Supplementary Materials: AesFA: An Aesthetic Feature-Aware Arbitrary Neural Style Transfer

Joonwoo Kwon[1*], **Sooyoung Kim**[1*], **Yuewei Lin**[2 †], **Shinjae Yoo**[2 †], **Jiook Cha**[1 †]

[1]Seoul National University
[2]Brookhaven National Laboratory
{pioneers, rlatndud0513, connectome}@snu.ac.kr, {ywlin, sjyoo}@bnl.gov

Ablation Studies

In this supplementary section, we present corroborating experiments and ablation studies that validate several of our design choices.

The Efficacy of Frequency-based Feature Decomposition Encoders. To verify the efficacy of frequency-based feature decomposition, we replaced all encoder architecture with VGG-19 (Simonyan and Zisserman 2014) and compared it to the default model. In order to further validate the efficacy of the number of Octave Convolutions (OctConvs) in a single layer, models with three and five OctConvs are also examined. Other architectural designs, such as a generator, retain their default configurations. Our feature decomposition encoders with three OctConvs in each layer perform the best at all spatial resolutions, as shown in Table 1. In this paper, we employed the AesFA model with two layers, each consisting of three OctConvs in both content and aesthetic feature encoders.

Resolution	Method (Encoder)	Style Loss (↓)	LPIPS (↓)	SSIM (↑)	Time (↓)
256^2	VGG-19	2.618	0.519	0.096	**0.016**
	AesFA (3 OctConvs)	0.692	**0.368**	**0.417**	**0.016**
	AesFA (5 OctConvs)	**0.672**	**0.368**	0.408	0.023
512^2	VGG-19	1.936	0.513	0.127	**0.015**
	AesFA (3 OctConvs)	**0.314**	**0.365**	0.371	0.017
	AesFA (5 OctConvs)	0.330	0.366	**0.374**	0.022
$1024^2(1K)$	VGG-19	1.615	0.575	0.136	**0.016**
	AesFA (3 OctConvs)	**0.283**	**0.392**	**0.405**	0.020
	AesFA (5 OctConvs)	0.292	**0.392**	**0.405**	0.025
$2048^2(2K)$	VGG-19	1.597	0.561	0.143	**0.017**
	AesFA (3 OctConvs)	**0.404**	**0.435**	0.378	0.020
	AesFA (5 OctConvs)	0.442	0.440	**0.384**	0.023

Table 1: The Effectiveness of Octave Convolutions (OctConvs) in encoder architecture. Our content and aesthetic feature encoders, which are comprised of multiple OctConvs, outperform encoders whose architecture was replaced by VGG-19.

Ablation studies on Octave Convolutions. Our key idea was to decompose the input image into two distinct components to better encode aesthetic features and synthesize the aesthetically enhanced image. As shown in Figure 2 and Figure 3, our method shows excellent decomposition ability and proves that AesFA have notable ability in the feature disentanglement, resulting in better stylization quality. In addition, compared to the results of the models without the octave convolutions, the models with the octave convolutions show higher image quality with less vertical artifacts behind (see Figure 2).

To find the optimal alpha value (α), three alpha values, 0.25, 0.5 and 0.75 are examined. The $\alpha = 0.5$ setting produces a well-balanced feature decomposition between the two frequency images and shows pleasing stylization results (see Figure 3 and Table 2). We observe that this trend was maintained at other resolutions as well.

Method	GFLOPS (↓)	Storage(GB) (↓)	Params (10^6) (↓)	Time (↓)
No Oct	1537.547	8.412	4.671	0.013
$\alpha = 0.25$	1146.603	6.424	3.601	0.020
$\alpha = 0.50$ (main)	800.424	3.438	3.221	0.020
$\alpha = 0.75$	575.045	3.330	3.601	0.019

Table 2: Efficiency comparison with no octave convolutions and octave convolutions in different alpha values. The $\alpha = 0.5$ setting produces a well-balanced output between the efficiency and the performance. Storage is measured in PyTorch model. GFLOPs and Time are measured when the content and style image are both 2K (2048×2048) images. All tests were conducted on a single NVIDIA A100(40G) GPU and under identical iterations.

Effect of Adaptive Octave Convolution (AdaOct). In Table 3, we show the quantitative comparisons between the model with and without AdaOct. The model with AdaOct outperforms in terms of stylizations.

