Learning Dynamic Style Kernels for Artistic Style Transfer –Supplemental Material–

Wenju Xu
OPPO US Research Center
Palo Alto, CA, USA
xuwenju123@gmail.com

Chengjiang Long
Meta Reality Lab
Burlingame, CA, USA
clong1@meta.com

Yongwei Nie
South China University of Technology
Guangzhou, Guangdong, China
nieyongwei@scut.edu.cn

Abstract

The supplementary material provides more implementation details of our network. We also provide more ablation study on different model variants of synthesizing stylized images, as well as more visualization comparison with state-of-the-art methods in terms of style consistency and structure similarity. Note that we could not include all the material in the main paper due to the space limit.

1. Implementation Details

Configuration of Our Networks. We take a pretrained VGG-16 network as the encoder, which is fixed during training. The activation map extracted from ReLU_4 of the VGG-16 network is utilized as the encoded feature representation. In style-alignment encoding module, we set the number of groups in multi-head attention module $G = 8$. For each attention head, we normalize the query and key features, and take three 3×3 convolution blocks to map the query, key and value features, respectively. The ψ comprising of three 3×3 convolution blocks maps the content feature $Z_c^{H_c \times W_c \times C}$ into scale parameter $\gamma^G \in \mathbb{R}^{H_c \times W_c \times G}$ and bias parameter $\beta^G \in \mathbb{R}^{H_c \times W_c \times G}$, from which we obtain the $\gamma \in \mathbb{R}^{H_c W_c}$ and $\beta \in \mathbb{R}^{H_c W_c}$ used in each attention head via splitting and reshaping. Z_{cs} is the concatenation of outputs from all attention heads. The φ consisting of two 3×3 convolution blocks takes in Z_{cs} to predict the filters $F \in \mathbb{R}^{H_c \times W_c \times C(2k+1)}$. For each spatial point in Z_c, we take the separable convolution to modulate the content feature with the predicted kernel. The modulated output $\bar{Z}_{cs}$ is fed into our decoder to generate the stylied image. The structure of our decoder is listed in Table 1. This decoder consists of three residual blocks, upsampling layers, and Convolution layers.

Table 1. The network architecture of our decoder.

Model	Block	Layer		Input Shape	Out Shape
Decoder	ResBlock	ReLu	ReLu	[H, W, C]	[H, W, C//2]
		Conv		[H, W, C//2]	[H, W, C//2]
		ReLu	Conv	[H, W, C//2]	[H, W, C//2]
		Conv		[H, W, C//2]	[H, W, C//2]
		Sum		[H, W, C//2]	[H, W, C//2]
	Up	Upsample		[H, W, C//2]	[2H, 2W, C//2]
	ResBlock	ReLu	ReLu	[2H, 2W, C//2]	[2H, 2W, C//4]
		Conv		[2H, 2W, C//4]	[2H, 2W, C//4]
		ReLu	Conv	[2H, 2W, C//4]	[2H, 2W, C//4]
		Conv		[2H, 2W, C//4]	[2H, 2W, C//4]
		Sum		[2H, 2W, C//4]	[2H, 2W, C//4]
	Up	Upsample		[4H, 4W, C//4]	[4H, 4W, C//4]
	ResBlock	ReLu	ReLu	[4H, 4W, C//4]	[4H, 4W, C//8]
		Conv		[4H, 4W, C//8]	[4H, 4W, C//8]
		ReLu	Conv	[4H, 4W, C//8]	[4H, 4W, C//8]
		Conv		[4H, 4W, C//8]	[4H, 4W, C//8]
		Sum		[4H, 4W, C//8]	[4H, 4W, C//8]
	Up	Upsample		[4H, 4W, C//8]	[8H, 8W, C//8]
		ReLu		[8H, 8W, C//8]	[8H, 8W, C//8]
		Conv		[8H, 8W, C//8]	[8H, 8W, 3]
		Tanh		[8H, 8W, 3]	[8H, 8W, 3]

Style Transfer with Grouped Shuffling. Due to the correlation within each channel-wise feature vector in content feature Z_c, directly forcing the output and style images to have similar global statistics (e.g., Gram matrices or covariance matrices) is challenging. As shown in the third and fourth columns in Figure 5 of the manuscript, even though the stylized results are visually close to the style reference, the differences (colors and style patterns) are obvious. To break down the correlation within the feature vector and improve the style consistency between stylized images and style references, we propose a simple but effective method, denoted as grouped shuffling, which is to group the content feature channel-wisely and randomly shuffle the order of groups. The grouping operation could be written as:

$$Z_c = \mathop{||}\limits_{gs}^{O} Z_c^{gs} \tag{1}$$

where O refers to the order of groups. $gs \in O$ is the index of each group and $||$ stands for channel-wise concatenation. $O = \{0, 1, 2, 3, 4, 5, 6, 7\}$ and $Z_c^{gs} \in \mathbb{R}^{H_c \times W_c \times C/8}$ given there are 8 groups. For shuffling, we randomly shuffle the order of groups, such as $O = \{2, 3, 5, 4, 6, 1, 7, 0\}$. The re-

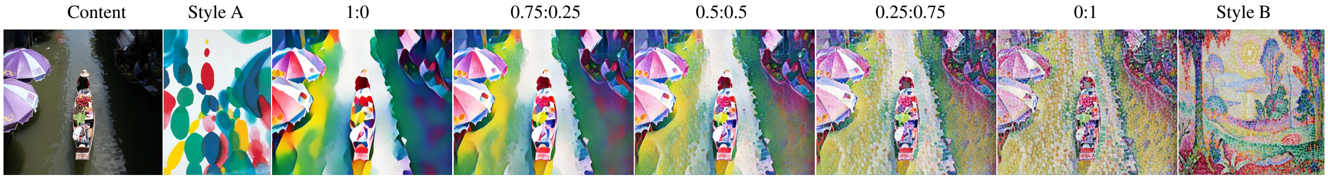

Figure 1. Performance of style interpolation.

