

WHISKER WONDERLAND

Members: Joong Ho Choi, Kyle Yang, Zun Wang, Aditti Ramsisaria



CONCEPT

Cats are universally adored - it's hard to deny their charm. Like art, cats come in various personalities and temperaments; some are playful and energetic, others are reserved and quiet. As such, we wanted to celebrate these diverse personalities with diverse art styles. In this project, we seek to showcase human's best friends as the ultimate embodiment of art, able to seamlessly merge with any style or medium. Through our experiments with text-to-image, image-to-image, and audio-to-image technologies, we've generated cats in a wide variety of art styles. As such, the techniques and models we've used are also diverse, including style transfer, stable diffusion, InstructPix2Pix, and DALL-E 2 APIs.

TECHNIQUES

Method 1: Neural Style Transfer

Neural style is an optimization method commonly used in the art domain. It takes in two image inputs, one for style and one for content. In layman's terms, it learns the artistic style of the 'style' image and embellishes the content image with the 'style' learned.

In order to achieve this effect, NST uses a pre-trained model trained on ImageNet- VGG in TensorFlow. To be specific, it uses the model for feature extraction and separation of content and style representations from an image. The architecture of the model performs the training using two loss terms: Content Loss and Style Loss. Content loss is calculated by measuring the difference between the higher-level intermediate layer feature maps. On the other hand, style loss can be measured by the degree of correlation between the responses from different filters at a level.

Method 2: Style Transfer by Relaxed Optimal Transport and Self-Similarity

Similar to the Neural Style Transfer method, STROTSS takes into two images as input, and tries to transfer the style of one image to the content of the other image. It also uses gradient descent to minimize content and style loss. The main difference between this approach and Neural Style Transfer is that this approach uses the Earth Mover's Distance to calculate and minimize loss. This means that instead of minimizing distance between individual patches, STROTSS minimizes the distance between the distribution of features. We could then adjust the content weight, which determines the tradeoff between stylization and content preservation.

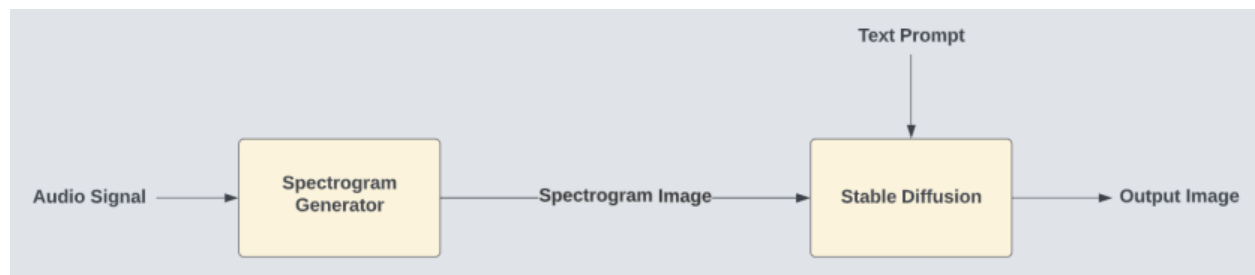
Method 3: Stable Diffusion (Text to Image)

Stable diffusion is a generative model based on diffusion models that gradually turns noise into high-quality images. The difference is that it stabilizes the training process through incorporating stability techniques, such as denoising score matching. The specific model here uses “Euler a” sampling method, which is a generative sampler whose output will keep changing as steps increase (generative samples in a forward rather than random manner). The hyper parameters sampling size and CFG scale are core to outputs of this method.

Method 4: InstructPix2Pix (Image Modification)

InstructPix2Pix is a diffusion-based algorithm for editing images from human instructions: given an input image and a written instruction that tells the model what to do, it follows these instructions to edit the image. The model uses a fine tuned GPT-3 model to generate instructions and edited captions. It then uses Stable Diffusion in combination with Prompt-to-Prompt to generate pairs of images from pairs of captions, to create a dataset of over 450,000 training examples. InstructPix2Pix is a conditional diffusion model trained on a paired dataset generated using GPT-3 and Stable Diffusion.