Method	Style Loss (↓)	LPIPS (↓)	SSIM (↑)	Time (↓)
No AdaOct	0.465	0.436	**0.399**	**0.019**
AdaOct	**0.427**	**0.435**	0.381	0.020

Table 3: Quantitative comparison between the model with and without Adaptive Octave Convolution (AdaOct) at 2K (2048×2048) resolutions.

*These authors contributed equally.
†Co-corresponding authors.

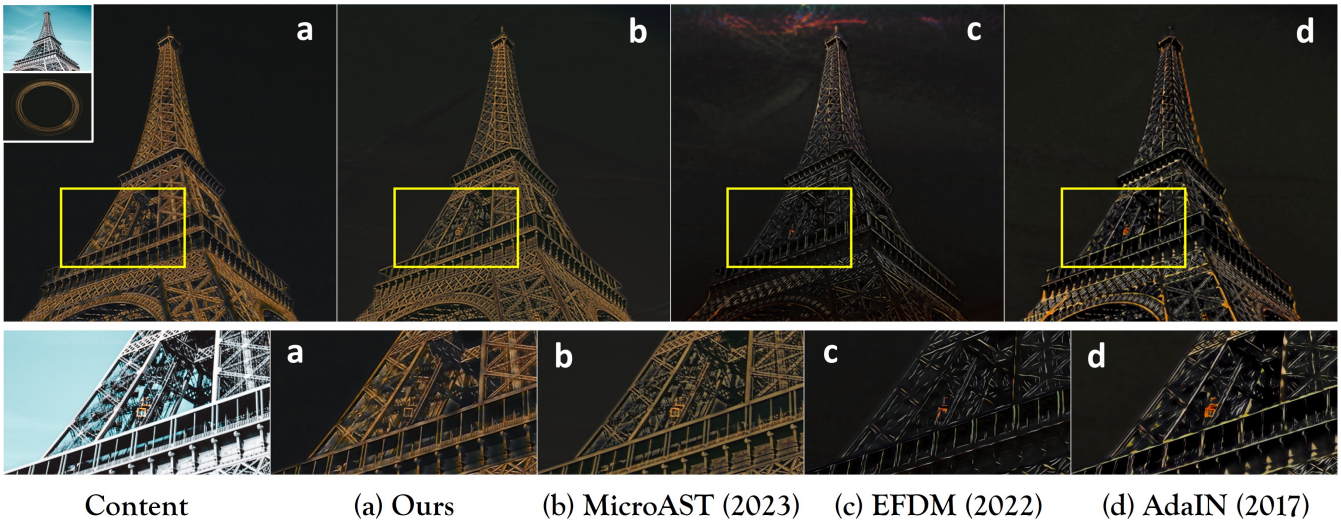


Figure 1: Additional qualitative comparison in 4K (4096×4096) resolution. Overall, our approach can generate aesthetically more realistic and pleasing results for arbitrary styles.

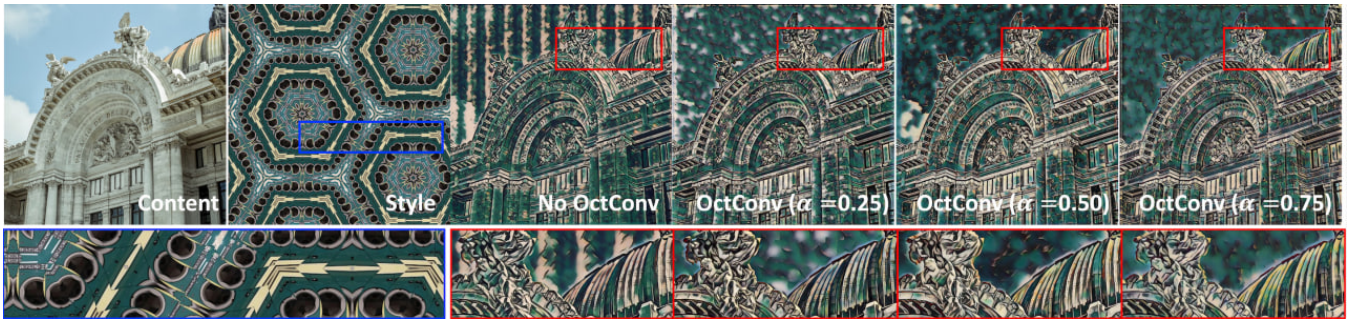


Figure 2: The stylization outputs with no octave convolutions and octave convolutions in different alpha values. Alpha values (α) denotes the ratio of the number of low-frequency channels to the total-frequency channels.

