

A. Pseudocode of CAP-VSTNet

Algorithm 1: Video style transfer process of CAP-VSTNet

Input: content frames $\{I_c^j\}_{j=1}^N$, style image I_s , the proposed new framework CAP-VSTNet which consists an unbiased style transfer module cWCT;

Optional: content semantic masks $\{M_c^j\}_{j=1}^N$, style semantic mask M_s ;

Result: stylized frames $\{I_{cs}^j\}_{j=1}^N$;

- 1 feed I_s to CAP-VSTNet and perform the forward inference of CAP-VSTNet to obtain the style feature f_s ;
- 2 **for** $j \leftarrow 1$ **to** N **do**
- 3 feed I_c^j to CAP-VSTNet and perform the forward inference of CAP-VSTNet to obtain the content feature f_c^j ;
- 4 **if** semantic masks M_c^j and M_s are provided **then** feed f_c^j, f_s, M_c^j, M_s to cWCT and obtain the stylized feature f_{cs}^j ;
- 5 **else** feed f_c^j, f_s to cWCT and obtain the stylized feature f_{cs}^j ;
- 6 perform the backward inference of CAP-VSTNet to obtain the stylized frame I_{cs}^j ;
- 7 **end**

B. Unbiased Transformation Module

We show that the whitening and coloring transforms in cWCT is an unbiased style transfer module. Suppose we have a style transfer module $f_{cs} = C(f_c)S(f_s)$, where C, S denote the content and style factor, and f_c, f_s denote the content and style feature. ArtFlow defines that f_{cs} is an unbiased style transfer module if $C(f_{cs}) = C(f_c)$ and $S(f_{cs}) = S(f_s)$.

Without loss of generality, we assume both f_c and f_s are centered. We have,

$$\begin{aligned} \text{Whitening} : \hat{f}_c &= L_c^{-1} f_c, \\ \text{Coloring} : f_{cs} &= L_s \hat{f}_c. \end{aligned} \quad (1)$$

where $f_c f_c^T = L_c L_c^T$, $f_s f_s^T = L_s L_s^T$ and L is a triangular matrix. Therefore,

$$f_{cs} = L_s L_c^{-1} f_c. \quad (2)$$

Here,

$$C(f) = L^{-1} f, S(f) = L. \quad (3)$$

Since,

$$f_{cs} f_{cs}^T = f_s f_s^T = L_s L_s^T, \quad (4)$$

We have,

$$C(f_{cs}) = L_s^{-1} f_{cs} = L_c^{-1} f_c = C(f_c), \quad (5)$$

$$S(f_{cs}) = L_s = S(f_s). \quad (6)$$

Therefore, cWCT is unbiased.

C. Style Interpolation

We investigate the linear interpolation of extracted style representations by CAP-VSTNet. Considering the feature covariance matrices Σ_A and Σ_B of style images I_A and I_B , the interpolated Σ_s should be:

$$\Sigma_s = (1 - \alpha)\Sigma_A + \alpha\Sigma_B \quad (7)$$

where α is the style ratio between the two. Figure 1 and 2 present the smooth transformation from one style image to another.

D. Comparison with ArtFlow

Figure 3 and 4 show the comparisons with ArtFlow.

E. Photorealistic Image Style Transfer

Figure 5 and 6 show the comparisons with advanced photorealistic image style transfer methods.

F. Video Style Transfer

Figure 7, 8, 9 and 10 show the comparisons with advanced photorealistic video and artistic video style transfer methods.

G. Ultra-Resolution Photorealistic Style Transfer

Figure 11 and 12 show the ultra-resolution (4K) photorealistic stylization results of CAP-VSTNet.



Figure 1. Style interpolation results of CAP-VSTNet on photorealistic style transfer.

