Fine-tuning ControlNet and Stable Diffusion for Few-Shot Style Transfer

Isabella Yu [*]
Massachusetts Institute of Technology
iyu@mit.edu

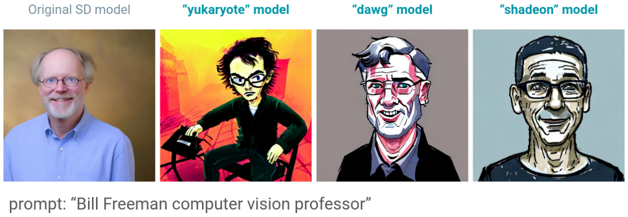

Figure 1. Stable Diffusion models fine-tuned on different artists' styles.

Abstract

Generative art models like Stable Diffusion, as well as conditioning models like ControlNet, are capable of creating art that can be indistinguishable from that of a human artist. This begs the question: how easily can these models mimic a certain artist's style? In this paper, we investigate this question by fine-tuning Stable Diffusion and ControlNet using Low-rank Adaptation (LoRA) and DreamBooth with varying amounts of data of different styles. We provide a new analysis of how dataset size and stylistic consistency contribute to a convincing style transfer using CLIP embeddings and conduct an informal survey to investigate the effect of truncating a dataset into highly stylized "clusters." We find that training a model on a cluster of just 10 images with high CLIP similarity may produce convincingly stylized images, but a more formal user study would be necessary to make concrete conclusions. We also find that ControlNet fails to produce faithfully colored images from small (around 20) datasets of lineart, so more research is needed to improve ControlNet's few-shot learning capabilities.

1. Introduction

Style transfer is a powerful technique in the field of computer vision that enables the transfer of artistic style from one image to another. It has found widespread applications in various domains, including photography, concept art, and fashion. With the advent of generative art models such as Stable Diffusion (SD) [8], as well as fine-tuning methods like Dreambooth [9], style transfer can be acheived with as little as three images. In addition, we can now condition generative art models on different conditions, e.g. depth, sketch, or edges, using ControlNet [11]. These tools give unprecedented control over the style and content of an AI-generated illustration.

However, these models are a double-edged sword; though they may lower the amount of work needed to produce artwork, some artists are fearful that generative art models may replace them due to their ability to mimic specific artists' styles [2]. Thus, analyzing how easily state-of-the-art models can perform style transfer is useful in providing transparency to artists about how their work is mimicked. Analyses of dataset size [4] and stylistic consistency [5] on style transfer performance have been done for older style transfer models, but to the authors' knowledge, there are currently no such published analyses for newer, diffusion-based models.

In this paper, we will study how dataset size and stylistic consistency affect the performance of LoRA/DreamBooth, Stable Diffusion, and ControlNet on style transfer. We gather art of different artists (with consent), label the artworks using CLIP, fine-tune a Stable Diffusion using DreamBooth with LoRA for each artist, and train a ControlNet on the fine-tuned model. For analysis purposes, we define stylistic consistency as the variance of each dataset's latent space vectors given by CLIP [7] embeddings. We will then analyze the qualitative performance of each fine-tuned Stable Diffusion model given the size and stylistic consistency of its training dataset, as well as the qualitative performance of the ControlNet model. Finally, we measure quantitative performance of the fine-tuned Stable Diffusion models via a user study that asked participants to rate how well each model adheres to an artist's style.

2. Related Work

One of the earliest and most popular approaches to style transfer is the neural style transfer proposed by Gatys et al. (2016) [1]. This method uses deep neural networks to sep-

[*] About 30% of the related work section was generated by chatGPT then edited by me; I verified its correctness by reading the papers it mentioned.

arate the content and style of an image and then combines them to generate a stylized output. Despite its effectiveness, this method suffers from slow convergence and instability, which limits its practical applicability.

