Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer

Supplementary Material

1. Appendix

Ablation study for color transfer capability. To validate the efficacy of the ablated methods for color transfer, we employ the RGB-uv histogram proposed in HistoGAN [1] to measure color transfer capability. Specifically, for a given input image I, we convert it into the log-chroma space. For example, choosing the R color channel as the primary and normalizing by G and B yields:

$$I_{uR}(x) = \log(\frac{I_R(x)+\epsilon}{I_G(x)+\epsilon}), I_{vR}(x) = \log(\frac{I_R(x)+\epsilon}{I_B(x)+\epsilon}) \tag{1}$$

where the I_R, I_G, I_B refer to the color channels of the image I, ϵ is a small constant for numerical stability, and x is the pixel index.

Then, they compute the intensity $I_y(x) = \sqrt{I_R^2(x)+I_G^2(x)+I_B^2(x)}$ for weighted scaling and differentiable the histogram. The final histogram follows:

$$\mathbf{H}(u,v,c) \propto \Sigma_x k(I_{uc}(x), I_{vc}(x), u, v) I_y(x), \tag{2}$$

where $I_{uG}, I_{vG}, I_{uB}, I_{vB}$ are R and B color channels which projected to the log-chroma space similar to Eq. 1, $c \in \{R, G, B\}$, and $k(\cdot)$ is a inverse-quadratic kernel.

We utilize the Histogram Loss [1] as a color similarity metric which measures the Hellinger distance between the histograms of stylized and style images.

$$C(\mathbf{H}_g, \mathbf{H}_t) = \frac{1}{\sqrt{2}} \| \mathbf{H}_{cs}^{\frac{1}{2}} - \mathbf{H}_s^{\frac{1}{2}} \|_2, \tag{3}$$

where $\mathbf{H}_{cs}$ and $\mathbf{H}_s$ are color histograms of stylized and style image, respectively, $\| \cdot \|$ is the standard Euclidean norm, and $\mathbf{H}^{\frac{1}{2}}$ denotes an element-wise square root. We adopt the default configuration of HistoGAN [1]. For a detailed description of the histogram loss, please refer to the original HistoGAN paper [1].

As a result, we evaluate the efficacy of Initial Latent AdaIN in color tone transfer. In Tab. 1, each proposed component contributes to transfer the color tone of the given style image. Especially, we confirm that the Initial Latent AdaIN prominently affects the for transferring of color tones.

Qualitative comparison with ablation of attention temperature scaling. To highlight the effects of attention temperature scaling, we provide some examples of style transfer results while ablating the attention scaling. As shown in Fig. 1, we validate that the attention scaling makes the model to synthesize sharp images and well-preserve the patterns in the given style image (e.g. stars in left example). This experimental result confirms the significance of the proposed attention temperature scaling method. Note that, we use $\gamma = 0.3$ for this experiment, to keep the strong effect of style transfer in visualization.

	Configuration	Histogram Loss [1] ↓
A	Ours ($\gamma = 0.75$, default)	**0.2804**
B	- Style Injection	0.4637
C	- Attention Scaling	0.3029
D	- Initial Latent AdaIN	0.5235

Table 1. Ablation study for color transfer capability.

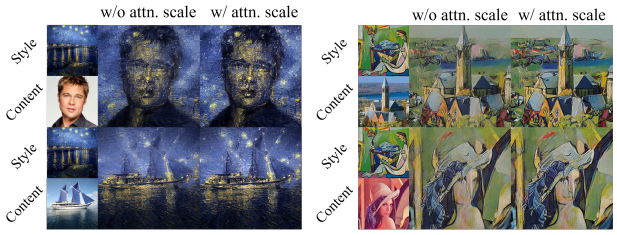

Figure 1. Qualitative comparison while ablating the attention temperature scaling. Attention temperature scaling prevents blurry results and helps to keep the local textures in the style image. We use $\gamma = 0.3$ for this experiment.

Quantitative comparison in the other set. In Tab. 2, we conduct quantitative experiments on a new set of style-content pairs (20 contents, 40 styles) randomly sampled without any overlap with original images. As reported, the performance enhancement of the proposed method still holds, confirming hyperparameters are well-generalized.

