An Artistic Art-ificial Intelligence: A More Intelligent Way to AI Art Style Transfer

Helen He, William Shabecoff, Regina Ta

Stanford University, Stanford, CA

{helenahe, wis23, rta}@stanford.edu

Abstract

Our project investigates cutting-edge methods in neural style transfer, analyzing the strengths and weaknesses of several approaches. We found that fine-tuning Dreambooth with semantic content capture yielded promising results in preserving both style and content. However, introducing source image residuals during diffusion, while maintaining composition, posed challenges in achieving high-quality outcomes due to limitations in current diffusion models. Notably, direct neural style transfer exhibited shortcomings in fidelity and adaptability, particularly with complex style-image combinations. Still, Dreambooth fine-tuning with semantic content capture emerged as the top-performing algorithm, offering superior control over style transfer and content preservation by learning strong semantic priors from image-caption pairs. Future research includes enhancing diffusion models with more advanced techniques for integrating source image information, exploring language-vision model integration for better style control, and refining evaluation metrics to be better tailored to style transfer tasks.

1. Introduction

In computer vision, neural style transfer is a powerful technique for transforming images by combining one image's content with the artistic style of another. This process is valuable for various applications, such as digital art, photo editing, and augmented reality. However, existing approaches have limitations, particularly with fidelity and flexibility. Right now, there are two primary paradigms for style transfer.

- The first paradigm involves directly combining textures and colors of a style image with content of another, usually with image convolution. This naive approach, while straightforward, often fails to achieve high-fidelity results [3].
- The second paradigm uses large, pre-trained models, such as Stable Diffusion. While these models can generate much more impressive results, they are constrained by the styles they have been explicitly trained on, and they offer limited usage control [6].

Our project aims to develop a high-fidelity style transfer method that overcomes the limitations of the above paradigms. **We propose a novel approach that uses a content image and a small set of style images to learn and apply styles more effectively.** Unlike the first paradigm, which focuses on features from a single image, our method leverages features from multiple images to capture a deeper and more nuanced understanding of style. Additionally, unlike the second paradigm, our approach allows transfer of novel, untrained styles with better precision and control.

Specifically, the input to our system is a single content image and a set of 5 style images. The output is a transformed (stylized) image that keeps the content image while incorporating the learned style from the style set. In our paper, we test out this method using two different techniques:

- Deep Style Transfer with Semantic Content Capture: We first train Stable Difusion using sets of images from different styles using the Dreambooth finetuning technique. We then use GPT-4V to caption images, combining these captions with style tokens. Our hypothesis is that this approach will yield better results than existing baseline methods.
- Deep Style Transfer with Source Image Residuals: In order to incorporate more visual content from the source image we introduce latent information from the source image during the diffusion process. This latent information is generated by encoding the source image with the variational autoencoder module of the latent diffusion model. We hypothesize that this technique will better preserve content and composition of the original image compared to methods relying only on captions.

By exploring these techniques, we hope to achieve style transfer with higher fidelity and push the envelope for what's possible in AI-driven image transformation.

2. Related Work

2.1. Image Generation & Style Transfer

Gatys et al. [3] illustrate the first paradigm of neural artistic style transfer. They separate the style and content of images with a deep CNN trained for object recognition. By analyzing the activations of the CNN's filters at different layers, they extract features representing image content (higher layers) and style (texture information across layers). They then use this separation to create new images by combining the content of one photograph with the style of famous paintings. This work is the first to achieve content and style separation in complex images and allows for controlling the balance between content and style by adjusting weights in the loss function). However, it fails to achieve high-fidelity results.

Johnson et al. [4] illustrates an attempt to create more visually appealing style transfer results. They build on Gatys's approach by introducing a perceptual loss function. This loss function relies on features extracted from pre-trained CNNs, and helps to improve image similarity between content and stylized images. However, they lack the improvements in content preservation and style control that later methods like Ruiz et al. [7] honed.

Rombach et al. [6] illustrate the second paradigm with latent diffusion, the architecture behind the Stable Diffusion model. While diffusion models traditionally operate in the pixel space, Rombach et al. notes that the pixel space is extremely high dimensional, meaning that training diffusion models is prohibitively expensive, taking hundreds of GPU days. Latent diffusion expedites diffusion model training by introducing a compression step where images are effectively downsampled using a variational auto-encoder. Unlike prior attempts at latent diffusion, they use a latent space that preserves the spatial information of image. This allows them to use a Unet architecture for the diffusion objective, which is advantageous since Unets can make use of spatial inductive biases for denoising, which are absent from unstructured 1-dimensional latent spaces. However, methods like Stable Diffusion do not enable fine-grained control over a generated image.

