# –Supplementary Material–
# Pluggable Style Representation Learning for Multi-Style Transfer

Hongda Liu[1], Longguang Wang[2], Weijun Guan[1], Ye Zhang[1], and Yulan Guo[1]

[1] The Shenzhen Campus of Sun Yat-Sen University, Sun Yat-Sen University
[2] Aviation University of Air Force
{liuhd36@mail2.,guoyulan@}sysu.edu.cn

We provide the following supplementary contents:

– Section 1: Detailed architecture of our SaMST.
– Section 2: More quantitative comparison results.
– Section 3: More model analysis.
– Section 4: Results and comparisons on video style transfer.
– Section 5: Additional stylization results produced by our SaMST and other methods.
– Section 6: More incremental training results produced by our SaMST and Stylebank [3].

## 1 Architecture Details

The detailed architecture of the encoder and decoder is summarized in Table 1. The encoder comprises 3 standard convolutional layers. The decoder is symmetrical to the encoder, but with 2 more upsampling layers.

The architecture of style-aware block (SAB) is shown in Table 2. SConv, SRAdaIN and SCM all contain a corresponding MLP. The style representation is fed to MLPs to predict dynamical model weights. After the 3 style-wise modules (SConv, SRAdaIN and SCM), the stylized image features are fed into a point-wise convolutional layer.

## 2 More Quantitative Comparison Results

To demonstrate the effectiveness of our method, We evaluate recent state-of-the-art methods on NVIDIA RTX 3090 GPU (24GB). The quantitative results are shown in Table 3. Note that, We obtain the results of the methods by following their official code with default configurations. It is clear that our SaMST achives comparable efficiency advantage. Moreover, our SaMST also achieves the best stylized quality according to the stylized quantitative metrics.

We also implement a user study. In the study, a single sample consists of a content image, a style image, and 10 corresponding stylization results generated by the 10 methods. We randomly select 25 content images and 25 style images to generate 25 samples for each user. For each sample, a user is asked to vote for the one that he/she likes the most. Finally, we collect 1000 votes from 40 users and calculate the percentage of votes

**Table 1:** Details of our encoder and decoder.

| Part | Layer | Kernel_size | Stride | Channel | Group | Activation |
|---|---|---|---|---|---|---|
| Encoder | Conv | 9 | 1 | 3→16 | 1 | - |
| | Instance Norm | - | - | 16 | - | ReLU |
| | Conv | 3 | 2 | 16→32 | 1 | - |
| | Instance Norm | - | - | 32 | - | ReLU |
| | Conv | 3 | 2 | 32→64 | 1 | - |
| | Instance Norm | - | - | 64 | - | ReLU |
| Decoder | Upsample | - | 1/2 | - | - | - |
| | Conv | 3 | 1 | 64→32 | 1 | - |
| | Instance Norm | - | - | 32 | - | ReLU |
| | Upsample | - | 1/2 | - | - | - |
| | Conv | 3 | 1 | 32→16 | 1 | - |
| | Instance Norm | - | - | 16 | - | ReLU |
| | Conv | 9 | 1 | 16→3 | 1 | ReLU |

**Table 2:** Details of SAB. 'IF' and 'OF' is short for 'input feature' and 'output feature', respectively.

| Part | Layer | Kernel_size | Stride | Channel | Group | IF | OF | Activation |
|---|---|---|---|---|---|---|---|---|
| SConv | MLP | - | - | - | - | 16 | 64×3×3 | ReLU |
| | DepthwiseConv | 3 | 1 | 64→64 | 64 | - | - | - |
| SRAdaIN | MLP | - | - | - | - | 16 | 64×2 | ReLU |
| | Instance Norm | - | - | 64 | - | - | - | ReLU |
| SCM | MLP | - | - | - | - | 16 | 4 | PReLU |
| | | - | - | - | - | 4 | 64 | Sigmoid |
| | Multiplication | - | - | 64 | - | - | - | ReLU |
| PointwiseConv | Conv | 1 | 1 | 64→64 | 1 | - | - | ReLU |

that each method received. The results are shown in Fig. 1. That demonstrates that our method produces results with better stylized quality.

In summary, both quantitative results and user study demonstrate that our method achieves the best overall performance.

