Supplementary Material

Bohai Gu[1,3] Heng Fan[2] Libo Zhang[1,3†]
[1] Institute of Software Chinese Academy of Sciences, Beijing, China
[2] Department of Computer Science and Engineering, University of North Texas, Texas, USA
[3] University of Chinese Academy of Sciences, Beijing, China

1. Image Style Transfer

1.1. Multi-Granularity Style Transfer

We manage to apply the UniST to multi-granularity style transfer for the first time. As shown in Fig. 1, we fixed the resolution of the content while changing the resolution of the style from high to low, we can see that the stylization results become more and more fabulous. But for other methods, they are either not supported by the source code or have poor performance as illustrated in Fig. 2, we leave it to the reader to judge for themselves.

More vivid results To better visualize UniST's amazing fine-grained style transfer, we fix the content to 1024x1024 and the style to 256x256, and zoom in on the stylization result as shown in Fig. 3. We can observe that the style is naturally integrated into the texture with the finest granularity. In order to demonstrate the robustness of our joint learning framework, we provide the stylization results of pair-wise combinations between 5 content images and 5 style images in Fig. 4. It can be seen that the joint learning framework achieves appealing image style transfer results with desirable style pattern details, while keeping the content structure well-maintained.

2. Video Style Transfer

2.1. Quantitative Results.

We provide the whole optical flow errors and *LPIPS* scores of five solutions in Table 2, as supplement to the Table 2 in main text. It can be seen that our model obtains the smallest average optical flow error and the lowest mean *LPIPS* score among the existing methods, having the superiority in the most styles. Therefore, our model maintains the best temporal consistency.

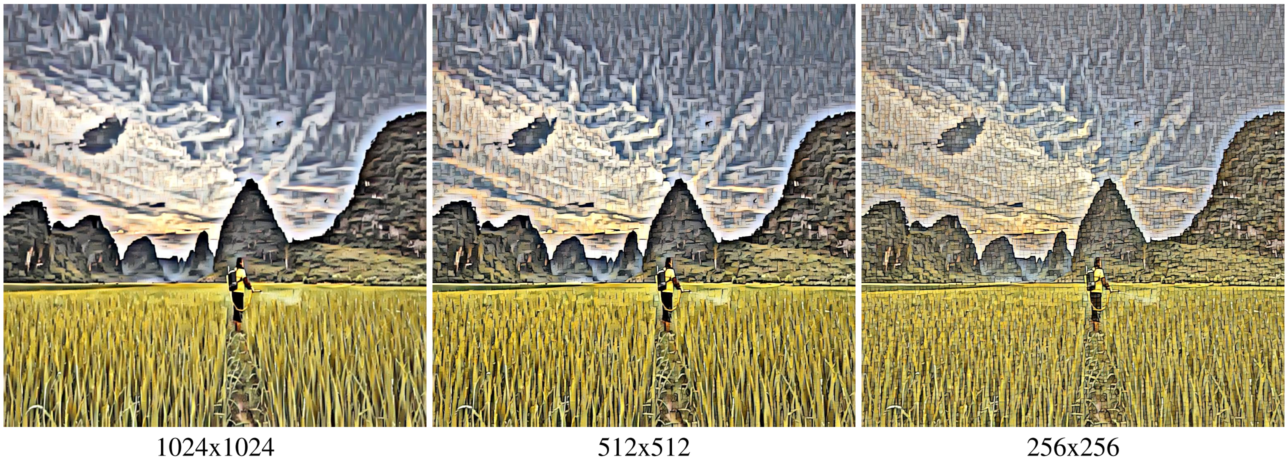
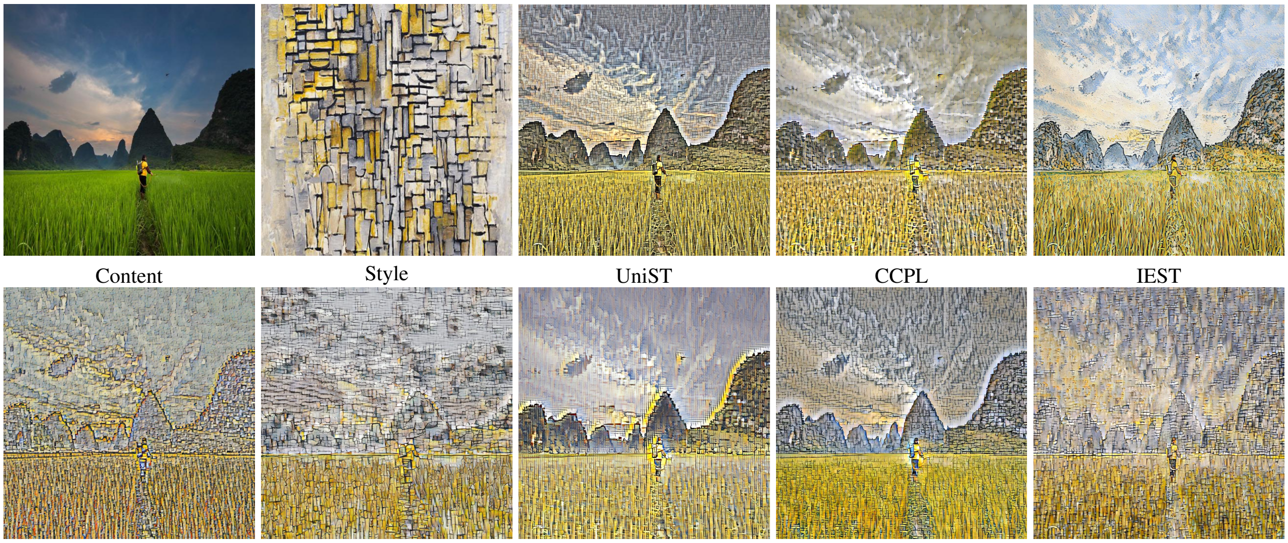