Aesthetic Feature Descriptor Dimensions. In Table 4, we show how changing the dimensions of *aesthetic feature descriptor* can affect the resulting stylizations.

More Experimental Results

Additional visual results at various resolutions. More experimental visual results are presented in Figure 1, Figure 4, and Figure 5.

Quantitative Comparisons at 512-pixel resolutions. Table 5 shows that our proposed method shows quantitatively promising results compared to the state-of-the-art NST techniques at 512 resolutions.

Quantitative Comparisons in terms of Content Perceptual Loss. Table 6 shows the quantitative comparison of content perceptual loss among various NST algorithms.

Video Style Transfer and Video Style Blending. Figure 6 shows qualitative comparisons on video style transfer at 2K resolutions. The first row shows several video frames and the style image. The rest of the rows show the stylization results by various algorithms. Our results can yield the best video results in terms of high consistency and aesthetic features (e.g., colors, and textures). Figure 8 shows video style blending results by AesFA. The tones from the low-style image and the texture and structure of the high-style image are well-transferred to the output image.

Societal Impacts

Positive impacts. This study may be useful to various types of people. For example, researchers with a neural style transfer interest might be motivated by our findings to create some innovative and effective techniques in the future. Also, artists can benefit from our model, as they can use the creative illustrations generated by our model as a springboard for their own ideas.

Negative impacts. Possible drawbacks include the possibility that the proposed method could replace some human tasks or be abused to produce an undesirable outcome.

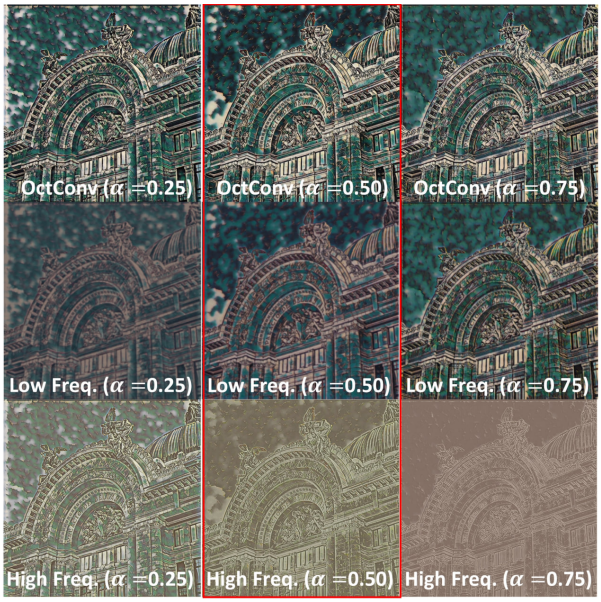


Figure 3: The feature decomposition experiments with different alpha values.