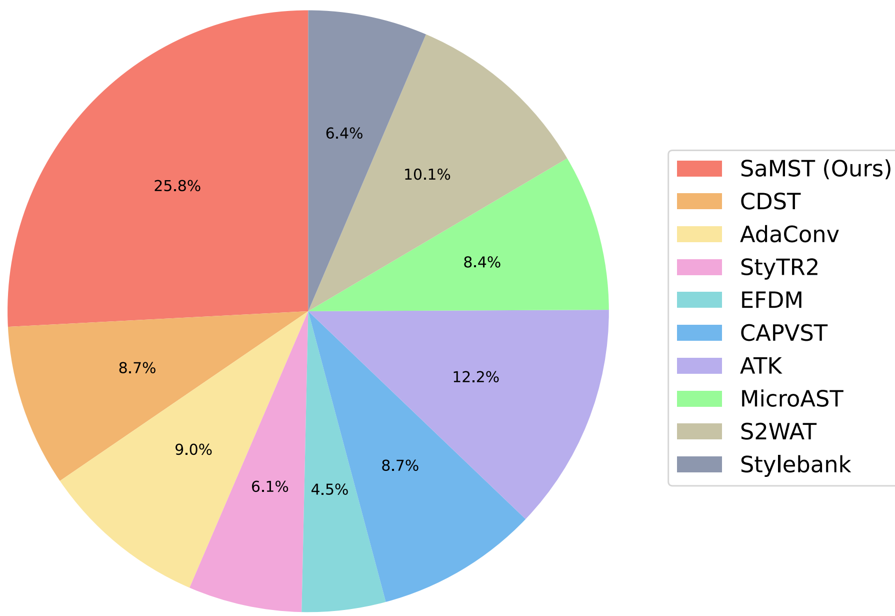


**Fig. 1:** User preference study of 10 methods.

**Table 3:** Quantitative comparison of the style transfer methods.Methods marked with * are MST based approaches, while other methods are AST based approaches. The best and second best results are in red and blue colors, respectively. Run time and FLOPs are evaluated on $512 \times 512$ images. "+" represents that the method expands new styles without forgetting. 'OIP' is short for 'once inference parameters', which refers to the number of parameters involved in one stylization inference.

| Methods | Efficiency | | | | | Metric | | |
|---|---|---|---|---|---|---|---|---|
| | Size (M)↓ | OIP (M)↓ | GFLOPs↓ | Time (ms)↓ | Capacity↑ | ArtFID↓ | CF↑ | GE+LP↑ |
| SaMST* (Ours) | 0.91 | 0.11 | 5.31 | 1.03 | 50k+ | 25.20 | 0.538 | 1.515 |
| Stylebank* [3] | 590.79 | 1.97 | 35.48 | 5.60 | 500+ | 64.02 | 0.296 | 1.089 |
| CIN* [8] | 161.68 | 1.68 | 40.58 | 6.23 | 50k | 54.38 | 0.435 | 1.253 |
| MSGNRT* [28] | 2.39 | 2.39 | 81.99 | 11.75 | 500 | 74.31 | 0.317 | 1.143 |
| ASN [9] | 11.00 | 11.00 | 54.98 | 17.69 | $\infty$ | 41.32 | 0.459 | 1.176 |
| AesUST [23] | 31.30 | 31.30 | 195.67 | 19.35 | $\infty$ | 60.79 | 0.392 | 1.345 |
| StyTR2 [6] | 35.39 | 35.39 | 1283.45 | 194.76 | $\infty$ | 46.78 | 0.475 | 1.337 |
| AdaAttN [15] | 13.63 | 13.63 | 293.21 | 26.81 | $\infty$ | 62.36 | 0.494 | 1.305 |
| IELCL [4] | 20.91 | 20.91 | 194.67 | 15.51 | $\infty$ | 42.70 | 0.424 | 1.473 |
| StyleFormer [26] | 19.90 | 19.90 | 172.14 | 13.45 | $\infty$ | 37.59 | 0.507 | 1.445 |
| AdaIN+ArtFlow [1] | 6.46 | 6.46 | 517.01 | 98.83 | $\infty$ | 58.43 | 0.385 | 1.234 |
| WCT+ArtFlow [1] | 6.46 | 6.46 | 517.01 | 102.07 | $\infty$ | 53.23 | 0.327 | 1.348 |
| MCCNet [5] | 17.76 | 17.76 | 259.09 | 30.28 | $\infty$ | 47.43 | 0.365 | 1.321 |
| MANet [7] | 19.60 | 19.60 | 278.62 | 29.44 | $\infty$ | 72.21 | 0.432 | 1.416 |
| TSPR [20] | 28.29 | 28.29 | 363.17 | 44.29 | $\infty$ | 83.28 | 0.393 | 1.289 |
| Linear [14] | 12.17 | 12.17 | 150.26 | 12.48 | $\infty$ | 68.58 | 0.381 | 1.248 |
| AvatarNet [19] | 7.01 | 7.01 | 142.38 | 229.31 | $\infty$ | 72.18 | 0.246 | 1.354 |
| AdaIN [13] | 7.01 | 7.01 | 142.38 | 11.63 | $\infty$ | 90.32 | 0.404 | 1.036 |
| CAST [31] | 10.52 | 10.52 | 142.38 | 11.68 | $\infty$ | 38.23 | 0.471 | 1.375 |
| UCAST [32] | 10.52 | 10.52 | 142.38 | 11.68 | $\infty$ | 47.99 | 0.426 | 1.258 |
| EFDM [30] | 7.01 | 7.01 | 63.30 | 14.92 | $\infty$ | 58.10 | 0.351 | 1.295 |
| AesPA [10] | 24.20 | 24.20 | 314.27 | 337.84 | $\infty$ | 40.20 | 0.362 | 1.439 |
| DAUST [11] | 7.01 | 7.01 | 142.36 | 221.06 | $\infty$ | 52.49 | 0.423 | 1.348 |
| ATK [33] | 11.18 | 11.18 | 291.44 | 24.01 | $\infty$ | 34.87 | 0.515 | 1.265 |
| S2WAT [27] | 64.96 | 64.96 | 582.62 | 94.88 | $\infty$ | 38.74 | 0.452 | 1.388 |
| QuantArt [12] | 112.35 | 112.35 | 1066.99 | 116.76 | $\infty$ | 58.54 | 0.503 | 1.304 |
| AdaConv [2] | 62.83 | 62.83 | 145.68 | 15.54 | $\infty$ | 52.31 | 0.363 | 1.482 |
| CAPVST [25] | 4.09 | 4.09 | 179.89 | 33.36 | $\infty$ | 53.97 | 0.397 | 1.184 |
| MicroAST [24] | 0.47 | 0.47 | 11.06 | 3.96 | $\infty$ | 59.46 | 0.335 | 1.312 |
| CDST [21] | 2.42 | 2.42 | 39.52 | 247.67 | $\infty$ | 57.02 | 0.320 | 1.125 |
| SANET [18] | 20.91 | 20.91 | 267.74 | 20.59 | $\infty$ | 57.63 | 0.481 | 1.425 |
| STTR [22] | 45.64 | 45.64 | 110.32 | 48.69 | $\infty$ | 69.21 | 0.387 | 1.034 |
| PAMA [16] | 35.39 | 35.39 | 359.32 | 28.39 | $\infty$ | 55.39 | 0.448 | 1.292 |