In recent years, several works have explored the use of diffusion-based methods for style transfer. These methods are based on the idea of iteratively diffusing the style information across the image to generate a stylized output. Rombach et al. [8] proposed the diffusion-based style transfer technique used by Stable Diffusion that addresses the instability and slow convergence issues of neural style transfer. Ruiz et al. introduced DreamBooth to fine-tune generative image models to mimc a certain style or subject from just 3-5 images [9], and Hu et al. proposed Low-rank Adaptation as a way to simplify the fine-tuned model by decomposing the fine-tuned weights into low-rank matrices [3].

Li et al. [4] investigated the impact of training data size on neural style transfer performance. They found that increasing the size of the training data improved the visual quality of the stylized outputs and reduced the artifacts and inconsistencies in the transferred styles. Similarly, Luan et al. [5] studied the effect of style consistency on neural style transfer. They found that using multiple reference style images and enforcing consistency between them improved the quality and consistency of the stylized outputs.

3. Methods

3.1. Dataset Preparation

We gather pieces of artwork from different artists–with consent–to form our datasets. We cropped each image to be 512×512. We labeled each artwork using CLIP Interrogator, then applied any grammatical or semantic corrections to the prompt. We appended “in the style of X” at the end of each label, where X is the artist’s name. The line art of each artwork for input to ControlNet was also matched to each color image; if no line art was present, we made a rough scribble of the artwork. Fig. 2 shows some examples of labels and artworks. In total, we collected 3 datasets from artists “dawg,” “yukaryote,” and “shadeon” of 18, 21, and 19 images each, respectively.

3.2. Fine-tuning with LoRA

After generating the dataset and prompts, we fine-tune Stable Diffusion to perform style transfer. Because fine-tuning a 857M parameter diffusion model is computationally expensive, we apply low-rank adaptation (LoRA) to hasten training. LoRA adds a pair of low rank matrices AB^T to the existing weights, which have much fewer parameters to train, and only updates these matrices. This allows for faster and less memory-intensive fine-tuning.

Specifically, we apply LoRA to Dreambooth, a method that fine tunes generative text-to-image models from as few as 3 images of a subject. We make two modifications to facilitate style transfer. First, since DreamBooth was created for learning a specific represenation of a *subject* instead of a style, we train a new text encoder, allowing the model to learn from multiple different prompts instead of a single “instance” prompt of a subject. This essentially performs naive fine-tuning. Second, we set the “class” prompt to be“in the style of [X]” instead of a class of objects to embed each artist’s style into the text encoding.

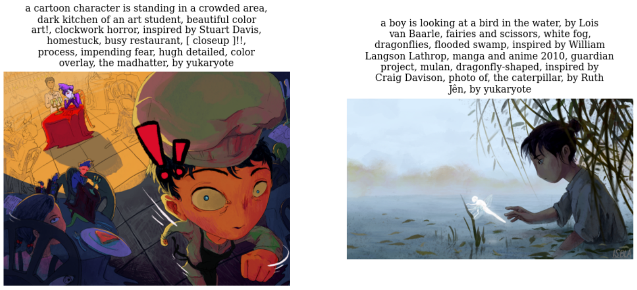
a cartoon character is standing in a crowded area, dark kitchen of an art student, beautiful color art!, clockwork horror, inspired by Stuart Davis, homestuck, busy restaurant, [closeup]!!, process, impending fear, hugh detailed, color overlay, the madhatter, by yukaryote
a boy is looking at a bird in the water, by Lois van Baarle, fairies and scissors, white fog, dragonflies, flooded swamp, inspired by William Langson Lathrop, manga and anime 2010, guardian project, mulan, dragonfly-shaped, inspired by Craig Davison, photo of, the caterpillar, by Ruth Jên, by yukaryote

Figure 2. Examples of images and their labels from one artist ”yukaryote” (the author’s pen name). The labels were generated by CLIP Interrogator, and ”drawn by yukaryote” was manually appended to each label to distinguish the artist.

