Generating Singaporean Art with Diffusion Models

ML Singapore 6

Contributions:
Edison Siow (A0233856X): Stable Diffusion, DreamBooth, LoRA, ControlNet
Kang Yue Ran (A0230366L): Stable Diffusion, DCGAN, Image Scraper, Inpainting
Germaine Lee (A0239523A): Stable Diffusion, VAEs, Image Repo, Upscaling

Abstract

To enhance the artistic outputs of amateurs and to expedite the creative process of artists, we propose a generative art AI application which will allow users to create stunning artwork given a textual description and an optional reference image. In order to tailor our application to the growing demand for Singaporean artwork, we also provide a proof-of-concept finetuning of an open-source Stable Diffusion model, depicting Singaporean icons and political figures in various user-requested styles.

The arts scene in Singapore is burgeoning, and Sigaporeans have a growing appetite for artwork depicting Singaporean culture. According to the latest statistics, the yearly contribution of the arts industry rose from 1.5 billion in 2013 to 1.8 billion in 2018. (Ministry of Culture, Community, and Youth, 2021) Furthermore, 89% of Singaporeans agreed that the arts gave them a better understanding of different cultures and backgrounds, while 82% agreed that the arts gave them a greater sense of belonging to Singapore. (National Arts Council, 2019)

Although there is clearly a demand for locally-produced art, the unfortunate reality is that the volume of local artists will always be limited, with employment in the arts remaining constant at about 26,000 from 2013 to 2018 despite the growing interest. (Ministry of Culture, Community, and Youth, 2021) Creating good art is time-consuming and often tedious; it takes years of practice for amateurs to achieve professional-level skills, and an artistic project can take a long time to complete, with famous Renassiance artists like Leonoardo da Vinci taking years to produce his paintings. (Encyclopedia Britannica, 2023)

However, with recent developments in generative AI art models becoming increasingly flexible and lifelike, we believe this could help novice and veteran artists alike expedite the art creation process by 1) allowing lightning-fast iteration of conceptual art and 2) granting artists access to a variety of artstyles, while 3) retaining the fidelity of the Singaporean subject-matter. Furthermore, we also consider commonly-cited ethical issues regarding the impact of AI art generators on artist livelihoods, and provide some suggestions to mitigate them.

Application details

Our application is an end-to-end AI image generation tool designed to be used by Singaporean artists to expedite the client-artist feedback loop. Specifically, given a textual description or reference images by the client, the artist should be able to quickly input them into the tool, produce a few varied concept artworks, refine them within our tool with their own skills, and return them to the client. When the client returns with the feedback, the artist may refine these artworks in our app based on the textual feedback, and so on. An example is shown in Figure 1.

Figure 1: The client asks for 'a portrait of Pritam Singh' (left) in the artstyle of a reference image (right). Our tool generates the requested image (middle) in minutes with high fidelity.

To tailor our application for the Singaporean market, we will pretrain our ML model to include Singaporean concepts such as Hawker Centres, Merlion, Garden by the Bay, and prominent Singaporean figures such as Lawrence Wong and Pritam Singh. Although such concepts are immediately recognizable for Singaporeans, based on our preliminary research in current commercial AI art products, we have found that they are unable to produce these with high fidelity, presumably due to the lack of text-image pairs of these subjects on the internet.

Casual users will be presented with a simple interface for producing images, only requiring to fill in a text field for their prompt and optionally the ability to add reference images. On the other hand, experts familiar with diffusion models can finetune our application to their needs as we provide

them an interface to tweak the control parameters, which can include the sampler and the number of diffusion steps. Finally, we will also allow high-quality upscaling and 'in-painting' (filling in only selected pixels in an image).wi

Comparison to commercial and open-source software

Given the plethora of generative AI tools flooding the market recently, we conducted some preliminary research to see if there is any existing software on the market that could meet our requirements of our application. As mentioned, we took a selection of Singaporean icons and Singaporean figures. We then proceeded to generate images using a set of free-for-use generative AI software on the internet, such as OpenAI's DALLE-2[1], Bing Image Creator[2], and Stable Diffusion base model 2.1[3] (link to repo for image generated). Out of all of these, we found DALLE-2's images to be the most aesthetically pleasing. However, its outputs for the hawker centres, Lawrence Wong and Pritam Singh, failed to capture the actual likeness of the subjects. For budgetary reasons, we were unable to compare with paid services like Midjourney[4] and Adobe Firefly[5], but as they likely use generalized datasets scraped from the internet similar to Stable Diffusion 1.5, it is unlikely that they would fair better. Furthermore, they lacked the ability to condition their outputs on reference images and only text. All generated images may be viewed in our image repository. [6] Hence, it was clear that we needed to train our own custom model to meet the requirements for our application.

