Supplementary Materials for "$\mathcal{Z}^*$: Zero-shot Style Transfer via Attention Reweighting"

Yingying Deng[◦,1], Xiangyu He[◦,1],Fan Tang[✉,2], Weiming Dong[1]
[1] MAIS, Institute of Automation, Chinese Academy of Sciences
[2] Institute of Computing Technology, Chinese Academy of Sciences

Contents

A. Qualitative Comparison with StyleDiff

Figure S1 presents a detailed comparative analysis conducted in conjunction with StyleDiff [7] to assess the performance of different methods. By utilizing images from the CelebA-HQ [9] dataset as content images, we observe in the first and second columns that the stylized output demonstrates a certain degree of resemblance to the input face. However, it falls short in capturing the intricate style patterns effectively.

Expanding our investigation beyond the limitations of the pre-trained dataset, as depicted in the third and fourth columns, we encounter a perplexing scenario where the stylized output loses its intended style representation. This outcome highlights the challenges associated with maintaining the desired style when confronted with unfamiliar or out-of-domain content.

B. Visualization of Content Style Correlation

Figure S3 showcases the visual depiction of the results obtained from the content-style correlation analysis, focusing specifically on the outcomes derived from Softmax($Q_c K_s$). The visualization provides evidence that a well-trained diffusion model [10] can effectively leverage the correlation between the content query and style key, which possess similar semantic information, without necessitating retraining. This observation, where the alignment of content and style

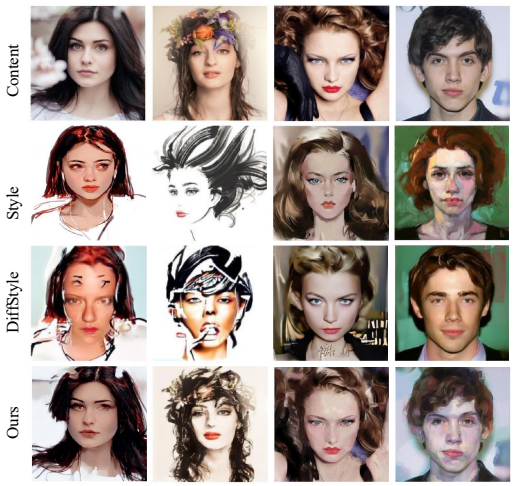


Figure S1. Visual comparisons with StyleDiff. As presented in the 1st and 2nd columns, StyleDiff demonstrates limitations in effectively handling images that deviate from the training domain. Nevertheless, our method exhibits consistent and satisfactory performance across all types of input images.

images is achieved through cross-attention, serves as a source of inspiration for our proposed dual-path network. This network incorporates an enhanced attention mechanism to facilitate the transfer of style to the content images, thereby improving the overall style transfer performance.

Our method exploits the inherent relationship between content and style by incorporating a more advanced attention mechanism, which enables the network to selectively focus on relevant style information and effectively transfer it to the content images. By leveraging the insights gained from the content-style correlation analysis, we anticipate that our proposed approach will lead to enhanced style transfer results, surpassing the limitations of existing methods.

C. Proof of Equation (14)

To enhance the clarity of the derivation process, the proof commences on a fresh page.

Proof.

$$\hat{f}_c = \frac{1}{2} \cdot \text{Attn}(Q_c, K_s, V_s) + \frac{1}{2} \cdot \text{Attn}(Q_c, K_c, V_c) = \frac{1}{2} \cdot \text{Softmax}(\frac{Q_c K_s^T}{\sqrt{d}})V_s + \frac{1}{2} \cdot \text{Softmax}(\frac{Q_c K_c^T}{\sqrt{d}})V_c \tag{1}$$

