StyTr^2: Image Style Transfer with Transformers – Supplementary Materials

Yingying Deng[1,2] Fan Tang[3*] Weiming Dong[1,2*] Chongyang Ma[4]
Xingjia Pan[2] Lei Wang[5] Changsheng Xu[1,2]

[1] School of Artificial Intelligence, UCAS [2] NLPR, Institute of Automation, CAS
[3] School of Artificial Intelligence, Jilin University [4] Kuaishou Technology [5] CIPUC

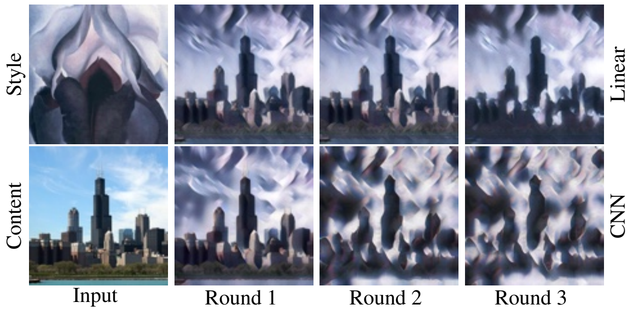


Figure 1. Comparisons of CNN based and our linear projection based layers through several rounds of image stylization.

1. Ablation Study

Linear projection layer. As described in Section 1 of our main text, CNN-based style transfer structures may have issues of content leak after several rounds of the stylization process due to the details missing caused by locality and spatial invariance. To show the impact of CNN structures, we replace the linear projection layer with a deep CNN projection layer to split the image into a sequence. Fig. 1 shows the stylization results after multiple rounds using different projection methods. When increase the number of stylization rounds, the content structures generated by the CNN projection layer blur out, while the content structures generated by the linear projection layer are still distinct.

Progressive generation. Thanks to the ability to capture long-range dependencies, transformer can be used to generate image in a progressive fashion without content leak. We use three layers of transformer encoder-decoder to simulate the process of art creation from coarse to fine. As depicted in Fig. 2, after remove the progressive layers and only use a single transformer layer, the results lose more structure details and the rendered style patterns become less similar to the reference style image (the square strokes in the sky).

In Table 1, we compute the content loss and style loss defined in Equations (9) and (10) of our main text. From this table we can see that our progressive generation scheme can increase both the content consistency and the style fidelity.

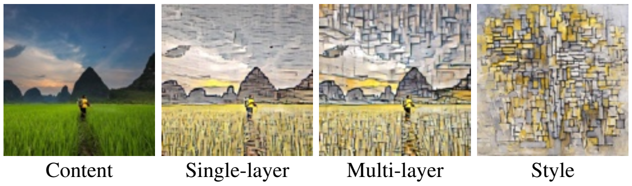


Figure 2. Impact of progressive generation of transformer decoder.

	Single-layer	Ours (multi-layer)
$\mathcal{L}_c \downarrow$	1.94	**1.91**
$\mathcal{L}_s \downarrow$	2.38	**1.47**

Table 1. Quantitative evaluation of progressive generation.

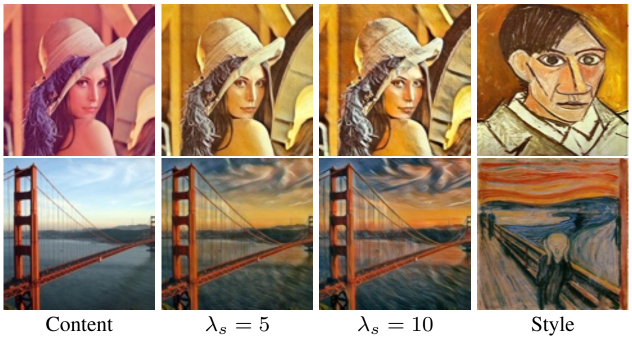


Figure 3. Impact of style weight.

Training weights. The loss weights λ_c, λ_s, λ_{id1}, and λ_{id2} are adopted to eliminate the impact of magnitude differences. We change λ_s from 10 to 5 to evaluate the impact of style weight value. Fig. 3 shows that reducing style weight leads to less obvious stylized patterns in the final output but the overall results still look reasonable.

Loss function. StyTr^2 is trained using a pretrained VGG model which is not free of CNNs completely. We try to use a pretrained VIT [3] model to extract features and construct content and style loss. We use the VIT-based loss to train a new model and show some results in Fig. 4.