We use a batch size of 2, learning rate of 0.00005, and train for 1000 epochs. We find that training for longer than 1000 epochs yield similar results, but we have not yet performed hyperparameter search to find the most optimal hyparameters.

3.3. Training ControlNet

The fine-tuned Stable Diffusion model is used to train a ControlNet conditioned on line art, i.e. ”scribble.” This is done by creating two copies of the fine-tuned SD model: a trainable copy and a “locked” copy with frozen weights, shown in Fig 7. “Zero convolution” layers connect the trainable copy to the locked copy. Zero convolutions are 1×1 convolutions with weights and biases both initialized to zero, and the final output of the model is the addition of the output of the locked copy and the convolved trainable copy. As ControNet is trained, the zero convolutions’ weights and biases change, allowing Stable Diffusion to be conditioned via the trainable copy.

We use a batch size of 2, learning rate of 0.00001, and train for at most 15 epochs due to the small amount of training data, but have not performed hyperparameter search to find the most optimal hyparameters.

4. Experimental results and discussion

4.1. Hardware

All models were trained on MIT SuperCloud using 2 Volta V100 GPUs.

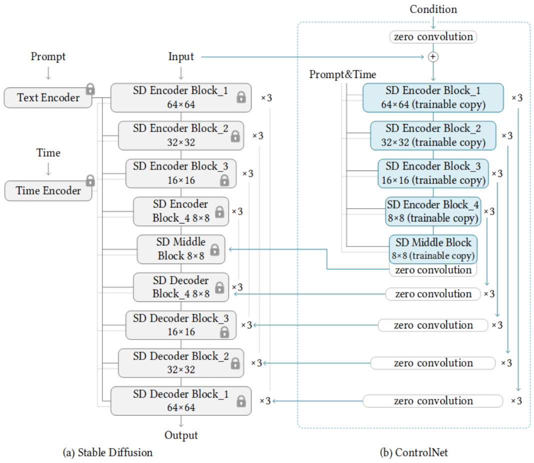

Figure 3. The ControlNet architecture, showing the locked SD copy (gray), the trainable copy (blue), and zero convolutions. For more detail see [11].

4.2. Results of LoRA fine-tuning

Qualitatively, we were able to successfully fine-tune Stable Diffusion v2.1 to perform style transfer for different artists. Fig 4 shows some images generated by each of the three fine-tuned models when each was trained on their respective artist's entire dataset. The first prompt, ”baby lion,” tests whether the model can learn each artist's *mechanical* style, i.e. the general pattern in which they place brush strokes. This is because a ”baby lion” was not present in any of the datasets. As shown, each artist's model deviates significantly from the original Stable diffusion, producing stylized, comic-like images that are representative of each artist's dataset. However, it fails to produce output representative of “yukaryote”'s dataset due to its low stylistic consistency, which we will analyze in Section 4.4. The next two prompts, ”a boy with a backpack running to school” and ”a woman,” test whether the model can learn the artist's *subject* style, or how they draw certain subjects (the prompt ”a woman” and ”a boy” was present in all the datasets). Here, generated images generally better resembled their training set because people were well-represented in each artist's dataset. In particular, “shadeon”'s model generated a crown when prompted to generated ”a woman” since one training example was that of a crowned queen.

However, qualitatively more faithful style transfer occurs when we train each model on subsets of each dataset with high style similarity. To measure style similarity, we pass each image x_i in the dataset $X = \{x_1, ..., x_n\}$ through CLIP to get its CLIP embedding $\hat{x}$, a 768-length vector. We measure style similarity in each dataset X by measuring its