**Table 4:** Quantitative ablation study of the number of SAB. The length of style representation is set to 16.

| #SAB | Efficiency | | | | Metric | | |
|---|---|---|---|---|---|---|---|
| | Size (M)↓ | OIP (M)↓ | GFLOPs↓ | Time (ms)↓ | ArtFID↓ | CF↑ | GE+LP↑ |
| 1 | 0.88 | 0.08 | 5.17 | 0.92 | 37.36 | 0.484 | 1.485 |
| 3 (Ours) | 0.91 | 0.11 | 5.31 | 1.03 | 25.20 | **0.538** | 1.515 |
| 5 | 0.95 | 0.15 | 5.45 | 1.15 | 24.46 | 0.516 | 1.573 |
| 7 | 0.98 | 0.18 | 5.60 | 1.29 | **20.87** | 0.502 | **1.604** |

**Table 5:** Quantitative ablation study of the length of style representation. The number of SAB is set to 3.

| Length | Efficiency | | | | Metric | | |
|---|---|---|---|---|---|---|---|
| | Size (M)↓ | OIP (M)↓ | GFLOPs↓ | Time (ms)↓ | ArtFID↓ | CF↑ | GE+LP↑ |
| 8 | 0.49 | 0.09 | 5.31 | 1.02 | 42.31 | **0.587** | 1.410 |
| 16 (Ours) | 0.91 | 0.11 | 5.31 | 1.03 | 25.20 | 0.538 | 1.515 |
| 32 | 1.76 | 0.16 | 5.31 | 1.03 | 22.19 | 0.503 | 1.563 |
| 64 | 3.45 | 0.25 | 5.31 | 1.05 | **19.34** | 0.511 | **1.596** |

## 3 More Model Analysis

In this section, we evaluation model variants with different number of SAB and style representation lengths. Note that, all of the model variants are trained on $50k$ style images.

### 3.1 Number of SAB

As shown in Table 4, complexity of SaMST increases linearly by adding more SAB. More SAB help the SaMST to achieves better ArtFID and GE+LP score. In Fig. 2, more SAB help SaMST learn more detailed texture patterns and more sufficient colors from style images. However, more SAB means significantly longer inference time and bigger computational volume.

### 3.2 Length of style representation

In our default setting, we set style representation length to 16. Then we propose 3 model variants with different style representation length. As shown in Table 5, when using longer style representation, we get better quantitative results. However, the model size increases significantly. As for visual results in Fig. 3, 16-dimension style representation already keeps good balance of style patterns and content preservation, which achieves competitive visual quality.

We prioritize model efficiency and complexity in our work for good application in real-world scenarios. So we use 3 SAB and 16-dimension style representation in our SaMST to make a trade-off. Users can adjust the architecture according to their practical requirements.