[*] Co-corresponding authors

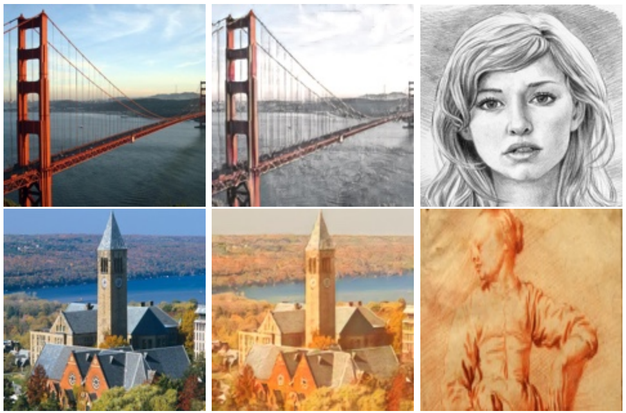


Figure 4. Style transfer results using model trained by VIT-based loss.

The content structures are well preserved but very few style patterns are transferred. Due to the global modeling ability of transformer, local texture patterns are difficult to be discovered. We leverage the pretrained VGG model to integrate some priors of local texture patterns. Therefore, our results can preserve both complete content structures and vivid style patterns.

2. Robustness Analysis

Solving the content leak problem proves from the side that our method has a strong ability of transferring the content information of I_c into I_{cs}. However, we argue that robust content feature extraction and representation are the foundation of image style transfer. From Figure 5 of our main text we can see that I_{cs}^1 and I_{cs}^{20} are almost the same by ArtFlow, which indicates that when the input contents are totally different (I_c for I_{cs}^1 while I_{cs}^{19} for I_{cs}^{20}), the generated results could be the same. Such phenomenon inspires us to further analyze the ability of content feature extraction for ArtFlow. To verify the robustness of different methods, we add small perturbation into the multi-round stylization process, that is:

$$I_{cs}^i = G_i(G \dots (G_1(I_c + \Delta, I_s), \dots) + \Delta, I_s), \qquad (1)$$

where $\Delta \sim \mathcal{N}(0, \sigma^2 I)$ is a small perturbation following a standard Gaussion distribution. We show the stylization results after the 1st, 2nd, 5th, and 10th rounds in Fig. 5. The small perturbation brings little impact to our results, but leads to significant content drop-out problem for ArtFlow. Such results prove that although ArtFlow is good at maintaining the content information, the content representation mechanism is not robust to small perturbation. Moreover, Fig. 5 demonstrates that flow-based model suffers from the problem of poor representation learning.

3. More Qualitative Comparison Results

In Fig. 6, we show our results in the resolution of 768 × 768. In the generated results, the outline of content structures is clear (e.g., the mountains and buildings in the first and second rows, the animals in the third and fourth rows), and details of style patterns are well preserved (the pencil strokes in the second and fourth rows). In Figs. 7, 8, 9, 10 and 11, we show more results to compare our method with AdaIN [4], AAMS [8], DFP [7], MCC [2], and ArtFlow [1].

4. Dataset License

Our model is trained on MS-COCO [5] and WikiArt [6]. According to license of MS-COCO[1], we are allowed to copy and redistribute the material in any medium or format, remix, transform, and build upon the material. WikiArt[2] can be used as historically significant artworks.

5. Broader Impacts

Compared to the CNN-based models, the transformer-based model brings some breakthroughs to style transfer. StyTr^2 uses the strong representation and relationship modeling ability of the transformer to avoid the image details missing and to improve the stylization effect. There can be some enlightenment for other vision tasks to use the specialties of transformers to enhance performance. Moreover, this work is proposed for promising painting creation using referenced content and style images. Although the application of this technique cannot threaten personal safety, deep fake could be used by crooks.

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 Computer Vision and Pattern Recognition (CVPR)*, pages 862–871, 2021. 2

[2] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. Arbitrary video style transfer via multi-channel correlation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1210–1217, 2021. 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[4] Xun Huang and Belongie Serge. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 2

