

**Project Title**

Automated Image Generation (CCDS24-0163)

**Project Summary**

This project will develop deep neural networks that is capable of generating various new images according to some pre-defined conditions or requirements.  
The generated images can be applied to train deep networks.

**FAQ with Professor Lu Shijian**

Q1: Your expectations and targeted scope of the project

*A1: It depends on your commitment. It can be reimplement, benchmark, and analyse existing methods, or you may develop something new that could mitigate some existing problem.*

Q2: Any potential resources I can start to research and read into before we start

A2: *You may google the Internet to read papers regarding Gaussian Splatting, download open-source codes, and see whether can reproduce some report results.*

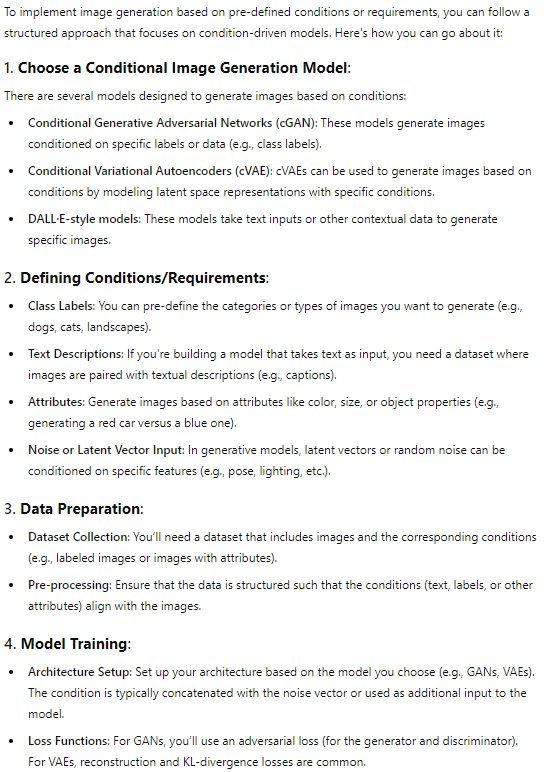
**Neural Networks for Image Generation**

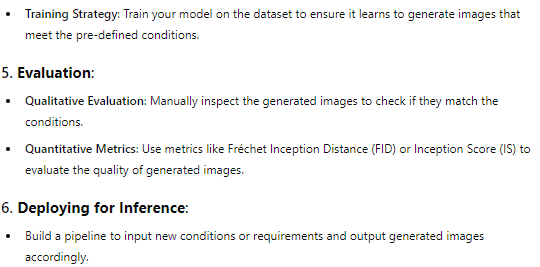
1. Generative Adversarial Networks (GANs):
   1. Architecture: GANs consist of two neural networks: a generator and a discriminator. The generator creates images, while the discriminator evaluates them against real images, learning to distinguish between real and generated images. Over time, the generator improves, producing increasingly realistic images.
   2. Applications: GANs are widely used in generating realistic images, deepfake creation, art generation, image super-resolution, and style transfer.
2. Variational Autoencoders (VAEs):
   1. Architecture: VAEs consist of an encoder that compresses images into a latent space and a decoder that reconstructs images from this latent space. Unlike traditional autoencoders, VAEs impose a probabilistic structure on the latent space, allowing for the generation of new images by sampling from this space.
   2. Applications: VAEs are used for image generation, especially in cases where the generated images need to have a degree of variation, such as generating new faces, objects, or even interpolating between different images.
3. Diffusion Models:
   1. Architecture: Diffusion models work by iteratively adding noise to the image and then learning to reverse this process to generate a new image from noise. They are trained to denoise images by modeling the distribution of the data.
   2. Applications: Diffusion models are known for generating high-quality images and have been used in state-of-the-art image generation models like DALL-E 2 and Imagen.
4. Autoregressive Models:
   1. Architecture: Autoregressive models generate images pixel by pixel, predicting the next pixel based on previous pixels. These models often use a sequential approach where the output of one step is fed into the next.
   2. Applications: Examples include PixelRNN and PixelCNN, used for generating images by predicting each pixel's value based on the context provided by already generated pixels.
5. Transformer-Based Models:
   1. Architecture: Transformers have been adapted for image generation by treating images as sequences of patches or tokens, similar to how text is processed. These models leverage self-attention mechanisms to capture relationships across different parts of the image.
   2. Applications: Models like DALL-E and ImageGPT use transformer architectures to generate images from text prompts or other input formats.