Dataset details

Due to the lack of datasets containing uniquely Singaporean images, we had to scrape for our data. Initially, we developed our own image scraper script using Python and ChromeDriver, then cropped and resized the scraped images using the Image Module from Python PIL. However, after settling on Stable Diffusion as our model of choice, we required high-quality images of the subject in varied angles and environments, which a simple scraper was unable to achieve. In order to achieve the best possible results, we opted to manually save images from their sources and crop them to a resolution of 512x512 pixels. Since this method was highly time-consuming, we were only able to collect a total of 120 images from a limited number of categories, including the Gardens By The Bay, Hawker Centres, the Merlion, Lawrence Wong and Pritam Singh. The set of images collected is linked in our image repository.

Different generative image AI techniques

To find out which model was best suited for our requirements, we did some preliminary testing on a few well-known deep-learing generative image AI models, such as Deep Convolutional Generative Adversarial Networks (DCGAN), Variational Autoencoders (VAEs), and Stable Diffusion.

Deep Convolutional Generative Adversarial Network (DCGAN)

Generative Adversarial Networks (GAN) are a powerful framework for teaching deep learning models to first capture the distribution of the training data then generate new data from this distribution. GANs consist of two models, a generator and a discriminator. The generator's role is to produce "fake" images that resemble the training images, while the discriminator's job is to differentiate between real training images and the fake images created by the generator (Goodfellow et al., 2014). A Deep Convolutional Generative Adversarial Network (DCGAN) is a type of GAN that employs convolutional and convolutional-transpose layers in the discriminator and generator respectively (Radford et al., 2016).

When we experimented on the PyTorch implementation of DCGAN[7], we found that the losses started to diverge. When using DCGAN on our relatively small set of training data, it is found that it is extremely easy for the Discriminator to overfit on our small training data, and distinguish between real and fake images (Karras et al. 2020). As a result, the Generator is unable to learn the underlying distribution of the data from the limited sample size, preventing us from modelling the adversarial nature of the problem. We have included the results of our experiment in the image repository.

Variational Autoencoder

VAEs are a form of data compression where the encoder compresses the data from its initial space to the latent space. Subsequently, the decoder decompresses the data. For our case, VAEs utilize convolutional layers as their encoder/decoders, and they attempt to minimize the Kullback-Leibler divergence between the true and generated distributions, which is fomulated as the evidence lower bound (ELBO) due to the former's intractability. Due to the regularization caused by this data encoding, the hope is that the VAE learns a robust decoding from the latent space which will allow it to generate varied images (Rocca, 2019).

When we attempted to train using the Pythae VAE implementation[8], we obtained poor results. The images generated were blurry and almost indistinguishable. One drawback of VAEs is that they have a tendency to produce blurry generative samples even if the images in the training distribution were sharp. Due to the asymmetry of the Kullback-Leibler divergence, the VAE model does not penalise heavily if the generated samples are more diverse than the original data. Hence, VAEs tend to suffer from blurry generated samples and reconstructions compared to the images they have been trained on. (Bredell et al., 2023)

Stable Diffusion

Stable Diffusion is a powerful latent diffusion model capable of producing stunning images with only a short user text

[1] https://openai.com/product/dall-e-2

[2] https://www.bing.com/create

[3] https://huggingface.co/spaces/stabilityai/stable-diffusion

[4] https://www.midjourney.com/app/

[5] https://firefly.adobe.com/

[6] https://github.com/germainelee02/CS3264-Project

[7] https://pytorch.org/tutorials/beginner/dcgan_faces_tutorial.html

[8] https://github.com/clementchadebec/benchmark_VAE

prompt (Rombach et al., 2022). Given some architectural extensions, it is also able to perform other tasks including producing an image based on another image and a short text prompt (i.e. image to image), image upscaling, and fine tuning on unseen text-to-image pairs.