LPIPS is affected by texture and color, as it is based on CNN features [2]. To evalute the content and color independently, we measure LPIPS-Grayscale and Histogram-loss in supplementary against the recent and lowest ArtFID baselines (AesPA-Net, InST, AdaAttN). As reported in Tab. 2, ours achieves lowest LPIPS-Grayscale, and highest color similarity.

Analysis on feature space of query preservation. Fig. 2 visualizes features of Q_t^c, Q_t^s, Q_t^{cs}, and $\tilde{Q}_t^{cs}$ for a style-content pair. As shown, interpolated features ($\tilde{Q}_t^{cs}$) are located in in-distribution nearby contents, since we gradually combine content query (Q_t^c) and stylized one (Q_t^{cs}) along

	Ours	AesPA-Net	AdaAttN	InST
ArtFID	**30.38**	34.55	31.87	39.11
FID	**18.87**	21.09	19.21	20.46
LPIPS	**0.528**	0.563	0.576	0.822
LPIPS-Gray	**0.417**	0.443	0.450	0.731
Histogram-loss	**0.303**	0.321	0.331	0.653

Table 2. Quantitative comparison in newly sampled test set.

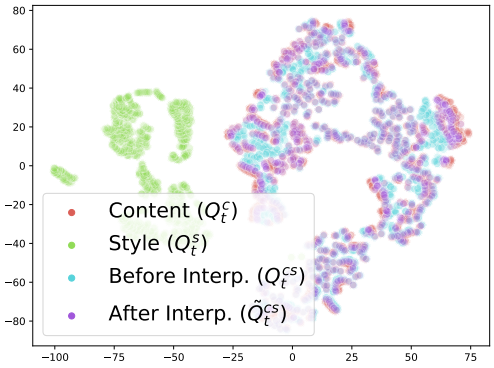

Figure 2. t-SNE visualization of query in SA for a style-content pair. Query of content, style, and stylized ones (Q_t^{cs} and $\tilde{Q}_t^{cs}$) at t=20 and 7$^{\text{th}}$ decoder layer are used for visualization.

	ArtFID	FID	LPIPS-Gray
Ours (w/ empty prompt, default)	34.9	21.2	**0.47**
Ours (w/ BLIP prompt)	**34.5**	**20.9**	**0.47**

Table 3. Ablation study of the null text token in the diffusion process.

with entire reverse process.

Furthermore, we compute the average distance of $\tilde{Q}_t^{cs}$ toward top-5 Nearest Neighbors (NNs) in (content, style, itself) and the number of them in NNs for all injected layers with t=[10, 20, 30, 40]. Distances and # NNs are (5.49, 9.06, 4.43), (1.24, 0.00, 3.76), implying $\tilde{Q}_t^{cs}$ residing in in-distribution nearby content.

Style transfer with text prompts. In this paragraph, We exploit text prompt, obtained by BLIP [3], for DDIM inversion instead of null text token. Images in 'data_vis' in the official repository are used, in which easy to caption as they mostly consist of single object. As a result, ours w/ text shows slight improvement as in Tab. 3.

User study. We compare ours with AesPA-Net and InST, the most recent conventional and diffusion methods, for 18 users and 10 examples per user. We observe that (57.2%, 76.7%) of users prefer the proposed method over (AesPA-Net, InST). Note that, ours has a much faster inference speed than InST.

Qualitative comparison with StyleDiffusion. As the implementation of StyleDiffusion [5] is unavailable, we compare ours with examples in supplementary of StyleDiffusion [5]. We obtain style-content pairs of StyleDiffusion in repositories of their baselines. We observe that ours is more suitable for transferring local textures, while StyleDiffusion tends to change the structure of the image significantly, as shown in Fig. 4. We hypothesize that optimizing the style in CLIP [4]'s semantically rich feature space forces StyleDiffusion to be trained in that manner.

Figure 3. Qualitative comparisons with diffusion-based baselines

Figure 4. Qualitative comparison with StyleDiffusion. * denotes cropped version of images. We use $\gamma = 0.5$ for visualization.

Additional qualitative results. We additionally compare the proposed method with the most recent baseline (AesPA-Net) and baseline with the lowest ArtFID (AdaAttN). Fig. 3 shows the additional qualitative comparison of ours with diffusion model baselines. Moreover, as shown in Fig. 5, 6, we observe that ours better-transfers the local texture of a given style into the content image.