Note that the matrix $\hat{f}_c$ can be expressed as a combination of individual elements. To simplify the derivation process, we focus on each element $\hat{f}_c^{i,j}$

$$\hat{f}_c^{i,j} = \frac{1}{2} \cdot \text{Softmax}(\frac{[Q_c]_{i,\cdot} K_s^T}{\sqrt{d}})[V_s]_{\cdot,j} + \frac{1}{2} \cdot \text{Softmax}(\frac{[Q_c]_{i,\cdot} K_c^T}{\sqrt{d}})[V_c]_{\cdot,j} \tag{2}$$

where $[Q]_{i,\cdot} \in \mathbb{R}^{1\times N}$ corresponds to the i-th row in matrix Q and $[V]_{\cdot,j} \in \mathbb{R}^{N\times 1}$ indicates the j-th column in matrix V. To facilitate the derivation, we introduce two intermediate variables:

$$x^{cs} = \frac{1}{\sqrt{d}} \cdot [Q_c]_{i,\cdot} K_s^T, \qquad x^{cc} = \frac{1}{\sqrt{d}} \cdot [Q_c]_{i,\cdot} K_c^T \tag{3}$$

where $x^{cs}, x^{cc} \in \mathbb{R}^{1\times N}$. Expanding the Softmax formulation, we express $\hat{f}_c^{i,j}$ as a weighted sum over N elements:

$$\hat{f}_c^{i,j} = \frac{1}{2} \cdot \sum_n^N \frac{\exp(x_n^{cs})}{\sum_m^N \exp(x_m^{cs})}[V_s]_{k,j} + \frac{1}{2} \cdot \sum_n^N \frac{\exp(x_n^{cc})}{\sum_m^N \exp(x_m^{cc})}[V_c]_{k,j} \tag{4}$$

$$= \frac{1}{2} \cdot \sum_n^N \frac{\exp(x_n^{cs})}{\sum_m^N \exp(x_m^{cs})}[V_s]_{k,j} + \frac{\exp(x_n^{cc})}{\sum_m^N \exp(x_m^{cc})}[V_c]_{k,j} \tag{5}$$

$$= \frac{1}{2} \cdot \sum_n^N \frac{\exp(x_n^{cs})[V_s]_{k,j} \sum_m^N \exp(x_m^{cc})}{\sum_m^N \exp(x_m^{cc}) \sum_m^N \exp(x_m^{cs})} + \frac{\exp(x_n^{cc})[V_c]_{k,j} \sum_m^N \exp(x_m^{cs})}{\sum_m^N \exp(x_m^{cc}) \sum_m^N \exp(x_m^{cs})} \tag{6}$$

$$= \sum_n^N \frac{\exp(x_n^{cs})[V_s]_{k,j} \sum_m^N \exp(x_m^{cc}) + \exp(x_n^{cc})[V_c]_{k,j} \sum_m^N \exp(x_m^{cs})}{2 \sum_m^N \exp(x_m^{cc}) \sum_m^N \exp(x_m^{cs})} \tag{7}$$

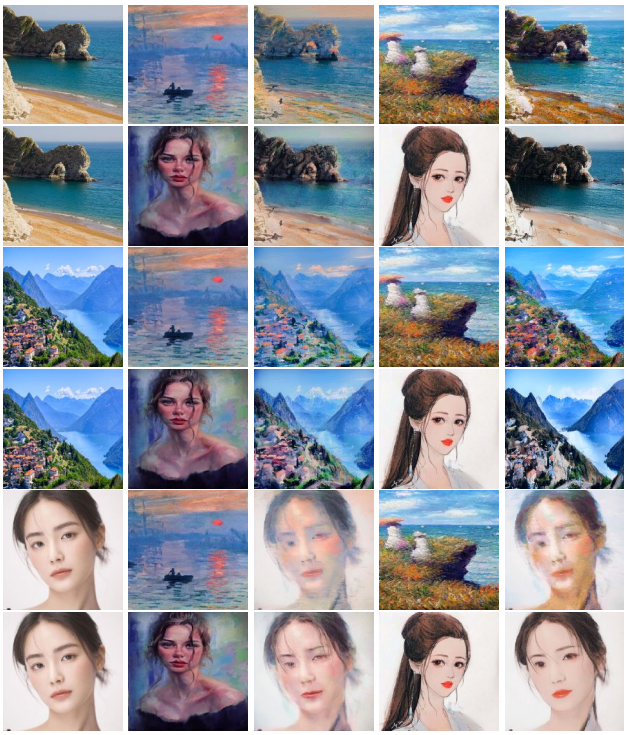
$$= \sum_n^N \frac{\exp(x_n^{cs})[V_s]_{k,j} \sum_m^N \exp(x_m^{cc}) + \exp(x_n^{cc})[V_c]_{k,j} \sum_m^N \exp(x_m^{cs})}{\sum_m^N \exp(x_m^{cc}) \sum_m^N \exp(x_m^{cs}) + \sum_m^N \exp(x_m^{cc}) \sum_m^N \exp(x_m^{cs})} \tag{8}$$