The text-to-image pipeline comprises three main components. The first component is the text encoder from the CLIPText model, a multimodal visual and language model trained in a contrastive fashion (OpenAI, 2021). Briefly, it attempts to maximize the cosine similarity of paired text and image embeddings, and minimize the unpaired embeddings. CLIPText is capable of zero-shot labelling of images, hence its text encoder is useful in generating latent vectors which are informative for image generation.

The other two components are the variational autoencoder (VAE) and denoising U-Net, which work in tandem to create the diffusion model. In the forward diffusion process, the image is first transformed into a lower-dimensional latent space by the VAE. Then, small amounts of Gaussian noise are applied in each successive step of a Markov Chain according to a variance schedule. Over enough steps, the image distribution approximates Gaussian noise.

The backwards diffusion process is also defined as a Markov chain where the U-Net predicts the Gaussian noise to subtract based on the latent image in the preceding step. The U-Net is trained to maximize the variational lower bound of the negative log likelihood between the target distribution (ground truth images) and its generated images. Using a reparameterization trick (Ho, Jain, Abbeel, 2020) the training loss can be simplified to:

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2]$$

The predicted noise by the U-Net is then subtracted from the latent image at every step by a sampler. The sampler controls the amount of noise removed from the image at every step, typically in a decreasing manner so that the image is allowed to converge. There are choice of many different possible samplers (eg. Euler, Heun, LMS etc.) and we elaborate on our choice of sampler in the 'Experimental Details' section.

During the reverse process, the UNet may also be conditioned on external data (text/images etc.) and incorporated into each step of the diffusion process via a cross attention mechanism. This allows it to condition its output on text embeddings from CLIPText (text-to-image), or latent representations of additional reference images (image-to-image) during the denoising.

Stable Diffusion was trained on subsets of the LAION-5B dataset, a collection of over 5 billion text-image pairs, making it infeasible for individuals such as ourselves to train it from scratch. Thankfully, StabilityAI has made the Stable Diffusion project open source and the final weights downloadable[9], making it the only open-source diffusion model amenable for personal use.

Additionally, several methods have been developed for finetuning Stable Diffusion models on a custom subject if the current text description does not produce the desired visual output. These include Dreambooth, textual inversion, LoRAs, hypernetworks. According to statistics gathered on the community Stable-Diffusion model-sharing platform CivitAI[10], the Dreambooth technique seems the most popular and effective, while LoRAs are well-suited for low-memory GPUs with nearly as good results. Hence, we experimented with these two techniques to yield promising results.

Dreambooth

Dreambooth (Ruiz et al., 2022) was a technique developed by Google Research for finetuning diffusion models, which was subsequently imported into Stable Diffusion by the open-source community. Given a small set of images (the authors used 5), Dreambooth is able to bind a unique identifier to the visual concept depicted in the images, which one can invoke in novel contexts not captured in the training data.

The main novelty of the Dreambooth technique is the class-specific prior preservation loss (PPL) used alongside the reconstruction loss when training the model. PPL penalizes the model from diverging from its prior heavily, which makes it highly effective at incorporating new visual-text concepts without losing the flexibity of the base model.

Low-Rank Adaptation (LoRA)

LoRA (Hu et al., 2021) was a technique created to finetune large language models without the prohibitive cost of retraining the entire model. All preretrained weight matrix $\mathbb{R}^{d\times k}$ are decomposed into two trainable lower-rank matrices low-rank matrices B $\in$R$^{d\times r}$ and A $\in$R$^{r\times k}$ where $r << min(d, k)$ while the pretrained weights are frozen. The modified forward pass is then given by:

$$h = W_0x + \nabla Wx = W_0x + ABx$$

where ∇W is further scaled by a hyperparameter.

Since the technique is model-agnostic, the open-source community adopted it for SD as well, with the output of LoRA (the trained low-rank matrices) to be distributed and injected into any generic SD model.

ControlNet

ControlNet (Zhang and Agrawala, 2023) is an auxilary architecture designed to augment large diffusion models in adhering to task-specific input conditions. Such conditions can comprise reference images as in Figure 1, or even depth maps or pose mappings.

Briefly, ControlNet works by training a separate copy of weights for every encoder block of the original diffusion model, while freezing the original layers during training. The outputs of these trainable weights are added to the skip connections and middle block in the U-Net, after passing through the ControlNet's 'zero-convolutions'. These 'zero-convolutions' are essentially 1x1 convolutional blocks which are intialized with weight and bias 0, since this initialization ensures that the untrained ControlNet simply gives the output of the diffusion model.