[1] https://creativecommons.org/licenses/by-nc-sa/2.0/

[2] https://www.wikiart.org/en/terms-of-use

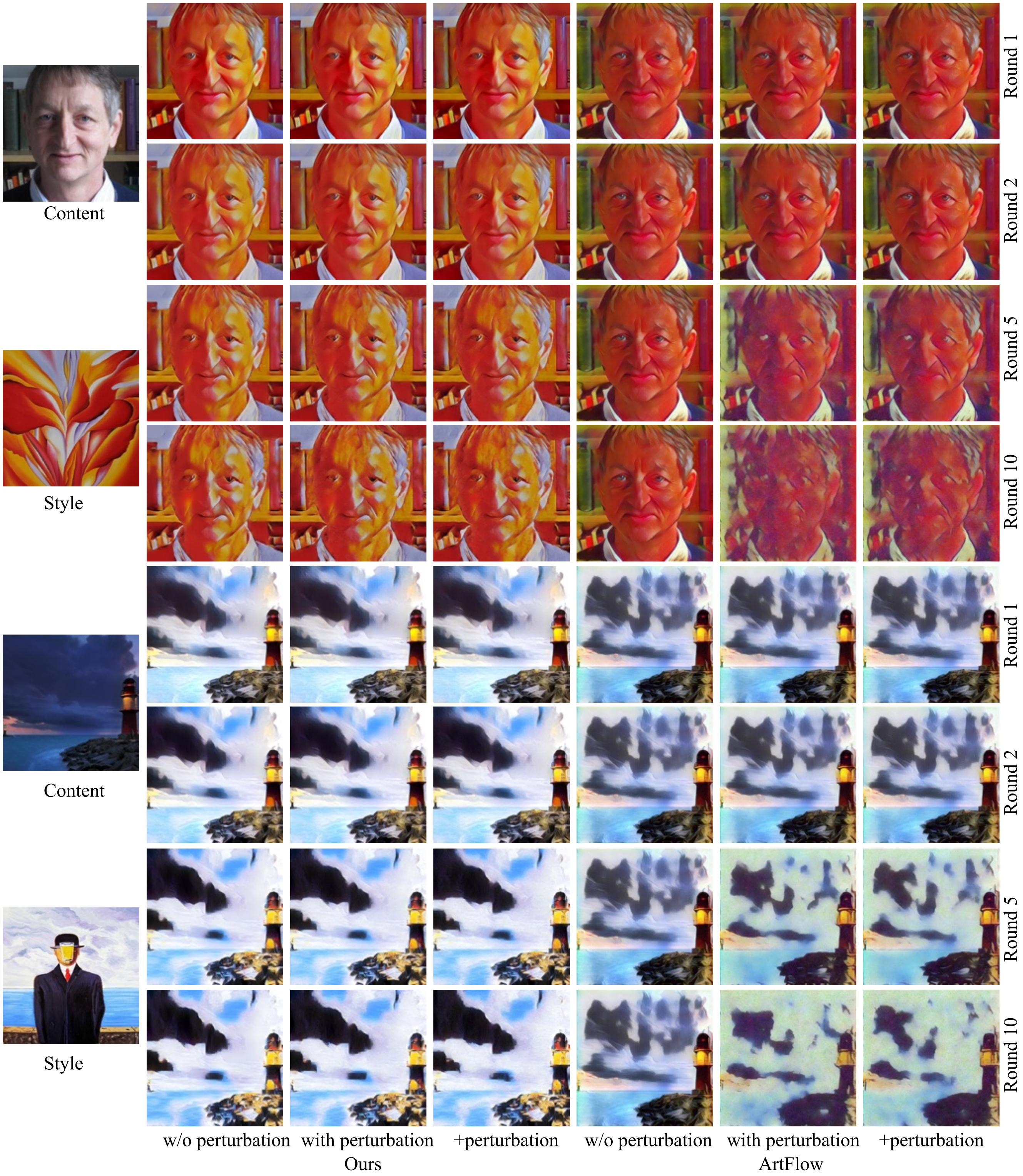


Figure 5. Comparisons of model robustness with ArtFlow. We visualize the impact of adding a small perturbation Δ following a Gaussian distribution on the stylized images in the 4th and 7th columns.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–

