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.

### Training weights
The loss weights $\lambda_c$, $\lambda_s$, $\lambda_{id1}$, and $\lambda_{id2}$ are adopted to eliminate the impact of magnitude differences. We change $\lambda_s$ 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.
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:

\[
I_{cs}^1 = G_i(G \ldots (G_1(I_c + \Delta, I_s), \ldots ) + \Delta, I_s),
\]

where \( \Delta \sim \mathcal{N}(0, \sigma^2 I) \) is a small perturbation following a standard Gaussian distribution. We show the stylization results after the 1\(^{st}\), 2\(^{nd}\), 5\(^{th}\), and 10\(^{th}\) 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


\(^{1}\)https://creativecommons.org/licenses/by-nc-sa/2.0/

\(^{2}\)https://creativecommons.org/licenses/by-nc-sa/2.0/
Figure 5. Comparisons of model robustness with ArtFlow. We visualize the impact of adding a small perturbation $\Delta$ following a Gaussian distribution on the stylized images in the 4th and 7th columns.

Figure 6. More style transfer results in a large resolution of 768 × 768.
Figure 7. Visualizations of the content leak issue.


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.