Casanova et al. [1] propose an instance-conditioned GAN (IC-GAN) to generate realistic variations of a target instance. Unlike prior methods, IC-GAN focuses on individual data points and their neighbors, which allows it to create images resembling the target distribution. This approach leads to subjectively better image generation on datasets like ImageNet, but it still struggles with replicating unique subjects and underperforms on preserving subject identity.

Gal et al. [2] propose Textual Inversion which allows users to submit concept images of a subject or style to preserve, and represents this concept with a ”pseudo-word” within the text-embedding space of a pre-trained text-to-image model. Then, they attempt to optimize the embedding vector linked to the ”pseudo-word” and minimize the loss between the generated images and the concept images, thus personalizing image generation without retraining the entire model. However, this approach does not perfectly capture details of the concept that are key to a human observer (e.g., shape or texture) and might not preserve subject identity across generated images.

Instead, Ruiz et al. [7] propose a new approach for image style transfer by customizing text-to-image diffusion models and expanding the language-vision dictionary to link words to particular subjects, so that model outputs preserve key features of the subject while contextualizing it in novel scenes. They aim to learn strong semantic priors from image-caption pairs to synthesize new outputs with specific subjects preserved across different contexts. By representing subjects with unique identifiers in the input prompt (followed by the subject's class name), model outputs can illustrate the semantic priors of the subject's class, while still preserving the subject itself. Next, they fine-tune a model on a small set of subject images. Finally, they compute an autogenous, class-specific, prior preservation loss, which is a loss function that helps to mitigate language drift and ensure diverse output scenes.

2.2. Evaluation Metrics

To measure the quality of our stylized images, Sanakoyeu et al. [8] propose the ”deception rate,” the fraction of stylized images that an artist classification network has assigned to the artist whose artworks are the style images. However, the deception rate has some key drawbacks: it requires training an artwork classifier network for every artist whom our style images are from, which is a problem if is that our style images come from a variety of artists. Plus, it doesn't account for content preservation.

Ledig et al. [5] introduce VGG (Visual Geometry Group) loss as an alternative to pixel-wise loss to measure similarity between images. VGG loss uses a pre-trained image recognition network (such as VGG itself) to analyze features in different image layers, focusing on human-perceived aspects like shapes and objects. However, VGG loss misses complex images details like textures that are important to human perception.

Zhang et al. [10] show LPIPS (Learned Perceptual Image Patch Similarity) as an improved way to measure content preservation. They use deep features, extracted from deep CNNs trained for image recognition, to measure perceptual similarity as a proxy for content preservation. Perceptual similarity is how ”semantically” similar two images appear to the human eye, which LPIPS quantifies by analyzing deep features between image patches. LPIPS outperforms other image comparison metrics like VGG loss when evaluated with large dataset of human-annotated perceptual similarity judgments.

Finally, Wright et al. [9] propose a more state-of-the-art metric, ArtFID, to considers two key aspects of neural style transfer: content preservation and style matching. First, they compute the distance between a content image and a stylized image using LPIPS, taking the mean distance for a batch of images. Next, style matching is how well the feature distribution of the style images match the stylized images. This is measured by training a large neural network classifier to learn image representations from a dataset of labeled artworks (artist and style). To quantify how well style is transferred, Wright et al. computes the distance between a style image's and a stylized image's feature distributions using Frechet distance. The final ArtFID score combines the content preservation and style matching measures into a single value.

3. Methods

With our selected of content and style images (detailed in Section 4: Dataset & Features), we conduct the following experiments:

3.1. Experiment 0: Gatys et. al's Neural Style Transfer

We implement neural style transfer from Gatys et. al's seminal ”Neural Style Transfer” paper as our baseline. This approach works by directly optimizing the pixel values of a generated image, such that the low level features are close to a style image and the high level features of the image are close to a content image. Like the original paper, we use VGG19 as our feature extractor using early layer for low level features and late layers for higher level features. We perform the optimization with gradient descent with Adam using a mean squared error loss between the sample and target features.

3.2. Experiment 1: Deep Style Transfer with Semantic Content Capture

Dreambooth finetuning allows for the creation of high-fidelity stylistic themes. These learned themes can be used for the style transfer task. However, applying style transfer to an image requires some way of representing the image to the tuned generator so that the new style can be applied. Thus we use **semantic content capture** where we leverage a vision-language model to describe the image, so that it can be recreated by our diffusion model. For captioning, we use GPT4-v with the prompt "please provide a very short caption of the image."