Method 5: Audio to Image



Audio to Image Pipeline Overview

The website <https://convert.ing-now.com/> converts the audio signal to a spectrogram with customizable colors and intensities using Fourier transforms. The image from the spectrogram then goes through Stability AI’s DreamStudio along with a text prompt for the final image generation. DreamStudio by Stability AI is an AI-powered platform that helps users create personalized dreamlike images and videos. It is based on Stable Diffusion and offers different versions of it with tunable hyperparameters. Stable Diffusion is a latent diffusion model conditioned on the (non-pooled) text embeddings of a CLIP ViT-L/14 text encoder. Stable Diffusion works by generating a sequence of random noise vectors, which are iteratively transformed to generate high-quality images.

Method 6: DALL-E 2 (Image Edit Endpoint)

DALLE-2's image edits endpoint is a diffusion based masking feature that allows for edits to be made to an artwork through a mask. It takes an input image and a mask (both 1:1 ratio), where the mask has a transparent area where the edits should be made. The model will edit the transparent areas of the image according to the text prompt, maintaining the same art style. DALL-E 2 uses the VGG19 network that has been pre-trained on the ImageNet dataset to extract style information and calculate style loss. While the optimization used to minimize style loss is unknown, some variant of gradient descent would be a good guess.

PROCESS

Josh:

I watched tutorial videos on NST that covered the model's architecture and loss functions involved. Afterward, I used git to clone an existing NST github repository. Initially, the original code used Adam and L-BFGS optimisers. Given the limited computation resources, I tested a different number of iterations with each optimiser to get a sense of the tradeoff between output quality and training time. After performing A/B testing with different content and style images along with a varying number of iterations, I explored different optimisers. However, from online resources and tutorials, it seems other optimisers, such as SGD, do not perform as well.

Because it needs to approximate the Hessian matrix and stores more search information, L-BFGS is more computationally expensive and requires more memory. Also, Adam's outputs with VGG16 have shown to be the most generally stable and reliable. Thus, I decided to go with Adam for the optimiser and VGG16 for the architecture.

Aditti:

I wanted to experiment with different input modalities for generating cat images, as I wanted to base my results on my own cats that I've fostered in the past. For this, I relied on diffusion-based image generators. Specifically, I focused on implementations of Stable Diffusion v1.5 and v2. I started by exploring natural language inputs as text prompts to Stable Diffusion over a few iterations to get a sense of how to tune hyperparameters. Based on my cat Katy's personality, I generated images using Stable Diffusion v1.5 until I was able to get something I liked:



(a) Katy the Cat



(b) Generated Image

Stable Diffusion v1.5 Night Cafe

Prompt: "grey cat as a samurai old painting" (weight = 1)

"detailed matte painting, deep color, fantastical, intricate detail, splash screen, complementary colors, fantasy concept art, 8k resolution trending on Artstation Unreal Engine 5" (weight = 0.9)

Seed: 436276

Guidance Scale: 10

Sampling Method: K_LMS

Then, I moved to image-to-image generation. I wanted to use images of my cats and turn them into different characters in line with our concept. For this, I used InstructPix2Pix which was introduced in class. Based on a picture of my cat Jax, I generated the following image:



(a) Jax the Cat



(b) Generated Image

InstructPix2Pix

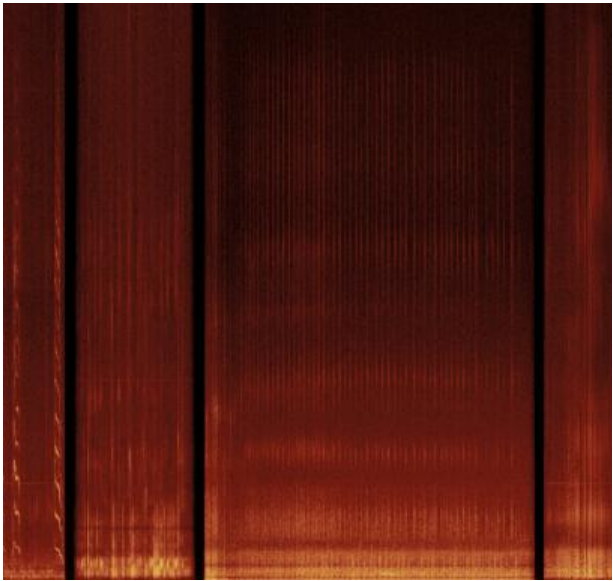
Prompt: "Turn the cat into an astronaut in outer space digital art"