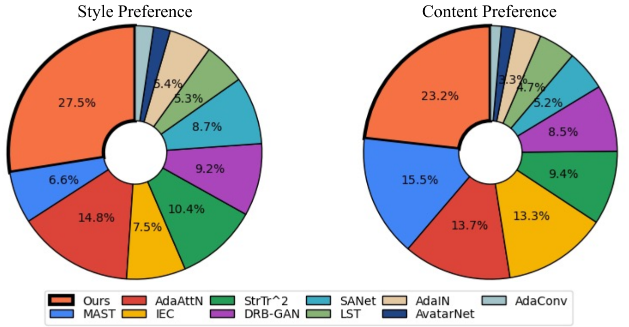

Figure 2. The quantitative comparison of the user study. We report the style preference and content preference of each compared method.

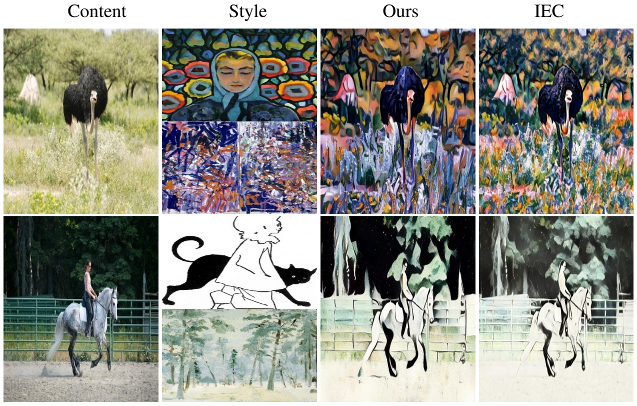

Figure 3. Performance of multi-style transfer, where the style images are created by stitching multiple style images into one image.

concatenated content feature will be fed into the decoder for stylization.

2. More Results

Style Interpolation. To make smooth transitions between different style images, we operate linear interpolation on feature map Z_{cs}. The interpolated weights $\tilde{Z}_{cs}$ are obtained as $\tilde{Z}_{cs} = \alpha Z_{cs}^{s_1} + (1 - \alpha) Z_{cs}^{s_2}$ where α is the interpolation factor between 0 and 1, s_1 and s_2 represent two different style images used to produce the feature map. Based on the $\tilde{Z}_{cs}$, we generate the dynamic kernels used to modulate the content image for style transfer. As shown in Figure 1, our model captures subtle variations between the two styles.

User Study. We perform a user study to further compare our method against other baseline methods in terms of content preference (*i.e.*, the structural similarity, artifacts, and texture details) and style preference (*i.e.*, the style consistency and distortion). 100 content images and 100 style images are selected to synthesize 100 stylizations. For each user, we sample 20 content-style pairs and present the stylized images by all methods in random order. The users are asked to select their favorite one from three aspects: content preservation, stylization degree, and overall preference. The style preference and content preference in Figure 2, which demonstrate that our stylized results are more appealing than competitors.

Table 2. Ablation study on model variants with different settings. SK represents Style Kernels. CGM stands for Content-based Gating Module. GS is the Grouped Shuffling.

Method	Base	$+\mathcal{L}_{adv}$	$+\mathcal{L}_{REMD}$	+ SK	+ CGM	+ GS (Ours)
Style Loss↓	2.92	2.14	1.48	1.11	1.08	0.98
LPIPS ↓	0.49	0.39	0.34	0.31	0.30	0.30

Ablation Study. The main contribution of this paper is the proposed network taking style kernels to generate stylized image maintaining structure similarity to the content image and reflecting the style characteristics of the style reference. To further demonstrate the effectiveness of different components in our network, we start with a simple Base method without involving the CGM, style kernels (SK) and GS (namely, it only use the attention mechanism to warp the style feature), and train this model without $\mathcal{L}_{adv}$ and $\mathcal{L}_{REMD}$. Then we keep adding different modulates into it until it is the same as our full model. We compare different model variants and report the quantitative performances in Table 2. One can see (1) The loss functions improve the performance of the base model in terms of both style and content preference scores; (2) **Style Kernels (SK)** works to inject the style characteristics into the output images, as shown in the fifth columns of Table 2; (3) **Grouped Shuffling (GS)** improves the style consistency to the style reference at a limited cost of distorting content structure; (4) **Content-Based Gating Module (CGM)** removes misaligned style patterns and preserve the completeness of semantic regions as shown in Table 2.