Our hypothesis is that this approach will yield better transfer results than the baseline.

3.3. Experiment 2: Deep Style Transfer with Source Image Residuals

Captioning enables the transfer of an image's subject to a new style, though it may not retain fine-grained details such as subject positioning within the composition. To address this, we suggest using the content image — the subject we aim to style transfer to — as the basis for our latents during the diffusion step, ensuring preservation of visual and compositional content. Our goal is to utilize the denoiser's ability to generate an image in a specific style, guided by the latent representation of the style image.

Our initial attempt involved forward diffusion before the backward pass, yet the erasure of image content by forward diffusion and the reliance of the backward pass on prompt guidance made this ineffective. Unfortunately, due to implementation constraints, executing a partial forward and backward diffusion pass was unfeasible.

In our second approach, we treated the content image as the noisy latent seed. However, this also failed to produce the desired outcome. The denoising Unet module expects pure noise at the initial latent, leading to instability if the input latent does not match its expected norm. Although scaling the input latent helps, it results in a blurred version of the source image, rather than a novel image.

To overcome this, we proposed feeding the denoiser a mixture of noise and source image latents. Yet, this approach also fell short, as the Unet denoiser either returned a blurred source image, or an unrelated image based on the noise-to-source latents ratio. The denoiser's failure highlights the need for a method that iteratively generates a novel image while favoring content similar to a source image. Thus, we introduce a new technique: **diffusion with source image residuals.**

Diffusion with source image residuals works by introducing source image latents to the diffusion process at multiple timesteps throughout the process. Rather than introducing the source latents to the seed noise of the diffusion process, where they will either be ignored or dominate the diffusion such that no novel content can get added to the image, we add a small "residual" at multiple early timesteps to gently influence the diffusion process. Just adding the source latent residual alone will lead to loss of noise too early in the diffusion process, leading to the same output as using just source latents as input noise. In order to smooth out the diffusion process, we will also add a small amount of additional noise between denoising steps. Since the Unet generates samples by removing noise, adding additional noise during the diffusion process is necessary to increase the expressiveness of the Unet for a better image.

Generally for diffusion models, the latent input to the next time step, l_{next}, is just the predicted latent matrix from the previous timestep l_{pred}. Our modification to the diffusion process can be formally defined as:

$$l_{\mathit{next}} = (1-a-b)l_{\mathit{pred}} + a \times l_{\mathit{src}}\left(\frac{||l_{\mathit{pred}}||_F}{||l_{\mathit{src}}||_F}\right) + b \times \epsilon\left(\frac{||l_{\mathit{pred}}||_F}{||\epsilon||_F}\right)$$

... with epsilon representing the added noise, l_{src} being the latent representation of the content image, and the Frobenius norm being used to fix the scale of the latents thus avoiding instability in the diffusion process. Our weight factors a and b for source image residual and noise are generated with exponential decay over the diffusion timesteps. Namely:

$$a = a_{\mathit{start}} \times \exp(\alpha_1 \frac{-t}{\mathit{timesteps}})$$

$$b = b_{\mathit{start}} \times \exp(\alpha_2 \frac{-t}{\mathit{timesteps}})$$

a_{start}, b_{start} control the impact of these factors weight at the start of the denoising process, while the α values control how fast these elements decay, t representing the current timestep of diffusion. Depending on (α_1, α_2), the effective source residual and additive noise quickly approach zero, allowing for later steps of the diffusion process to function as standard latent diffusion steps. We note the observed behavior of this technique is heavily dependent on these hyperparameters.

As with Experiment 1, we use generated captions with appended and prepended style tokens produced from Dreambooth finetuning runs when generating samples.

3.4. Experiment 3: Compositional Style Transfer

We attempted to replicate Experiment 1 using multiple learned style tokens. Empirically, the model completely fails to incorporate elements from multiple learned styles, revealing a limitation in the finetuning method we employed. Due to extremely poor empirical results, we did not formally evaluate this approach any further.