Also, in Fig. 7, 8, we visualize the style transfer results of various pairs of content and style images.

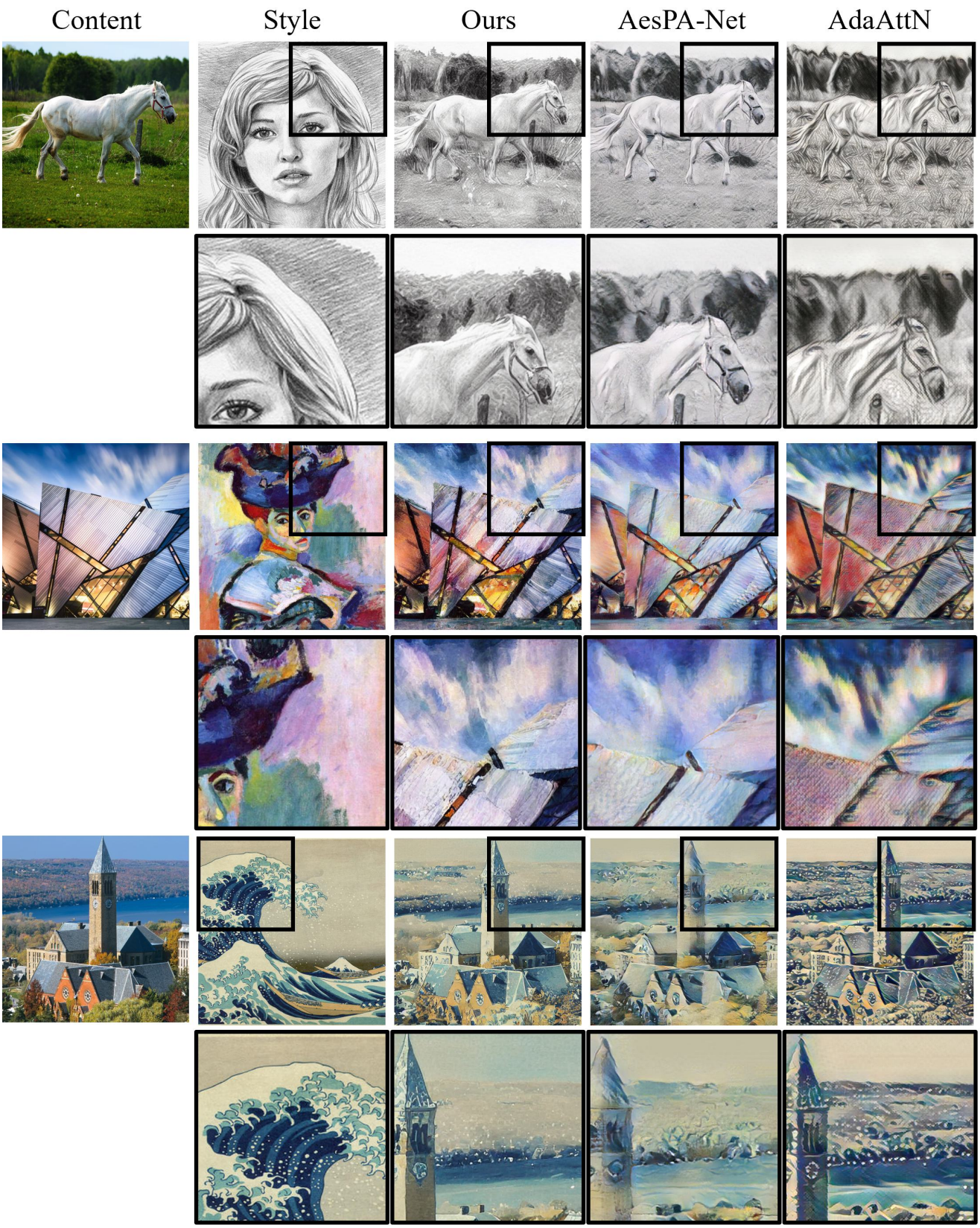

Figure 5. Qualitative comparison with baselines (AesPA-Net, AdaAttN). For visualizing the detailed textures, we provide the cropped version of the style image and its stylized counterparts in the second row of every content-style pair. Zoom in for viewing details.