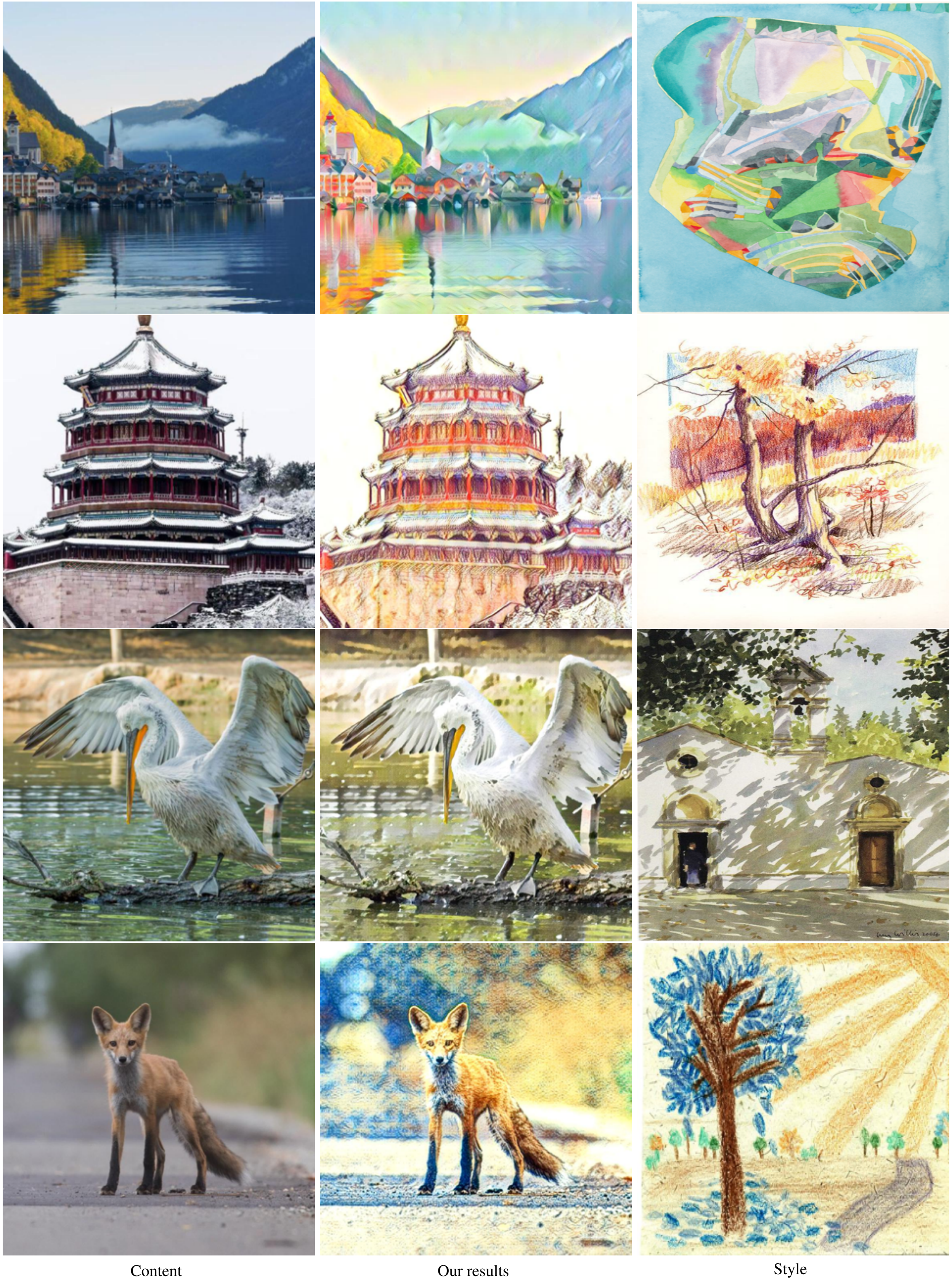


Figure 6. More style transfer results in a large resolution of 768 × 768.