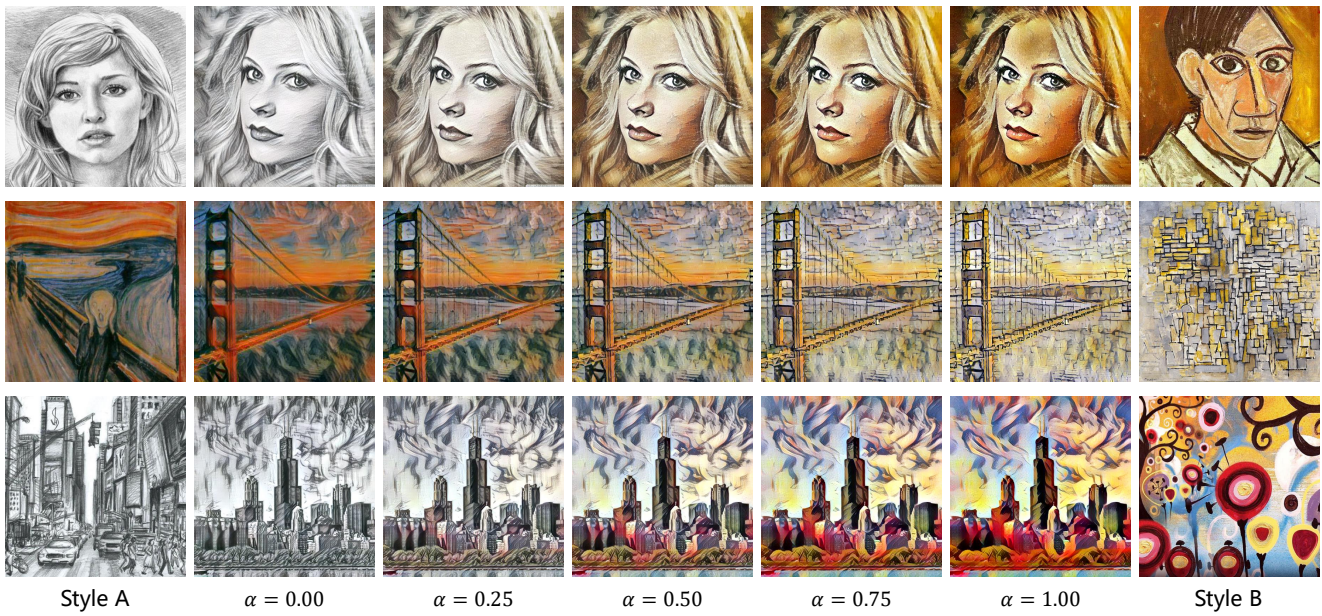


Figure 2. Style interpolation results of CAP-VSTNet on artistic style transfer.

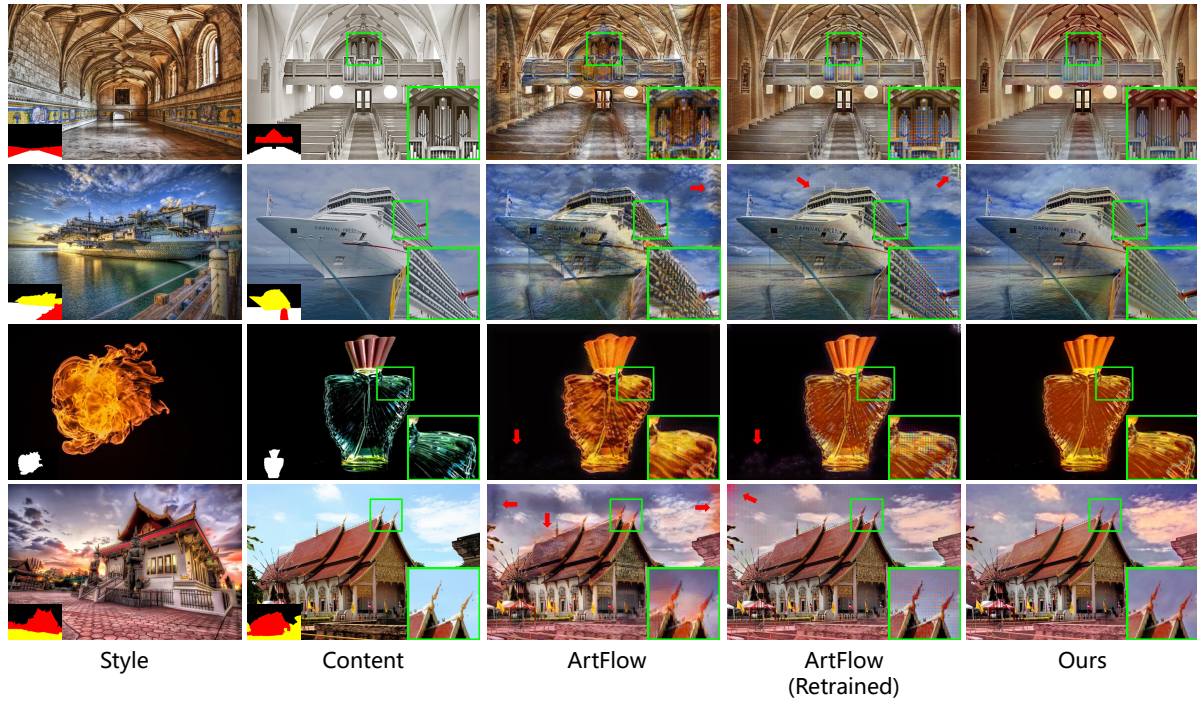


Figure 3. Visual comparisons with ArtFlow on photorealistic style transfer. We retrain ArtFlow on the Microsoft COCO (photorealistic) dataset, in order to avoid the problem of domain gap. We only increase the content loss weight to preserve more content.

