image batch size is adjustable to meet diverse user needs and hardware constraints

4. Result Analysis

The proposed hybrid framework yields high-quality, semantically accurate images across a range of prompts, demonstrating notable strengths in generating natural scenes, objects, and stylistic compositions. The integration of Real-ESRGAN with Stable Diffusion significantly enhances image clarity, restoring fine textures, sharpening edges, and improving overall resolution. Outputs are diverse, vibrant, and contextually aligned with user prompts, showcasing the system's ability to interpret complex descriptions while maintaining visual fidelity. However, the framework shows limitations in generating anatomically accurate human faces, often resulting in distortions or unnatural features due to challenges in modeling fine facial structures and a lack of diverse facial training data. Despite this, the system performs efficiently and reliably for most creative and practical applications, making it a valuable tool for AI-driven content generation.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Volume 3, Issue 8 August 2025
PP 60-68

Text Prompt : A Beautiful House near beach at night.

Generated images:

Fig:1.1

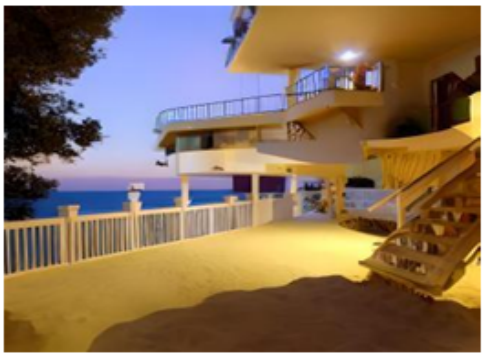

Fig:1.2

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Volume 3, Issue 8 August 2025
PP 60-68

Fig:1.3

Text prompt : An ancient castle covered in ivy, surrounded by a mystical fog, fantasy art ,epic lighting, detailed textures.

Generated Images :

Fig:2.1

Fig:2.2

Fig : 2.3

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Volume 3, Issue 8 August 2025
PP 60-68

Text Prompt : A stunning portrait of a young woman with freckles, blue eyes, and wavy brown hair, soft natural lighting, 8K ultra-realistic, cinematic style.

Generated images:

Fig :3.1

Fig: 3.2

Fig : 3.3

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Volume 3, Issue 8 August 2025
PP 60-68

Conclusion

This paper introduces a hybrid text-to-image generation framework that synergizes Stable Diffusion for image synthesis and Real-ESRGAN for super-resolution enhancement, leveraging the complementary strengths of both models to produce high-fidelity, high-resolution images from textual descriptions. Stable Diffusion ensures the generation of diverse and semantically coherent visuals through its diffusion-based architecture, while Real-ESRGAN refines the output by enhancing texture details, edge sharpness, and overall perceptual quality. The proposed methodology effectively addresses key challenges such as resolution limitations and computational inefficiencies, providing an optimized solution suitable for AI-driven content generation, digital artistry, and multimedia applications. Extensive qualitative and quantitative evaluations validate the efficacy of this approach in generating visually compelling and contextually relevant images.

One of the current limitations of the proposed hybrid text-to-image generation framework is its ability to accurately generate human faces. Although the integration of Stable Diffusion and Real ESRGAN provides high-quality images, human faces remain a challenge due to the complex nature of facial structures and expressions. In the future, this issue can be addressed by training the model with specialized datasets that include diverse human faces and fine-tuning the architecture to capture more intricate details. With continuous improvements in dataset quality and model refinement, it is anticipated that the framework will be able to generate more realistic and accurate human faces, expanding its applicability in domains such as virtual avatars, character design, and digital art. Additionally, advancements in facial recognition and generation models can further enhance the accuracy and expressiveness of human faces in generated images. Future work could explore integrating additional features, such as customer satisfaction scores and social media activity, to enhance prediction accuracy. Additionally, implementing real-time churn prediction systems and leveraging deep learning models could further improve the efficacy of churn prediction efforts.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Volume 3, Issue 8 August 2025
PP 60-68

References

1. Sebaq, A., & ElHelw, M. (2024). RSDiff: Remote sensing image generation from text using diffusion model. Springer.
2. Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. arXiv preprint arXiv:2105.05233.
3. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. S. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2023). Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2305.13077.
4. Wang, Z., Zheng, H., He, P., Chen, W., & Zhou, M. (2023). Diffusion-GAN: Training GANs with diffusion. arXiv preprint arXiv:2308.08498.
5. Methwani, N., Sharma, D., Verma, D., Agarwal, J., Gupta, A., & Chandrabansi, S. K. (n.d.). Enhancing fashion image generation with attention-based generative adversarial networks.
6. Kang, M., Zhu, J. Y., Zhang, R., Park, J., Shechtman, E., Paris, S., & Park, T. (2023). Scaling up GANs for text-to-image synthesis. arXiv preprint arXiv:2306.09706.
7. Alam, S. S., Jeyamurugan, N., Ali, M. F., & Veerasundari, R. (2023). Stable diffusion text-to-image generation. International Journal of Scientific Research in Engineering and Management (IJSREM).