Multiple Style Transfer. In Figure 3, we show the result of applying our model on multiple style transfer, where the style reference is a concatenation of two different style images. We can see our model can flexibly create a new styl-

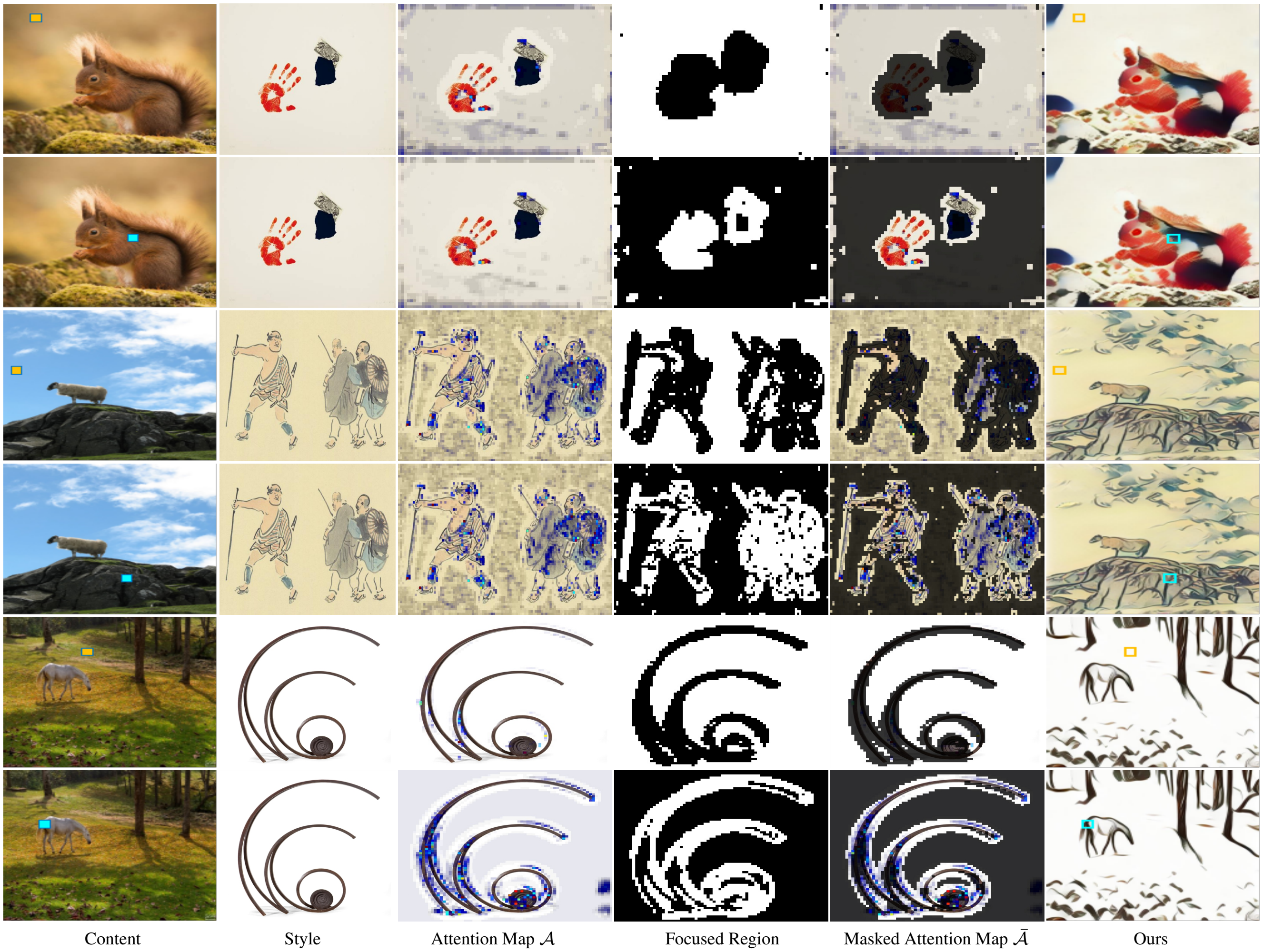
Content
Style
Attention Map $\mathcal{A}$
Focused Region
Masked Attention Map $\bar{\mathcal{A}}$
Ours

Figure 4. Visualizations of the focused region. Given a query point in the content image (yellow and Cyan rectangles), the corresponding focused region is learned to filter out points with low correlation. The focused region contains a complete semantic area, enabling our method to create style-consistent stylizations with well-preserved content structure. Please zoom in for better visualization. The attention distributed over the image is highlighted in colors.

ization reflecting style characteristics in different style references.

Visualizations on Focused Region. In Figure 4, we demonstrate more visualizations on focused regions. Given a query point in the content image (yellow and Cyan rectangles), the corresponding focused region is learned to filter out points with low correlation. The focused region contains a complete semantic area, enabling our method to create style-consistent stylizations with well-preserved content structure. This is because our CGM aims to select focusing regions, where query points can aggregate style information from fairly correlated nodes.

Arbitrary Style Transfer. In Figure 5 we demonstrate stylization results of all the combinations between 7 content images and 9 style images. The first row and column are the content images and style images, respectively. Others are the stylization results created by our method. One can see from the stylization results that, in each row, our method can robustly produce stylizations with consistent style characteristics sharing by the style image, and the structure similarity is well maintained as seen in each column.

More Qualitative Comparisons with State-of-the-art methods. In Figure 6, we demonstrate additional qualitative comparisons with state-of-the-art methods. The main advantage as we can observe is that our method is able to maintain the completeness of semantic regions in the stylized images. While other methods manipulate the content structure and introduce artifacts in the generated images.