Seed: 10205

Image CFG: 1.5

Text CFG: 12

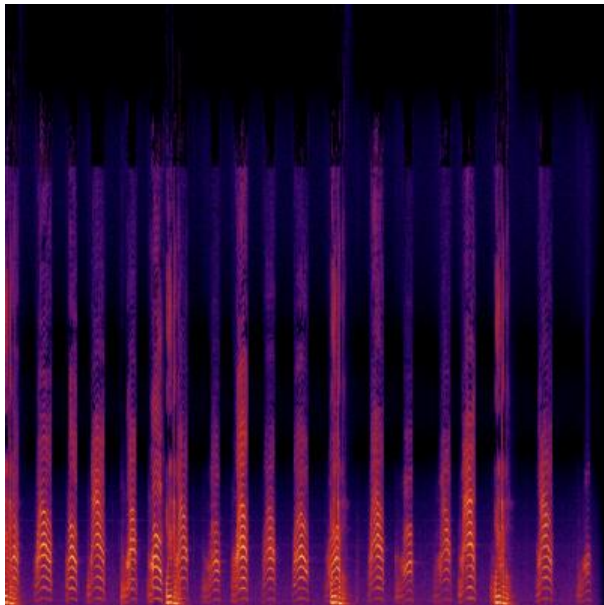
Lastly, I wanted to try audio-to-image generation. Spectrograms were generated from audio clips of cat meows and purrs. These spectrograms were then used as base images in an image-to-image generator which modified them using a text prompt.



(a) Audio Spectrogram



(b) Generated Image 1



(c) Audio Spectrogram




(d) Generated Image 2

S. No	Prompt	Steps	CFG	Version
1	Generate a cat witch using this photo, 4k, oil painting	30	10	Stable Diffusion v2.0 with CLIP Guidance
2	Add a tuxedo cat to this image	50	11	Stable Diffusion v1.5 with K_LMS sampling

Zun:

I wanted to replicate the art styles of artists such as Van Gogh as closely as possible while maintaining the theme of cats. I first tried tuning the CFG scale of stable diffusion to make it generate images as close to the “Cat Painting in Van Gogh Art Style” prompt as possible.

I tested with various CFG scales and discovered that the range of 7-10 yielded the best results. But while the artworks generated were notably similar to Van Gogh’s style, they were a touch too vibrant in color.

Prompt	Steps	CFG	Version	Result
Cat paintings in Van Gogh style	20	7	Stable Diffusion v2.1	

Then, I thought it’d be better to go from image to image. So I used DALL-E 2’s image edit endpoint and provided PNGs of Van Gogh’s paintings and a mask indicating where I envisioned the cat to appear. This turned out to be a much more effective process for what I wanted to create.

Prompt:	A Cat next to the Sunflowers	A Tortoiseshell Cat Next to the Sunflowers	A Purple Cat Sniffing the Sunflowers
---------	------------------------------	--	--------------------------------------