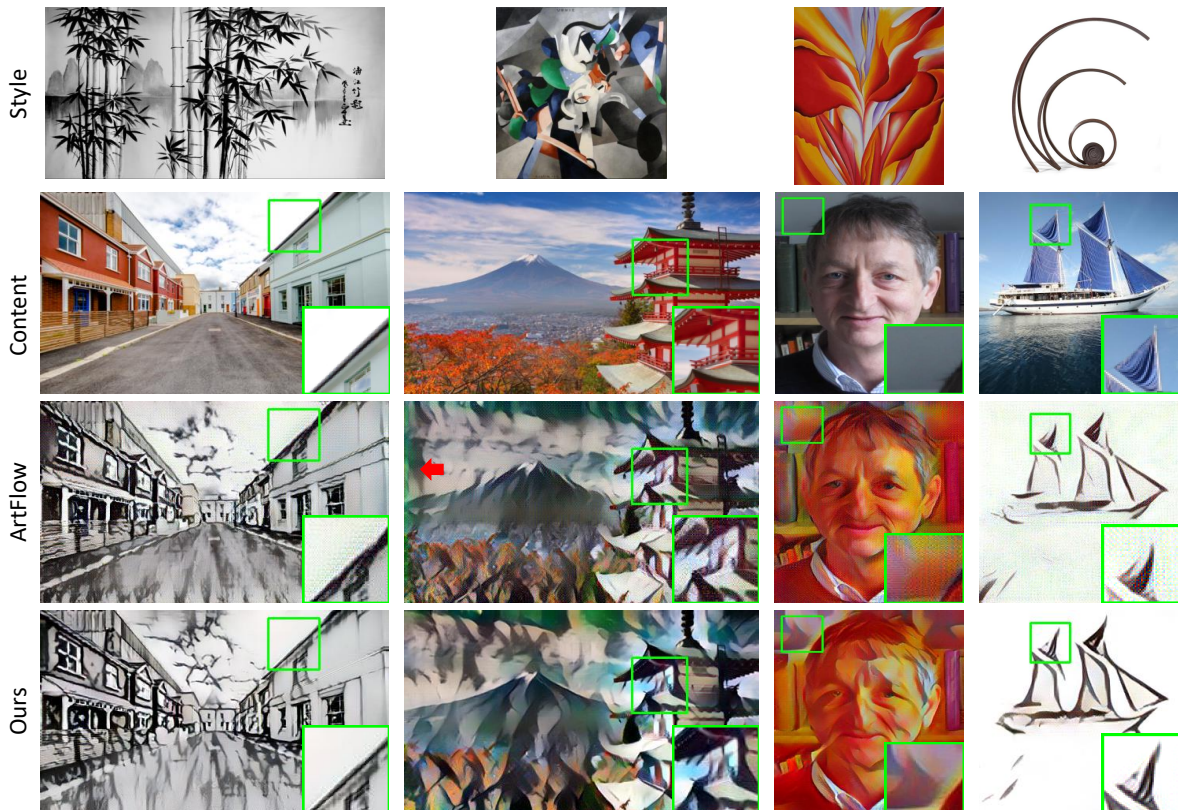


Figure 4. Visual comparisons with ArtFlow on artistic style transfer.