Video Style Transfer. We compare the performances on image sequences style transfer in Figure 7, provide the real-world video style transfer result in Figure 8, which proves the effectiveness of video style transfer and demonstrates that our model outperforms the current SOTA methods in

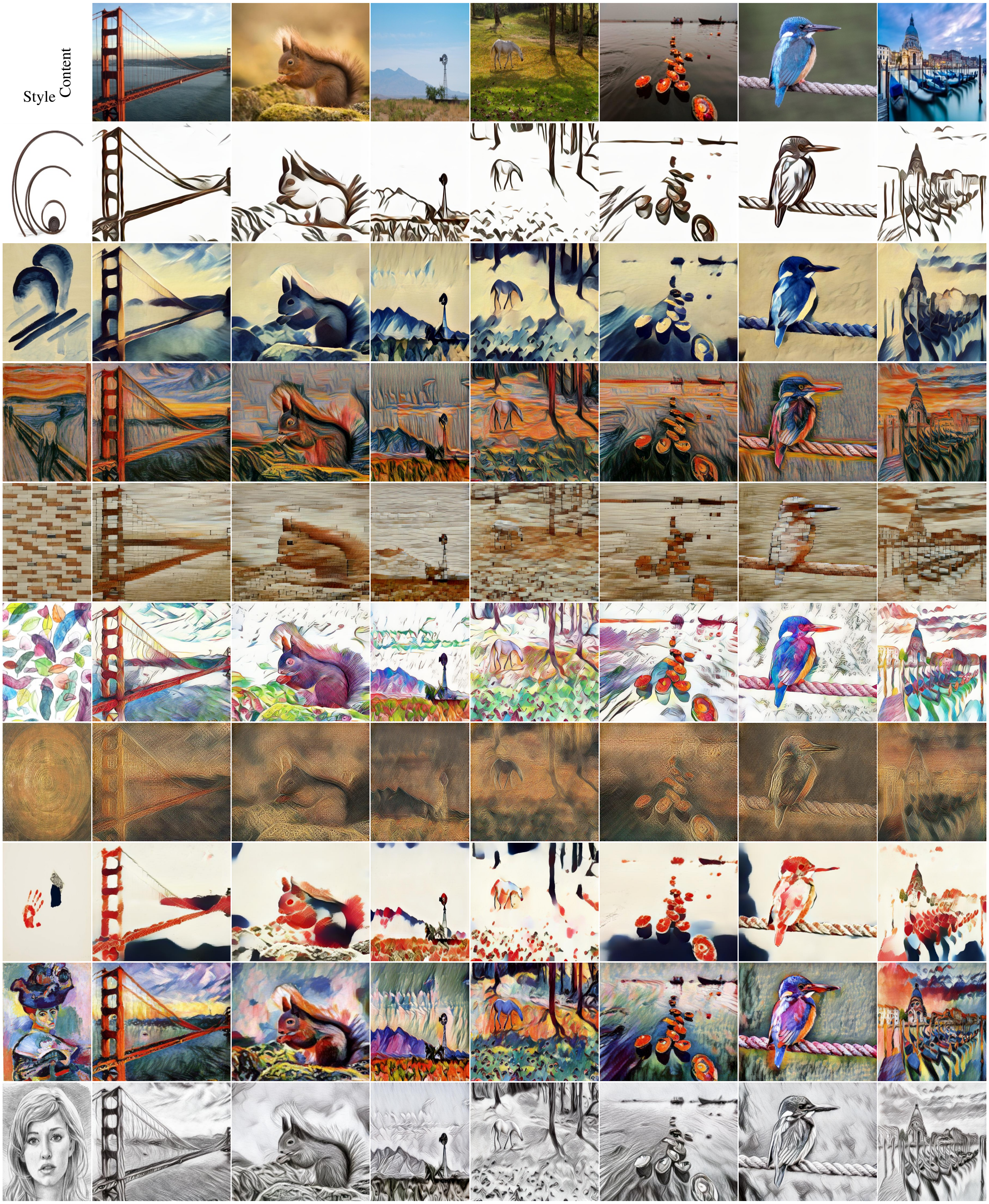

Figure 5. More qualitative performance on arbitrary style transfer. The first row and column are the content images and style images, respectively. Others are the stylization results created by our method.