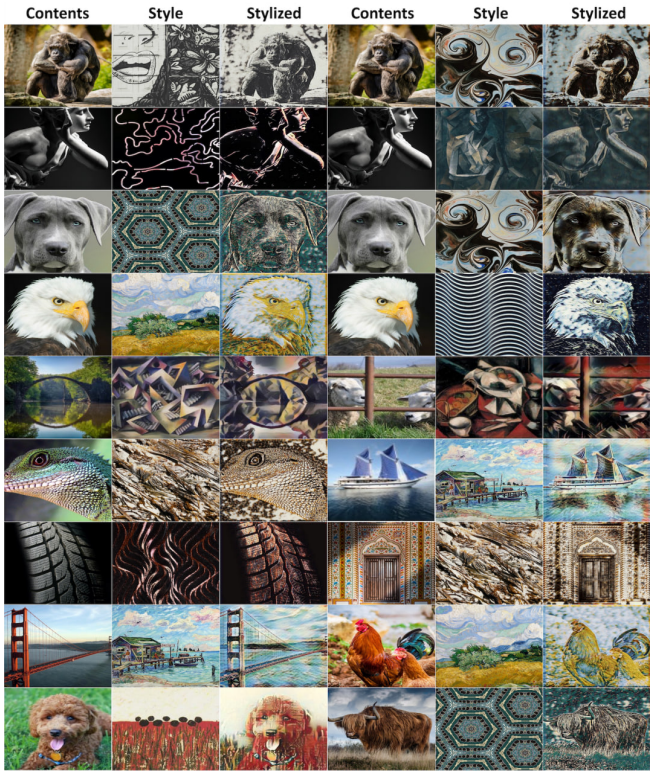
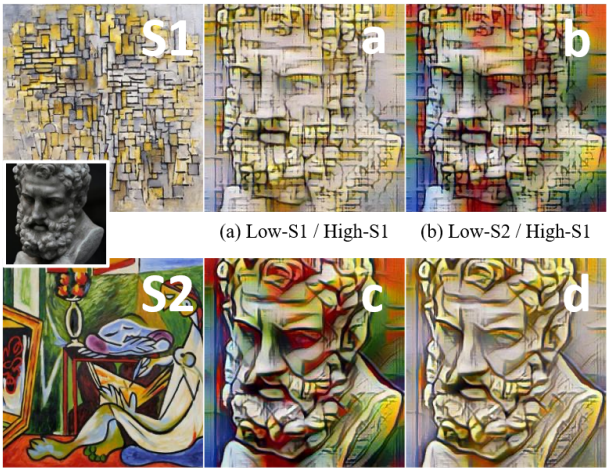


Figure 4: Additional visual results generated by AesFA.



(a) Low-S1 / High-S1 (b) Low-S2 / High-S1

(c) Low-S2 / High-S2 (d) Low-S1 / High-S2

Figure 5: Additional Style Blending images in 256 image resolution.

Resolution (px)	Dimensions ($C \times H \times W$)	Style Loss (↓)	LPIPS (↓)	SSIM (↑)	Time (↓)
256^2	(256, 3, 3)	**0.692**	**0.368**	**0.417**	**0.016**
	(256, 5, 5)	0.742	0.371	0.407	**0.016**
	(256, 7, 7)	0.736	0.369	0.384	0.019
512^2	(256, 3, 3)	**0.314**	0.365	**0.371**	0.017
	(256, 5, 5)	0.346	0.365	0.364	**0.016**
	(256, 7, 7)	0.364	**0.358**	0.348	0.018
$1024^2 (1K)$	(256, 3, 3)	**0.283**	0.392	**0.405**	0.020
	(256, 5, 5)	0.319	0.395	0.386	**0.018**
	(256, 7, 7)	0.362	**0.384**	0.371	0.020
$2048^2 (2K)$	(256, 3, 3)	**0.404**	**0.435**	**0.378**	0.020
	(256, 5, 5)	0.462	0.437	0.361	**0.019**
	(256, 7, 7)	0.552	0.438	0.345	0.020

Table 4: Quantitative comparison with different *aesthetic feature descriptor* dimensions at 2K (2048×2048) resolutions. C, H, and W represent the channel, height, and width of the feature, respectively. The *aesthetic feature descriptor* with dimensions of (256, 3, 3) performs the best.

Resolution (px)	Method	Style Loss (↓)	LPIPS (↓)	SSIM (↑)	Time (↓)	Pref. (↑)
512^2	AdaConv (2021)	N/A	N/A	N/A	N/A	-
	AdaIN (2017)	0.355	**0.364**	0.253	**0.012**	8.54
	MicroAST (2023)	0.624	0.372	0.381	0.009	13.27
	EFDM (2022)	0.340	0.371	0.240	**0.012**	4.41
	AdaAttn (2021)	0.582	0.405	**0.429**	0.041	4.72
	Aes-UST (2022)	0.411	0.386	0.372	0.022	16.46
	IECAST (2021)	0.592	0.390	0.349	0.020	8.84
	StyTr^2 (2022)	**0.306**	0.378	0.423	0.097	11.08
	AesFA (Ours)	0.314	**0.365**	0.371	0.017	**29.49**

Table 5: Quantitative comparison with the various state-of-the-art NST algorithms at 512 pixel resolutions. "N/A" represents "Not Applicable at this resolution" and the unit for time is second/image.