$$= \sum_n^N \frac{\exp(x_n^{cs})[V_s]_{k,j} \frac{\sum_m^N \exp(x_m^{cc})}{\sum_m^N \exp(x_m^{cs})} + \exp(x_n^{cc})[V_c]_{k,j}}{\sum_m^N \exp(x_m^{cs}) \frac{\sum_m^N \exp(x_m^{cc})}{\sum_m^N \exp(x_m^{cs})} + \sum_m^N \exp(x_m^{cc})} \tag{9}$$

Upon rewriting the equation $\ln \frac{\sum_m^N \exp(x_m^{cc})}{\sum_m^N \exp(x_m^{cs})} = C$, we obtain the resulting expression

$$\hat{f}_c^{i,j} = \sum_n^N \frac{\exp(x_n^{cs} + C)[V_s]_{k,j} + \exp(x_n^{cc})[V_c]_{k,j}}{\sum_m^N \exp(x_m^{cs} + C) + \sum_m^N \exp(x_m^{cc})} = \text{Softmax}([x^{cs} + C, x^{cc}]) * \begin{bmatrix} [V_s]_{\cdot,j} \\ [V_c]_{\cdot,j} \end{bmatrix} \tag{10}$$

By substituting the derived expression for $\hat{f}_c^{i,j}$ into the matrix form, we obtain the formulation presented in Equation (8). □



(a) Content (b) Style I (c) Result I (d) Style II (e) Result II

Figure S2. We present a visual representation of the style transfer outcomes achieved by applying landscape image styles to both landscape and portrait images, as well as vice versa. While our method demonstrates promising results in the context of style transfer within the same domain (as evidenced by the 1st, 3rd, and last rows), there is room for improvement when transferring portrait styles to landscape images, as illustrated by the 2nd and 4th rows.

D. Impact of λ in Equation (8)

Unlike the resilient nature of the λ parameter in Attention Reweighting, achieving a suitable balance between the style and content features in Equation (8) proves to be challenging. To elucidate this issue, we present the outcomes in Figure S4, employing a pair of style transfer images that exhibit varying levels of difficulty in style transfer.

E. Quantitative Analysis

For quantitative comparison with state-of-the-art methods, we employ content loss and style loss as evaluation metrics to assess the preservation of content and rendering of style, respectively. Content loss and style loss are commonly adopted in style transfer tasks, defined as:

$$\mathcal{L}_c = \frac{1}{N_l}\sum_{i=0}^{N_l}\|\phi_i(I_o) - \phi_i(I_c)\|_2, \tag{11}$$

where $\phi_i(\cdot)$ denotes features extracted from the i-th layer in a pre-trained VGG19 and N_l is the number of layers. The style perceptual loss $\mathcal{L}_s$ is defined as

$$\begin{aligned}\mathcal{L}_s = \frac{1}{N_l}\sum_{i=0}^{N_l}&\|\mu(\phi_i(I_o)) - \mu(\phi_i(I_s))\|_2 \\ &+ \|\sigma(\phi_i(I_o)) - \sigma(\phi_i(I_s))\|_2,\end{aligned} \tag{12}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and variance of extracted features, respectively. The content loss is computed using the features extracted from the 5-th layer, while the style loss is calculated using the features from the 1-st to 5-th layers.