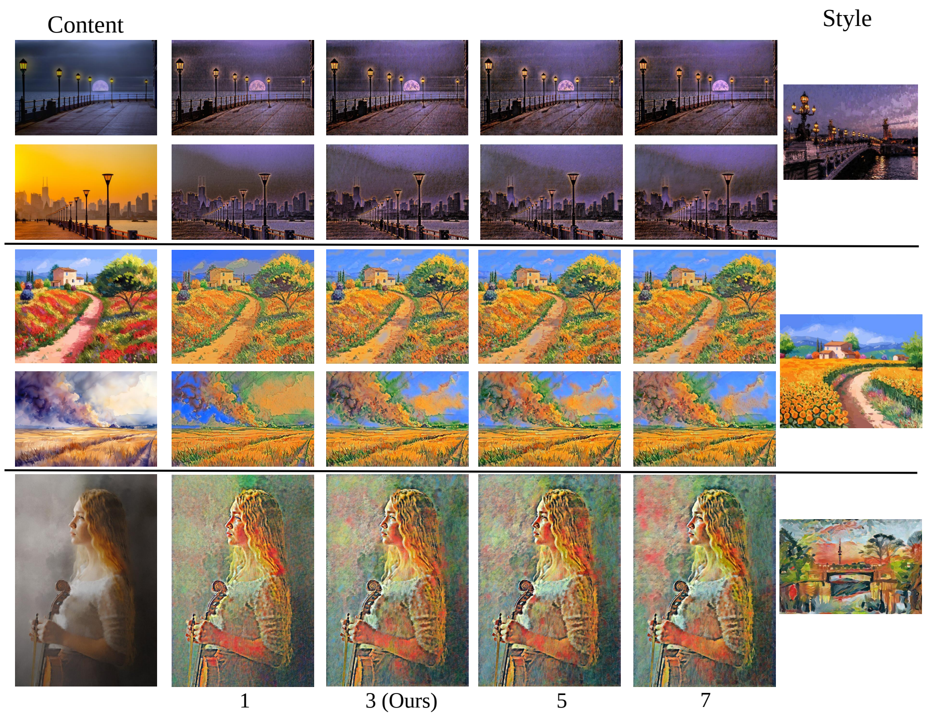


**Fig. 2:** Qualitative ablation study of the number of SAB.

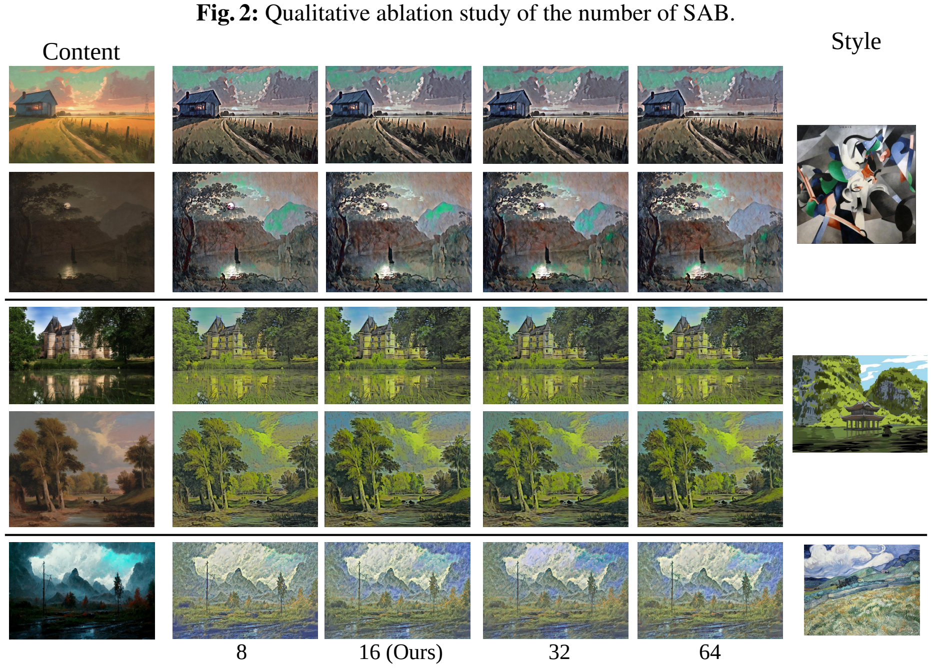


**Fig. 3:** Qualitative ablation study of the number of style representation length.

**Fig. 4:** Visualization of style representations learned from different style images.

### 3.3 Style Visualization

We further visualize the style representations learned from style images using the t-SNE method [17]. The result is shown in Fig. 4. It is demonstrated that our SaMST has good locality, which gathers visually similar styles into discriminative clusters. For example, the styles in the orange boxes contain black, white and grey colors, which look like the sketches.


**Fig. 5:** Comparison results on video style transfer.

# 4 Results and Comparisons on Video Style Transfer

Here we also provide results of video style transfer. As shown in Fig. 5, it generalizes well to video content. Our method preserves keep great balance in content details, style textures and style colors. In contrast, some methods are good at preserving scene content, but weak at extract style information (*e.g.*, MicroAST [24] and CAPVST [25]). Moreover, Stylebank [3] and AdaConv [2] pay more attention on local structures of the style image, which results in severe video distortion. PAMA [16] achieves relatively better visual quality, but contains scene distortion in results.

To show the stable results produced by our model, we further compute the difference between neighboring frames to show the smoothness between frames. As shown in Fig 6, the difference generated by our method keep stable structure similar to the difference from the input frames, especially in scene details. Moreover, the style textures keep relatively stable. It is because our method could well preserve the image content structure and style image textures.

Similar to [24], we employ LPIPS (Learned Perceptual Image Patch Similarity) [29] distance to quantitatively measure the stability and consistency of rendered scenes by computing the average perceptual distances between neighbor frames. We produce 50 videos for each method and report the average LPIPS distances in Table 6. And our SaMST obtains competitive score among including methods.

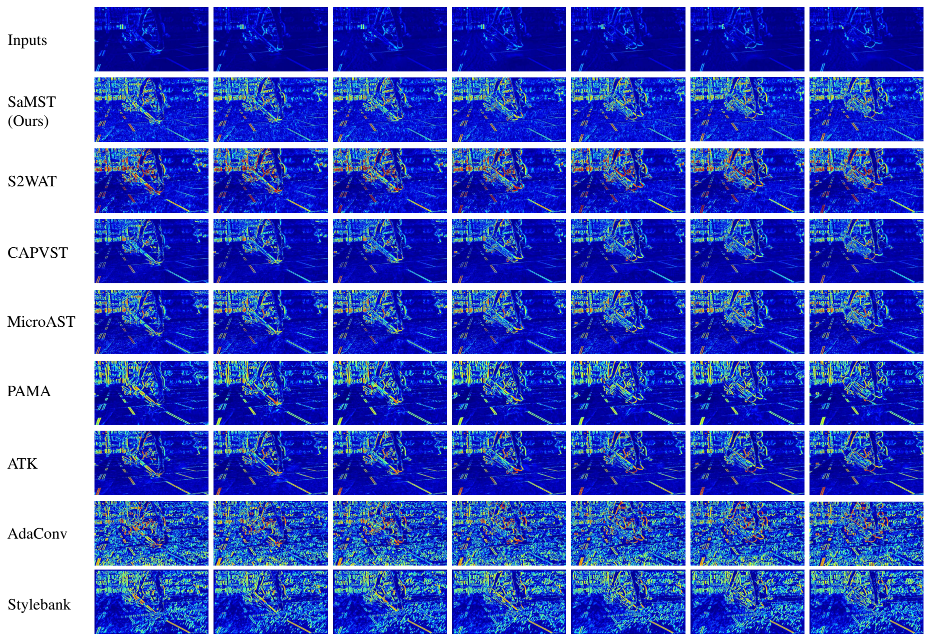


**Fig. 6:** Difference between neighboring frames.

**Table 6:** The average LPIPS [29] distances for different methods on video style transfer. The best and second best results are in red and blue colors, respectively.