Results:			
----------	---	--	---

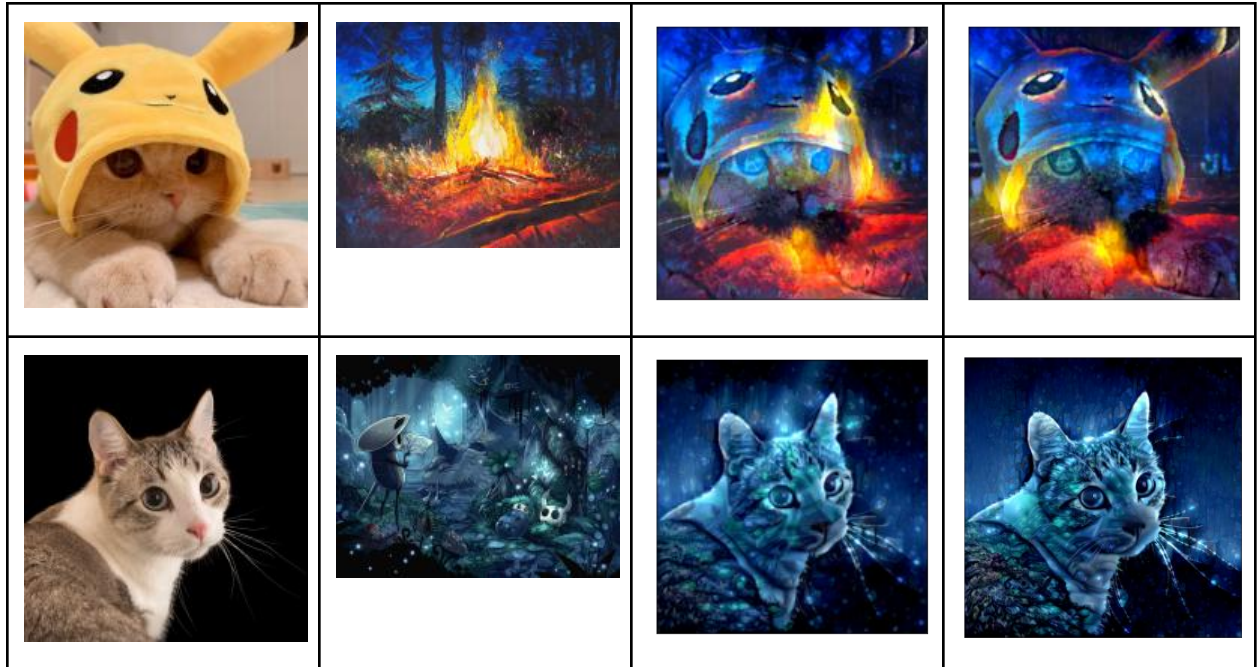
After testing a few different masks, I realized the shape of the mask didn’t influence the resulting image much other than guaranteeing the area where changes would be made. So I chose a mask with a large area of the input made transparent to give the model more freedom in generating different looks of cats. It was also clear that the prompt was essential to the result, having a vision and specifying the prompt accordingly was crucial to obtaining good results.

Kyle:

I wanted to try multiple modes of getting new images that fit our theme. I generated an image Deep Dream generator, with the prompt “cat in a coat in the mountains” and the following hyper parameters: High quality, face enhanced, 0 upscale and enhancement, landscape.

Then, I started testing with STROTSS. I went through several combinations of content images, style images, and content weight in order to find a good balance between content and style.