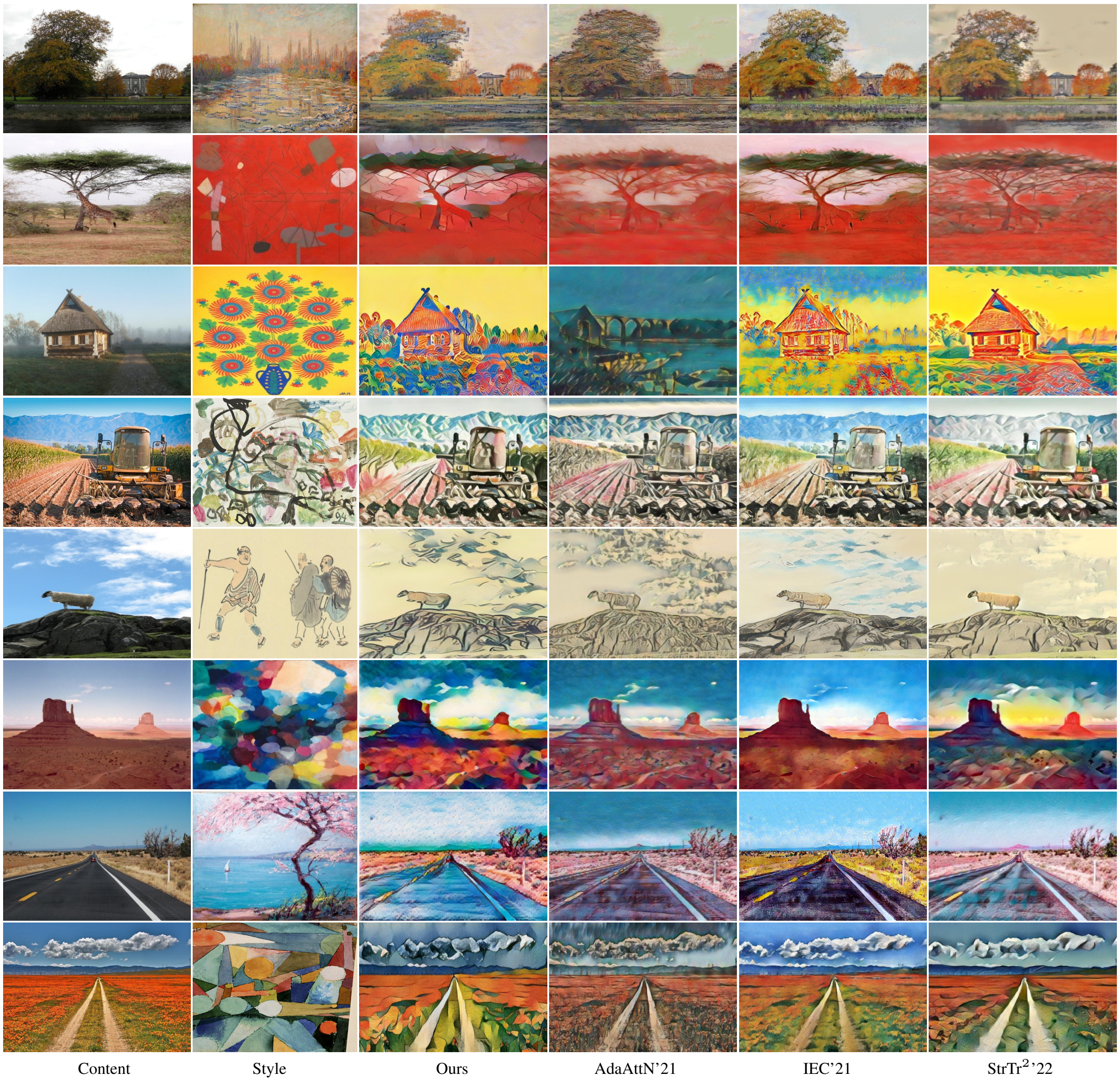

Figure 6. Qualitative performance comparison on stylized results.

terms of style consistency and structure similarity.

High Resolution Style Transfer. Our model is also robust to perform style transfer on images of high resolutions. To illustrate the influence of image resolutions, we show the qualitative comparison in Figure 10, Figure 11 and Figure 12. It demonstrates that our model creates consistent stylizations on different content images with slight variations. A lot of fine details and brushstrokes are visible.