| | Inputs | SaMST (Ours) | S2WAT [27] | CAPVST [25] | MicroAST [24] | PAMA [16] | ATK [33] | AdaConv [2] | Stylebank [3] |
|---|---|---|---|---|---|---|---|---|---|
| LPIPS ↓ | 0.263 | 0.332 | 0.365 | 0.317 | 0.348 | 0.401 | 0.385 | 0.466 | 0.475 |

## 5 Additional Stylization Results

We provide additional stylized results produced by our SaMST, as shown in Fig. 7. Moreover, we also provide stylized results of Stylebank [3] (Fig. 8), MicroAST [24] (Fig. 9) and AdaConv [2] (Fig. 10) for comparison. Our SaMST produces stylized images with sufficient content details and more accurate style patterns.

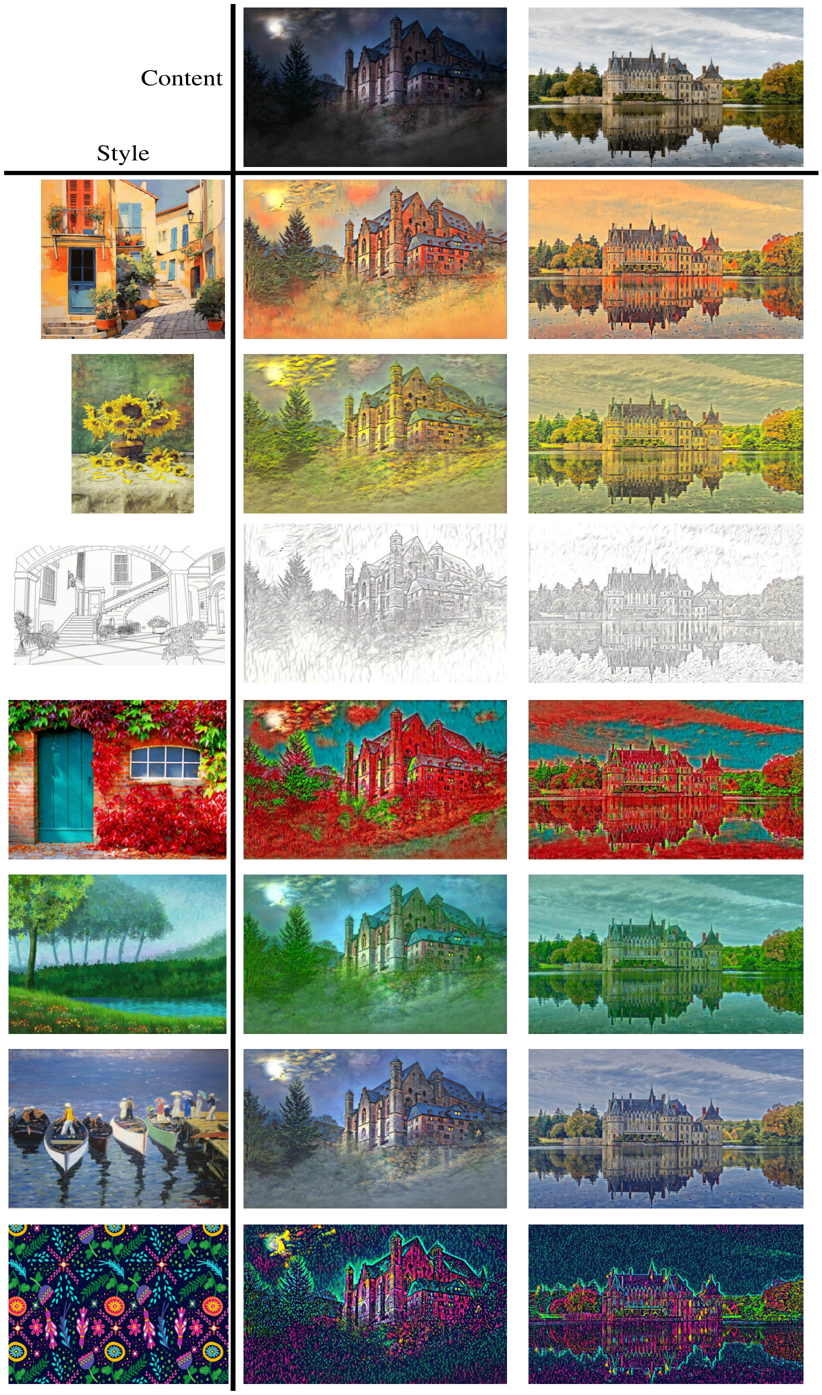


**Fig. 7:** Additional stylization results produced by our SaMST.


**Fig. 8:** Additional stylization results produced by Stylebank [3].


**Fig. 9:** Additional stylization results produced by MicroAST [24].

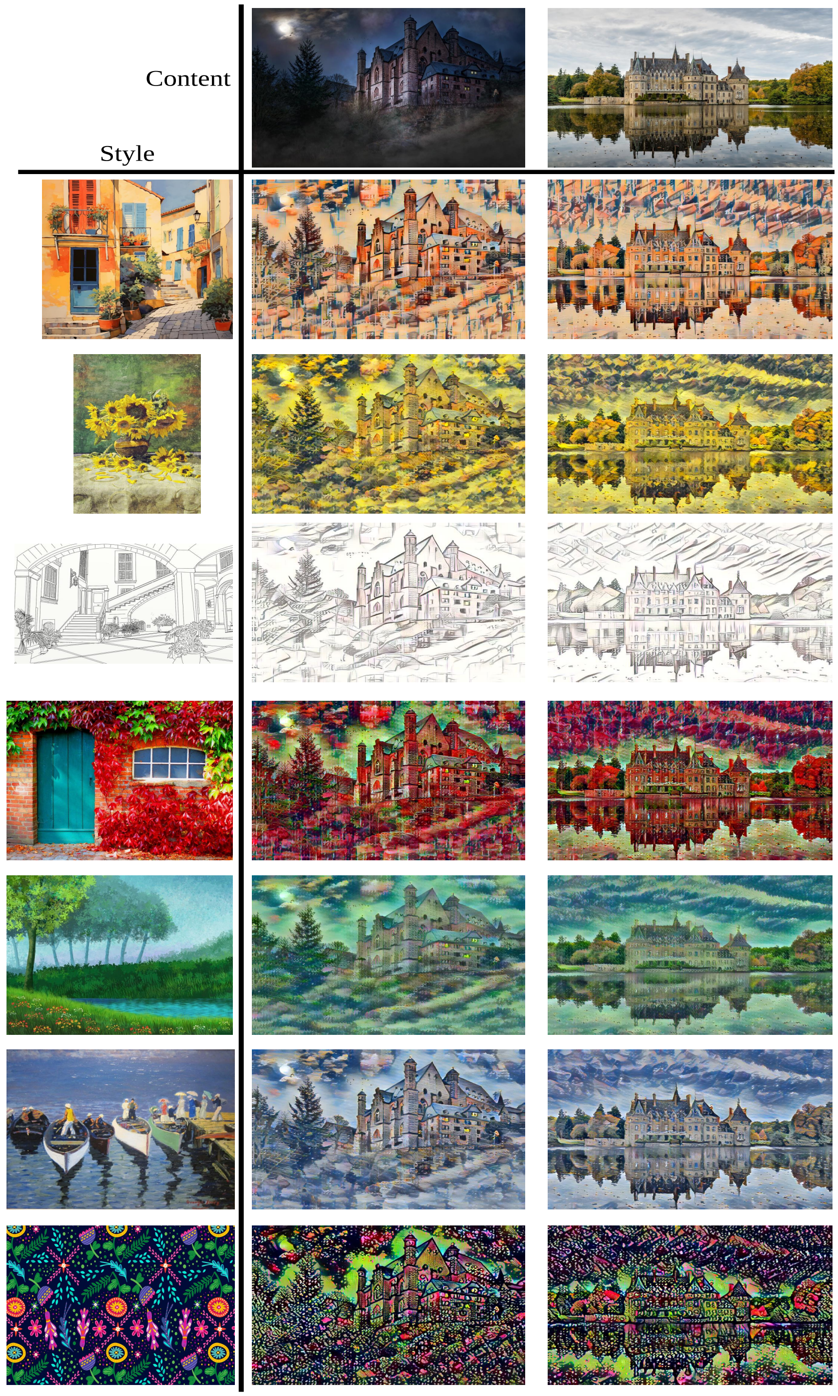


**Fig. 10:** Additional stylization results produced by AdaConv [2].

## 6 More Incremental Training Results

To validate the effectiveness of our incremental training scheme, we randomly select unseen styles to do style expansion. For each new style, we finetune corresponding style representation on only $1k$ content images for $3k$ iterations. It costs around 60s on a single NVIDIA RTX 3090 GPU. For comparison, we add new styles to Stylebank [3] in Table 3. Figure 11 shows the stylized results with new styles by our SaMST. And Fig. 12 shows the stylized results with new styles by Stylebank [3]. The results indicate that our incremental training scheme also achieves competitive visual quality, while Stylebank [3] produces stylized results with severe image distortion and artifacts.

