Training ControlNet for Cubism Art

Lucas Goh
lucasgoh@mit.edu

Jeremy Susanto
jsusanto@mit.edu

Abstract

ControlNet is a neural network architecture designed to integrate spatial conditioning controls into large, pretrained text-to-image diffusion models. By locking the production-ready diffusion models, ControlNet leverages their deep and robust encoding layers, which are pretrained with billions of images, as a strong foundation for learning a diverse set of conditional controls. The architecture employs "zero convolutions" (convolution layers initialized to zero), which allow the parameters to grow progressively from zero, ensuring that no harmful noise impacts the fine-tuning process.

This report focuses on training a specialized version of ControlNet on a small dataset, specifically targeting the manipulation of image synthesis in the context of Cubism using sketches as the control mechanism. Given the real-world limitation of available images in many domains, particularly for historical art forms like Cubism, our initiative aims to demonstrate the feasibility of using limited data effectively. Cubism art, created by deceased artists Pablo Picasso and Georges Braque, presents a unique challenge due to its finite number of images. By fine-tuning ControlNet to interpret and synthesize Cubist art guided by rudimentary sketches, we address these constraints and demonstrate the potential and limitations of utilizing a small, limited dataset.

1. Introduction

The evolution of image synthesis via artificial intelligence has led to the development of various models capable of generating highly realistic and artistic images. ControlNet enhances this capability by incorporating control images and text prompts to guide the synthesis process, allowing for more precise outputs based on predefined conditions. This project seeks to extend the utility of ControlNet by training it to generate images in the style of Cubism, controlled by input sketches. Cubism, an early 20th-century art movement led by Pablo Picasso and Georges Braque, is distinguished by its fragmented and abstracted forms that represent subjects from multiple viewpoints to convey a greater context. Figure 1 below shows an example of a Cubist image and its rough sketch. Many art enthusiasts, like myself, are captivated by Cubism but find the limited number of available images to be a challenge. Imagine being able to transform a real-life image into a Cubist masterpiece. While text-to-image diffusion models can create visually stunning images from text prompts, controlling the spatial composition through text alone is difficult. This limitation is particularly evident when attempting to recreate complex art forms like Cubism, which rely on precise shapes and forms. To address this, we explore training ControlNet on a small dataset of Cubist art, using sketches as a control mechanism. This project aims to fine-tune ControlNet to interpret and synthesize Cubist art, demonstrating the potential of using limited data effectively. By leveraging a small but specialized dataset, we hope to show that it's possible to create controlled, stylized artworks even with the constraints of limited image availability.

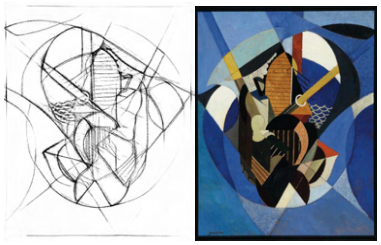

Exhibit 1. (Left) Generated Sketch and (Right) Example Cubist Art

2. Related Work

2.1. Stable Diffusion

Stable Diffusion is a powerful framework for generating high-quality images from text prompts using diffusion models. It operates by iteratively refining a noisy image to match a target distribution defined by the input prompt. The architecture is based on the U-Net structure, which

includes downsampling and upsampling layers to capture and reconstruct image details across multiple scales.

2.2. Transfer Learning

Transfer learning is a technique where a pretrained model is adapted to a new, often smaller, dataset. By leveraging knowledge from a large, general-purpose dataset, the model can quickly learn and perform well on specific tasks with limited data. In the context of our work, transfer learning enables the adaptation of pretrained Stable Diffusion models to generate Cubist art, significantly reducing the required training time and computational resources.