The outcomes of the comparative analysis are presented in Table S2. Our proposed method achieves the lowest style loss, indicating its superior ability to render style. On the other hand, StyTr^2 [4] achieves the lowest content loss, highlighting its effectiveness in preserving content. Notably, our method demonstrates relatively low values for both content and style loss, suggesting a favorable balance between content preservation and style rendering.

We also performed further quantitative analysis for the ablation study, detailed in Table S1, which supports the findings discussed in the ablation study.

F. Visual Comparison with SOTA

In order to provide further evidence of the effectiveness of our proposed method, we present additional results in Figures S5 and S6. These figures demonstrate the performance of our method across a range of image genres, including portraits and landscapes, as well as concrete and abstract styles. The results consistently indicate the strong performance and versatility of our method in handling various image styles.

G. Limitations

Like a coin with two sides, the attention-based zero-shot style transfer method presents a nuanced evaluation with both positive and negative aspects. The favorable outcomes of this approach benefit the utilization of a pre-trained stable diffusion model, which is also widely acknowledged to yield varying quality results based on the input data. Through our rigorous experimentation, we discovered a potential limitation of our method when transferring style information from portraits to landscape images. To facilitate a comprehensive understanding, we have included a visual comparison of the results in Figure S2.

References

[1] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. ArtFlow: Unbiased image style transfer via reversible neural flows. In *IEEE/CVF Conferences on*

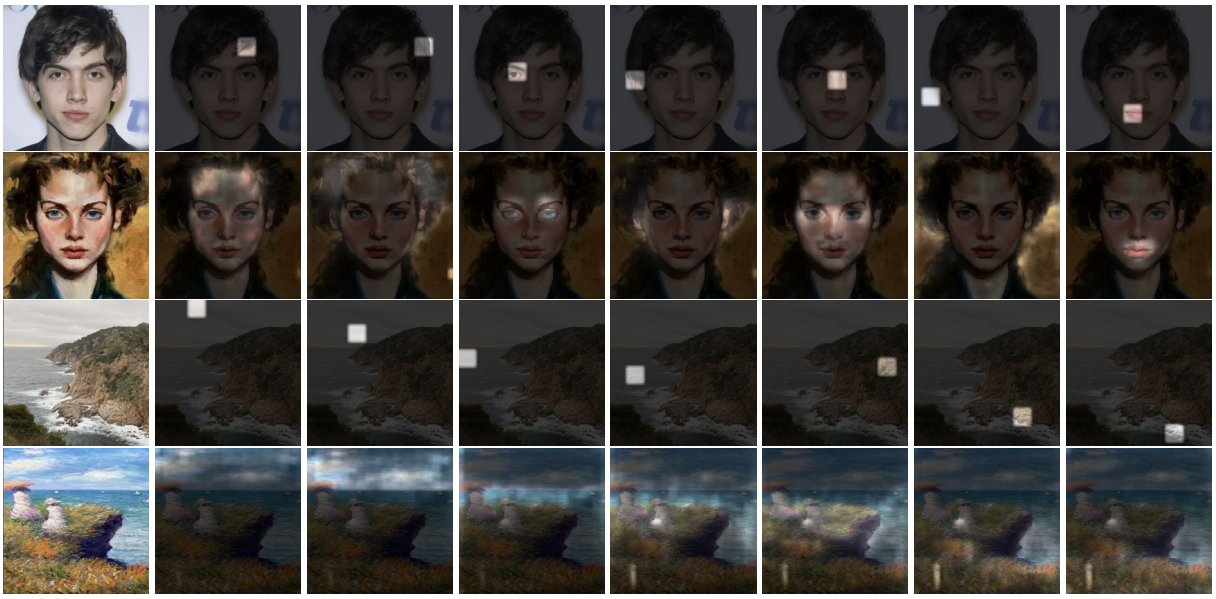


Figure S3. Visualization of cross-attention results. The first and third rows pertain to the content images, whereas the second and last rows pertain to the style images. The highlighted blocks in the content images denote the content queries, while the highlighted regions in the style images correspond to the relevant regions of the corresponding content query blocks, as determined via the Softmax function in cross-attention with high values.