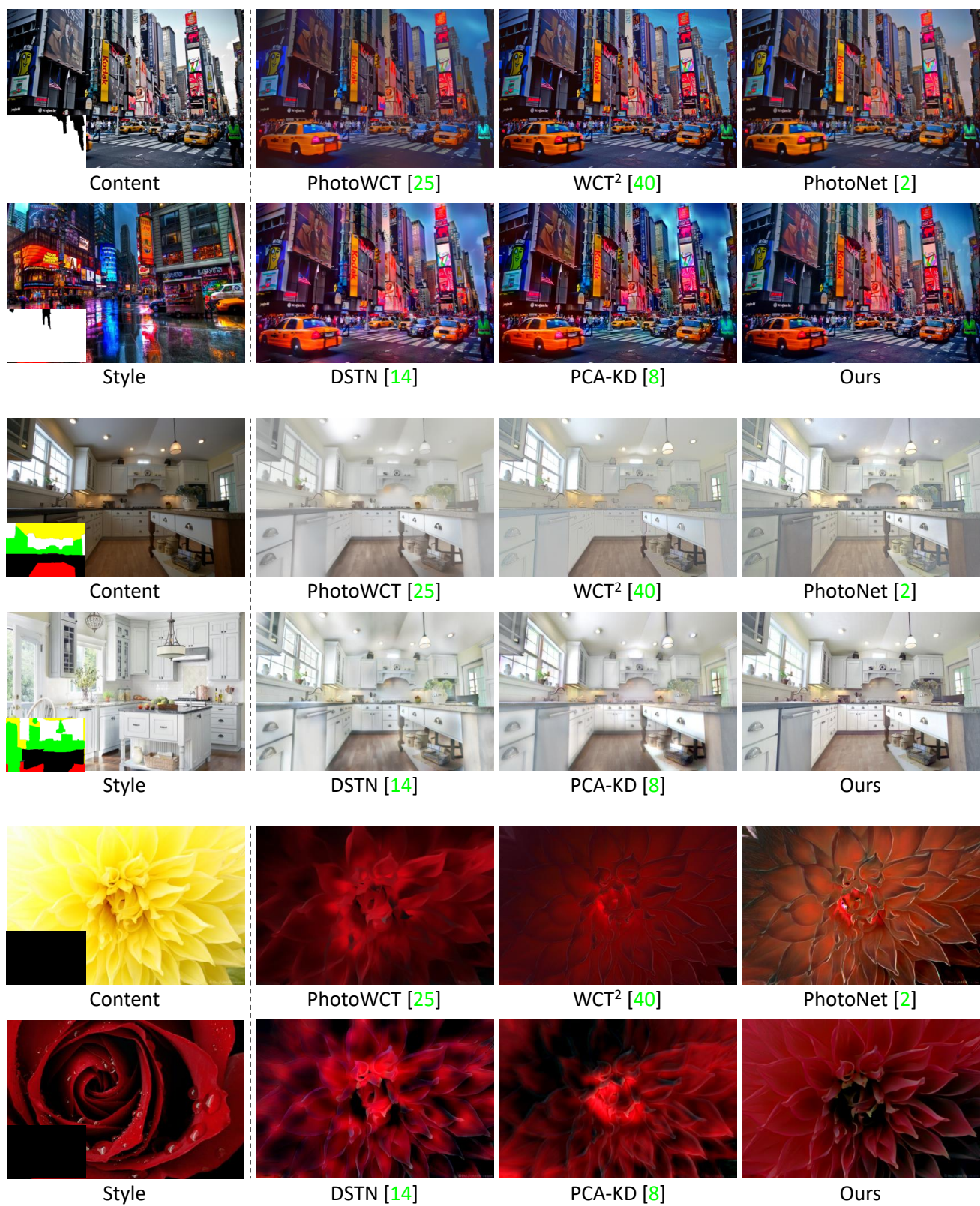


Figure 5. Visual comparisons of photorealistic image style transfer.