[9]https://github.com/CompVis/stable-diffusion

[10]https://www.youtube.com/watch?v=dVjMiJsuR5ot=4s

Using various ControlNet models trained on separate curated datasets, the authors were able to achieve precise and attractive imagery using a range of input conditions. For the sake of this project, we will only mention 3:

- *Canny* - Images were processed with a Canny edge detector and the corresponding ControlNet model was trained on over 3 million text-image pairs. The Canny model is useful for replicating fine details in the reference image, for instance Pritam Singh's stubble in Figure 1.
- *User Sketch* - Human sketches were synthesized by using HED boundary detection (Xie and Tu, 2015) and data augmentations, and the model was trained on 500,000 sketch image pairs. This is useful for rendering rough sketches, as in Figure 4.
- *Depth (large-scale)* - Authors used the Midas (Lasinger et al., 1998) technique to obtain 3 million depth-image-caption pairs from the internet to train the model. this is useful for exactly replicating the layout of a room, eg. Figure 5

Suitability of Stable Diffusion for our application

To test Stable Diffusion's ability to generate Singaporean subject matter in dynamic contexts, we trained Stable Diffusion 1.5 using Dreambooth technique on our dataset, and asked it to produce "a mural of {subject} on a wall". We managed to gather a sample of (in our opinion) aesthetically pleasing renditions which can serve as an inspiration for artists to refine or iterate upon. (Figure 2)

Figure 2: Collage of murals of Singaporean subject-matter produced by SD finetuned with Dreambooth

Additionally, our application will also be able to exploit Stable Diffusion's ability to generate images of varying styles. Besides the base SD model's ability to mimic various styles from its pretraining on the LAION datasets, we can further hone in on a specific artstyle by including outputs of Dreambooth/LoRA/textual inversion/hypernetworks. Figure 1, for instance, was generated with a Chinese-artwork style LoRA found on CivitAI. This will give the artist more flexibility and freedom to vary their outputs for inspiration, and allow us to capture a larger market.

Finally, our proposed application exploits the ability of Stable Diffusion to scale vertically and horizontally. Stable Diffusion's inference, which is essentially the reverse denoising process, can be done within the manner of minutes per image. This is hundreds if not thousands of times of magnitude faster than human-generated art, allowing artists to quickly iterate through a multitude of concept artworks at a lightning-fast pace.

To scale horizontally, given sufficient investment of a million dollars, we would be able to invest in a hundred NVIDIA Tesla P100 GPUs, which should be more than sufficient to serve Singapore's 26,000 artists and beyond with a guaranteed limit of a few minutes latency. (For reference, Midjourney uses a cluster of 9000 cloud GPUS for inference to serve a total of 16 million users.) Compared to the hundreds of millions invested by the government into the arts yearly, this would be a drop in the bucket for potentially hundred-fold returns in artist productivity. Furthermore, these productivity gains could hyper-accelerate the industry and bring Singapore notoriety in the global art scene, increasing the contribtuion of the arts industry to the Singaporean economy significantly.

Experimental Set-up and Outcomes

Finetuning models on Singaporean subject matter

For finetuning, we used the Dreambooth technique with either the base Stable Diffusion 1.5 model for non-human subjects or Stable Diffusion 2.1 model for human subjects due to its better photorealism. We used our collection of 512x512 images as training inputs, with the following hyperparameters for all inputs: 1500 UNet training steps, learning rate of 2e-6, 350 Text Encoder training steps with a learning rate of 1e-6. Training was done on a Colab notebook[11] since none of our GPUs had sufficent VRAM for Dreambooth. The model weights were saved to our Google Drives which we could download and save to our local Stable Diffusion repositories to perform inference.

Inference on Stable Diffusion

All inference was done locally on our computers with NVIDIA RTX 2070 and 3060 GPUS respectively. We utilized the Automatic1111 Stable Diffusion WebUI interface, due to its high volume of innate features and versatility in adding extensions.

Inference on Stable Diffusion also involved a great deal of variable parameters. For brevity, we will list only the most prominent:

- *Sampler*: We used the default Euler ancestral sampler, since from our plots (Figure 3) it seemed to generate the

[11] https://colab.research.google.com/github/TheLastBen/fast-stable-diffusion/blob/main/fast-DreamBooth.ipynbscrollTo=LC4ukG60fgMy

best images in a reasonable amount of time. The Euler sampler formulates the denoising process as a probability flow Ordinary Differential Equation (ODE) problem, and uses the Euler method to solve for it. (Karras et al., 2022) The ancestral variant increases the amount of subtracted noise per step, but also adds a small amount of random noise to introduce stochasticity within the process. This gives varied and more creative results, but makes the process non-deterministic for every run.