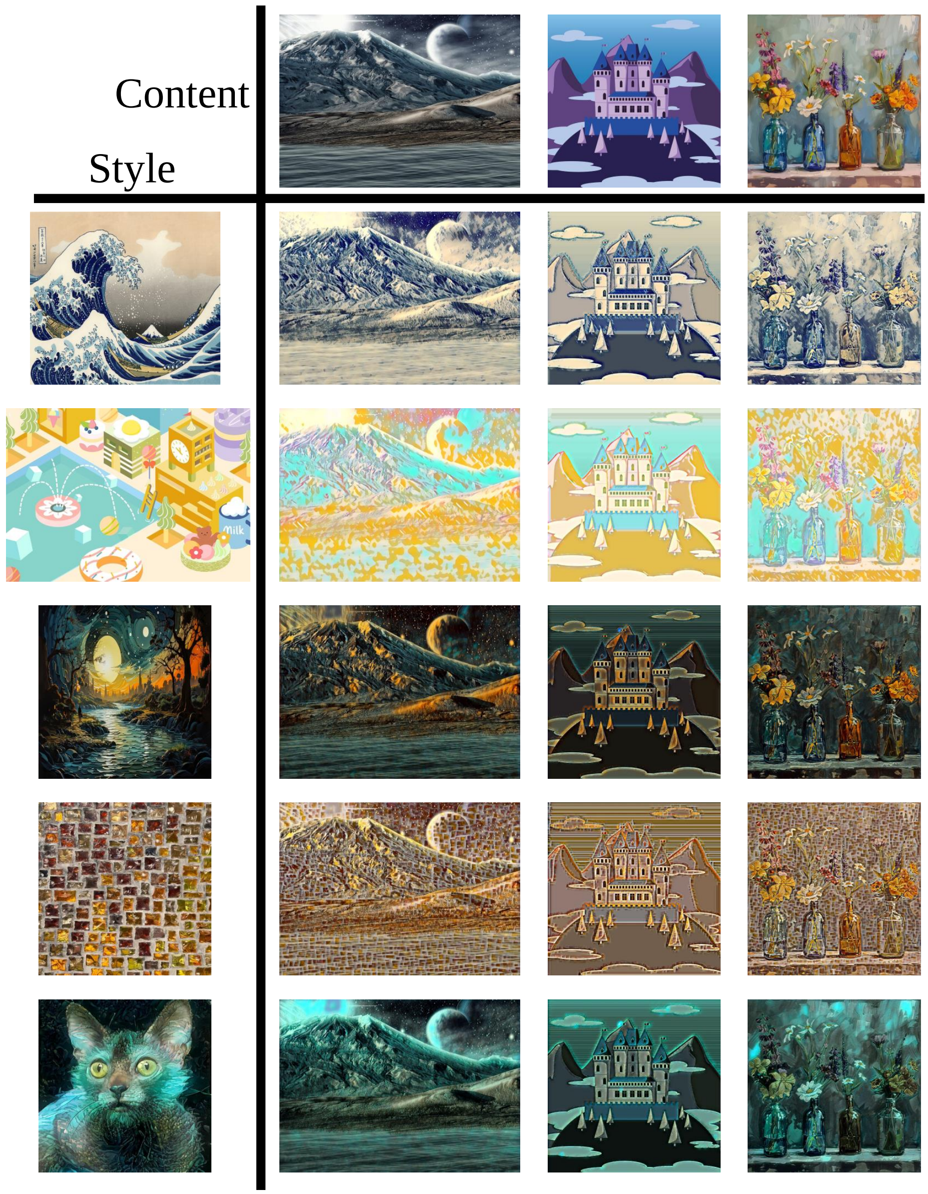


**Fig. 11:** Styles for incremental training and corresponding stylized results.

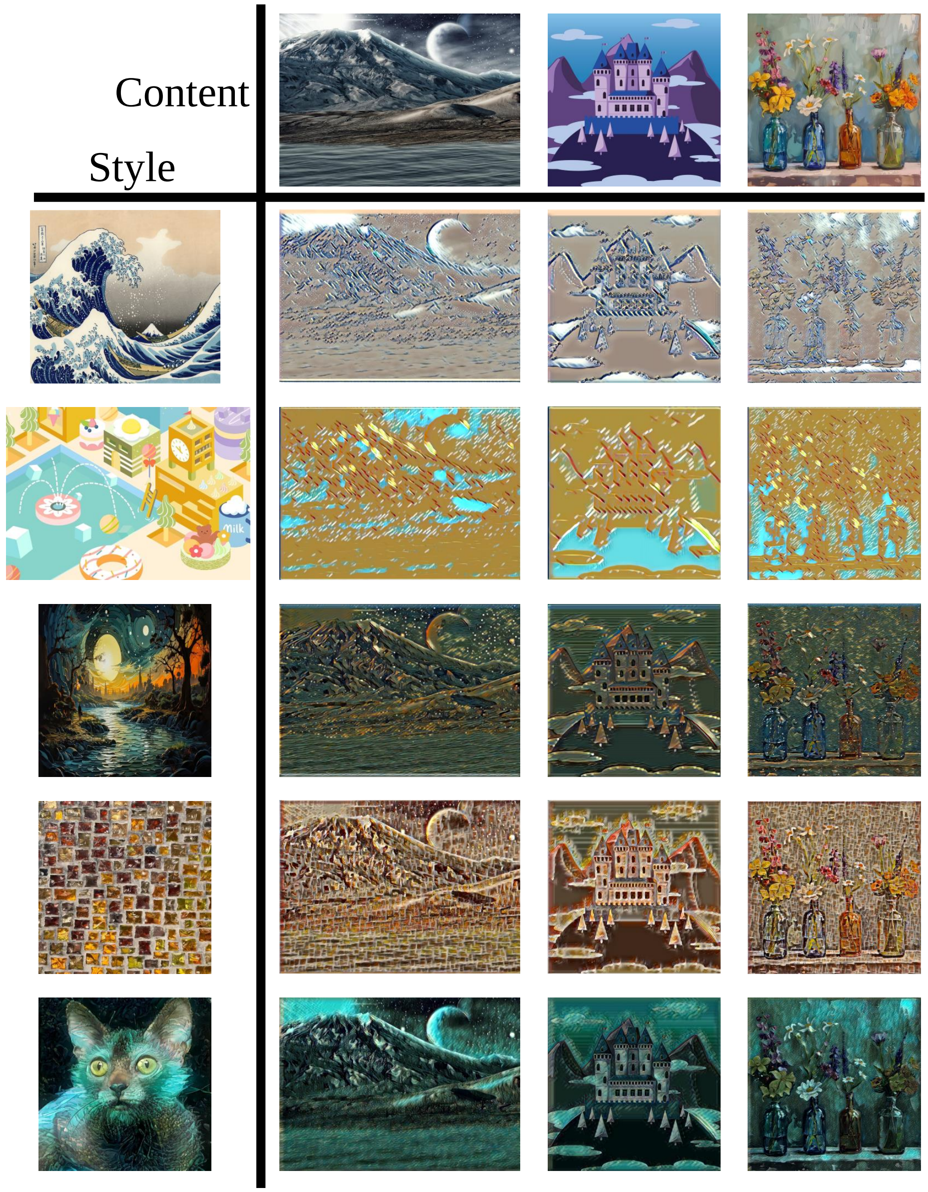


**Fig. 12:** Styles for incremental training and corresponding stylized results. Note that, the results are produced by Stylebank [3].