Limitations

Despite its impressive performance, AesFA has certain limitations. To begin, the results by AesFA are sensitive to the weighting hyper-parameters for each loss, often resulting in over-stylized output (e.g., the repetitive style patterns on the backgrounds). However, this could be mitigated by carefully adjusting the weighting hyper-parameters. Besides, the vertical line-shape artifacts alongside the images are often observed (see Figure 7). These artifacts appear in the AdaConv

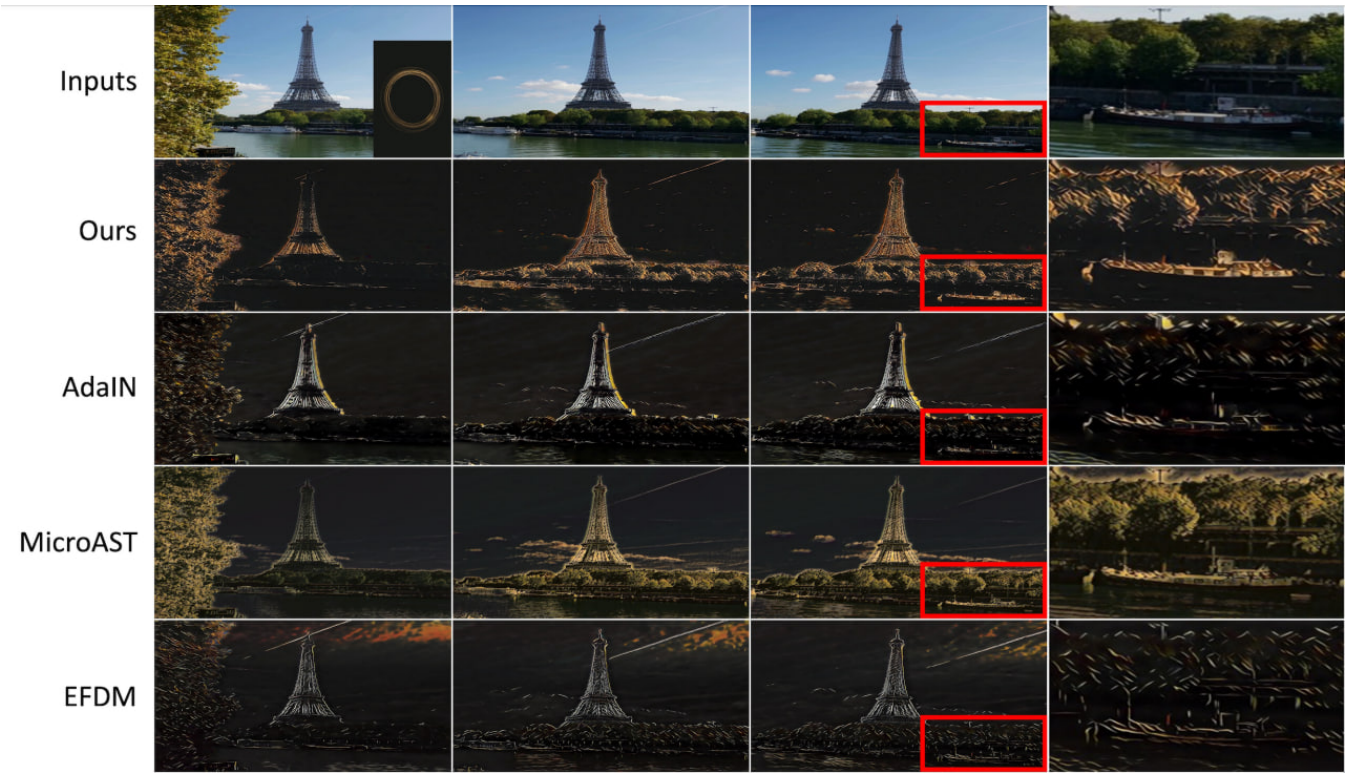


Figure 6: The qualitative video style transfer comparisons with various SOTA NST algorithms.

Resolution (px)	AdaConv	AdaIN	MicroAST	EFDM	AdaAttn	Aes-UST	IECAST	StyTr	AesFA (Ours)
256	5.747	6.466	5.895	6.849	5.941	5.517	5.681	6.106	6.241
512	N/A	4.207	3.492	4.453	3.519	3.259	3.478	3.703	4.132
1024 (1K)	N/A	3.223	2.807	3.494	2.945	2.606	2.670	2.679	3.277
2048 (2K)	N/A	3.267	2.590	3.504	OOM	2.874	OOM	OOM	3.247
4096 (4K)	N/A	2.470	2.071	2.652	OOM	OOM	OOM	OOM	2.434

Table 6: Quantitative comparison in terms of content perceptual loss among various NST algorithms. "N/A" and "OOM" represent "Not Applicable at this resolution" and "Out of Memory", respectively. The unit for time is second/image.