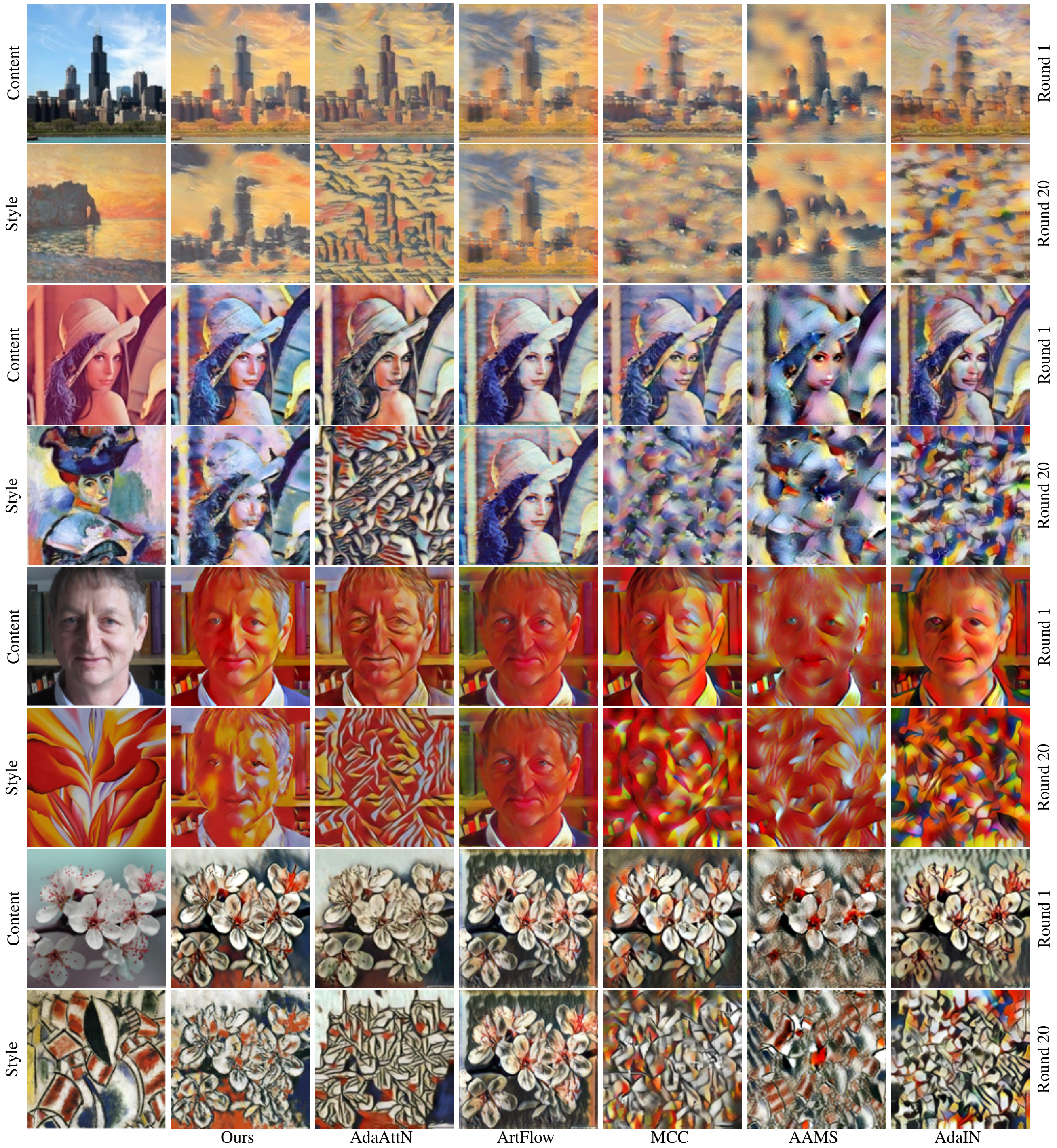


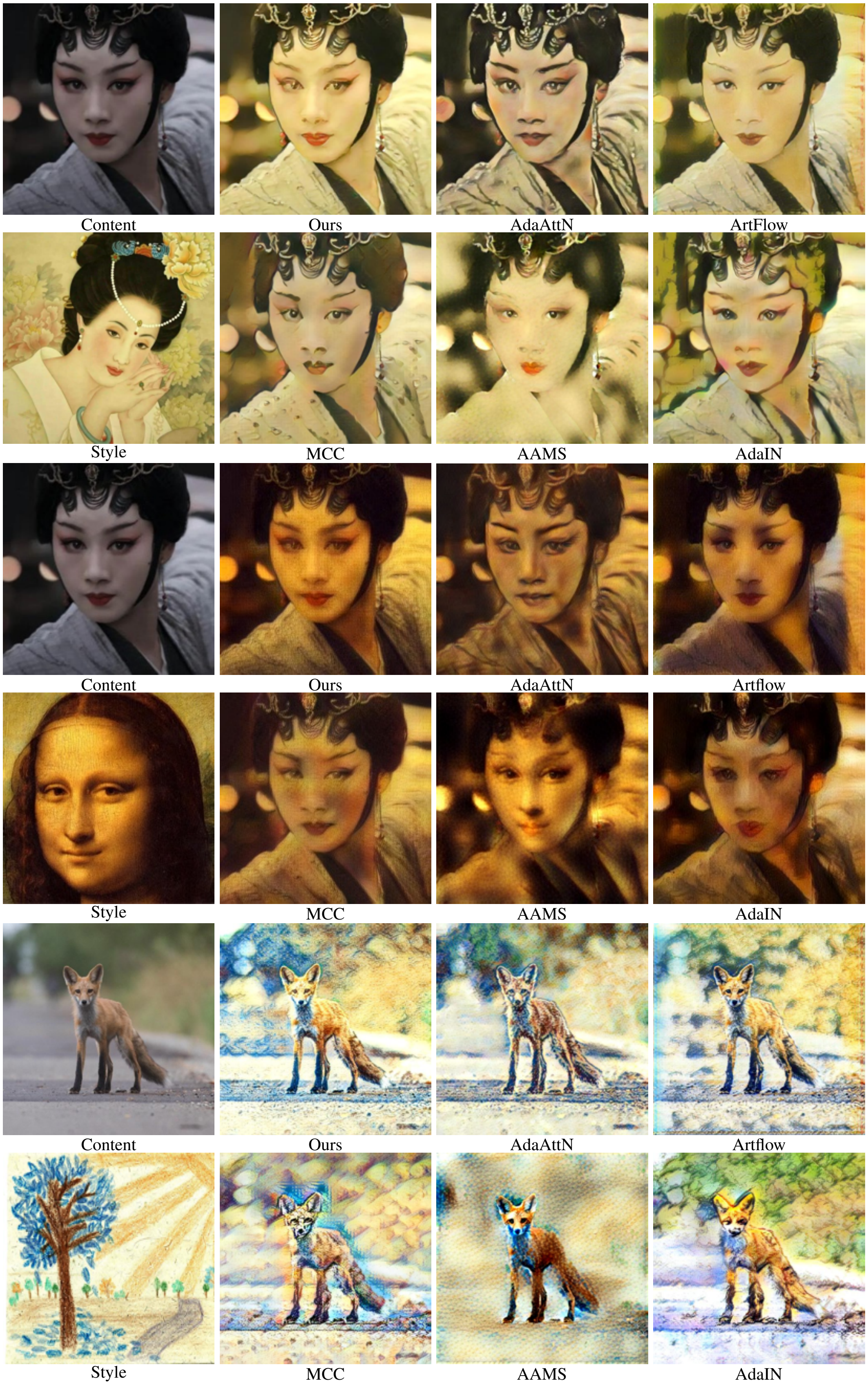
Figure 7. Visualizations of the content leak issue.