3.5. Evaluation

We use ArtFID to evaluate performance of all three experiments:

$$ArtFID(X_g, X_c, X_s) = \\ (1 + \frac{1}{N}\sum_{i=1}^{N} d(X_c^{(i)}, X_g^{(i)})) \cdot (1 + FID(X_s, X_g)$$

ArtFID combines two aspects of style transfer, where $d(X_c^{(i)}, X_g^{(i)})$ measures content preservation (via taking the LPIPS metric between each content image and stylized image), and $FID(X_s, X_g)$ represents style matching (via taking the Frechet distance between the feature distributions of the style images X_s and the stylized images X_g).

4. Dataset & Features

4.1. Selected Content Images

We select the five images displayed in Table 1 as our designated content to test various style transfer methods. All selected content images are in the public domain. Criteria for selection include:

- Variety of mediums: Images include photography, painting, animation, and a woodblock print.

- Variety of subjects: Subjects include animals, people, and environments.

(And as you can see, we like the color blue, a lot.)

4.2. Selected Styles

We turned to HuggingFace datasets to find comprehensive image sets that represent unique art styles. The four styles chosen for this study are listed in Table 2: Impressionism, Chinese Landscape, Basquiat, and Skribbl (yes, you read that right). We chose these styles because they represent a broad spectrum of artistic expressions, so they form a robust set of styles to test our style transfer techniques.

5. Experiments, Results, & Discussion

5.1. Experiments

For Experiment 0 (our baseline), we used an Adam optimizer with a learning rate of 0.03 to update the generated image, because these hyperparameters allowed us to balance convergence speed and image quality. Image results are in Table 3.

For Experiment 1, we used a small learning rate of 5e-06, a batch size of 2 with gradient accumulation steps of 2, as well as precision training with gradient checkpointing to lower memory usage. The learning rate and batch size were selected based on task complexity and hardware constraints, and we used of mixed precision training and gradient checkpointing to optimize memory efficiency. Image results are in Table 4.

For Experiment 2, we used the same training runs as Experiment 1, only updating the inference method. For our hyperparameters, we used $a_{start} = 0.018$ and $b_{start} = 0.009$, which are the starting weights for source image residual and noise components respectively. For decay factors, we used $\alpha_1 = 6.5$ and $\alpha_2 = 4.5$, so that the added image residual decays faster than the added noise. Last but not least, image results are in Table 5.

5.2. Results

The following table contains our ArtFID scores for each style we transferred. Before each style transfer, we first computed the average style image from all training images of that style. After transfer, we used each average style image as a point of reference and compared content images against their corresponding stylized images.

Experiment → **Style ↓**	0	1	2
Impressionism	328.05	360.6	354.14
Chinese Landscape	282.35	391.62	306.3
Basquiat	615.65	511.55	463.37
Skribbl	481.89	599.64	520.39

5.3. Discussion

5.3.1 Experiment 0

Our results with the neural style transfer baseline show clear weaknesses with the technique. The output images for the most part represent somewhat naive superpositions of the content and style images. In the case where the low level features of an image are **not** homogenous, the method fails completely. For instance, for the Skribbl art style, there only exist low-level features to transfer for a small part of the image, which leads to an awkward superposition. For images with richer low-level details though, the neural style transfer results look much more satisfying.

5.3.2 Experiments 1 and 2

We observe clear failure modes for these experiments too. Some learned styles could be transferred to some content, but not other content. For instance, for Experiment 1 the generation for sheep (Content Image 1) in the learned Basquiat (<BAS>) style shows no sign of any influence from Basquiat. This is in spite of the qualitative observation the trained model is quite capable of creating Basquiat-style images. When prompted with "**<BAS>** art" the model produced the following image which is very high-fidelity in terms of style as shown in fig 1. This speaks to the strengths and weaknesses of the Dreambooth finetuning method. The method works by modifying how the Stable Diffusion models responds to certain prompts, namely those with the special token associated with the training theme. However, some input prompts will be too dissimilar to the training prompts to elicit the learned behaviour from the diffusion model even when the special token is included in the prompt. Further work includes modifying the Dreambooth training regime such the output model is more sensitive to the learned special token across inputs prompts.

On the positive side, we observe that incorporating source image latent residuals during the diffusion process successfully maintains source image composition while allowing novel image generation in a new style. We note that some variation in the output image is the result of a random crop transform we used when encoding the content image as latents. For instance, generated variants based on the Studio Ghibli content image (Content Image 5) often only include one of the subjects — the boy facing the window or the girl facing us — which is a result of this random crop transform. Replacing this random crop with an interpolation-based re-scaling transform would fix this issue for square input images.