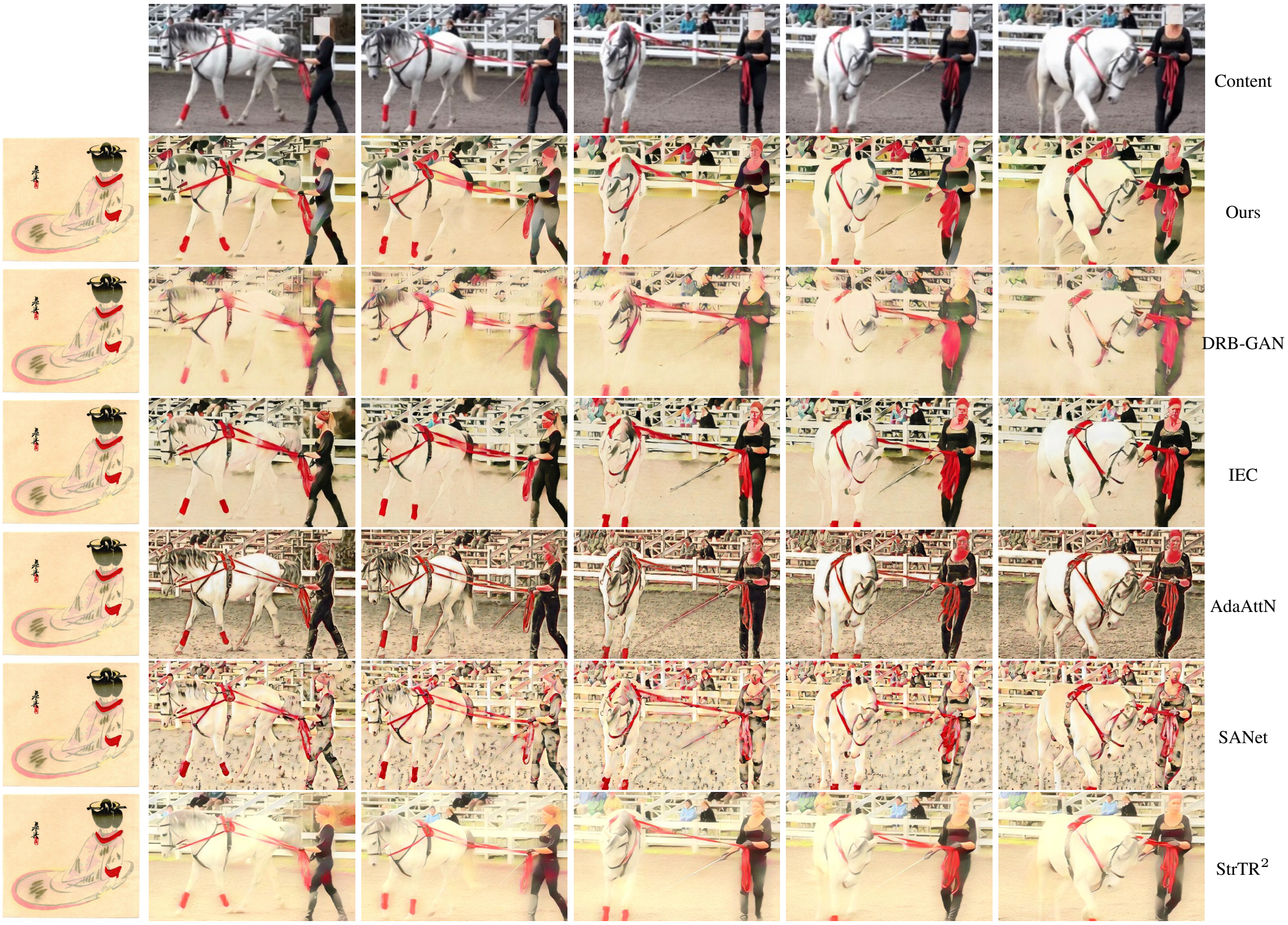

Figure 7. Comparisons on video style transfer. The 1st row lists 5 frames from a video clip as content images. The style images are listed in the first column.

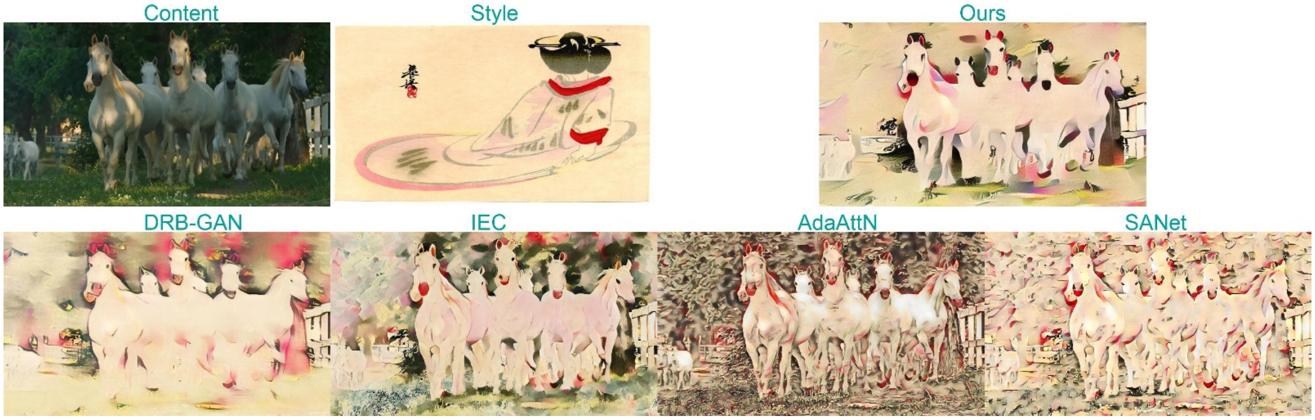

Figure 8. Snapshot of video style transfer task. Full video with more detainls are in the supplemental video.

