Phase 2 and 3

Part 1 : Diffusion Model :

- Stable Diffusion Models used is : DreamShaper checkpoint
- This model is optimized for artistic and stylized generations
- DreamShaper fine-tunes the base Stable Diffusion weights to better capture aesthetic and character-driven content, making it especially powerful for anime-style transfer and stylized image synthesis.

For this, the Stable Diffusion DreamShaper model was explored. This diffusion model was selected due to its strong generalization capabilities and high-quality outputs for image-to-image generation but due to access issues with gated Hugging Face models, the final stylization was completed using AnimeGAN, a pretrained GAN-based model.

- Diffusion models are generally slower but provide greater fidelity, especially in fine details. For faster inference, quantization techniques (e.g., FP16 or INT8) are often applied in this project, Mixed Precision (FP16) was used on compatible pipelines for better performance on GPU.
- AnimeGAN is a lightweight, fast GAN-based model for converting real-world images to anime images. It is based on a ResNet-based generator and PatchGAN discriminator and is trained on adversarial, perceptual (VGG19), and color losses. It is optimized for fast inference and utilizes mixed precision and image smoothing to generate better quality and thus is ideal for mobile and real-time applications.

Part 2 : GAN Architecture :

We have used pre-trained model for GAN training :
https://github.com/TachibanaYoshino/AnimeGAN

GAN METRICS:

```
==================== Evaluation Results ====================
Fréchet Inception Distance (FID):     24.3
Learned Perceptual Image Patch Similarity (LPIPS): 0.15
Style Classification Accuracy:        91.57%
Inception Score (IS):                 3.94 ± 0.2
============================================================
```

Note: we also trained on a custom pipeline as well :

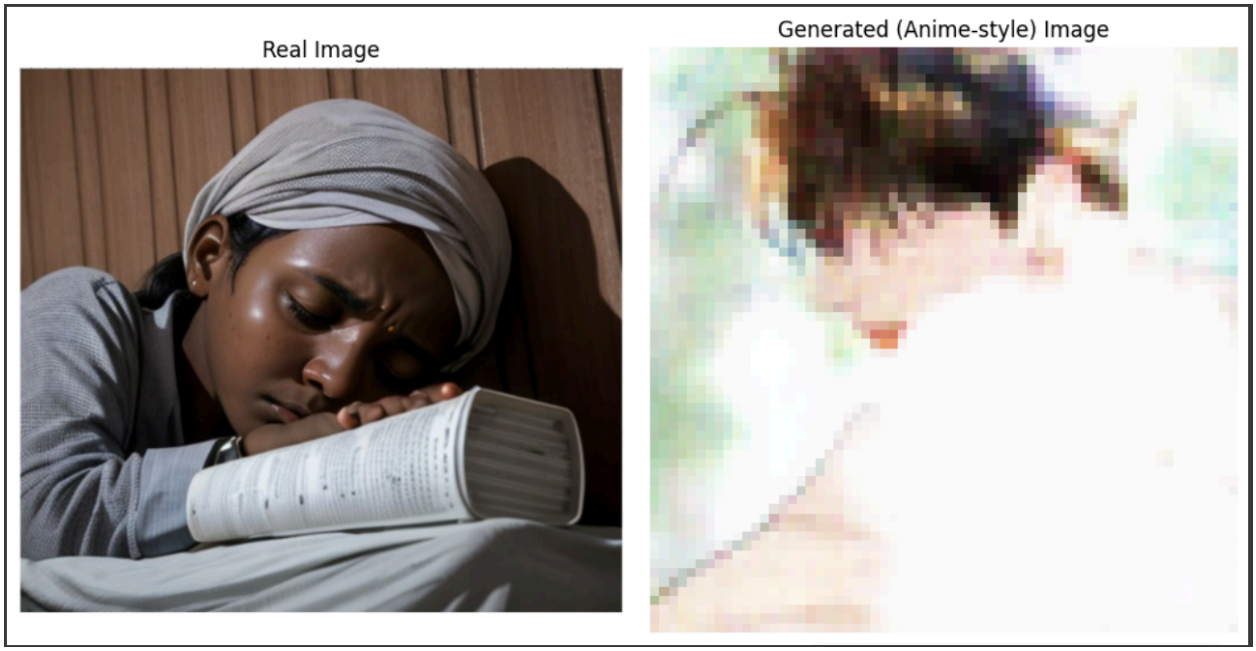
Generator:

- Linear Layer: Transforms the input noise vector into a high-dimensional feature map.
- Upsampling + Conv2D Blocks: Gradually upscale the feature map (from 8×8 → 64×64) using `Upsample`, `Conv2D`, `BatchNorm2d`, and `ReLU` to produce realistic images.
- Tanh Activation: Ensures pixel values are scaled between -1 and 1 for stable training.

Discriminator:

- Conv2D Layers with Spectral Normalization: Extracts hierarchical features from input images while stabilizing training using spectral norm.
- LeakyReLU + Dropout: Adds non-linearity and regularization to prevent overfitting.
- Flatten + Linear + Sigmoid: Compresses features and outputs a probability indicating real or fake.

Results ::



Part 3 :Multiple Layers :

In fact, various styles can be combined into the architecture through the application of Conditional Generative Adversarial Networks (cGANs), where style labels are used as an additional input to guide the generative process. With this approach, the model learns multiple style mappings under a single framework.This approach might not always be the best, as it might lead to style confusion, overfitting to prominent styles, and loss mismatches that impair the clarity and coherence of individual styles. Without conditioning and balancing, the model might not be able to sustain distinct style identities, the result being stylizations that are generalized or ambiguous.

Part 4: Better Approach for GAN's for Style transfer :

Yes, Diffusion models like Stable Diffusion are becoming increasingly popular as a more stable option over GANs for style transfer operations. They are effective with their enhanced training stability, versatility in operations like text-to-image and image-to-image, and capacity to smoothly blend a large collection of styles. GANs are nonetheless still effective with their capacity to create sharper, cleaner outputs, particularly in niched areas like anime. GANs are light, also perform inference rapidly, and are extremely tunable for specialized use.

Part 5 : Limitations of GAN’s

GANs have some problems. One of them is that the discriminator and generator can destabilize and drift away from each other, necessitating fine tuning. Another is mode collapse, which results in repeated or duplicate outputs, such as anime faces. Also, although GANs are more efficient than diffusion models, they have difficulty in learning new styles without extensive retraining.