1024x1024 512x512 256x256

Figure 1. The multi-granularity style transfer application of UniST. We fix the content to 1024x1024 and change the style from 1024x1024 to 256x256.



Content Style UniST CCPL IEST

StyleFormer MAST MCCnet Linear SANET

Figure 2. Qualitative comparison of the multi-granularity style transfer. Zoom in for the better judgement.

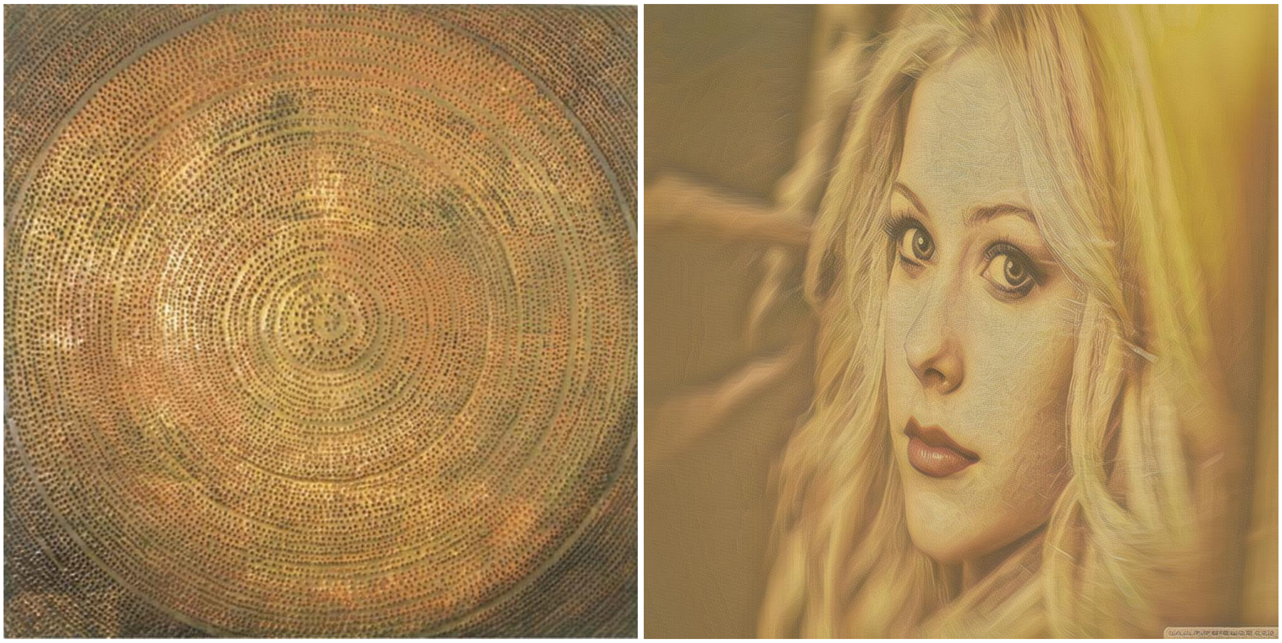


Figure 3. Zoom in on the single fine-grained result for better view.

2.2. Qualitative Results.

We provide more video style transfer examples in Fig. 5. We also recommend you to zoom in, because the results of our joint learning framework is fine-grained enough with vivid style patterns. Full animations can be found in the attachments. The video transfer results not only maintain excellent temporal consistency, but also guarantee the overall stylization effect.