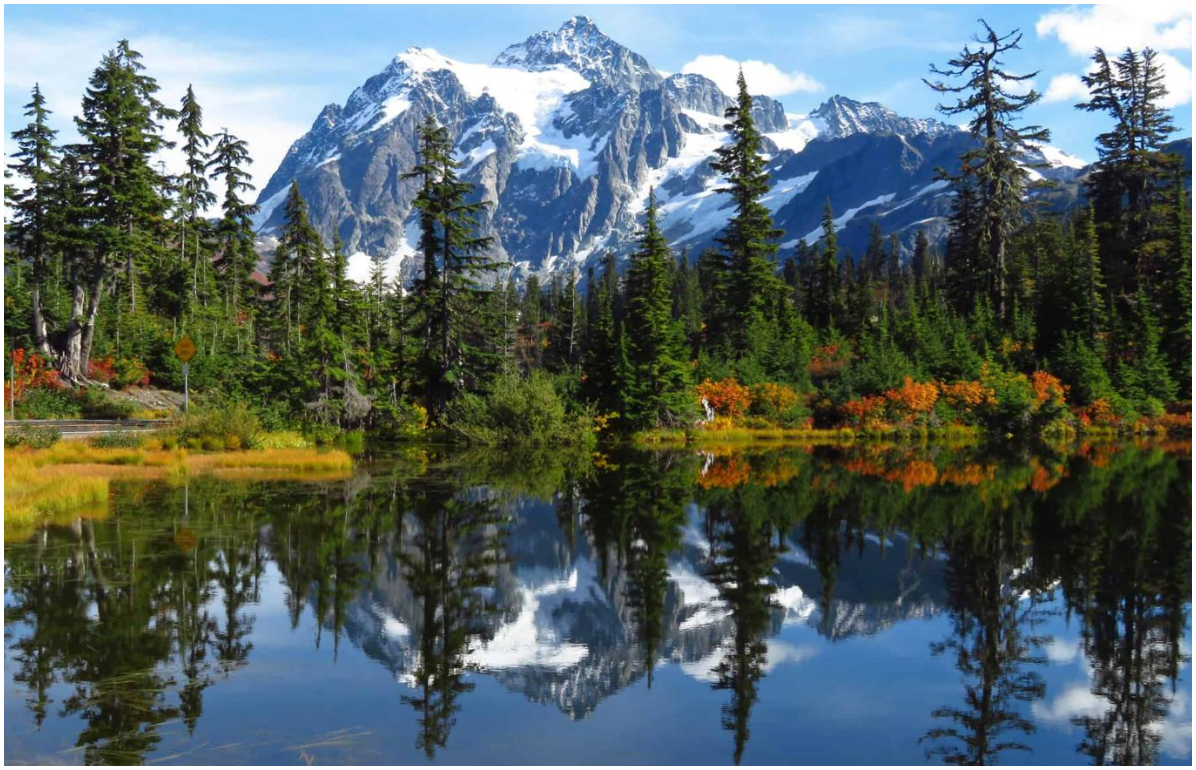

Figure 9. Content image of high resolution .

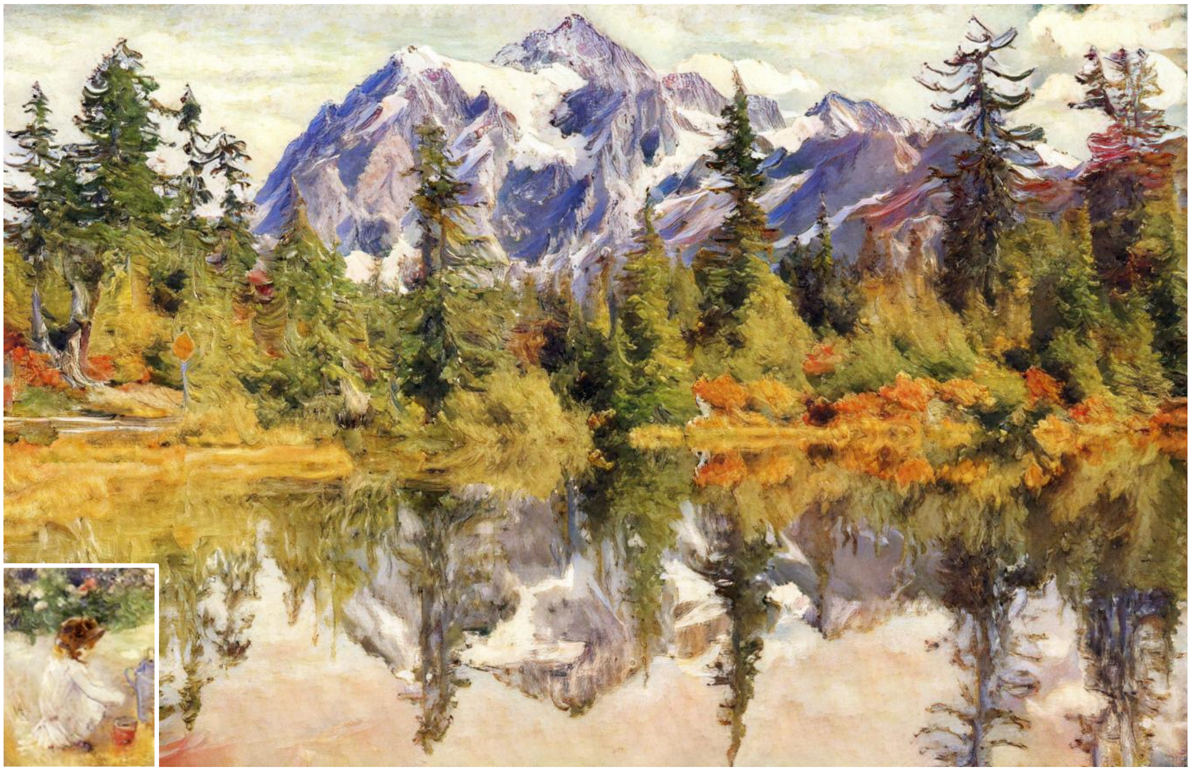

Figure 10. Stylized image. The style image is at the bottom left.