755, 2014. 2

[6] Fred Phillips and Brandy Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3):593–608, 2011. 2

[7] Zhizhong Wang, Lei Zhao, Haibo Chen, Lihong Qiu, Qihang Mo, Sihuan Lin, Wei Xing, and Dongming Lu. Diversified arbitrary style transfer via deep feature perturbation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7789–7798, 2020. 2

[8] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1467–1475, 2019. 2



Content Ours AdaAttN ArtFlow

Style MCC AAMS AdaIN

Content Ours AdaAttN Artflow

Style MCC AAMS AdaIN

Content Ours AdaAttN Artflow

Style MCC AAMS AdaIN

Figure 8. Comparisons of style transfer results using different methods.

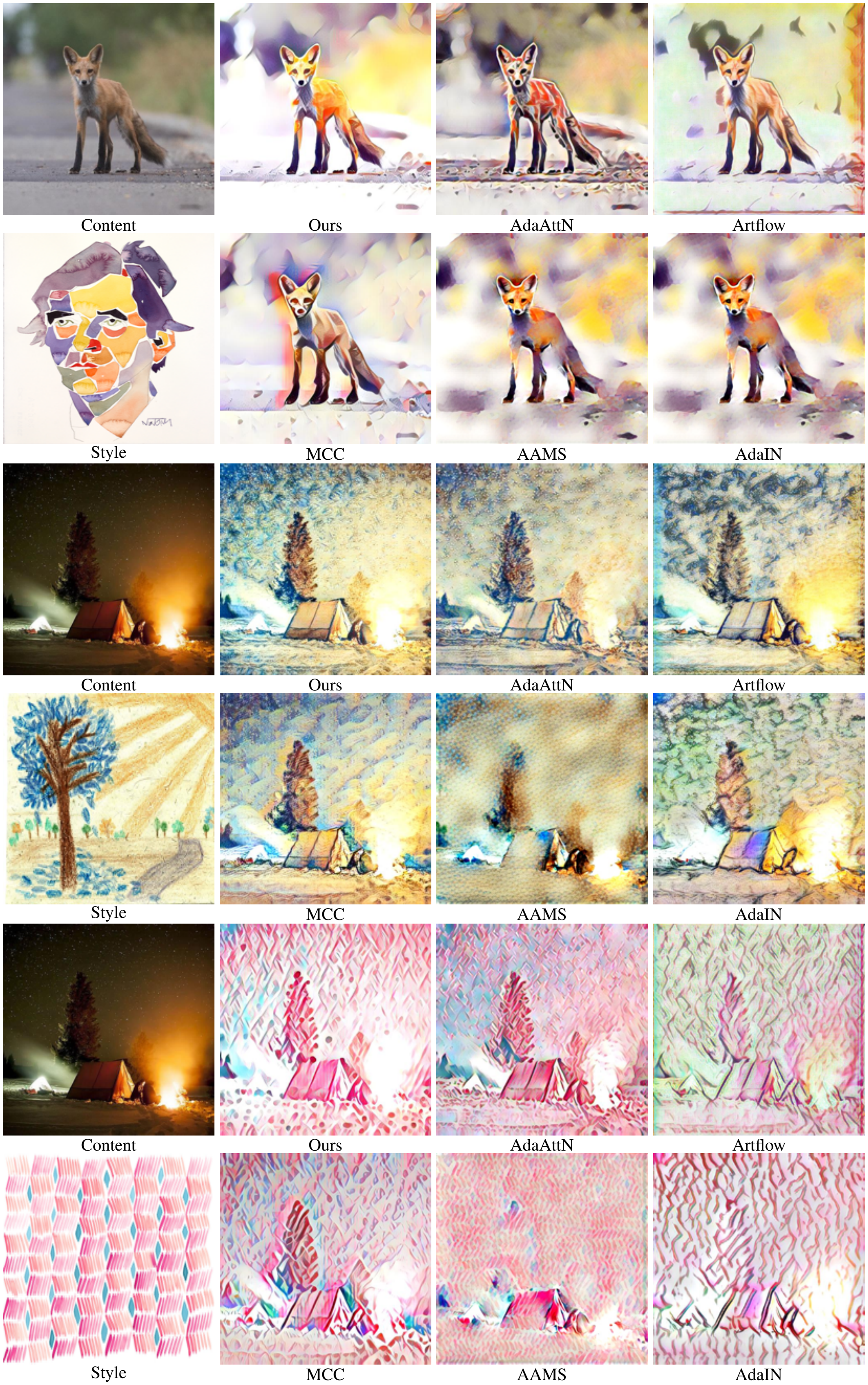


Figure 9. Comparisons of style transfer results using different methods.

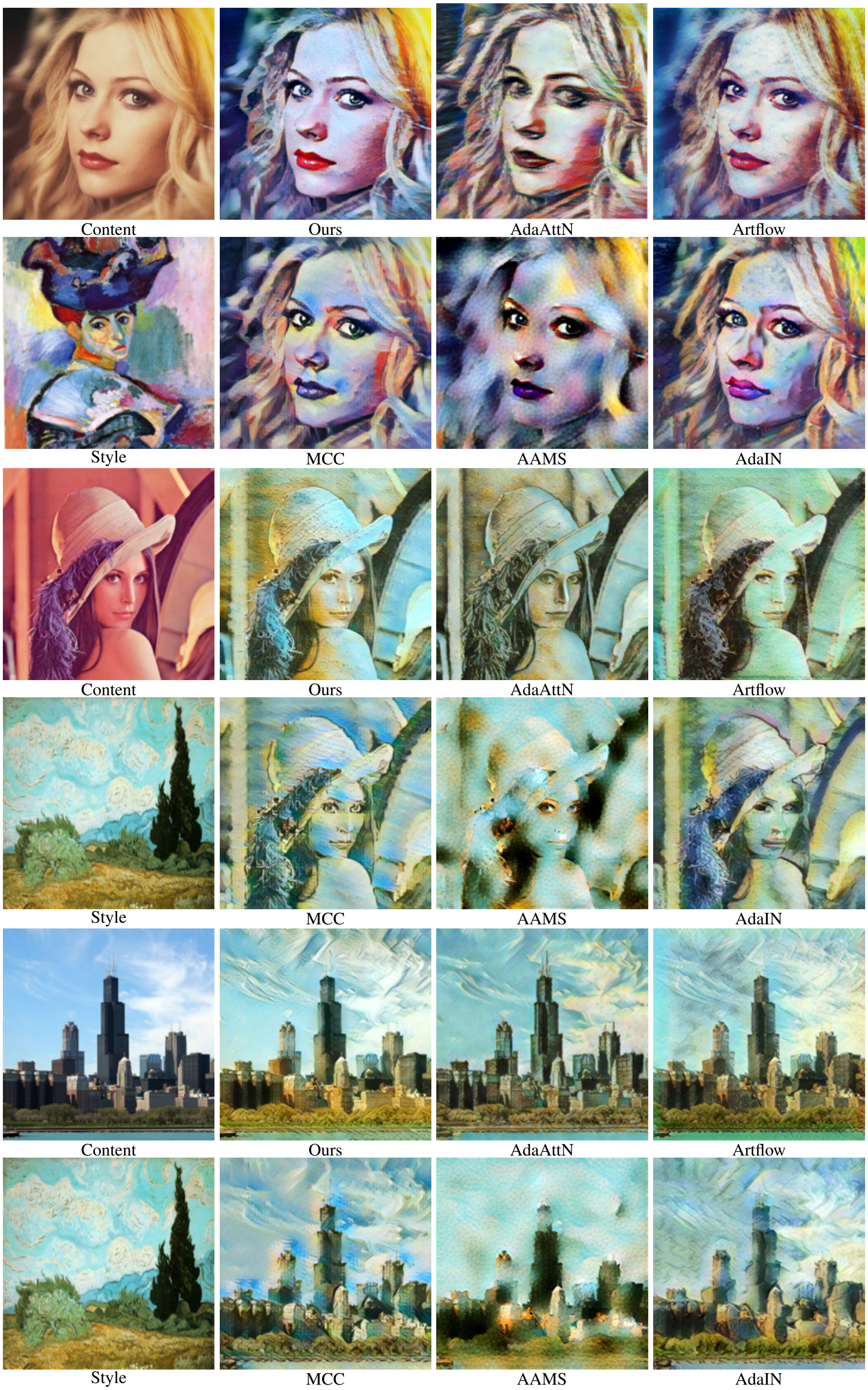


Figure 10. Comparisons of style transfer results using different methods.