- *Classifier Free Guidance (CFG) scale*: CFG Scale is proportional to how much the generated image adheres to the textual prompt and extra conditions. It does so by weighting an unconditional model against a conditionally trained model in the calculation of gradients during denoising (Ho and Salimans, 2022). CFG Scale increases adherence to the prompt, but overly high values often result in image decgradation, hence we stuck to the default value of 7.5.
- *Sampling steps*: We used 50 sampling steps for all images, as we felt it was a good tradeoff between time and quality.

To decide on the best set of hyperparameters for our images, we created several 'XY-plots' of hyperparameters, which displays a grid of sample images for every combination of hyperparameters. This is useful in helping to eliminate hyperparameters with clearly poor convergence and a high number of artifacts. This option will also be available to expert users of our application giving them greater clarity on how to tweak these parameters.

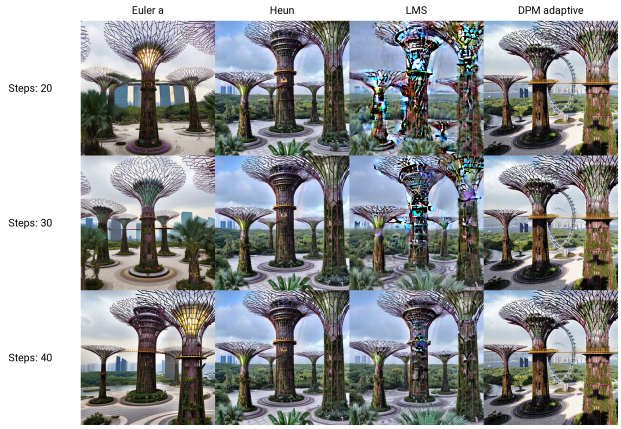

Figure 3: An 'XY plot' created to allow experts to vary hyperparameters in our tool.

Experiments on model flexibility

Using our trained model, we devised 3 levels of difficulty to test its versatility. First, was to recreate the subject matter faithfully using the subject name as prompt, which our model was easily able to achieve (images in repo). Next, if it could create the subject matter in various artstyles, where it largely succeeded after several attempts (eg. see Figure 2). Finally, if it could create the subject matter in dynamic and unprecedented contexts. For human subjects and backgrounds, it was largely successful (eg. Lawrence Wong cooking), but for other subjects there were mixed results (eg. person riding the Merlion), all of which can be found in our image repository.

Experiments on Stable Diffusion with ControlNet

We obtained the most successful results with ControlNet. Inspired by the examples given in the ControlNet Github repository, we devised 3 use-cases of our application by real-world artists. Firstly, an artist may want to generate some concept art given a preliminary sketch of a subject. Using a rudimentary sketch of the Merlion we found on the internet, we created generated 2 x 4 sets of images using the Sketch Controlnet paired with our Dreambooth-trained SD model. The prompts are listed in the Figure 4 captions.

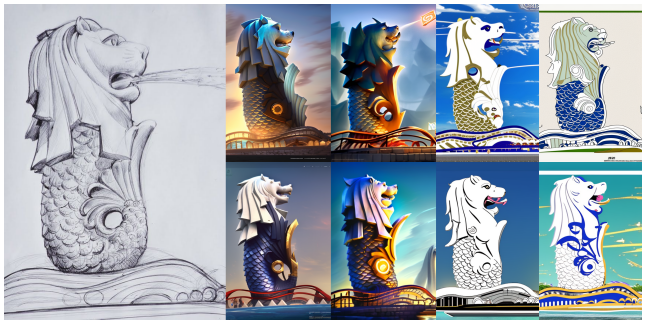

Figure 4: Concept art generated with Sketch ControlNet. First 4 images is described as 'epic, fantasy, highly detailed'. For the next 4 images, 'traditional, mural'

Next, we imagined a designer tasked with creating a futuristic version of a classroom following the specific layout shown in a reference image. While this would be impossible to describe in words, it is no problem with the depthmap ControlNet.

