Supplementary Materials for "HSI: A Holistic Style Injector for Arbitrary Style Transfer"

Shuhao Zhang[1,2], Hui Kang[1,3], Yang Liu[1], Fang Mei[1,3], Hongjuan Li[1]
[1] College of Computer Science and Technology, Jilin University, Changchun, 130012, China
[2] Academic Affairs Office, Jilin University of Arts, Changchun, 130021, China
[3] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, China

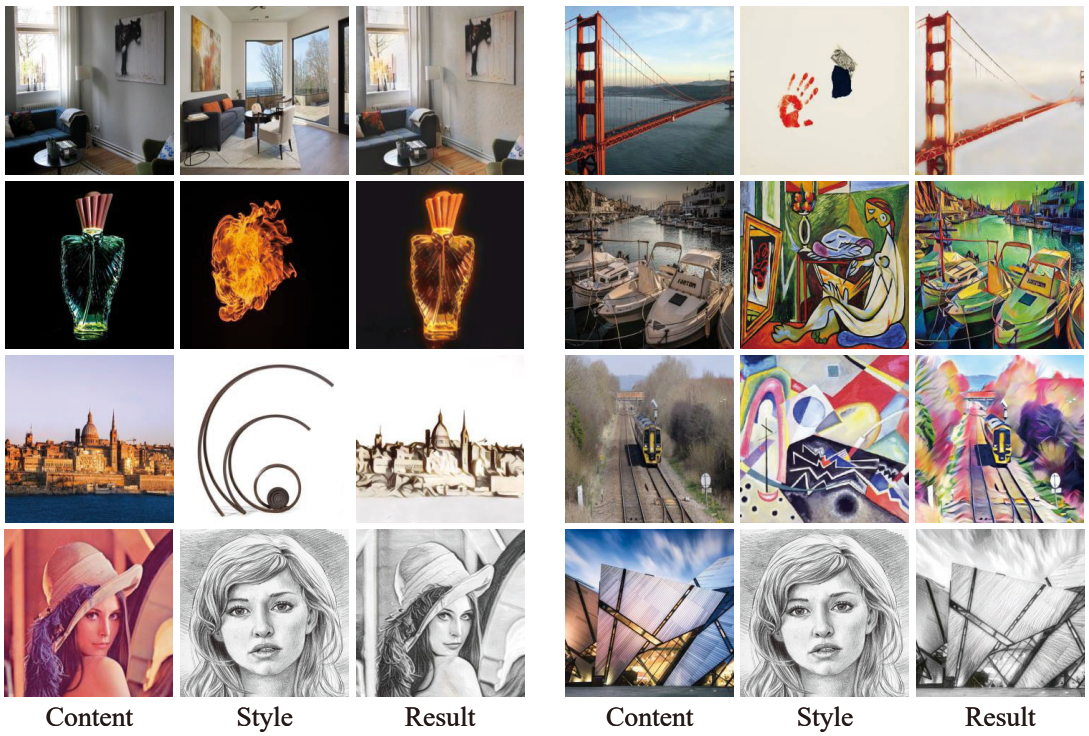

Figure S1. Stylization with diverse styles.

A. More Stylization Results

Figure S1 presents more stylized images with real photos, abstract styles and artistic paintings, which demonstrates the powerful and flexible stylization capability of our method.

B. More Implementation Details

We use 4 HSI modules to perform feature transformation between encoder and decoder. In the HSI, the global content feature Q_c is obtained through global average pooling.

C. Matrix Multiplication "⊗" VS Element-Wise Multiplication "⊙".

Compared with self-attention module, a key improvement of HSI is replacing matrix multiplication "⊗" with element-wise multiplication "⊙". Note that "⊗" or "⊙" is only a small component of the transform module. The "⊗" can establish a dense point-to-point mapping relationship between content and style features for stylization. However,

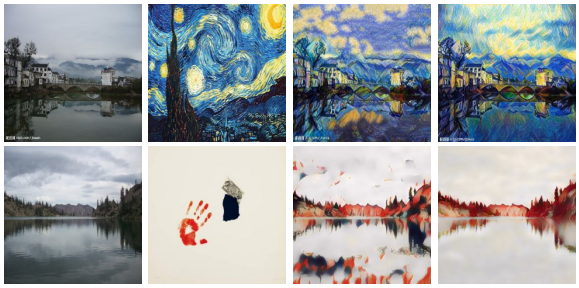

Figure S2. Comparison between self-attention (SA) and HSI.

Figure S3. Comparison with diffusion based methods.

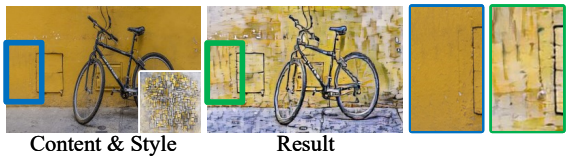

Figure S4. Limitation in "middle" styles.

a single-point feature is hard to convey clear style patterns or may even introduces noises. In contrast, HSI adopts representative style statistical features (mean, variance, *etc*.) to establish flexible connections with local and global content through "$\odot$" and dynamic networks. This is a more flexible and powerful style-matching paradigm than self-attention, which can adaptively match local or global styles for different content regions. To validate this, we conduct experiments by replacing HSI with the self-attention in our framework, and the results are shown in Figure S2. HSI is significantly better than self-attention in capturing local starry patterns (1st row) and maintaining global style distribution (2nd row). In addition, HSI only has a linear computational complexity that is more suitable for real-time style transfer than self-attention.

D. Comparison with Diffusion based Models

Most existing diffusion-based arbitrary style transfer (AST) models focus on improving the aesthetics of stylization results. To achieve this goal, they always distort the content structure and even semantics, which significantly differs from the traditional style transfer. DreamStyler [1] is a typical representative of this class of methods. Recently, a new diffusion-based style transfer method called StyleID [2] was proposed. Compared with DreamStyler, StyleID not only leverages the powerful generation performance of the diffusion model but also can maintain the consistency of input content and style. We compare our model with DreamStyler and StyleID, and the results are shown in Figure S3. In terms of visual effects, our method outperforms DreamStyler in content consistency and StyleID in capturing brush lines (e.g., the clouds in the sky). Additionally, the slow inference speed and high computational cost make it challenging for diffusion-based methods to meet the real-time requirement. StyleID takes nearly 30 seconds / 20 GB GPU memory to render a 512×512 image, whereas our method only takes about 0.008 seconds / 0.6 GB. Nevertheless, we think that users may prefer diffusion-based methods due to their flexibility and controllability for stylization.

E. Limitation

Although our method can achieve satisfactory results, it tends to capture relatively soft patterns when dealing with some style images with dense and repetitive textures. As shown in Figure S4, the square patterns in the resulting image are not as apparent as in the input style, which may not meet the stylization requirements of some users for the prominent styles. This limitation might be due to HSI is prefer to obtain more "middle" styles for highly textured styles. In the future, we will explore the feasibility of enhancing the HSI module to capture prominent styles.

References

[1] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 674–681, 2024. 2

[2] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8795–8805, 2024. 2