(Chandran et al. 2021) as well. We reason that these appear because the content features are being convolved directly with the predicted *aesthetic feature-aware kernels and biases*. Also, the upsampling operations could be the ones that create these artifacts. The weighting hyper-parameters could be fine-tuned to solve this problem.

Involved Assets

Existing assets that we used in this work mainly include: 1) the codes of AdaConv (Chandran et al. 2021), AdaIN (Huang and Belongie 2017), MicroAST (Wang et al. 2023), EFDM (Zhang et al. 2022), AdaAttn (Liu et al. 2021), AesUST (Wang et al. 2022), IECAST (Chen et al. 2021), StyTr^2 (Deng et al. 2022) and 2) the MS-COCO dataset (Lin et al. 2014), WikiArt dataset (Phillips and Mackintosh 2011) and the images from *pexels.com*. We report their URLs and licenses in the following,

- AdaConv: https://github.com/RElbers/ada-conv-pytorch, MIT License.
- AdaIN: https://github.com/naoto0804/pytorch-AdaIN, MIT License.
- MicroAST: https://github.com/EndyWon/MicroAST, MIT License.
- EFDM: https://github.com/YBZh/EFDM, MIT License.
- AdaAttn: https://github.com/Huage001/AdaAttN, Apache-2.0 License.
- AesUST: https://github.com/EndyWon/AesUST, MIT License.
- IECAST: https://github.com/HalbertCH/IEContraAST, MIT License.
- StyTr^2: https://github.com/diyiiyiii/StyTR-2.git, we were unable to find its license.
- MS-COCO: https://cocodataset.org/#download, we were unable to find its license.
- WikiArt: https://www.kaggle.com/c/painter-by-numbers, we were unable to find its license
- *pexels.com*: https://www.pexels.com/, Pexels License.

Note that MS-COCO, WikiArt and *pexels.com* have been widely used in a lot of existing works, and *pexels.com* are only used to get ultra-high resolution images (e.g., 4K). To the best of our knowledge, they do not include any personally identifiable information or offensive content. In this study, a total of 60,000 images from MS-COCO and 26,689