Despite the success in maintaining image composition, the residual-controlled generations are generally inferior in quality to the just-text samples. They are a bit blurry, have less detail, and often do not achieve the target style. Since the samples generated tend to be a bit flat, the method does work well for reproducing the Skribbl style. While these issues could potentially be fixed by finding better hyperparameters for the residual and noise magnitude schedules, the decrease in sample quality seems like a reasonable result as we are tampering with the diffusion process. Generating quality samples with this method would likely require an additional trained component, since taking a weighted average between predicted and content latents is a somewhat naive approach to controlling generation by visual features. One direction for improving this method would to be to introduce a learned model which can combine these latent matrices intelligently. This model could then be trained with a reward model that represents a pseudo-label for image quality.

Another issue with this approach is the nature of the latents produced by the VAE. These latents are more of a condensed version of the image pixel values rather than a high-level semantic representation of the image. For this reason, introducing the residual for the content latent has the effect of introducing low-level features from the source image, rather than the high-level semantic features we desire. Introducing these latents therefore

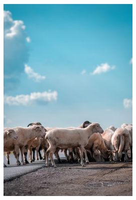
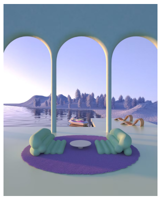
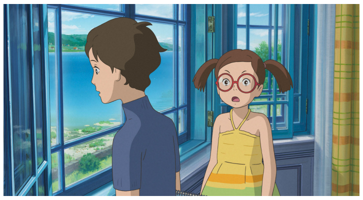
				
Content Image 1	Content Image 2	Content Image 3	Content Image 4	Content Image 5

Table 1: Our dazzlingly designated content images. Sourced linked in labels.

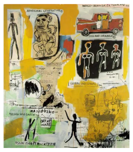
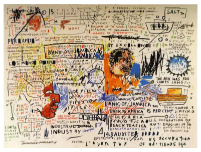
Style Name	Example Images
Impressionism	
Chinese Landscape	
Basquiat	 
Skribbl	

Table 2: Our stylishly selected styles. Sources are linked in labels.

consigns the generated image to have the same colors palette as the original image, which is not desired for the style transfer task. Directly introducing high-level features and not the low level features is not possible with the latent representations from the Stable Diffusion VAE. We do essentially capture these high level features in the semantic capture setup where these desired features are represented with a textual prompt.

6. Conclusion

Our project explores state-of-the-art techniques for neural style transfer, highlighting strengths and weaknesses of each approach. Fine-tuning Dreambooth with semantic content capture demonstrated promising results in preserving both style and content; language-vision models guide the generation process effectively. However, the introduction of source image residuals during diffusion, while maintaining composition, reveals challenges in achieving high-quality results due to inherent limitations in current diffusion models. Notably, the direct application of neural style transfer showcases limitations in fidelity and adaptability, particularly evident in more complex style-image combinations.

Still, fine-tuning Dreambooth with semantic content capture emerges as the highest-performing algorithm, offering better control over style transfer and preservation of content. This success can be attributed to the model's ability to learn strong semantic priors from image-caption pairs, enabling the synthesis of outputs with specific subjects preserved across different contexts. The challenges encountered with other techniques, such as diffusion with source image residuals, underscore the importance of further research to address limitations in current methodologies.

For future work, exploring enhancements to diffusion models, such as incorporating more sophisticated methods for incorporating source image information, could lead to significant improvements in style transfer fidelity. Additionally, investigating the integration of language-vision models with diffusion-based approaches may offer new avenues for enhancing content preservation and style control. Moreover, continued research into evaluation metrics tailored for style transfer tasks could provide more nuanced insights into model performance, guiding the development of more effective algorithms. Overall, the report lays a solid foundation for future research directions in AI-driven image transformation, emphasizing the need for interdisciplinary approaches to tackle complex challenges in this domain.

References

[1] A. Casanova, M. Careil, J. Verbeek, M. Drozdzal, and A. Romero-Soriano. Instance-conditioned GAN. *CoRR*, abs/2109.05070,

