Manifold Alignment for Semantically Aligned Style Transfer Supplementary Material

1. Details of the transformation of the objective function

As the transformation of the objective function $J(P)$ in Eq. (2) of the main paper into matrix form in Eq. (4) is omitted in the main paper due to limited space, details of the transformation are thus given here as follows. All the notations are the same as the main paper. Firstly, as $F_{cs} = P^T F_c$, we can replace $\phi_i(F_{cs})$ with $P^T\phi_i(F_c)$, which leads to:

$$\begin{aligned} J(P) =& \frac{1}{N}\sum_{i=1}^{W_c\times H_c}\sum_{j=1}^{W_s\times H_s} A_{ij}^{cs}\|\phi_i(F_{cs}) - \phi_j(F_s)\|_2^2 \\ =& \frac{1}{N}\sum_{i=1}^{W_c\times H_c}\sum_{j=1}^{W_s\times H_s} A_{ij}^{cs}\|P^T\phi_i(F_c) - \phi_j(F_s)\|_2^2. \end{aligned} \tag{1}$$

Then by expanding the term $\|P^T\phi_i(F_c) - \phi_j(F_s)\|_2^2$, we can obtain to following equation:

$$\begin{aligned} J(P) =& \frac{1}{N}\sum_{i=1}^{W_c\times H_c}\sum_{j=1}^{W_s\times H_s} A_{ij}^{cs}\text{tr}(P^T\phi_i(F_c)\phi_i(F_c)^T P \\ & - P^T\phi_i(F_c)\phi_j(F_s)^T - \phi_j(F_s)\phi_i(F_c)^T P \\ & + \phi_j(F_s)\phi_j(F_s)^T), \end{aligned} \tag{2}$$

where tr() denotes the trace operation. With the property of trace that $\text{tr}(A) = \text{tr}(A^T)$, the equation can be further transformed into:

$$\begin{aligned} J(P) =& \frac{1}{N}\sum_{i=1}^{W_c\times H_c}\sum_{j=1}^{W_s\times H_s} A_{ij}^{cs}\text{tr}(P^T\phi_i(F_c)\phi_i(F_c)^T P \\ & - 2P^T\phi_i(F_c)\phi_j(F_s)^T + \phi_j(F_s)\phi_j(F_s)^T). \end{aligned} \tag{3}$$

By further writing the above equation into matrix form and defining $U_{cs} = \frac{1}{N}A^{cs}$, the second term can be transformed as:

$$\begin{aligned} & \frac{1}{N}\sum_{i=1}^{W_c\times H_c}\sum_{j=1}^{W_s\times H_s} A_{ij}^{cs}\text{tr}(-2P^T\phi_i(F_c)\phi_j(F_s)^T) \\ & = \text{tr}(-2P^T F_c U_{cs} F_s^T). \end{aligned} \tag{4}$$

Besides, define $D_c \in \mathbb{R}^{(W_c\times H_c)\times(W_c\times H_c)}$ a diagonal matrix with its diagonal element as $D_c(i,i) = \sum_{j=1}^{(W_s\times H_s)} U_{cs}(i,j)$. Similarly, define $D_s \in \mathbb{R}^{(W_s\times H_s)\times(W_s\times H_s)}$, which is also a diagonal matrix and $D_s(j,j) = \sum_{i=1}^{(W_c\times H_c)} U_{cs}(i,j)$. The the first term can be transformed into:

$$\begin{aligned} & \frac{1}{N}\sum_{i=1}^{W_c\times H_c}\sum_{j=1}^{W_s\times H_s} A_{ij}^{cs}\text{tr}(P^T\phi_i(F_c)\phi_i(F_c)^T P) \\ & = \sum_{i=1}^{W_c\times H_c} D_c(i,i)\text{tr}(P^T\phi_i(F_c)\phi_i(F_c)^T P) \\ & = \text{tr}(P^T F_c D_c F_c^T P). \end{aligned} \tag{5}$$

Similarly, the third term can be transformed into:

$$\begin{aligned} & \frac{1}{N}\sum_{i=1}^{W_c\times H_c}\sum_{j=1}^{W_s\times H_s} A_{ij}^{cs}\text{tr}(\phi_j(F_s)\phi_j(F_s)^T) \\ & = \sum_{j=1}^{W_s\times H_s} D_s(j,j)\text{tr}(\phi_j(F_s)\phi_j(F_s)^T) \\ & = \text{tr}(F_s D_s F_s^T). \end{aligned} \tag{6}$$