2.3. Dream Booth

Dream Booth is a novel approach that utilizes diffusion models to create personalized and high-quality images based on user-defined concepts and images. By fine-tuning diffusion models with specific inputs, Dream Booth allows for the generation of images that are tailored to the user's requirements, providing a high level of customization and detail. This method emphasizes the importance of user input and customization in the image synthesis process, aligning closely with the goals of our work

3. Overview

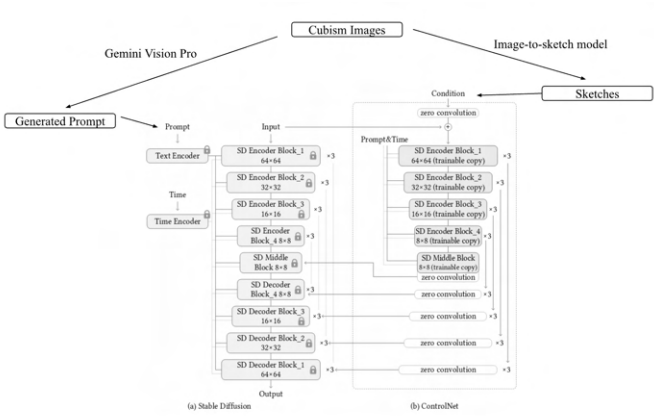

Exhibit 2. Cubism Images as input to pre-trained ControlNet

We utilized the U-net architecture of Stable Diffusion, integrating a ControlNet with the encoder blocks and the middle block. The locked, gray blocks represent the structure of Stable Diffusion V1.5. To build the ControlNet, the SD Encoder Blocks labeled with 'trainable copy' are trained along with the white zero convolution layers.

4. Method

4.1. Hypothesis

We hypothesize that a ControlNet can be effectively trained using a very small dataset of 227 Cubist images by leveraging the robust pretrained encoding layers of a large diffusion model like Stable Diffusion. By fine-tuning ControlNet with these 227 images, we expect the model to learn and adapt to the unique features and styles of Cubist art. This will enable the generation of new Cubist-style images guided by sketches as control mechanisms. Despite the limited dataset, the strong foundational knowledge embedded in the pretrained model should facilitate the successful adaptation to Cubist art, demonstrating that high-quality, stylized image synthesis is achievable with minimal data.

4.2. Data Collection and Synthesis

Cubism is a complex art style, so we initially used the fill50k dataset to familiarize ourselves with ControlNet's functionality and set up the necessary environment, including all dependencies and practical model usage. This dataset was downloaded from HuggingFace.

Next, since a Cubism dataset designed specifically for ControlNet applications was not available, synthesizing our own data was essential for training ControlNet tailored for Cubism. We leverage a comprehensive art dataset from WikiArt, accessible via Hugging Face datasets, specifically filtering the collection to retrieve artworks classified under the Cubism genre. This dataset provides a rich variety of Cubist images, offering a broad perspective on the style's characteristic abstract and fragmented forms. In total, we obtain a total of 227 Cubist images.

To complement the Cubist images, we generate corresponding sketches that serve as control images for the training process. This is achieved by utilizing a image-to-sketch model on Hugging Face Spaces, which converts our selected Cubist artworks into line drawings. These sketches are critical as they define the structural outlines

and distinctive features that our model will learn to recognize and interpret in the context of Cubism.

To obtain the image descriptions required for training, we employ Gemini Vision Pro, a LLM model by Google, to generate textual annotation from the target images. These prompts describe the artworks in terms of visual elements and thematic content, providing a textual layer that assists in aligning the generated images with the stylistic nuances of Cubism.

The images in our cubism dataset varied significantly in size, which could potentially affect the consistency of the model's performance. To address this, all images were resized to uniform dimensions of 512x512 pixels. This standardization helps in normalizing the input data and simplifies the model's computational requirements.

Finally, given the limited size of our dataset, each generated textual prompt was manually reviewed. This step ensures the quality and relevance of the prompts, verifying that they accurately reflect the visual content and stylistic elements of the corresponding images.

"Colorful painting of a town on a hill with trees in front in cubism style."