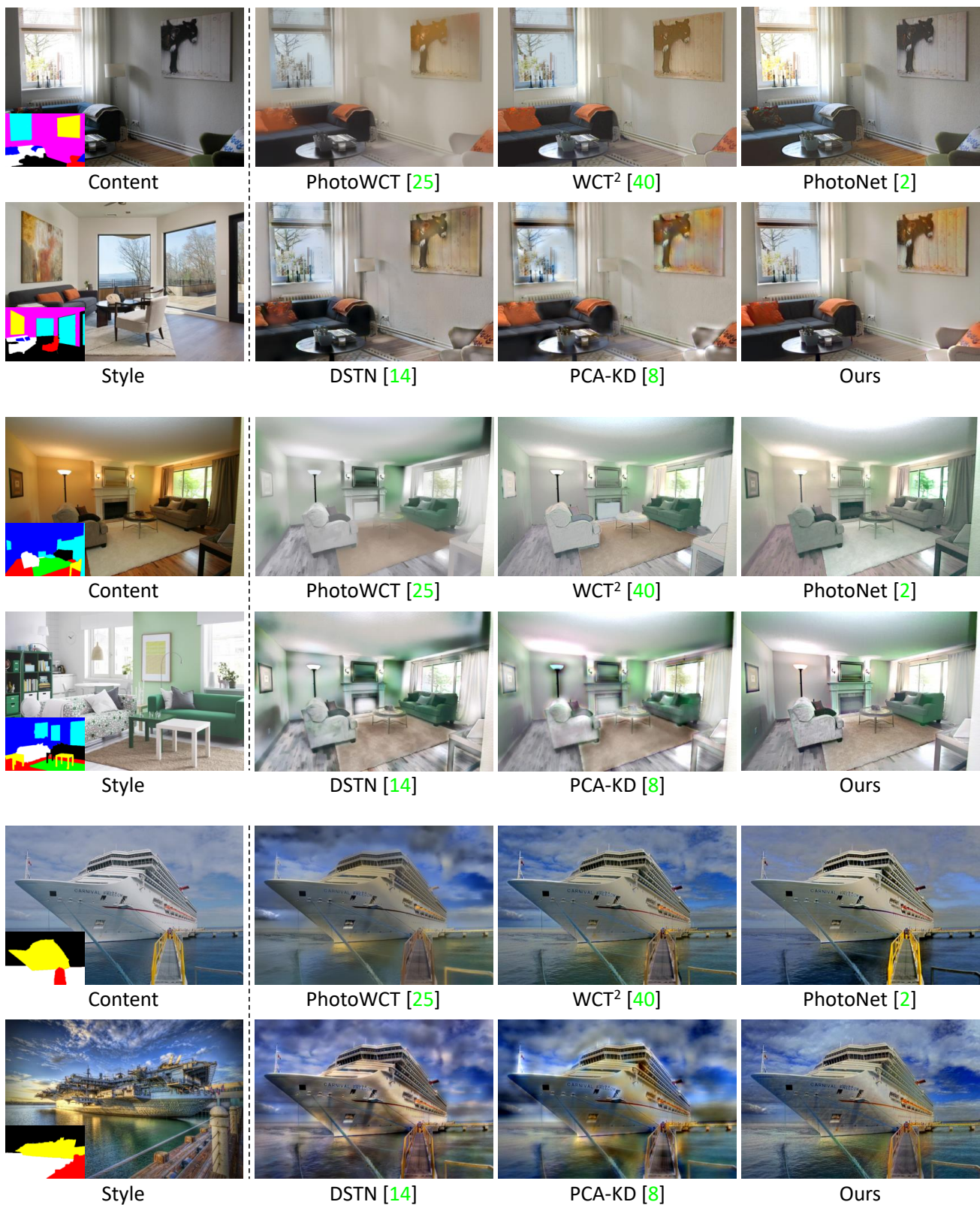


Figure 6. Visual comparisons of photorealistic image style transfer.

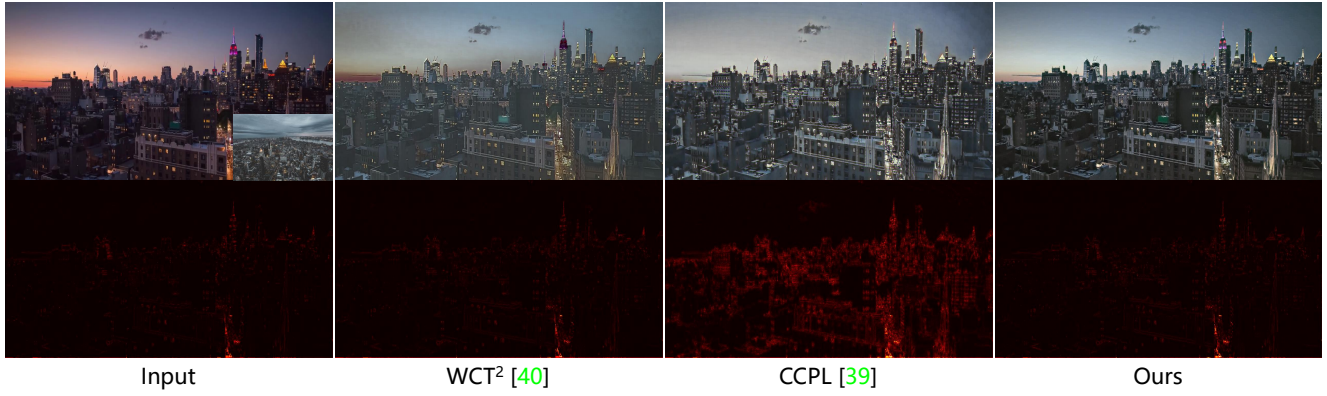


Figure 7. Visual comparison of photorealistic video style transfer. The odd rows show the stylization effect. The even rows show the temporal error heatmap of adjacent frames.