**17/10/2024**

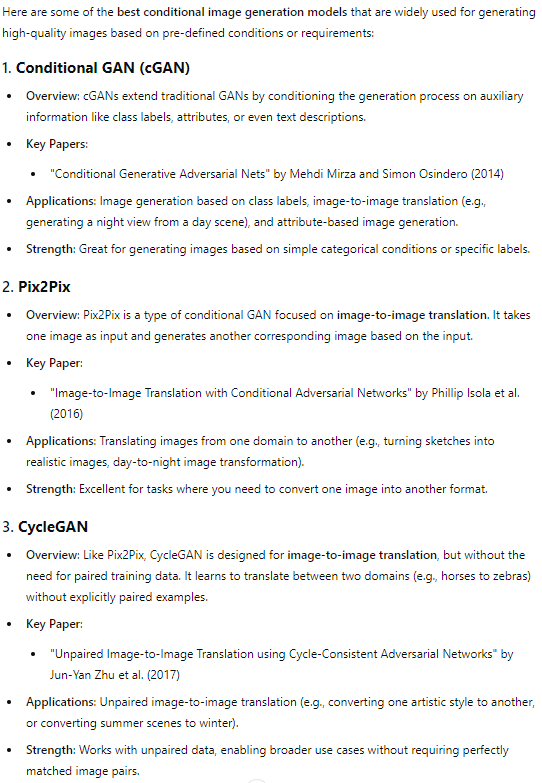
**From Project Summary:** This project will develop deep neural networks that is capable of generating various new images according to some pre-defined conditions or requirements.

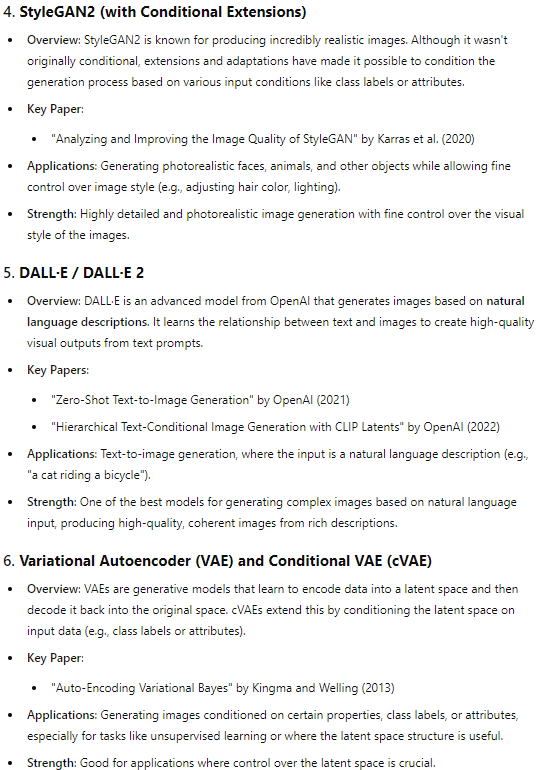
1. **Tackling “generation according to some pre-defined conditions”**

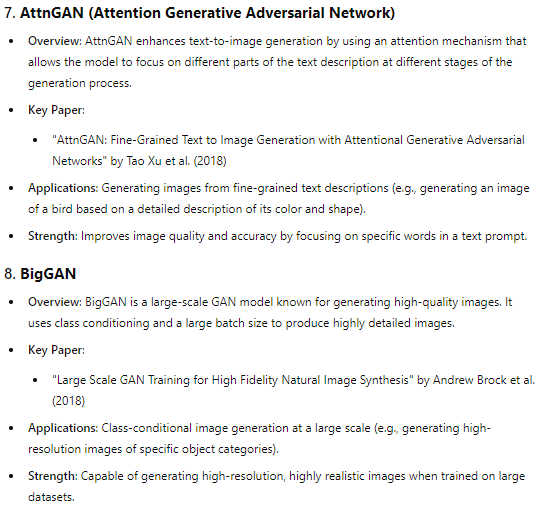
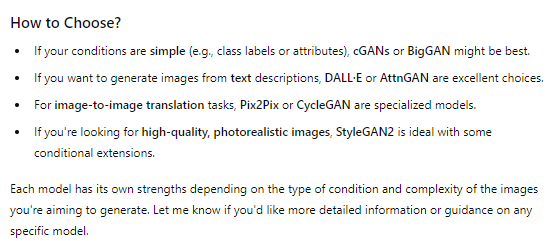