Exhibit 3. Image annotation example

4.3. Fill50k

The fill50k dataset consists of 50,000 images that are much simpler in style compared to Cubist art. This simplicity makes it an ideal starting point for familiarizing ourselves with ControlNet's functionality and setting up the necessary environment, including all dependencies and practical model usage. Using the fill50k dataset, which has been successfully fine-tuned on ControlNet in previous examples from Illayvsiel's GitHub, helps us ensure our implementation is correct before moving on to the more complex Cubism dataset.

With the fill50k dataset, we conducted a key experiment to gain insights into how our small Cubism dataset might affect the quality of outputs. We created three separate datasets from the fill50k dataset. The small dataset contains 300 images, mimicking the size of the Cubist dataset. The medium dataset consists of 3,000 images, and the large dataset utilizes all 50,000 images. Our goal was to compare these three dataset sizes fairly in terms of training time, but achieving a successful model was our priority. It's important to note that epochs and steps are not the same; there can be more steps in one epoch if the training size is larger. We used a batch size of 4 for all experiments. The small dataset was trained for approximately 3 hours over 70 epochs, the medium dataset for 5 hours over approximately 15 epochs, and the large dataset for 10 hours over about 8 epochs.

4.4. Cubism

With the Cubism dataset, our primary goal is to ensure the model works effectively and to monitor its performance across different epochs. We aimed to reach the sudden convergence point where the model's performance significantly improves. To achieve this, we tried two different approaches: (Model 1) using a base ControlNet model with only the pretrained Stable Diffusion 1.5 weights, and (Model 2) A pretrained ControlNet model that had been trained on sketch data, which is similar to our input.

We set up an experiment to evaluate the model's performance at 5, 10, 20, 40, and 80 epochs, and this was designed to understand how the model learns and adapts to the unique features of Cubist art over time. By systematically increasing the number of epochs, we hoped to identify the point at which the model begins to produce high-quality, stylized images.

4.5. Contributions

Lucas was responsible for gathering and preparing the Cubism dataset, ensuring that it was ready for use in our experiments. He managed the collection process, curated the images, and ran the experiments with the Cubism dataset. Jeremy focused on setting up the environment, which involved installing and configuring all necessary dependencies, troubleshooting compatibility issues, and optimizing the setup for efficient model training. He also conducted the experiments using the fill50k dataset, and assisted with the Cubism experiments.

5. Experimental Results & Discussion

5.1. Dataset Size Experiment on Fill50k

The results of the Dataset Size Experiment are shown in Exhibit X. below. The test set was used with the following prompts in descending order: "*light purple circle with black background*," "*blue circle with purple background*," and "*maroon circle with light blue background*."

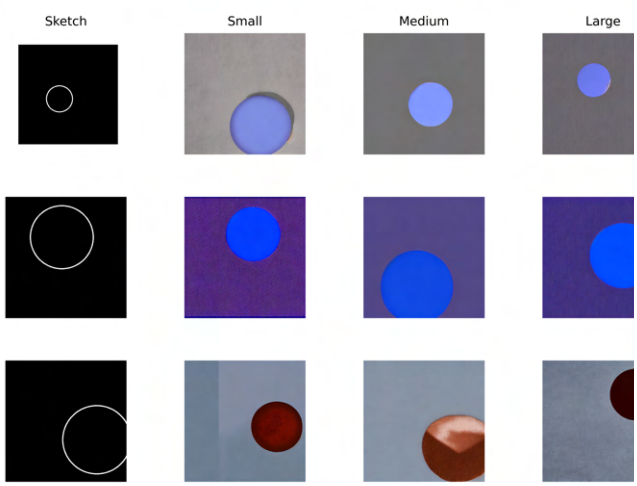

Exhibit. 4. Results of Fill50k Experiment

From this experiment, we observed that the performance across the small, medium, and large datasets was quite similar. To evaluate the results, we focused on two key qualitative aspects: how closely the resultant images followed the sketch (control) and how accurately the colors of the resultant images matched the prompt.

Across all dataset sizes, ControlNet demonstrated a strong ability to adhere to the color specifications provided in the prompts, indicating effective integration of color information from the textual input during image generation. However, we noted some discrepancies in the positional accuracy of the generated images relative to the control sketches. The resultant images did align perfectly with the specified positions.