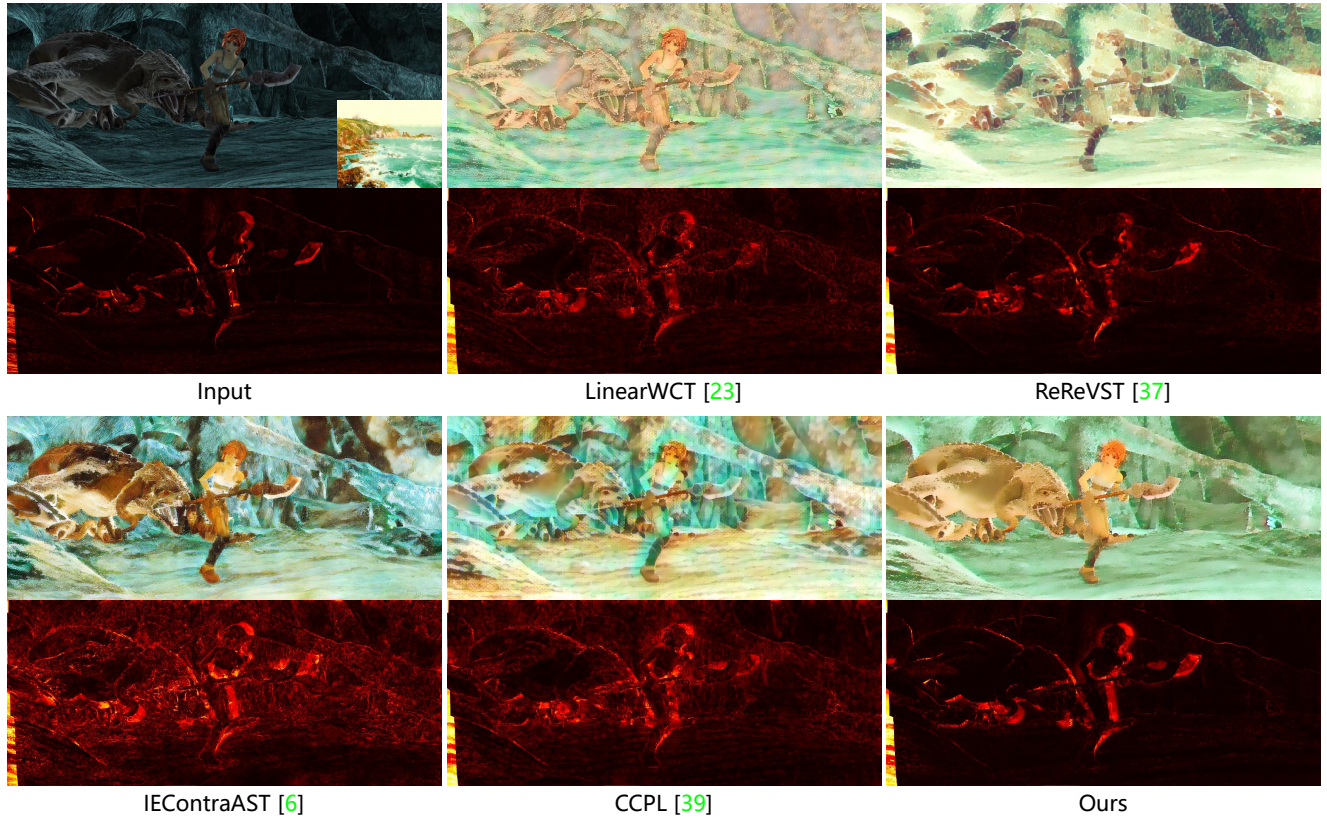


Figure 8. Visual comparison of artistic video style transfer. The odd rows show the stylization effect. The even rows show the temporal error heatmap of adjacent frames.