## References

1. An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J.: Artflow: Unbiased image style transfer via reversible neural flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 862–871 (2021) 4
2. Chandran, P., Zoss, G., Gotardo, P., Gross, M., Bradley, D.: Adaptive convolutions for structure-aware style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7972–7981 (2021) 4, 8, 9, 10, 14
3. Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G.: Stylebank: An explicit representation for neural image style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1897–1906 (2017) 1, 4, 8, 9, 10, 12, 15, 17
4. Chen, H., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., Lu, D., et al.: Artistic style transfer with internal-external learning and contrastive learning. Advances in Neural Information Processing Systems **34**, 26561–26573 (2021) 4
5. Deng, Y., Tang, F., Dong, W., Huang, H., Ma, C., Xu, C.: Arbitrary video style transfer via multi-channel correlation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1210–1217 (2021) 4
6. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr2: Image style transfer with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11326–11336 (2022) 4
7. Deng, Y., Tang, F., Dong, W., Sun, W., Huang, F., Xu, C.: Arbitrary style transfer via multi-adaptation network. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2719–2727 (2020) 4
8. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. arXiv preprint arXiv:1610.07629 (2016) 4
9. Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V., Shlens, J.: Exploring the structure of a real-time, arbitrary neural artistic stylization network. arXiv preprint arXiv:1705.06830 (2017) 4
10. Hong, K., Jeon, S., Lee, J., Ahn, N., Kim, K., Lee, P., Kim, D., Uh, Y., Byun, H.: Aespa-net: Aesthetic pattern-aware style transfer networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22758–22767 (2023) 4
11. Hong, K., Jeon, S., Yang, H., Fu, J., Byun, H.: Domain-aware universal style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14609–14617 (2021) 4
12. Huang, S., An, J., Wei, D., Luo, J., Pfister, H.: Quantart: Quantizing image style transfer towards high visual fidelity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5947–5956 (2023) 4
13. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017) 4
14. Li, X., Liu, S., Kautz, J., Yang, M.H.: Learning linear transformations for fast image and video style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3809–3817 (2019) 4
15. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6649–6658 (2021) 4
16. Luo, X., Han, Z., Yang, L.: Progressive attentional manifold alignment for arbitrary style transfer. In: Proceedings of the Asian Conference on Computer Vision. pp. 3206–3222 (2022) 4, 8, 9

17. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 7
18. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5880–5888 (2019) 4
19. Sheng, L., Lin, Z., Shao, J., Wang, X.: Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8242–8250 (2018) 4
20. Svoboda, J., Anoosheh, A., Osendorfer, C., Masci, J.: Two-stage peer-regularized feature recombination for arbitrary image style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13816–13825 (2020) 4
21. Wang, H., Li, Y., Wang, Y., Hu, H., Yang, M.H.: Collaborative distillation for ultra-resolution universal style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1860–1869 (2020) 4
22. Wang, J., Yang, H., Fu, J., Yamasaki, T., Guo, B.: Fine-grained image style transfer with visual transformers. In: Proceedings of the Asian Conference on Computer Vision. pp. 841–857 (2022) 4
23. Wang, Z., Zhang, Z., Zhao, L., Zuo, Z., Li, A., Xing, W., Lu, D.: Aesust: towards aesthetic-enhanced universal style transfer. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1095–1106 (2022) 4
24. Wang, Z., Zhao, L., Zuo, Z., Li, A., Chen, H., Xing, W., Lu, D.: Microast: Towards super-fast ultra-resolution arbitrary style transfer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2742–2750 (2023) 4, 8, 9, 10, 13
25. Wen, L., Gao, C., Zou, C.: Cap-vstnet: content affinity preserved versatile style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18300–18309 (2023) 4, 8, 9
26. Wu, X., Hu, Z., Sheng, L., Xu, D.: Styleformer: Real-time arbitrary style transfer via parametric style composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14618–14627 (2021) 4
27. Zhang, C., Xu, X., Wang, L., Dai, Z., Yang, J.: S2wat: Image style transfer via hierarchical vision transformer using strips window attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7024–7032 (2024) 4, 9
28. Zhang, H., Dana, K.: Multi-style generative network for real-time transfer. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018) 4
29. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 8, 9
30. Zhang, Y., Li, M., Li, R., Jia, K., Zhang, L.: Exact feature distribution matching for arbitrary style transfer and domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8035–8045 (2022) 4
31. Zhang, Y., Tang, F., Dong, W., Huang, H., Ma, C., Lee, T.Y., Xu, C.: Domain enhanced arbitrary image style transfer via contrastive learning. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–8 (2022) 4
32. Zhang, Y., Tang, F., Dong, W., Huang, H., Ma, C., Lee, T.Y., Xu, C.: A unified arbitrary style transfer framework via adaptive contrastive learning. ACM Transactions on Graphics **42**(5), 1–16 (2023) 4
33. Zhu, M., He, X., Wang, N., Wang, X., Gao, X.: All-to-key attention for arbitrary style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23109–23119 (2023) 4, 9