Figure 5: Concept art of futuristic classroom generated with Depthmap ControlNet.

Finally, we attempted to perfectly replicate a person's likeness in a particular style. We found a LoRA on CivitAI trained on Chinese-style artwork[12], and used this in conjunction with Canny ControlNet on an image of Pritam Singh. This resulted in Figure 1 as shown in 'Application details'. The resulting image is notable for its color fidelity, since none of the training data contained images of Sikh men, and were mostly pale-faced women as is customary for this artstyle.

[12]https://civitai.com/models/12597?modelVersionId=20143

Experiments with Upscaling

For particularly nice images, artists may want to upscale them to edit or present to clients. Automatic1111's Stable Diffusion WebUI comes integrated with several upscalers, which we will omit here for brevity. Four our experiments, the different AI upscalers are used on an image of HDB Flats which were generated by our finetuned model. All these images are generated with GFPGAN visibility at maximum strength. The images being upscaled are linked in our repository. Of these, we found ESRGAN and R-ESRGAN to produce the most consistent results. In short, ESRGAN is a SRResNet-based architecture with residual-in-residual blocks, and consists of a mixture of context, perceptual and adversarial losses. (Wang et al., 2018). R-ESRGAN, compared to ESRGAN, is trained with pure synthetic data (Wang et al. 2021).

Experiments with Inpainting

During inpainting, the sampling method we used was Euler a, which seems to produce the best results for blending realistic faces into pictures. We used a CFG scale of 15, to respect the prompt provided for the inpainting process. Then, we chose the 'restore faces' feature to produce more realistic features in the face, and chose to fill the masked content and inpaint only the masked area, to blend our image into the original image better.

We attempted inpainting of faces onto the Merlion, and obtained satisfactory results with animals and generic human faces. However, when we used our Singapore specific data such as Lawrence Wong, the faces were not able to blend well, which we hypothesized to be due to a lack of training images, especially from different angles. All these results can be viewed in our image repository.

Interesting Technical Insights

Looking upon Stable Diffusion's text2img and img2img pipelines as a whole, we found it truly wondorous how many state-of-the-art models and techniques they comprised of. CLIPText embeddings, diffusion models, VAEs, LoRA and ControlNet were all seamlessly integrated to accomplish the seemingly impossible task of sampling from the subspace of attractive images conditioned on text/images. On our part, it allowed us to appreciate the necessary intersection of groundbreaking ML theory and ML engineering.

Conclusion and Ethical Concerns

Through our exploration, we have found that diffusion models can produce high quality images under specific requirements. However, going forward, we do have to address the ethical issues raised by generative image AI models. A large criticism lobbied against image AI models is that the training data was scraped from the internet from artists who did not provide their consent. To avoid this, in our application we hope to train a Stable-Diffusion type model from scratch, but only using data of consenting artists and compensating them fairly.

Another danger that Rombach et al. raised was the creation of images promoting discriminatory/dehumanizing content harming communities or individuals (eg. propaganda against certain religions, deepfakes of politicians etc.). Going forward, we will need to enforce strong barriers against the generation of such content with our app, likely by disallowing sensitive terms relating to race, religion, and public figures.

References

1. Bond-Taylor, S., Leach, A., Long, Y., & Willcocks, C. G. (2021). Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models.
2. Encyclopædia Britannica, (2023). Leonardo da Vinci. Encyclopædia Britannica.
3. Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014). Generative Adversarial Nets.
4. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic models.
5. Ho, J., Salimans, T. (2022, July 26). Classifier-free diffusion guidance.
6. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. (2021). Lora: Low-rank adaptation of large language models.
7. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data.
8. Karras, T., Aittala, M., Aila, T., &; Laine, S. (2022). Elucidating the design space of diffusion-based Generative Models.
9. Lasinger, K., Ranftl, R., Schindler, K., & Koltun, V. (2019). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer.
10. Ministry of Culture, Community and Youth. (2021, May 11). Government Funding for the Arts Sector.
11. Ministry of Culture, Community and Youth. (2023, March). Singapore Cultural Statistics 2
12. National Arts Council. (2020). Population Survey on the Arts 2019.
13. OpenAI. (2021). CLIP: Connecting text and images.
14. Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks.
15. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models.
16. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., Tang, X. (2018). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks.
17. Wang, X., Xie, L., Dong, C., & Shan, Y. (2021). Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data.