(a) Content (b) Style (c) $\lambda = 0.1$ (d) $\lambda = 0.3$ (e) $\lambda = 0.5$ (f) $\lambda = 0.7$ (g) $\lambda = 0.9$

Figure S4. We investigate the impact of the λ parameter in Equation (8), specifically in the context of Simple Addition. We present a visual representation of the outcomes obtained when the style and content images exhibit a lack of correlation (i.e., a challenging sample) in the first row, and when the content and style images are similar (an easier sample) in the second row. Notably, we observe that the second row consistently yields favorable visualization results when λ is set to 0.9. However, the first row fails to achieve a satisfactory balance between the content and style. For instance, when λ is low, the content is well preserved at the expense of losing the style, and vice versa.

Table S1. Quantitative results for ablation study

	$\mathcal{L}_c$	$\mathcal{L}_s$
ours	1.61	3.70
step 0-10	2.70	3.56
step 10-20	1.17	4.56
step 20-30	0.11	5.03
step 5-20	1.37	4.13

	$\mathcal{L}_c$	$\mathcal{L}_s$
step 15-30	1.25	4.53
Simple Add	1.86	4.64
Naïve Set	1.92	3.96
layer 0-5	2.16	4.81
layer 5-10	0.96	6.17

	$\mathcal{L}_c$	$\mathcal{L}_s$
layer 0-15	2.46	4.04
λ = 1.0	1.04	4.80
λ = 1.1	1.56	4.22
λ = 1.3	2.08	3.47
λ = 1.4	2.18	3.30

Table S2. Quantitative comparisons. To evaluate the preservation of input content and style, we calculate the average values of content and style loss for the results obtained through various methods. The most favorable outcomes are highlighted in **bold**.

	ours	VCT [3]	StyleDiff [6]	InST [12]	QuanArt [5]	CAST [11]	StyTr^2 [4]	IEST [2]	AdaAttN [8]	ArtFlow [1]
$\mathcal{L}_c$	1.61	1.73	3.55	2.85	1.71	1.66	**1.59**	1.82	1.89	1.93
$\mathcal{L}_s$	**3.70**	3.98	4.01	4.70	5.50	4.18	3.79	3.86	3.78	4.32

Computer Vision and Pattern Recognition (CVPR), pages 862–871, 2021. 5

[2] Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. Artistic style transfer with internal-external learning and contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5

[3] Bin Cheng, Zuhao Liu, Yunbo Peng, and Yue Lin. General image-to-image translation with one-shot image guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22736–22746, 2023. 5

[4] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11326–11336, 2022. 3, 5

[5] Siyu Huang, Jie An, Donglai Wei, Jiebo Luo, and Hanspeter Pfister. Quantart: Quantizing image style transfer towards high visual fidelity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 5

[6] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free style transfer emerges from h-space in diffusion models. *arXiv preprint arXiv:2303.15403*, 2023. 5

[7] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649*, 2023. 1

[8] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6649–6658, 2021. 5

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 1

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[11] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022. 5