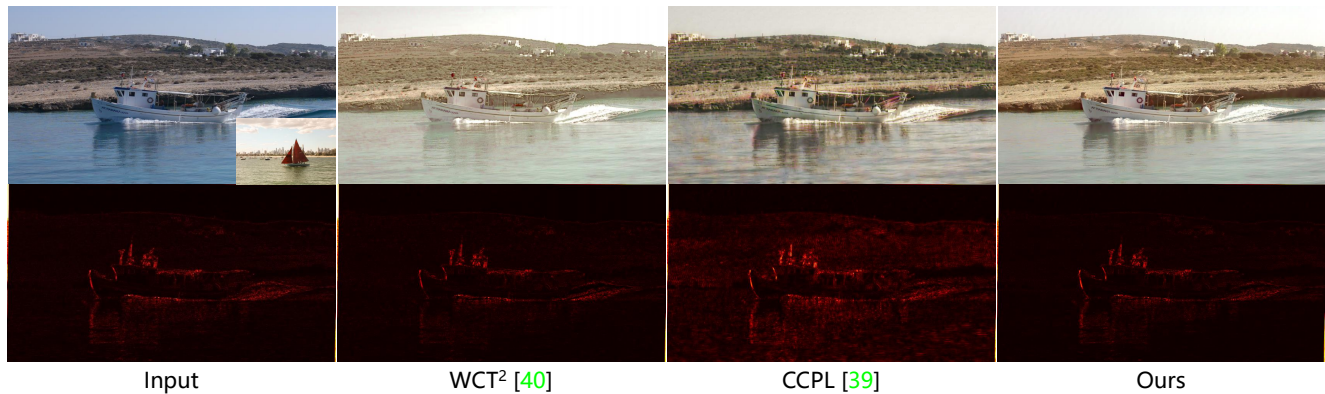


Figure 9. Visual comparison of photorealistic video style transfer. The odd rows show the stylization effect. The even rows show the temporal error heatmap of adjacent frames.

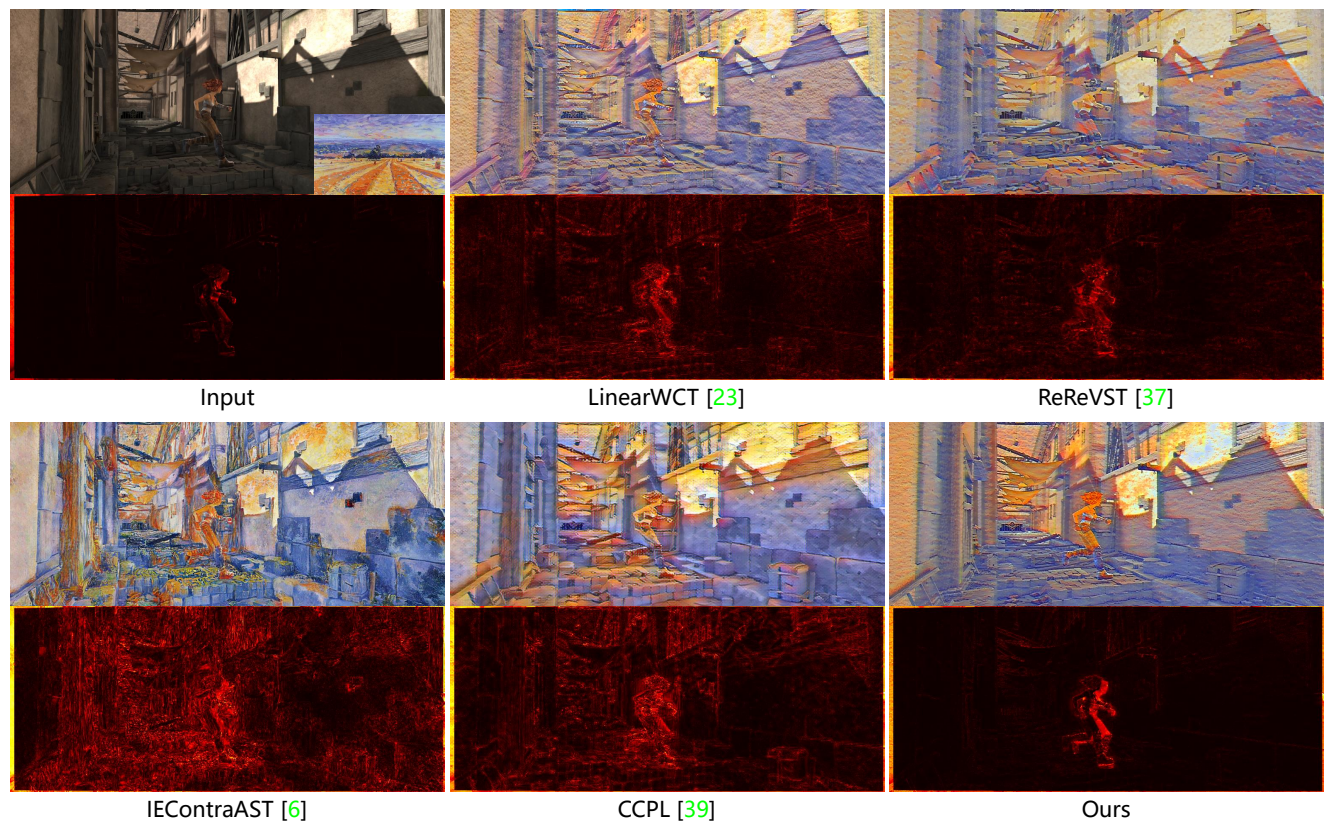


Figure 10. Visual comparison of artistic video style transfer. The odd rows show the stylization effect. The even rows show the temporal error heatmap of adjacent frames.