Additionally, we observed instances where the generated images contained undesired shadows and textures. We believe these issues resulted from insufficient training, potentially not reaching the 'sudden' convergence point. Nonetheless, this is promising as it suggests that even a small training set may be sufficient to obtain reasonable results from ControlNet.

5.2. Epoch Variation Experiment on Cubism

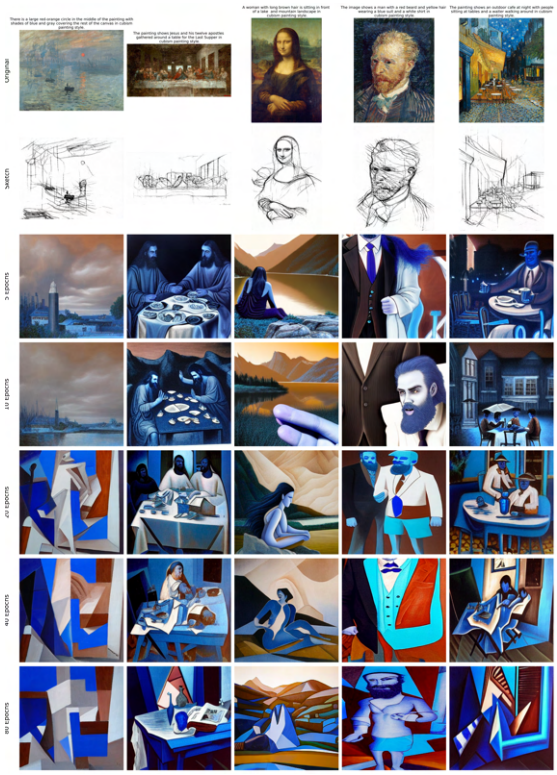

Exhibit 5. Results of Model 1

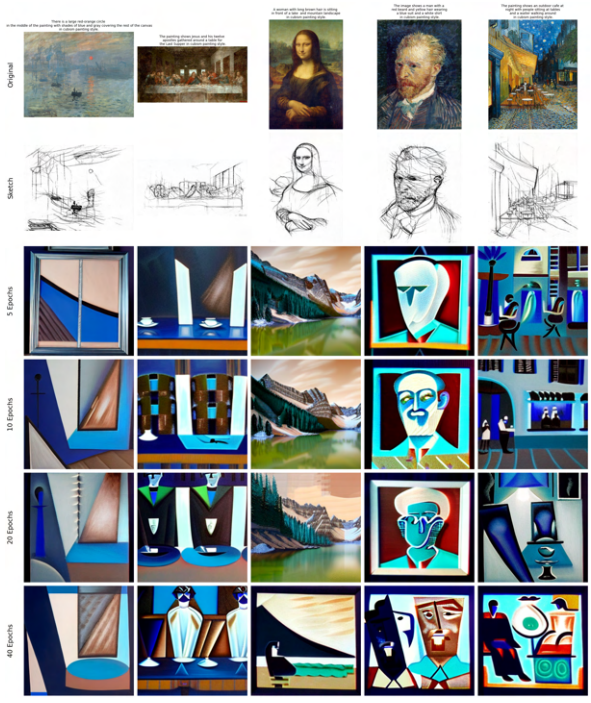

Exhibit 6. Results of Model 2

As seen in Exhibits 5 and 6, the two ControlNet models gradually begins to capture the nuances of Cubism as we increase the number of epochs from 5 to 80.

Initially, at 5 epochs, the generated images show a basic adherence to the prompts and control sketches, but they lack the distinct characteristics of Cubist art. The colors are present, but the forms and structures are still somewhat rudimentary and do not fully align with the control sketches.

By 10 and 20 epochs, there is noticeable improvement in the structural representation of Cubism. The images begin to exhibit more geometric shapes and abstract forms, which are key features of Cubist art. The model shows better integration of the control sketches, but there are still discrepancies in spatial accuracy and some presence of unwanted textures.