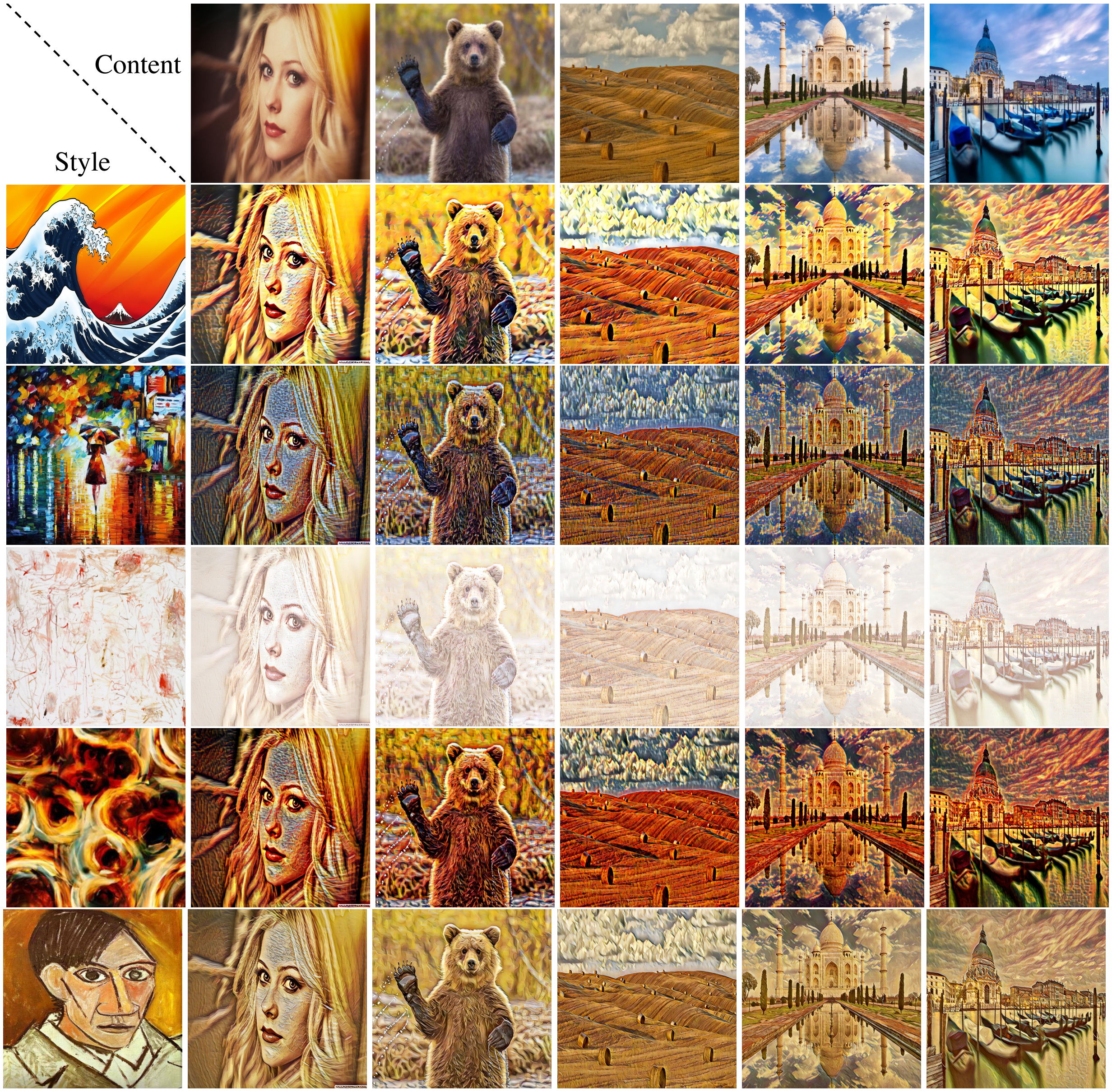


Figure 4. More image style transfer results.