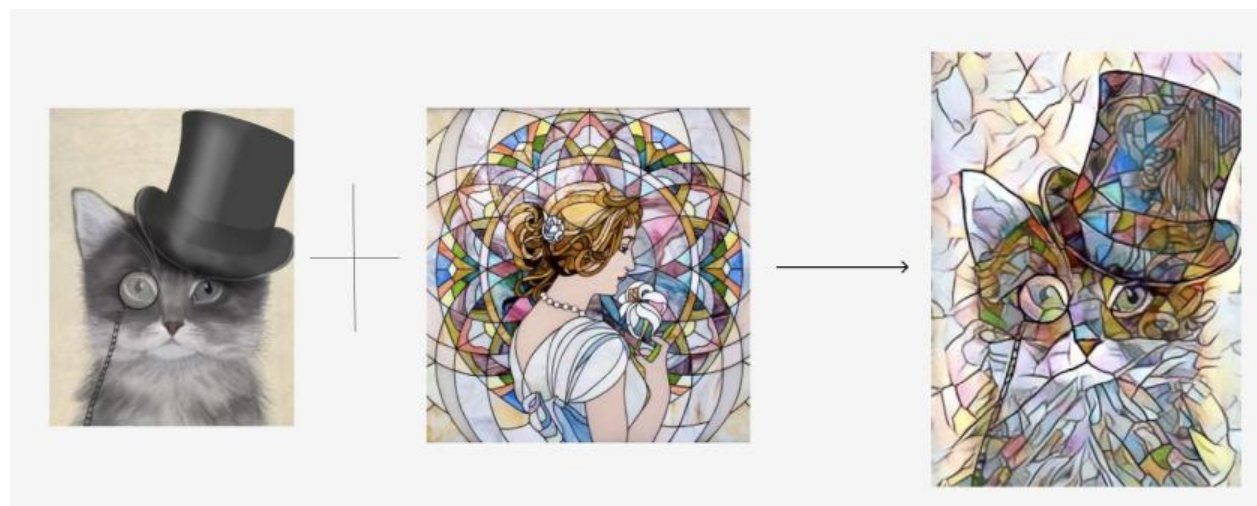
		Content Weight	
Content	Style	0.5	1
			

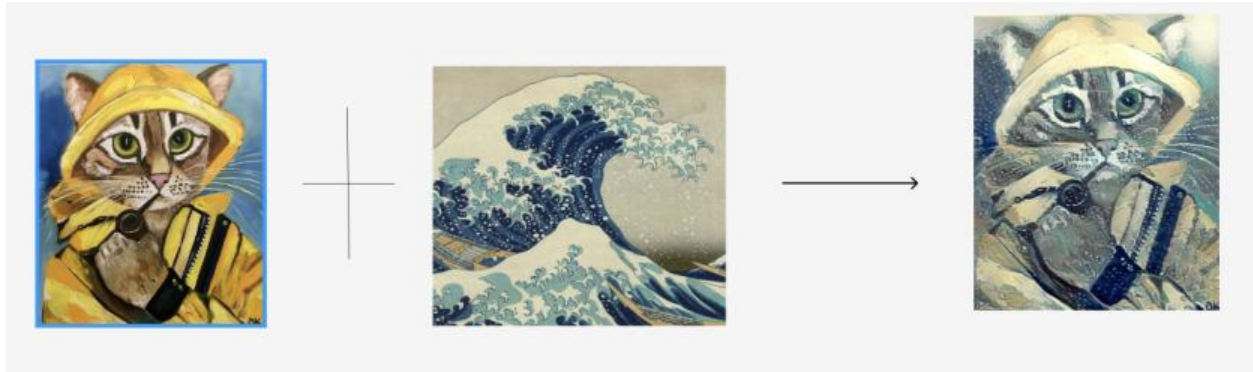


RESULTS

Josh:

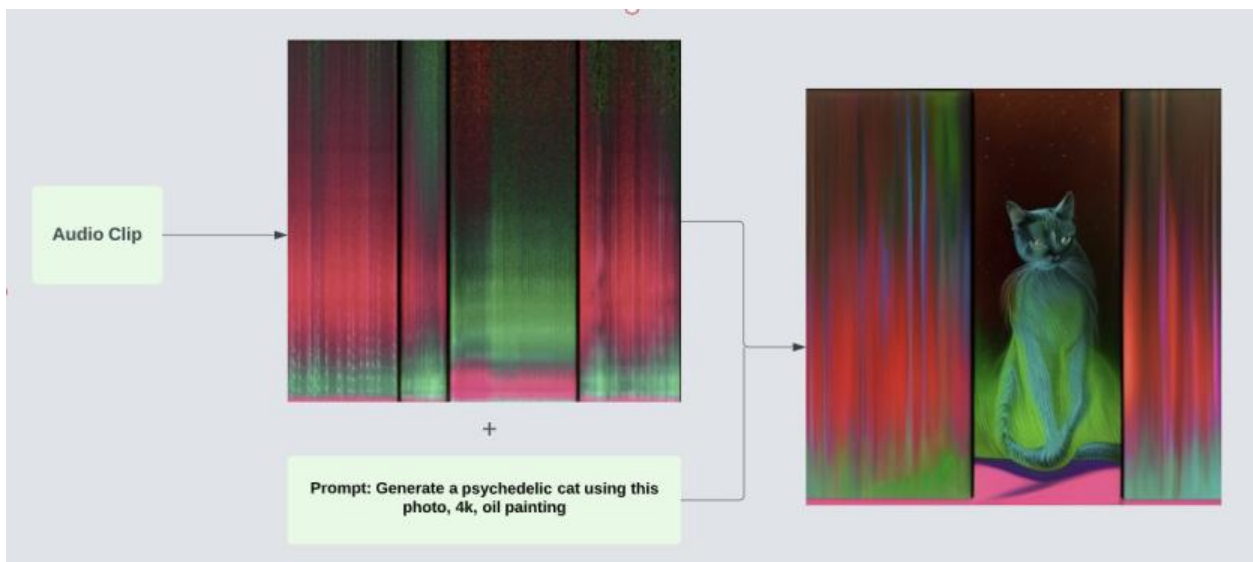
Using Adam as optimiser and VGG16 as the model for neural style transfer, I performed 800 training epochs with the following style images and content images in order to produce these two outputs. Any more training would overwhelm the original content images with the style, and I felt these two outputs were the most balanced results out of all my attempts. In addition, they suited the surrealism of our theme 'Whisker Wonderland'. Hence, I decided to choose these two as my final choices.





