STAT3007/7007 Project Proposal

Group G11
Malaika Vaz (s4699270) · Lauren Rouse (s4742379) · Xuran Wang (s4838862)
Ganesh Channaiah (s4919002) · Mats Martinussen (s4946842)

1 Topic

Our project explores combining two different deep learning architectures, specifically Generative Adversarial Networks (GANs) and diffusion models, to enhance image generation. This approach will draw on the strengths of the two models to generate images that are diverse and creative, while also being realistic. We have decided to limit the scope of our project to face images initially and will use pre-trained open-source GAN and diffusion models to develop our hybrid model, rather than training the models from scratch. We will experiment with a variety of strategies to integrate the two architectures to obtain the best performing model. Based on the outcome and performance of our preliminary model, we can then expand the scope of our project to include other image domains.

2 Significance

The combination of diffusion models and generative adversarial networks (GANs) represents a cutting-edge direction in the field of image generation, which can effectively solve the limitations of the two methods while giving full play to their advantages. Diffusion models are excellent at capturing complex data distributions and generating diverse content, but the generation process is often time-consuming and may lack realism in detail (Salimans and Ho, 2022). In contrast, GANs are good at generating highly realistic images and can present exquisite details at high resolution, but their training process is unstable and prone to pattern collapse, resulting in insufficient diversity of generated samples (Allahyani et al., 2023). By combining these two models, we hope to develop a hybrid framework that takes advantage of the diversity and creativity of diffusion models while enhancing the realism of generated images with the power of GAN discriminators.

Moreover, recent research has highlighted that current generative models often fail to fully capture fidelity and diversity, which are critical for real-world applications (Betzalel et al., 2022a). Therefore, our model aims to combine the advantages of the two methods and explore more reasonable evaluation criteria while improving the quality of generation. In the future, this direction may provide a better reference for the practical application of the generated model.

3 Feasibility

The feasibility of the project is supported by three main factors: the availability of pre-trained, open-source models for both diffusion and GAN architectures; the maturity of the field, with similar hybrid approaches already successfully explored; and a well-scoped work plan that enables us to iteratively build, test, and refine our prototypes.

The project benefits from the widespread availability of pre-trained, open-source diffusion and GAN models. Libraries such as Hugging Face's diffusers provide access to models like Stable Diffusion (Rombach et al., 2022) and ControlNet (Zhang and Agrawala, 2023), while the GAN model StyleGAN2 (Karras, Laine, and Aila, 2020), and many more, are publicly available with pre-trained weights and modular PyTorch implementations. These resources allow us to focus on experimenting with hybrid mechanisms, rather than training large models from scratch.

The field has also seen increasing research interest in combining diffusion models with GAN-based techniques. Recent work has shown that these two families of generative models offer

complementary strengths, exemplified by the CVPR 2024 paper (Kim et al., 2024), which maps semantic features from a diffusion model into the latent space of a pre-trained GAN to leverage the diffusion model's conditioning capabilities alongside the GAN's ability to synthesize realistic images. While our initial approach is to use a GAN discriminator to provide adversarial feedback to a diffusion model, we are generally interested in exploring a broader range of ways these architectures can be combined—whether through loss functions, sampling guidance, or latent-space interaction. The success of related hybrid approaches demonstrates the viability of this research direction, while our goal is to expand the design space and investigate underexplored combinations.

We will begin by establishing a baseline using a pre-trained diffusion model and evaluating its image quality when generating images of faces. From there, we will incorporate adversarial components, starting with a pre-trained GAN discriminator and a dataset of faces from Kaggle (Arnaud58, 2020), to investigate whether feedback from such a model can improve perceptual quality. We will initially apply the adversarial loss to the final output and backpropagate through the full denoising process, allowing the discriminator to guide the refinement of the diffusion model's generative behavior. This approach treats the diffusion model as a generator in a GAN-like setup, where its parameters—such as those within the denoising U-Net—are fine-tuned to better fool the discriminator. In later experiments, we may restrict adversarial feedback to specific denoising steps, which would require adapting the discriminator by training it on real images corrupted with corresponding levels of noise. To support this setup, we will explore noise-aware training, experiment with alternative loss functions, and apply selective fine-tuning to the diffusion model—for instance, by only updating layers closer to the output or injecting learnable adapter modules at key blocks in the network. We may also explore additional interaction mechanisms between the two architectures. Progress will be evaluated using perceptual metrics (e.g., FID, IS (Betzalel et al., 2022b)) and qualitative inspection. The modular structure of our setup allows us to adapt the scope and focus of the project based on early outcomes while keeping development manageable.

References

Allahyani, M. et al. (2023). "DivGAN: A diversity enforcing generative adversarial network for mode collapse reduction". In: *Artificial Intelligence* 317, p. 103863. DOI: 10.1016/j.artint.2023.103863.

Arnaud58 (2020). *Flickr-Faces-HQ Dataset (FFHQ)*. https://www.kaggle.com/datasets/arnaud58/flickrfaceshq-dataset-ffhq. Accessed: 2025-04-14.

Betzalel, E. et al. (2022a). "A study on the evaluation of generative models". In: *arXiv preprint arXiv:2206.10935*. URL: https://arxiv.org/abs/2206.10935.

Betzalel, O. et al. (2022b). *A Study on the Evaluation of Generative Models*. arXiv: 2206.10935 [cs.CV]. URL: https://arxiv.org/abs/2206.10935.

Karras, T., S. Laine, and T. Aila (2020). "Analyzing and improving the image quality of StyleGAN". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119.

Kim, T. et al. (2024). "Diffusion-Driven GAN Inversion for Multi-Modal Face Image Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rombach, R. et al. (2022). "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695.

Salimans, T. and J. Ho (2022). "Progressive distillation for fast sampling of diffusion models". In: *arXiv preprint arXiv:2202.00512*. URL: https://arxiv.org/abs/2202.00512.

Zhang, L. and M. Agrawala (2023). "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1836–1846.