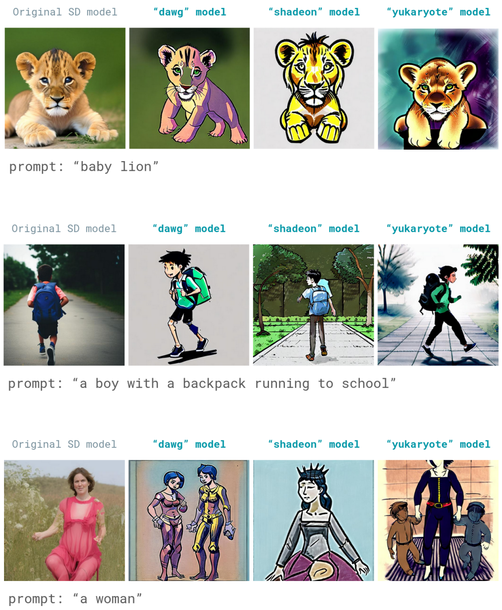

Figure 4. Results from training 3 different stylized Stable Diffusion models. The leftmost column is the output of the original Stable Diffusion model, while the rest are ones fine-tuned on each artist's dataset.

Dataset	Largest 3-means cluster dispersion (dataset size)	Entire dataset dispersion (dataset size)
“dawg”	3.451 (10)	4.139 (18)
“shadeon”	3.099 (10)	3.988 (21)
“yukaryote”	4.115 (7)	4.257 (19)

Table 1. Dispersion rates for whole datasets and for the largest 3-means clusters of dispersion.

CLIP dispersion $d(X)$, which we define as

$$d = \frac{1}{N}\sum_{i=1}^{N} ||\tilde{x} - x_i||$$

where $\tilde{\hat{x}} = \frac{1}{N}\sum_{i=1}^{N}\hat{x}_i$ is the mean CLIP embedding of the dataset. This is essentially the average distance that each image embedding in the dataset is from the mean embedding.

We perform k-means clustering for $k = 3$ in the embedding space to find the largest cluster, then train each model with only that cluster as the training set. This generally leads to better style transfer, as each cluster has lower dispersion, i.e. higher CLIP embedding similarity, allowing the model to learn a more specific style. Table 1 summarizes the difference in dispersion between the largest 3-

means cluster and the entire dataset; as expected, the dispersion is smaller for the cluster.

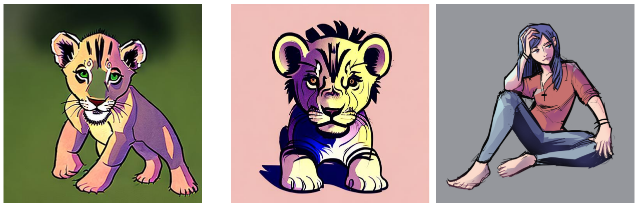

Figure 5. Results from training "dawg"' model on a cluster with high CLIP embedding similarity, i.e. low dispersion. Left is the output of the model trained on the entire dataset, middle is the output of the model trained on the low dispersion cluster, and right is an example of "dawg"'s artwork.

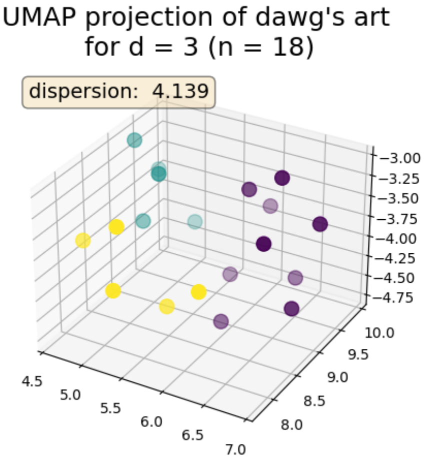

Figure 6. Visualization of CLIP embeddings for "dawg"'s dataset. Here, here we project the 768-dimensional CLIP embedding into 3D space using Universal Manifold Approximation and Projection (UMAP) [6]. The colors represent different clusters calculated from 3-means clustering.