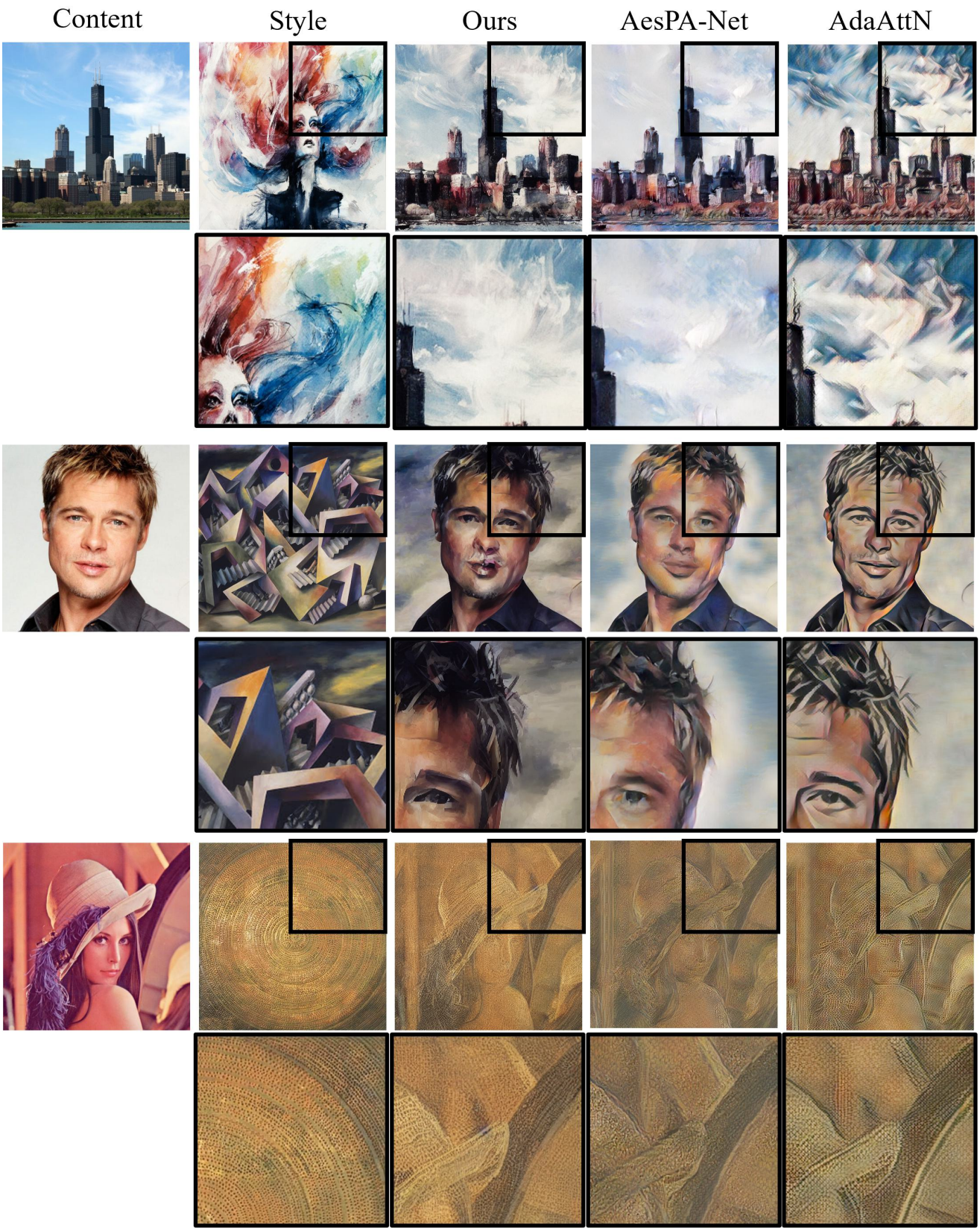

Figure 6. Qualitative comparison with baselines (AesPA-Net, AdaAttN). For visualizing the detailed textures, we provide the cropped version of the style image and its stylized counterparts in the second row of every content-style pair. Zoom in for viewing details.

Figure 7. Style transfer results of style and content image pairs. Zoom in for viewing details.

Figure 8. Style transfer results of style and content image pairs. Zoom in for viewing details.

References

[1] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7941–7950, 2021. 1

[2] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1

[3] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[5] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023. 2