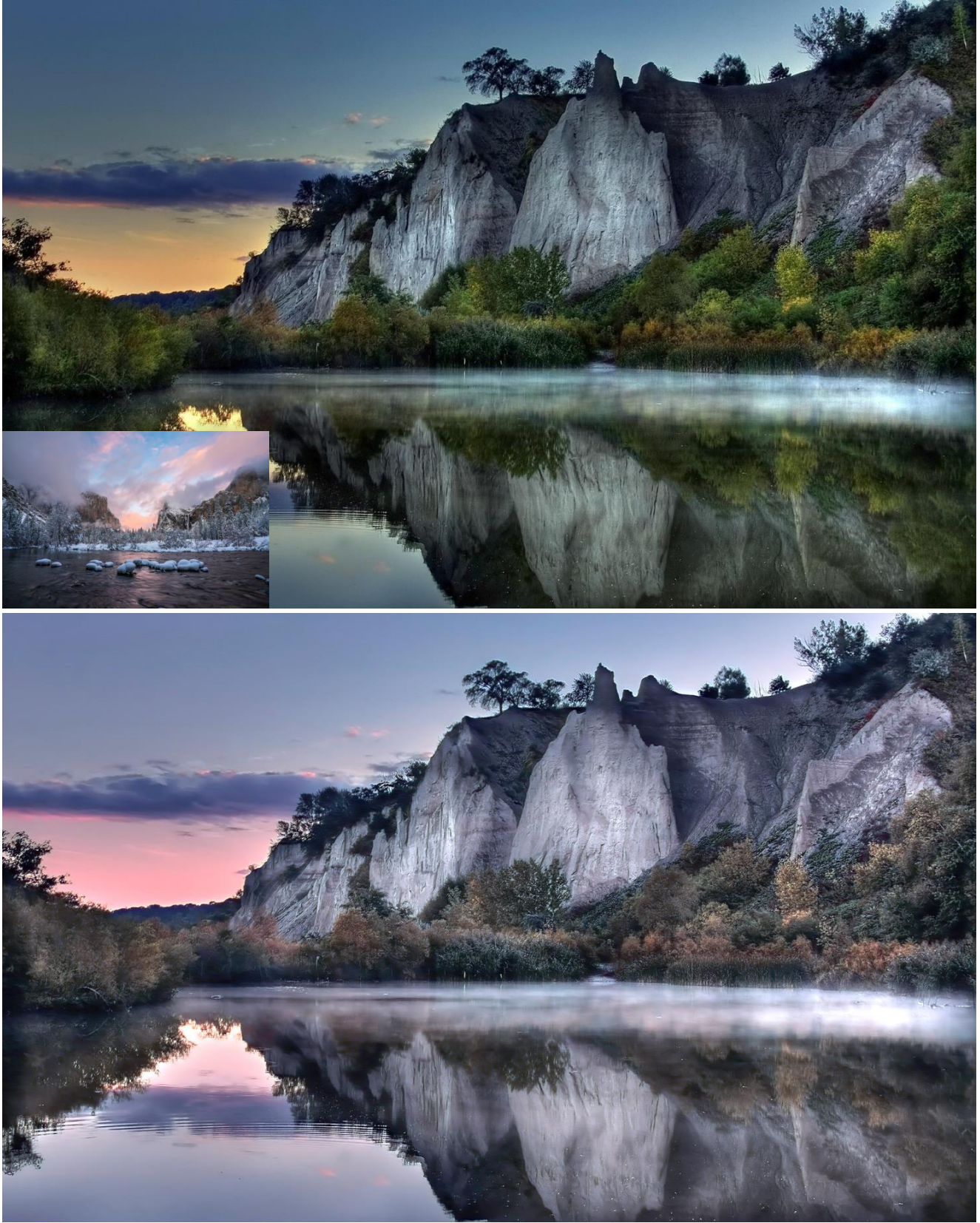


Figure 11. An example of ultra-resolution (4K) photorealistic style transfer generated by CAP-VSTNet.



Figure 12. An example of ultra-resolution (4K) photorealistic style transfer generated by CAP-VSTNet.