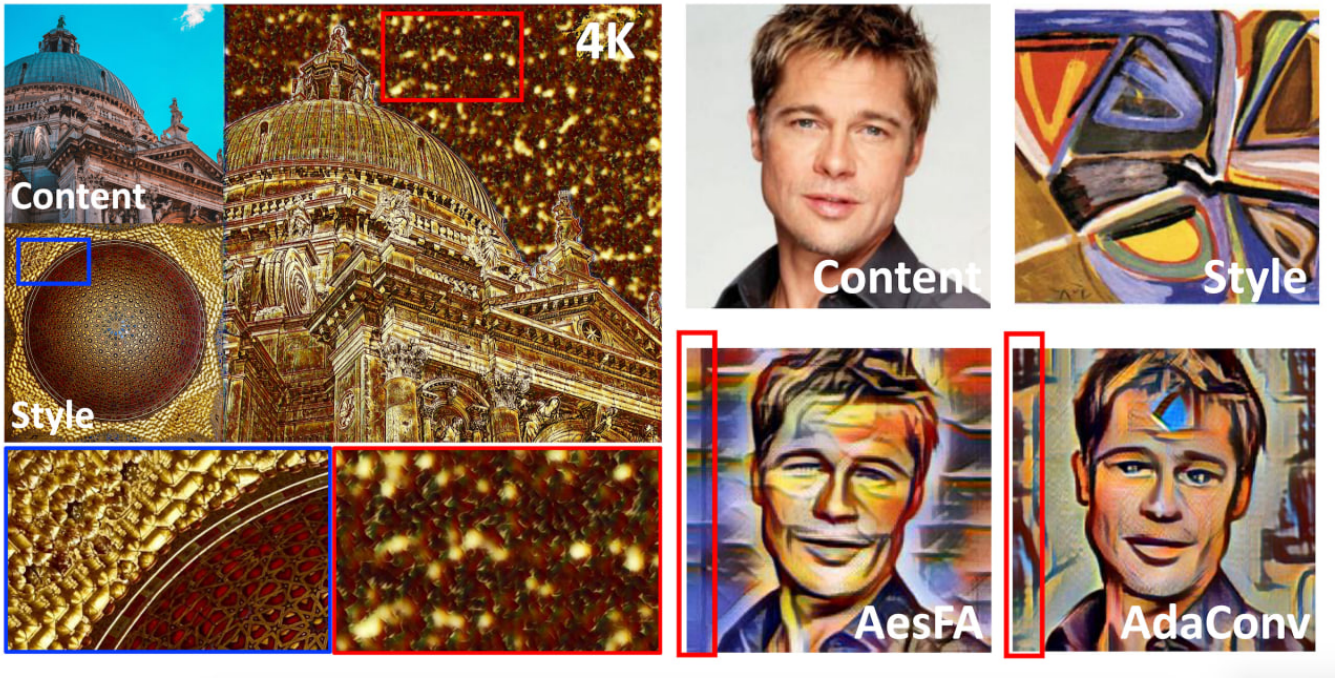


Figure 7: Over-stylized output and the artifacts examples generated by AesFA and AdaConv.

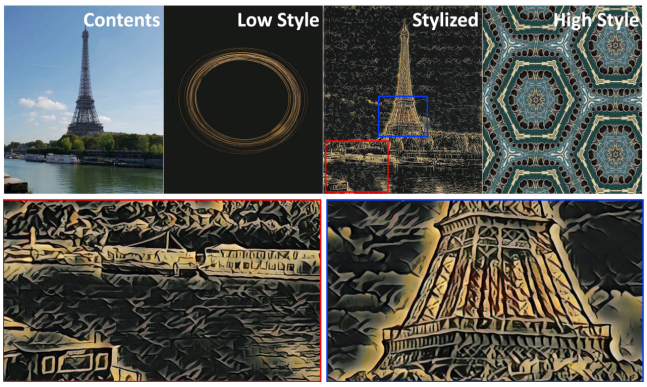


Figure 8: The video style blending result by AesFA.

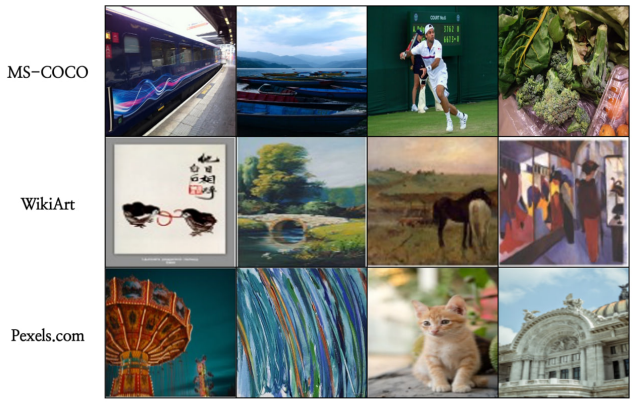


Figure 9: A representative subset of MS-COCO, WikiArt, and *pexels.com*.

images from WikiArt are used for training. A representative subset of each dataset is illustrated in Figure 9. For fair comparisons, all existing algorithms are re-trained using these datasets with the respective author-released codes and default configurations.

References

Chandran, P.; Zoss, G.; Gotardo, P.; Gross, M.; and Bradley, D. 2021. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7972–7981.

Chen, H.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; Lu, D.; et al. 2021. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34: 26561–26573.

Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; and Xu, C. 2022. StyTr^2: Image Style Transfer with Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzer-*

land, September 6-12, 2014, Proceedings, Part V 13, 740–755. Springer.

Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; and Ding, E. 2021. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6649–6658.

Phillips, F.; and Mackintosh, B. 2011. Wiki Art Gallery, Inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3): 593–608.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wang, Z.; Zhang, Z.; Zhao, L.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2022. AesUST: towards aesthetic-enhanced universal style transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1095–1106.

Wang, Z.; Zhao, L.; Zuo, Z.; Li, A.; Chen, H.; Xing, W.; and Lu, D. 2023. MicroAST: Towards Super-Fast Ultra-Resolution Arbitrary Style Transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, Y.; Li, M.; Li, R.; Jia, K.; and Zhang, L. 2022. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8035–8045.