[12] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In

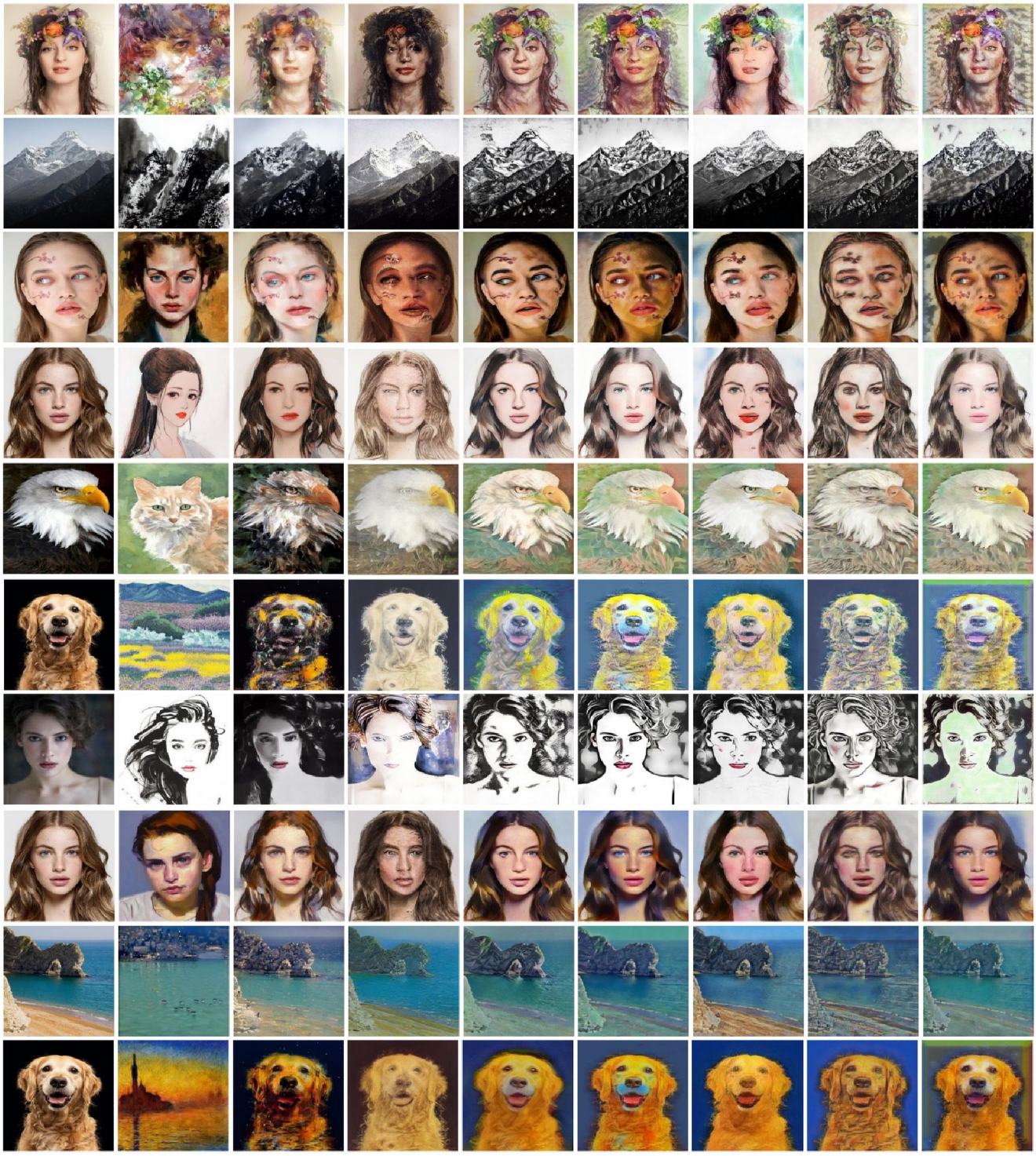


Figure S5. Compared with other style transfer methods. The content images are presented in the first column, the style images are presented in the second column, and the stylized results generated by different methods are presented in the rest.