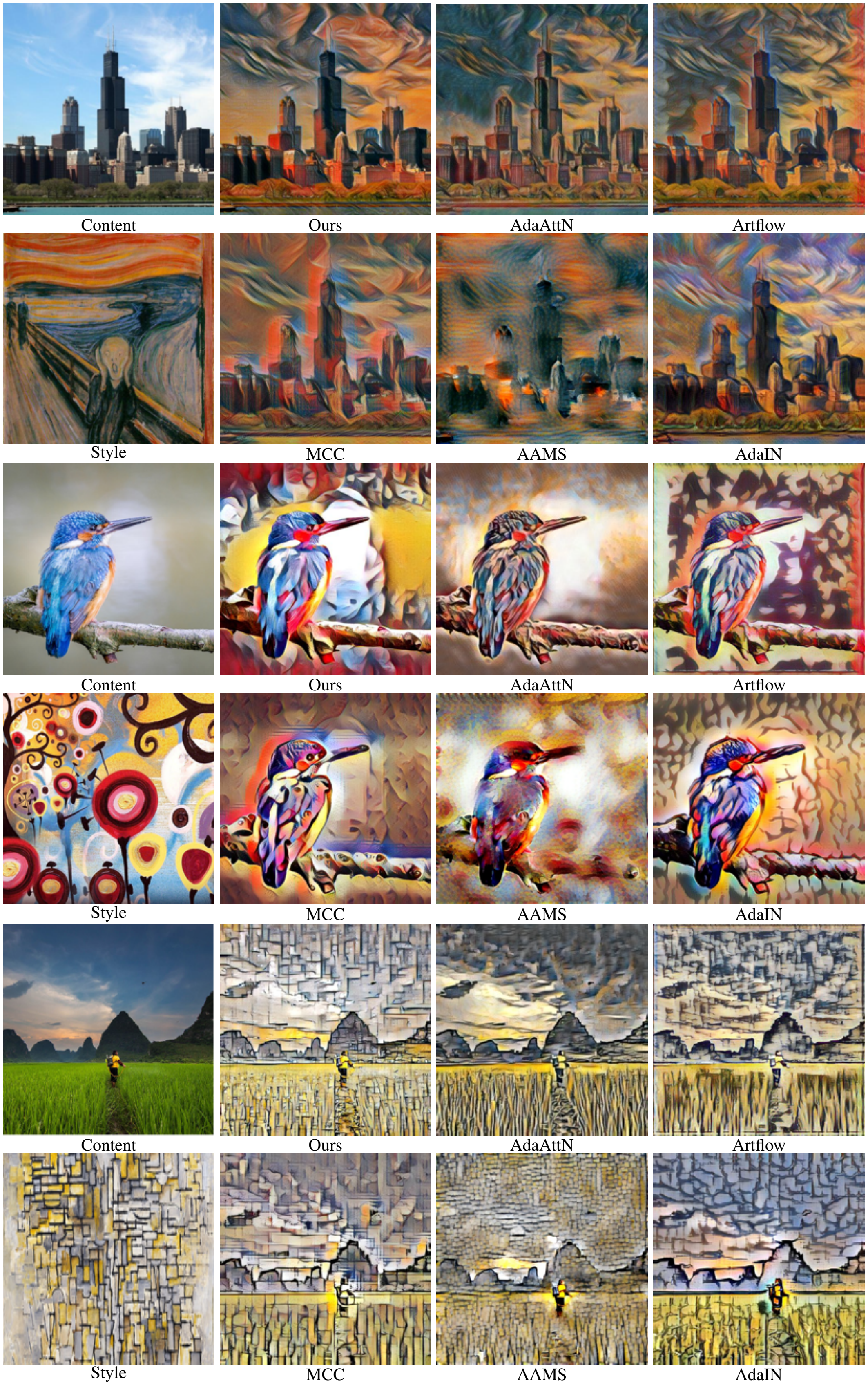


Figure 11. Comparisons of style transfer results using different methods.