4.3. Results of ControlNet training

Our trained ControlNet does not seem to successfully color an image faithfully conditioned on lineart; the resulting image almost always deviates from the lineart. However, making the SD decoder trainable for further finetuning does seem to yield better-conditioned results. Fig 7 shows an example of this phenomenon. We believe this is primarily due to the very small (about 20 images each) datasets we use; training on fill50k, a dataset of 50 thousand colored circles and their edge maps, for 4000 epochs produces reasonable results. Refer to the appendix for results of ControlNet trained on fill50k.

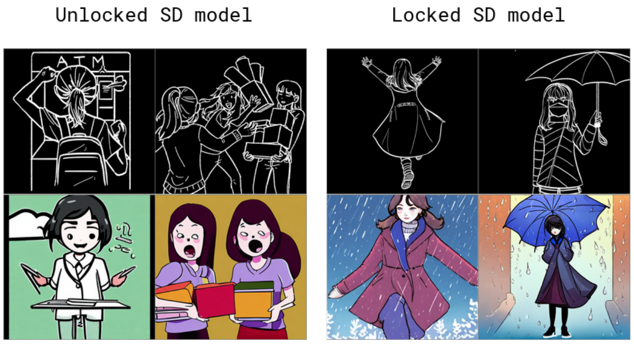

Figure 7. Results of ControlNet training after 15 epochs (we found that training for more epochs, up to 100, yielded similar results). On the left is the image produced by locking the entire SD model, and on the right is the image produced by allowing the decoder of the SD model to be trained.

4.4. Survey on stylistic consistency

To quantify the relationship between the human notion of "stylistic consistency" and dispersion, we conducted a survey that asked participants to rate how well each model mimicked each artist's style. Participants were first shown examples of each of the 3 artists' work to familiarize themselves with each style. Then, they were shown 6 images generated by each artist's model, 3 from the one trained on the entire artist's dataset and 3 from the one trained on the low-dispersion cluster. The participants did not know which dataset that the model was trained on. They were asked to rate each image from 1 to 7, with 1 being "not at all like the artist's style" and 7 being "indistinguishable from the artist's style." In total, 5 participants responded.

Fig. 8 analyze the survey results. Fig. 8a shows the ratio of the average style consistency ratings between art generated by the low-dispersion cluster compared to those generated via the entire dataset. Although participants favored the models trained on the low-disperion clusters for "dawg" and "shadeon," there was a slight negative preference for the images generated by the low-disperion model for "yukaryote." This may be because the change in dispersion between "yukaryote"'s clustered data and their entire dataset was not significant compared to that of the other artists', as shown in Table 1. Fig. 8b shows that there may be a slight negative correlation ($\rho = -0.34$). between models trained on higher-dispersion datasets and their ability to perform faithful style transfer. This sounds reasonable, as high-dispersion style may have patterns that are harder for a diffusion model to learn. However, on the more philosophical side, a high-enough dispersion style might not classify as a "style" anymore, so a diffusion model may not even be performing style transfer in the first place.