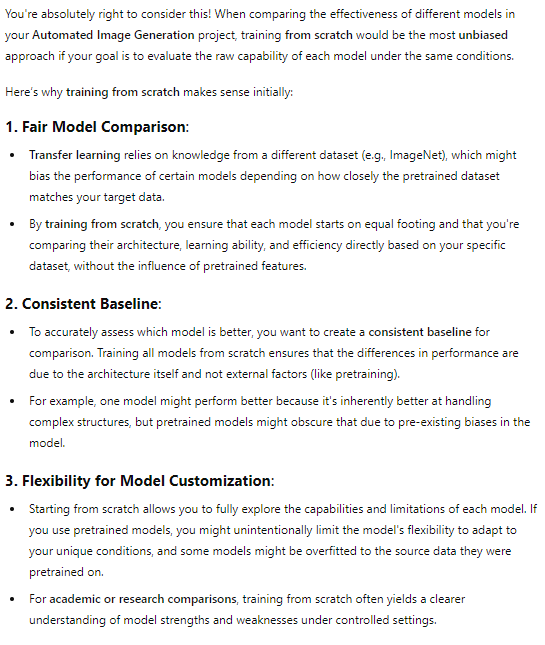
1. **Best Conditonal Image Generation Models**

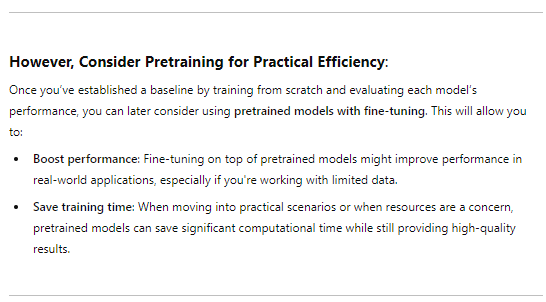


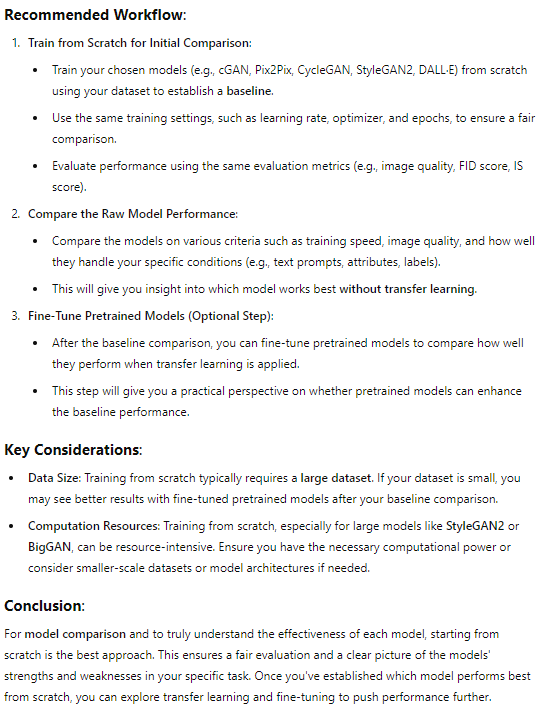


1. **WorkFlow**







Here's a breakdown of the neural networks used by the tools mentioned, highlighting the specific architectures and techniques behind each:

**1. DALL-E 2 by OpenAI**

* **Neural Network:** Diffusion Models and Transformers
* **Details:** DALL-E 2 uses a combination of a diffusion model for generating images and a transformer model for understanding and processing text prompts. The diffusion model iteratively refines images from noise, while the transformer helps in translating the text description into a visual representation.

**2. Midjourney**

* **Neural Network:** Diffusion Model (specifics not publicly detailed)
* **Details:** Midjourney also utilizes a diffusion-based approach, similar to DALL-E 2, to create images from text. The exact architecture details are proprietary, but it is optimized for generating artistic and stylized visuals.