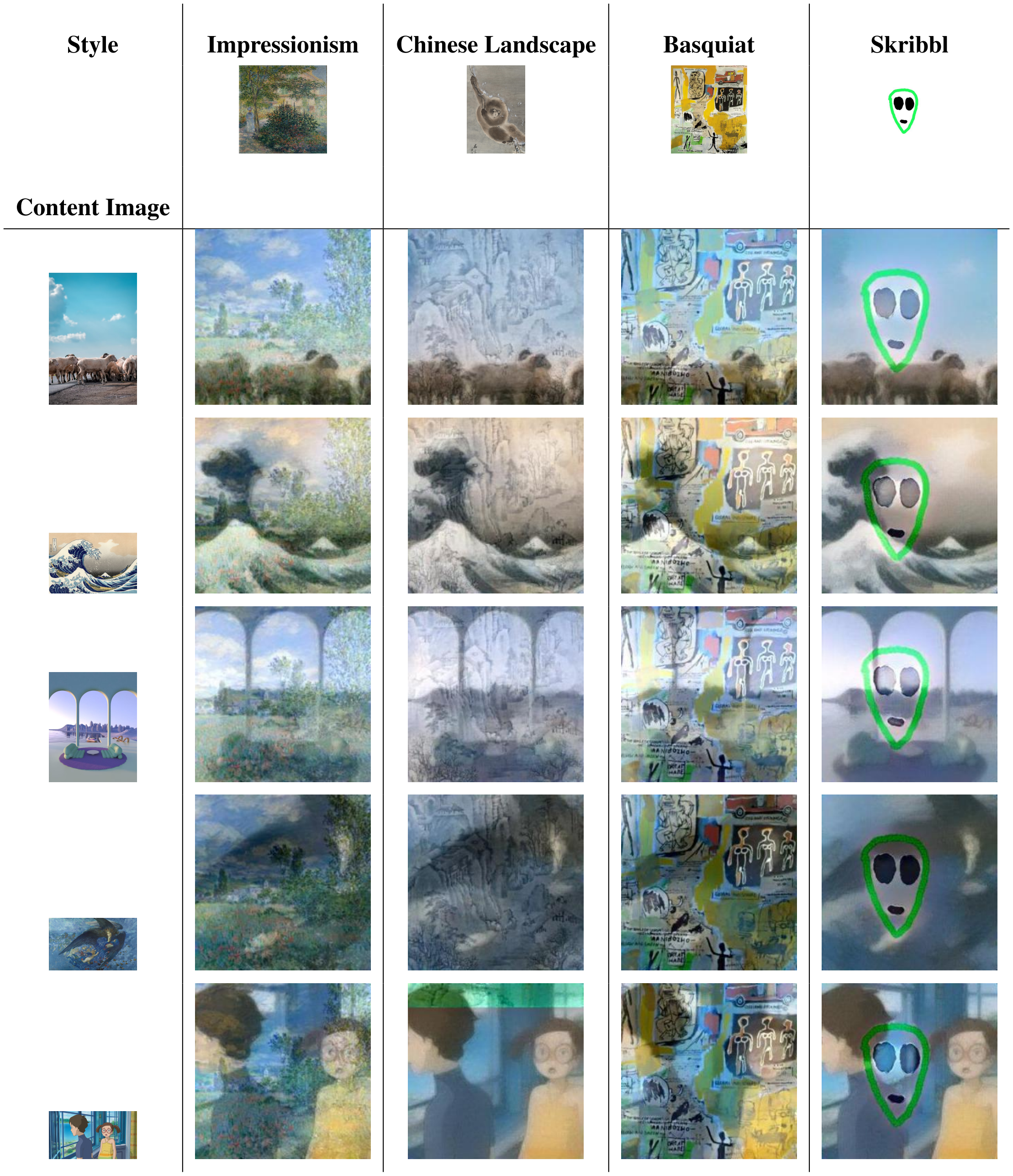


Table 3: Baseline image results. As you can see, content and style images are basically just stacked on top of each other.

2021.

[2] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[3] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[4] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016.

[5] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[7] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

[8] A. Sanakoyeu, D. Kotovenko, S. Lang, and B. Ommer. A style-aware content loss for real-time HD style transfer. *CoRR*, abs/1807.10201, 2018.

[9] M. Wright and B. Ommer. Artfid: Quantitative evaluation of neural style transfer, 2022.

[10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018.