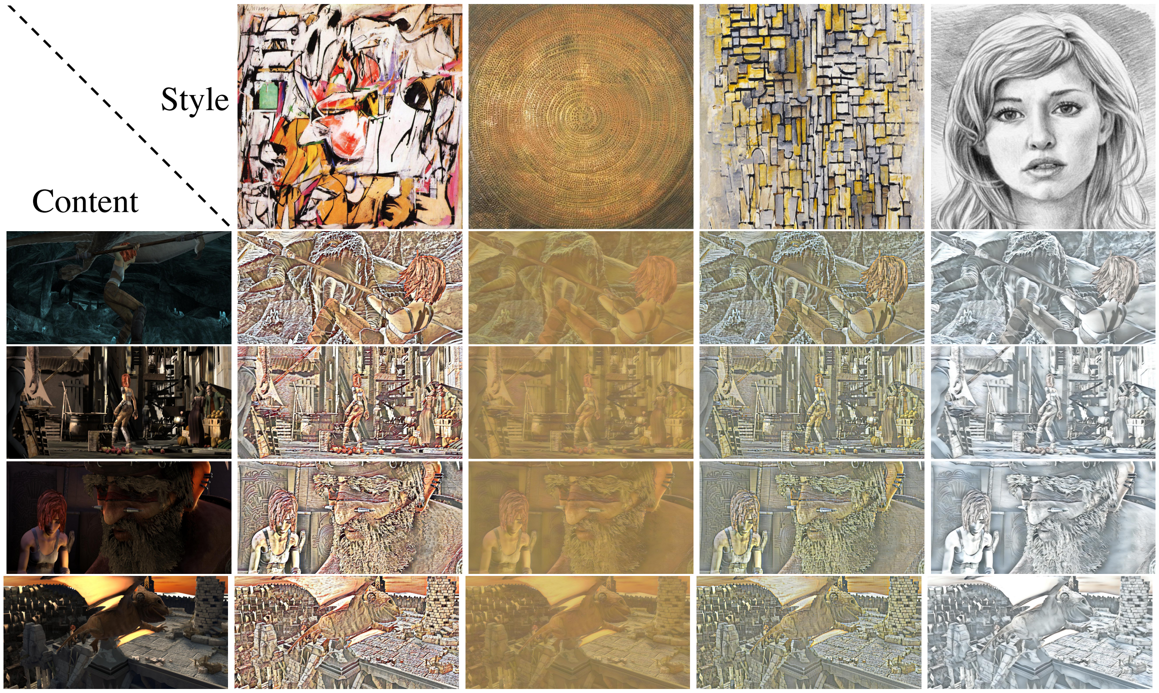


Figure 5. More video style transfer results.

Method	Style1	Style2	Style3	Style4	Style5	Style6	Style7	Style8	Style9	Style10	Style11
AdaATTN	2.04	2.43	2.04	2.36	1.92	1.70	1.95	1.63	2.32	2.49	2.46
MCCNet	2.16	2.48	2.04	2.40	1.71	1.83	1.73	1.51	2.46	2.36	2.34
Linear	2.09	2.31	2.03	2.25	1.82	1.79	2.01	1.57	2.23	2.27	2.26
CCPL	2.11	2.41	2.07	2.34	1.64	1.82	1.99	1.66	2.34	2.33	2.30
Ours	**1.78**	**2.10**	**1.95**	**2.13**	**1.42**	**1.61**	1.95	**1.35**	**2.08**	**2.05**	**2.04**
Method	Style12	Style13	Style14	Style15	Style16	Style17	Style18	Style19	Style20	Mean	/
AdaATTN	1.68	2.23	2.01	1.96	2.26	1.79	1.88	1.77	2.21	2.05	/
MCCNet	1.96	2.23	2.11	1.85	2.13	2.10	1.90	1.86	2.30	2.07	/
Linear	1.86	2.22	1.97	2.00	2.08	1.87	1.89	1.83	2.13	2.02	/
CCPL	1.91	2.23	2.02	1.99	2.10	1.99	1.96	1.87	2.21	2.06	/
Ours	1.70	**2.05**	**1.56**	**1.63**	**1.73**	1.81	**1.61**	**1.53**	**1.88**	**1.79**	/

Table 1. The whole *LPIPS* scores ($\times 10^{-2}$) of SOTA methods. Smaller values mean better temporal consistency.

Method	Style1	Style2	Style3	Style4	Style5	Style6	Style7	Style8	Style9	Style10	Style11	Style12	Style13	Style14	Style15	Style16	Style17	Style18	Style19	Style20	Mean
AdaATTN	3.73	5.88	4.13	4.62	2.60	2.21	6.12	2.25	4.50	7.09	6.71	1.66	4.04	3.33	3.98	4.26	2.37	2.69	2.48	3.66	3.91
MCCNet	4.95	6.88	4.62	5.59	2.40	2.54	8.61	3.02	5.89	6.83	7.10	1.99	4.38	3.89	4.19	4.60	3.28	3.20	2.96	4.59	4.57
Linear	4.49	6.10	4.49	4.84	2.44	3.11	6.25	2.90	4.64	6.81	6.49	2.78	4.47	3.80	4.57	4.23	2.74	3.02	2.92	3.90	4.25
CCPL	5.20	7.29	5.49	5.81	2.45	3.30	7.42	3.05	5.89	7.65	7.57	2.41	5.14	4,13	4.65	5.14	3.39	3.73	3.44	4.87	4.90
Ours	4.11	**5.85**	4.75	4.88	**1.93**	2.30	6.61	**2.11**	4.58	**6.16**	**5.78**	1.88	4.29	**3.09**	**3.31**	**3.64**	2.77	2.75	2.53	3.80	**3.86**

Table 2. The whole optical flow errors ($\times 10^{-2}$) of SOTA methods. Smaller values mean better temporal consistency.