Aditti:

Using the convert.ing-now spectrogram generator on an audio clip of [cat meows](#) with Stable Diffusion v1.5, I generated this image of a psychedelic cat with the following parameters.



Stable Diffusion v1.5 DreamStudio (Image-to-Image)

Prompt: "Generate a psychedelic cat using this photo, 4k, oil painting"

CLIP Guidance: Enabled

Steps: 50

CFG: 12

Using InstructPix2Pix, I generated the following images:



(a) Original Image



(b) Generated Image 1



(c) Generated Image 2

S. No	Prompt	Seed	Steps	Text CFG	Image CFG
1	Turn the cat into an astronaut	63503	50	9	1.65
2	Turn the cat into an astronaut in outer space	71710	50	10	1.5

Zun:

Using Adobe Photoshop, I made transparent large portions of the original input where I imagined a cat should be. I refined the prompt to be as specific as possible for the best results. Then DALL-E 2's image edit endpoint was used to generate a response with input image, mask, and prompt. The following is one of the more satisfactory results.

Prompt: "A Cute Tortoiseshell Cat Laying Beside the Sunflowers"



Kyle:



The image to the left is from Deep Dream generator with the prompt “cat in a coat in the mountains,” and the image to the right is from STROTSS with content weight 1 and max width of 1024.

REFLECTION

While getting familiar with these different methods to generate cat characters and caricatures, we noticed a few different trends and tried to use our intuitions of the mathematical modeling behind these methods to tune hyperparameters.

In the case of experimentations with NST and STROTSS, we learned that what really matters is finding a good combination of style image and content image. While increasing the number of training epochs and adjusting content weight does help, if the style present in style images is not too clear, the output—even with simple and clear content images—can end up being subpar, as shown below.



One should choose style images, in which there is one unequivocal style, and content images, which do not have noises or distractions in the background. It also helps when content images have clear edges.

During the process of choosing different Stable Diffusion models and tuning hyperparameters, we noticed that Stable Diffusion v2.0 requires a different way of prompting than v1.5. After reading up on the differences between the two models, it is likely that this is due to the shift from CLIP to OpenCLIP in text encoding. Overall, v2.0 seemed to be less intuitive than v1.5 in prompt engineering and more difficult to generate images in different art styles.

While using InstructPix2Pix and Stable Diffusion, it was very important to get a good result tuning Image CFG and Text CFG, along with rephrasing the prompts sometimes. Having too high of a Text CFG would sometimes completely warp the original image and distort it beyond the point of recognition. On the other hand, having too high of an Image CFG would result in little to no changes in the original image. We found that a value between 7 - 14 was a good ballpark for the Text CFG, and a value between 1.2 - 2 was good for Image CFG.

In the case of DALL-E 2's image edit endpoint, we first overestimated the importance of the mask and we assumed the shape of the transparent area may influence the result. However, it became clear that this was not the case and the mask was simply to indicate to the AI where to make the changes generally. The prompt, on the other hand, was far more influential to the

result we yielded. As such, when using this model, it is important to be specific and detailed with the prompt.

By taking all of this information into consideration as we worked with each modality, we were able to generate better output. We could then pick the best results in order to create our final collage. This gives us better intuition for what makes good input and how to make the best use of the different modalities in future projects.

Code

<https://github.com/JC-78/artML/tree/main>

References

NST:

<https://www.youtube.com/watch?v=S78LQebx6jo&list=PLBoQnSflObcmbfshq9oNs41vODgXG-608>