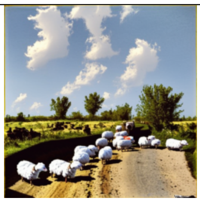
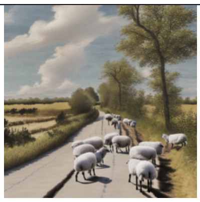
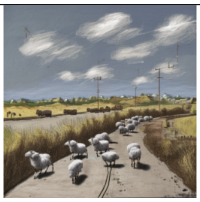
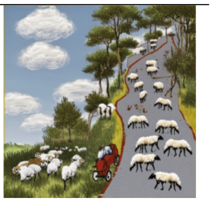
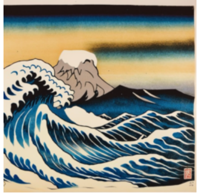
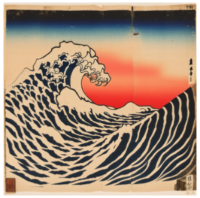
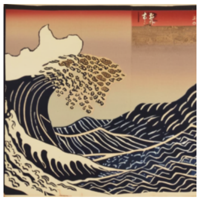
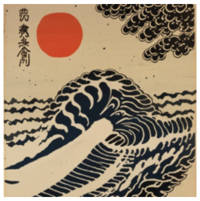
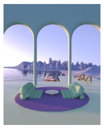
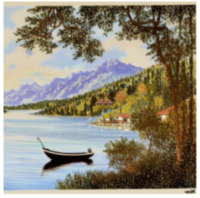
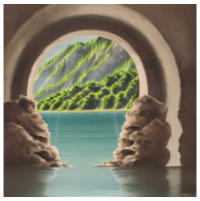
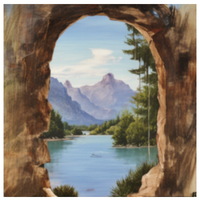
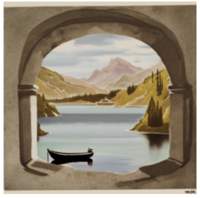
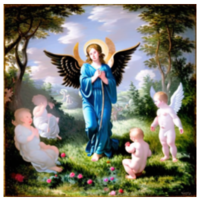
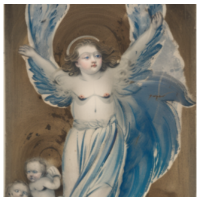
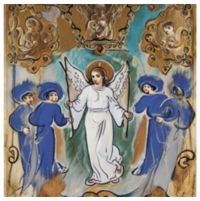
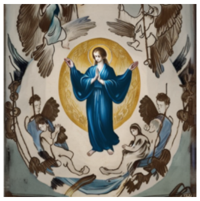
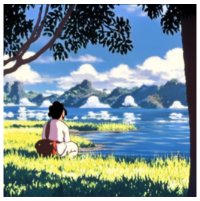
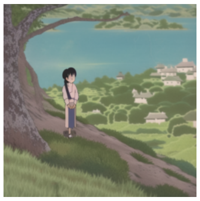
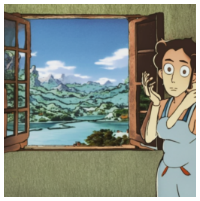
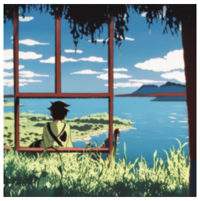
Style Content Image	Impressionism 	Chinese Landscape 	Basquiat 	Skribbl
				
				
				
				
				

Table 4: Finetuned model image results. Note that style images are now the average image of all style images, since we are now training on multiple style images.