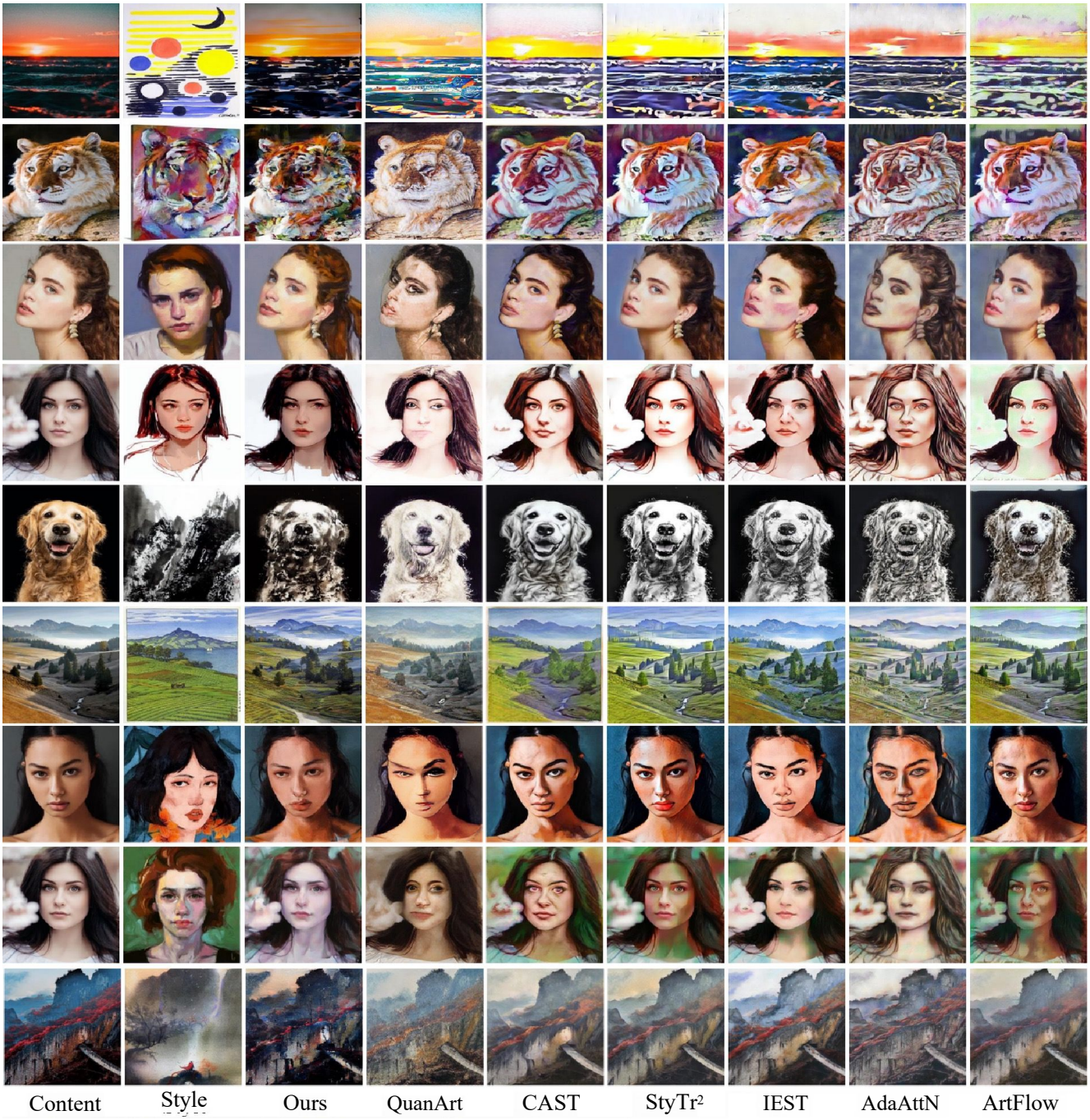


Figure S6. Compared with other style transfer methods. The content images are presented in the first column, the style images are presented in the second column, and the stylized results generated by different methods are presented in the rest.

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10146–10156, 2023. 5