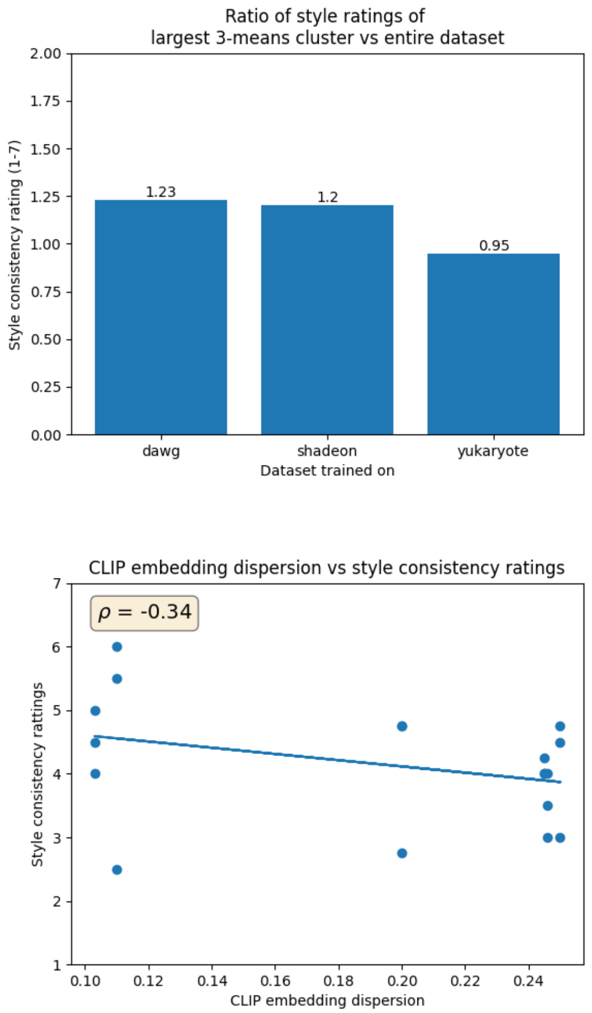

Figure 8. Top: analysis of the ratio of stylistic consistency rating of the model trained on the clustered dataset vs the model trained on the entire dataset for all artists. Bottom: scatterplot showing how the dispersion of a dataset a model is trained on relates to its style consistency rating.

5. Conclusion

In this report, we studied the effect of dataset size and stylistic consistency on the quality of Stable Diffusion-based style transfer and ControlNet training. We defined a metric for stylistic consistency–the dispersion of a dataset's CLIP embeddings–and showed that there may exist small, low-dispersion clusters of a dataset that can be used to fine-tune a Stable Diffusion model. We conduct and analyze the results of an informal user survey that measured the style consistency of models trained on such low-dispersion clusters and find that low dispersion may be correlated with better style transfer, but more participant data is needed to make a concrete conclusion.

For fine-tuning Stable Diffusion, future work can include finding methods for convincing style transfer for artists whose styles are diverse, like "yukaryote"'s, as well as methods for zero or one-shot style transfer. More work is certainly needed in developing methods to train a ControlNet from small amounts of data. Finally, perhaps a more ethically pressing research problem than fine-tuning diffusion models is preventing diffusion models from copying artists' work. To prevent generative models from mimicking certain artists' styles, Shan et al. introduced *Glaze*, a tool that enables artists to apply "style cloaks", or barely perceptible perturbations to images [10].

References

[1] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. 1

[2] Kashmir Hill. This tool could protect artists from a.i.-generated art that steals their style, Feb 2023. 1

[3] 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, 2021. 2

[4] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer, 2017. 1, 2

[5] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer, 2017. 1, 2

[6] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. 4

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 2

[9] 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. 2022. 1, 2

[10] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by text-to-image models, 2023. 5

[11] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 3

6. Appendix

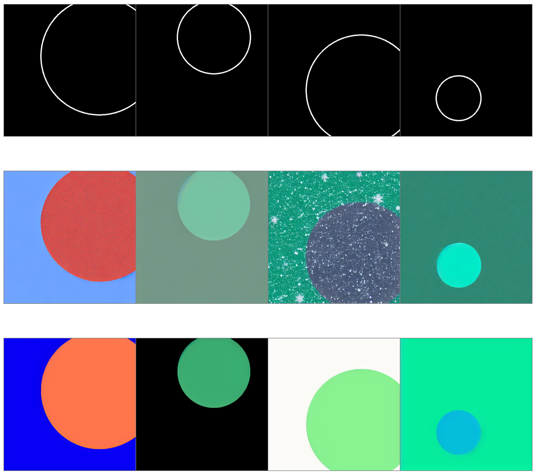

Figure 9. Results from training ControlNet on the fill50K dataset for 5000 epochs. These results seem qualitatively better than those of the ControlNet trained on the artists' datasets, which is presumably because of the dataset's large size and simple structure. Top: Canny edge control. Middle: generated image. Bottom: ground truth.