At 40 and 80 epochs, the images display further refinement and complexity, closely mimicking the style and elements of Cubist paintings. The forms are more cohesive, and the color application is more consistent with the prompts. However, even at these higher epochs, the outputs do not follow the control sketches precisely, indicating that we have not trained enough to reach the 'sudden convergence' point. More training would be beneficial, but 80 epochs already utilized significant computational resources and time, making it difficult to train further. There is also no guarantee that we will reach the convergence point within a reasonable number of epochs.

Another observation is the seemingly unnatural prevalence of the color blue. This may indicate some bias in the training set towards this color, suggesting that the dataset might not be perfectly balanced in terms of color diversity.

Additionally, examining the Mona Lisa example, it appears that the generated images are dominated by mountains and lakes rather than the figure of the Mona Lisa. This issue could be attributed to the prompt overwhelming the image generation process or a bias in the training set towards landscape elements. This suggests that the model might be overly sensitive to certain types of content in the prompts or that the dataset used for training included a disproportionate number of landscape images, thereby influencing the model's output.

Moreover, this observation highlights a broader challenge in fine-tuning models with small, specialized datasets: ensuring balanced and representative training data. To address this, future experiments could involve more careful curation of the training set to include a more diverse range of subjects, thereby helping the model to generalize better across different prompts and control sketches. This approach could mitigate the dominance of unintended elements and improve the accuracy and relevance of the generated images to the desired artistic style.

All in all, while models 1 and 2 generated distinctly different images, neither approach closely followed the control sketches. This indicates that despite the different training strategies and initial conditions, both models struggled with accurately capturing the spatial configurations outlined by the control inputs.

6. Conclusion

In this report, we explored the training of a specialized version of ControlNet on a small dataset to manipulate image synthesis in the context of Cubism using sketches as the control mechanism. Our experiments focused on understanding how varying dataset sizes and the number of training epochs affect the model's ability to produce accurate and stylistically faithful Cubist art.

We conducted two key experiments. First, we used the fill50k dataset to assess how dataset size influences output quality. Our findings indicated that the performance across small, medium, and large datasets was quite similar, with ControlNet showing strong color adherence but some discrepancies in spatial accuracy. This suggests that even small datasets can yield reasonable results, though further training might improve precision.

Second, we experimented with the Cubism dataset, evaluating model performance across different epochs. As training progressed from 5 to 80 epochs, the ControlNet model began to capture the nuances of Cubism more effectively. However, the outputs still did not perfectly follow the control sketches, indicating that additional training could be beneficial.

References

[1] Lllyasviel, "ControlNet/Docs/train.md at main LLLYASVIEL/ControlNet," GitHub, https://github.com/lllyasviel/ControlNet/blob/main/docs/train.md (accessed Apr. 30, 2024).

[2] H. Community, "Huggan/wikiart · datasets at hugging face," huggan/wikiart · Datasets at Hugging Face, https://huggingface.co/datasets/huggan/wikiart (accessed Apr. 29, 2024).

[3] Image-to-line-drawings - hugging face, https://huggingface.co/spaces/awacke1/Image-to-Line-Drawings (accessed Apr 29, 2024).

[4] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242, 2022.

[5] Andrey Voynov, Kfir Abernan, and Daniel Cohen-Or. Sketchguided text-to-image diffusion models. 2022.

[6] Sadia Afrin. Weight initialization in neural network, inspired by andrew ng, https://medium.com/@safrin1128/weightinitialization-in-neural-network-inspired-by-andrew-nge0066dc4a566, 2020.

[7] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scenebased text-to-image generation with human priors. In European Conference on Computer Vision (ECCV), pages 89–106. Springer, 2022.

[8] Stability. Stable diffusion v1.5 model card, https://huggingface.co/runwayml/stable-diffusion-v1-5, 2022

[9] Stability. Stable diffusion v2 model card, stable-diffusion2-depth, https://huggingface.co/stabilityai/stable-diffusion-2-depth, 2022

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.