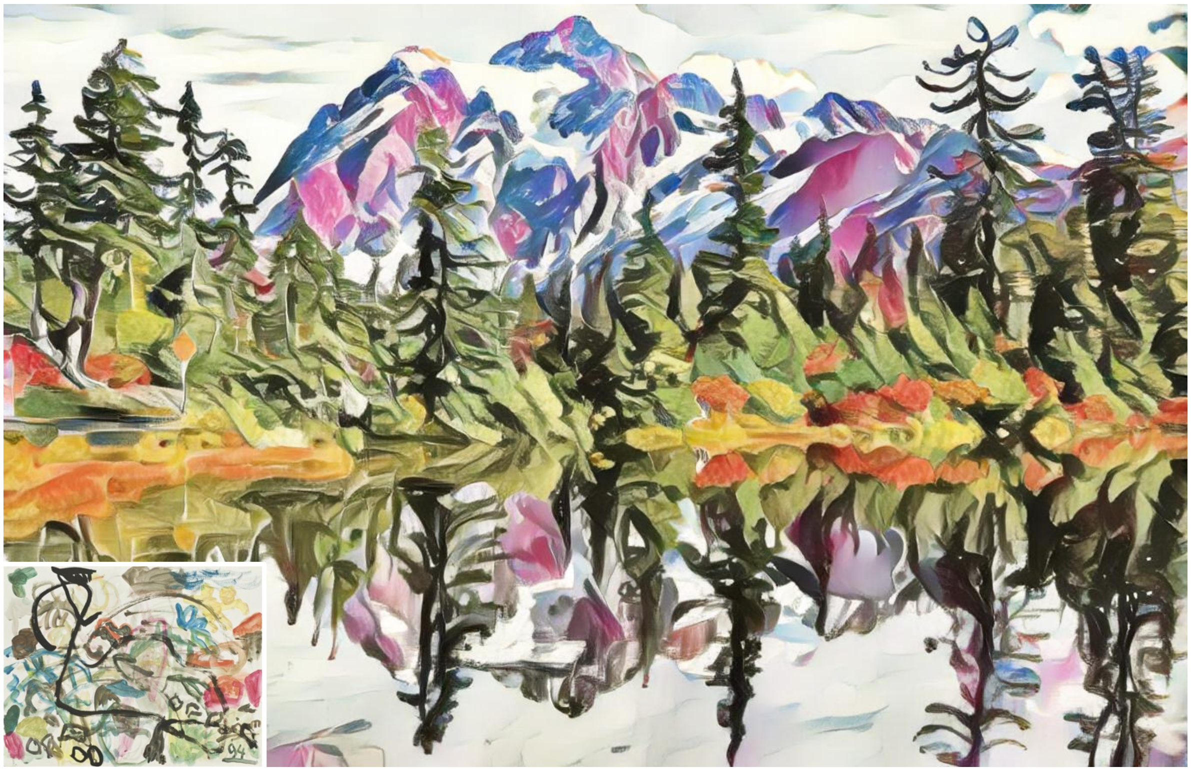

Figure 11. Stylized image. The style image is at the bottom left.

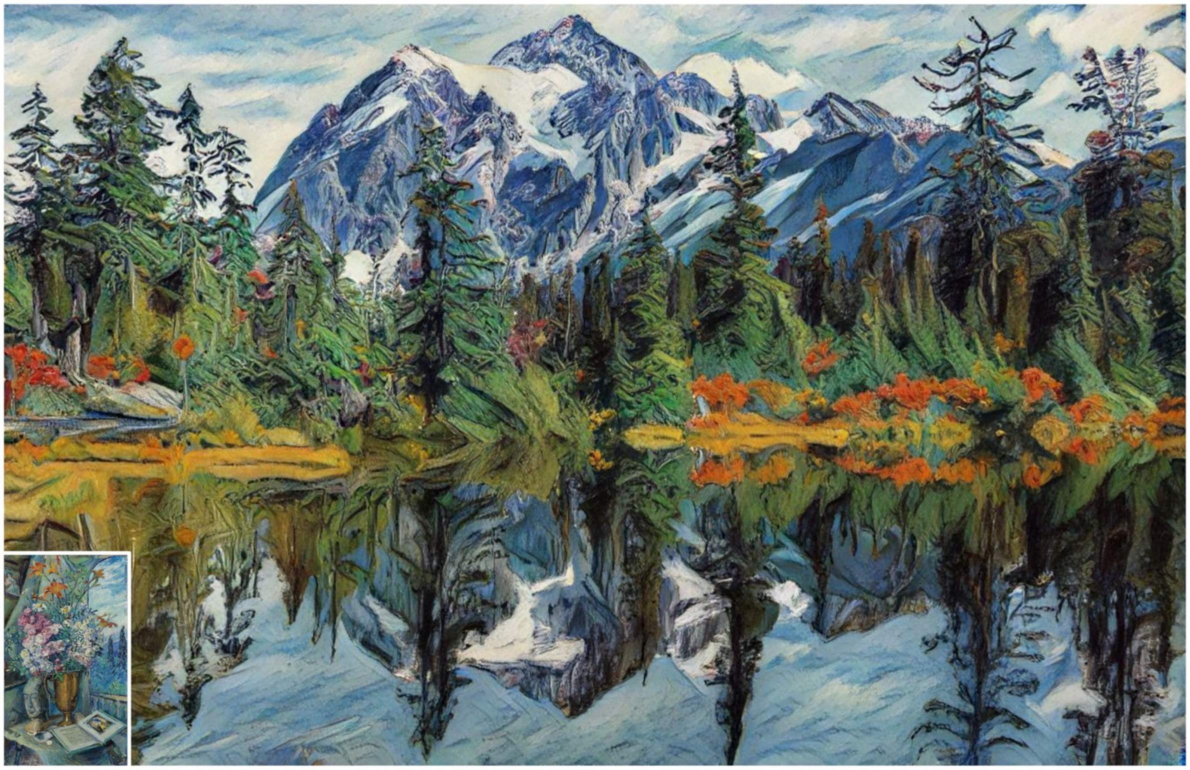

Figure 12. Stylized image. The style image is at the bottom left.