**3. Stable Diffusion**

* **Neural Network:** Latent Diffusion Model (LDM)
* **Details:** Stable Diffusion uses a Latent Diffusion Model, which operates in the latent space (compressed representation) to efficiently generate high-quality images. LDMs make the diffusion process more computationally feasible, enabling high-resolution image generation.

**4. Artbreeder**

* **Neural Network:** BigGAN and StyleGAN
* **Details:** Artbreeder primarily leverages BigGAN for generating photorealistic images and StyleGAN for controlling the styles and features of the generated images. Users can manipulate these latent spaces to blend and evolve images collaboratively.

**5. Runway ML**

* **Neural Networks:** BigGAN, StyleGAN, CLIP, among others
* **Details:** Runway ML offers a variety of models, including BigGAN (Generative Adversarial Network for generating high-quality images), StyleGAN (for creating highly controllable images), and CLIP (Contrastive Language-Image Pretraining, used for text-to-image tasks).

**6. DeepArt**

* **Neural Network:** Convolutional Neural Networks (CNNs) with Style Transfer
* **Details:** DeepArt uses CNNs for neural style transfer, where it takes the style of one image and applies it to the content of another image. This process involves matching the style statistics of the style image with the content image's features.

**7. NightCafe**

* **Neural Networks:** VQGAN+CLIP, Style Transfer
* **Details:** NightCafe employs VQGAN (Vector Quantized Generative Adversarial Network) in combination with CLIP to generate images from text prompts. It also uses neural style transfer techniques for applying artistic styles to images.

**8. Craiyon (formerly DALL-E mini)**

* **Neural Network:** Variational Autoencoder (VAE) and Transformers
* **Details:** Craiyon uses a simplified version of the architectures found in models like DALL-E, involving a VAE for image generation combined with a transformer model to process text input.

**9. DeepDream**

* **Neural Network:** Convolutional Neural Networks (CNNs)
* **Details:** DeepDream operates by using pre-trained CNNs (often using models like Inception) to enhance patterns within an image. It works by amplifying features that the network recognizes, creating surreal, dream-like visuals.

**10. Lensa AI**

* **Neural Network:** Convolutional Neural Networks (CNNs)
* **Details:** Lensa AI uses CNNs for various photo-editing tasks, including face retouching, background removal, and applying artistic filters. These networks are optimized for recognizing and enhancing facial features and other key elements in images.

**Model Analysis:**

**1. Fréchet Inception Distance (FID)**

* **Purpose**: Measures the distance between the feature distributions of real and generated images.
* **How it works**: Uses the Inception network to extract features from the real and generated images and then computes the Fréchet distance between these feature distributions.
* **Interpretation**: Lower FID scores indicate better performance, meaning the generated images are closer to the real images in terms of feature distribution.

**2. Inception Score (IS)**

* **Purpose**: Evaluates both the quality and diversity of generated images.
* **How it works**: The Inception model is used to classify generated images, and the score combines the confidence of these classifications (for quality) and the diversity of the classified labels (for diversity).
* **Interpretation**: Higher IS scores indicate better performance, with high-quality and diverse images.

**3. Kernel Inception Distance (KID)**

* **Purpose**: Similar to FID, but uses a different approach to compare the distributions of features between real and generated images.
* **How it works**: Measures the squared Maximum Mean Discrepancy (MMD) between Inception features of real and generated images.
* **Interpretation**: Lower KID values indicate better alignment between real and generated image distributions.

**4. Precision and Recall**

* **Purpose**: Measures the fidelity (precision) and diversity (recall) of generated images.
* **How it works**: Precision measures the proportion of generated images that are similar to real images (fidelity), while recall measures the proportion of real images that have close generated counterparts (diversity).
* **Interpretation**: Higher precision and recall indicate better performance. A trade-off might be necessary between precision and recall.

**5. Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR)**

* **Purpose**: Evaluate the pixel-level differences between generated and real images, primarily used in tasks like image reconstruction.
* **How it works**: MSE calculates the average squared difference between corresponding pixels, while PSNR measures the ratio of the maximum possible pixel value to the noise.
* **Interpretation**: Lower MSE and higher PSNR values indicate better image quality.