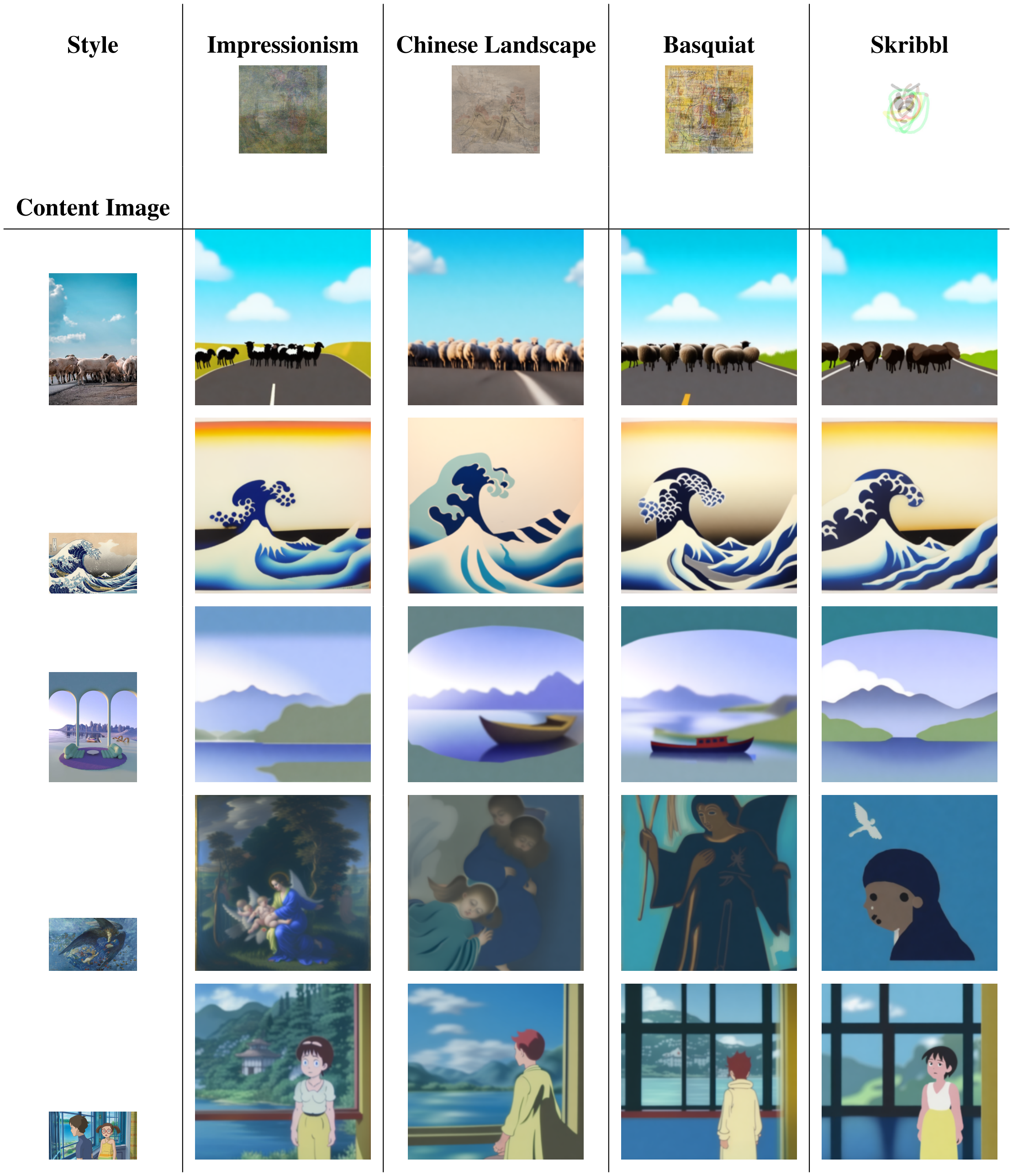


Table 5: Reresolved model image results. Similar to the table of finetuned images, note that style images are now the average image of all style images, since we are now training on multiple style images.

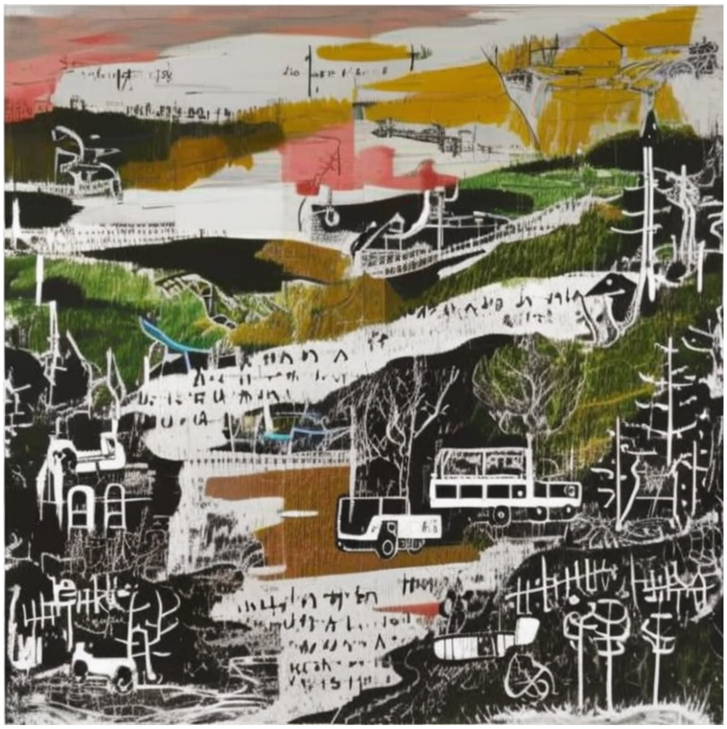


Figure 1: This sample — generated with prompt ”<**BAS**> art” — shows high fidelity to the Basquiat style.