Supplementary Materials: Leveraging Diffusion Models for Stylization using Multiple Style Images

Dan Ruta Abdelaziz Djelouah Raphael Ortiz Christopher Schroers
DisneyResearch|Studios
dan.ruta@disney.com abdelaziz.djelouah@disney.com

In this document we provide some additional details and visualizations.

1. Details of the two-stage pipeline

As describe in the main paper, to achieve high resolution and high quality style transfer results, we use a two-stage pipeline. This is illustrated in Figure 1. First a lower resolution version of the image is stylized. Once done, the result is resized and used as starting point for the second stage, where the same stylization technique is used, this time using higher resolution controlnet guidance maps. The improvement in terms of quality of the style transfer are clearly visible between the first and second stage.

2. Results on Different Diffusion Models

The technique can be applied to other latent diffusion models. We provide here initial results with SD-XL [2]. Some additional work is required to focus on most relevant layers [1], and push the quality.

3. Post-processing for Better Results

Although our results are state-of-the-art, there are some remaining issues. For example, in Figure 4, the brightness contrast that the neon style (style 15 from our evaluation) is not well rendered. However the over quality is sufficiently good that some basic color matching with the original style can improve the results.

4. Benefits of Fine-tuning

In our experiments with multiple style images, we fine-tune the IPAdapter model with the set of style images. Figure 3 illustrates the effect of this fine tuning that helps better match the style.

5. Gram Metric for Style Evaluation

Despite widespread use in early years of style transfer literature, we do not use the Gram metric for quantifying the quality of style transfer in our work. This metric measures correlation of mostly textural information, and does not accurately capture higher level style features and semantics. We include some examples in Figure 5, showing the Gram values of some of our own results, as well as ones from another method (here, IPAdapter). We sorted the images by the Gram value, and show some of the lowest scoring ones on the top row, and some of the highest scoring ones on the bottom row. On the top row, we show the same pair of results. Here, despite having similar gram values, the results from our method are evidently much more faithful to the source style. On the bottom row, we show further such pairs, where our results have higher Gram values (therefore worse) than those from another method. However, again, the results are clearly higher quality. On both of these rows, this perceptual quality difference is not reflected in the Gram value.

References

[1] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora, 2024. 1

[2] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 1

[3] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 2

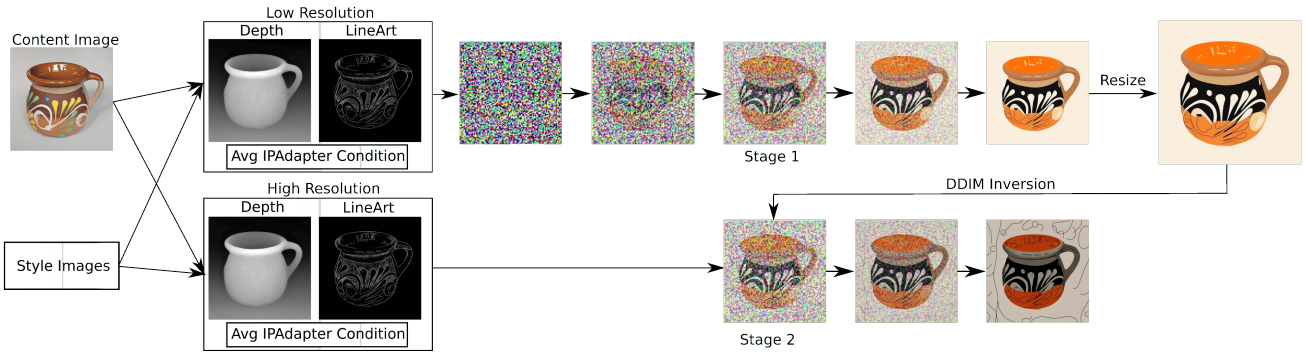

Figure 1. Visualization of our two-stage pipeline. We first produce a lower resolution image, with 512px on the shortest side - a resolution closer to the training data of Stable Diffusion. Next, we upscale the image and refine it with an additional few iterations of our pipeline, following a partial inversion. We use ControlNet and IPAdapter inputs of matching resolutions for each stage.

Figure 2. Representative example of applying our technique to a different base model, SDXL, testing out its generalization. We use the Canny ControlNet adapter for SDXL, as a LineArt version is not currently available.

Figure 4. Exposure issues with our method when stylizing with Style 15, which contains very dark areas, alongside neon lights. We show some stylization examples, and the same results with some simple color adjustment, to better match the range of the style.

Figure 3. Comparing stylization results with and without adapter fine-tuning. A small amount of finetuning can be beneficial to achieve better style transfer results.

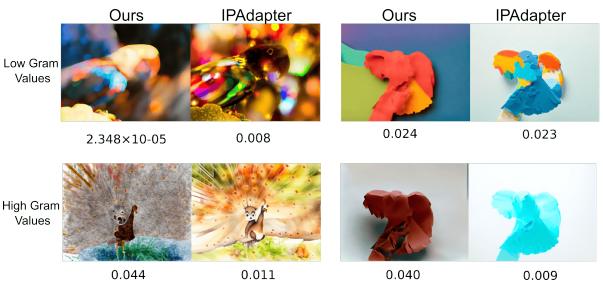

Figure 5. High Gram metric value is supposed to indicate worse results. However we note that at both ends of the spectrum, low and high, our method obtains better quality compared to IPAdapter [3]. Additionally, the metric itself does not align with the quality of the stylization with IPAdapter obtaining lower (hence better) values in most cases.