**6. Structural Similarity Index Measure (SSIM)**

* **Purpose**: Evaluates the structural similarity between generated and real images, considering luminance, contrast, and structure.
* **How it works**: Compares local patterns of pixel intensities that have been normalized for luminance and contrast.
* **Interpretation**: SSIM values range from -1 to 1, with values closer to 1 indicating higher structural similarity.

**7. Diversity Metrics**

* **Purpose**: Assess the variety of generated images, ensuring that the model doesn't produce repetitive or limited types of images.
* **Common Methods**:
  + **LPIPS (Learned Perceptual Image Patch Similarity)**: Measures the perceptual distance between pairs of generated images.
  + **Mode Score**: Similar to Inception Score but penalizes models that generate images concentrated in fewer modes (classes).

**8. Human Evaluation**

* **Purpose**: Captures qualitative aspects that might not be fully represented by quantitative metrics.
* **How it works**: Human evaluators rate the generated images on aspects like realism, aesthetic appeal, and diversity.
* **Interpretation**: Human evaluations provide subjective insights that complement quantitative metrics.

**9. Perceptual Path Length (PPL)**

* **Purpose**: Measures the smoothness of the latent space in generative models like GANs.
* **How it works**: Computes the perceptual difference between images generated from linearly interpolated latent vectors.
* **Interpretation**: Lower PPL values indicate smoother transitions and more consistent image generation.

**10. Entropy**

* **Purpose**: Measures the randomness and diversity of the generated images.
* **How it works**: Calculates the entropy of the predicted label distribution for generated images.
* **Interpretation**: Higher entropy indicates greater diversity, while lower entropy suggests mode collapse or less diversity.

**Combination:**

**1. FID + Inception Score (IS)**

* **Use Case**: Widely used in evaluating GANs, especially when both image quality (realism) and diversity are important.
* **Reason**:
  + FID captures the similarity between the distribution of real and generated images, providing a measure of how close the generated images are to the real ones.
  + IS evaluates both the quality and diversity of generated images, ensuring that the model produces high-quality images across different categories.

**2. FID + Precision and Recall**

* **Use Case**: Used when a more detailed analysis of image fidelity (precision) and diversity (recall) is required.
* **Reason**:
  + FID provides a holistic measure of quality.
  + Precision and Recall give insights into the trade-offs between generating images similar to real ones (precision) and covering the diversity of real data (recall).

**3. KID + FID**

* **Use Case**: Suitable for scenarios where robustness in evaluation is needed, as KID is unbiased and can be more reliable in some cases than FID.
* **Reason**:
  + FID is a well-known and accepted standard.
  + KID offers a complementary perspective, using different assumptions and avoiding some of the biases in FID.

**4. SSIM + MSE/PSNR (for Image Reconstruction Tasks)**

* **Use Case**: Commonly used in evaluating image reconstruction tasks or VAEs, where pixel-level accuracy and structural similarity are important.
* **Reason**:
  + MSE and PSNR provide a straightforward measure of pixel-level differences.
  + SSIM adds a perception-based assessment, focusing on structural similarity and visual quality.

**5. LPIPS + FID**

* **Use Case**: Effective when perceptual quality and image diversity are the main concerns, often used in GAN evaluations.
* **Reason**:
  + LPIPS measures perceptual differences between images, providing a human-like assessment of quality.
  + FID remains the go-to for overall quality and realism comparison.

**6. Human Evaluation + FID/IS**

* **Use Case**: Necessary when the final application of the images involves human judgment, such as in creative industries.
* **Reason**:
  + Human evaluation captures qualitative aspects that automated metrics might miss.
  + FID or IS provide the necessary quantitative backing.

**7. PPL + FID/IS (for Latent Space Exploration)**

* **Use Case**: Common in applications where smooth transitions and latent space consistency are important, such as style transfer or interpolations in GANs.
* **Reason**:
  + PPL assesses the smoothness of the latent space, ensuring that small changes in the latent vector lead to smooth changes in the generated images.
  + FID or IS assess overall quality and diversity.

**8. FID + Entropy (for Mode Collapse Analysis)**

* **Use Case**: Effective in evaluating models where mode collapse (i.e., the model generating limited types of images) is a concern.
* **Reason**:
  + FID ensures overall quality.
  + Entropy measures the diversity, helping to detect mode collapse by observing the distribution of generated categories.