Combining the above results can conclude that:

$$J(P) = \text{tr}(P^T F_c D_c F_c^T P + F_s D_s F_s^T - 2P^T F_c U_{cs} F_s^T). \tag{7}$$

2. Bidirectional Transfer

The proposed orthogonal constrained manifold alignment method can not only align content features to style features' subspace, but is also feasible to project style features to content features' subspace. This is because with the orthogonal constraint, $F_{cs} = P^T F_c$ and $F_c = PF_{cs}$. This means P^T can be used to project content features to style features' subspace and P can be used to project the F_{cs} back to the content features' subspace. Therefore, P should also be able to project style features to content features' subspace, leading to $F_{sc} = PF_s$. We show some bidirectional transfer results in Figure 1. The 3rd column shows results of using the images in the 1st column as content and images in the 2nd column as style. The 4th column shows results of reverse transfer. As can be seen, good transfer results

are obtained for both directions. This shows the orthogonal constrained manifold alignment method is feasible for bidirectional style transfer.

Figure 1. Results of bidirectional style transfer.

3. Visualization of feature distributions

In order to clearly verify the effectiveness of the proposed manifold alignment method in aligning content features' distribution and style features' distribution, we use t-SNE [2] to visualize content features, style features, and the aligned features, as shown in Figure 2. The original high-dimensional features are extracted at layer Conv_4_1 in VGG-19 [1] with input images of 128×128 and are projected to a 2-dimensional subspace via t-SNE. It can be seen from the fourth column of Figure 2 that there is a large distribution gap between the content features and the style features. From the fifth column, the difference between the aligned feature distribution and the style distribution is significantly reduced. This shows that our proposed manifold alignment method can effectively align two distributions for effective style transfer.

4. Influence of k nearest neighbors

In the experimental part of the main paper, we have discussed the setting of parameter k. In this supplementary document, we further provide visual results to show the influence of k nearest neighbors defined in Eq. (8). We have set the parameter of k to $1, 5, 50, 100, 500, 1000$. Results are given in Figure 3. As can be seen, with the increase of k, the results tend to be smooth or blurry. This is because, in the objective function Eq. (9), larger k will force more features, and in turn more pixels, at different spatial locations to become similar, leading to blurry results. However, no significant influence is observed when setting k to a small number around 5. Therefore, we set k as 5 for all the experiments.

$$A_{ij}^{cs} = \begin{cases} 1 & \phi_i(F_c) \in \mathcal{N}_k(\phi_j(F_s)) \text{ or } \phi_j(F_s) \in \mathcal{N}_k(\phi_i(F_c)) \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

$$\min_P J(P) = \frac{1}{N} \sum_{i=1}^{W_c \times H_c} \sum_{j=1}^{W_s \times H_s} A_{ij}^{cs} \|\phi_i(F_{cs}) - \phi_j(F_s)\|_2^2 \tag{9}$$

5. Influence of stylization weight

We have also provided the results of varying different stylization weight α, which are given in Figure 4. α is used to control the contribution of the original content feature and the transformed content feature $F_{cs} = (1-\alpha)F_c + \alpha F_{cs}$. As can be seen, with the increase of stylization weight, the stylization level becomes stronger. However, the content structure may be corrupted. Therefore, the stylization level is set to 0.6 to balance between the content preservation and stylization level.

6. Additional user controlled style transfer results

In Figure 5, we show more user drawn editing results. Comparing the fifth and sixth columns of Figure 5, we can see that by drawing in the corresponding areas on content and style images, the style transfer results have more desired appearance as specified by the user, *e.g.*, the style of sky in the first, second and third rows becomes more like the style defined by users on the corresponding style image. From Figure 6, which provides more segmentation guided style transfer results, results using segmentation maps as guidance are more semantically aligned compared with the original results. By introducing the segmentation map, the style of the ship and the flower in the second and third rows becomes more like the style of the corresponding objects in the style images.

7. Solution to style transfer with high resolution input

To address the problem of style transfer with high resolution input, we can add sampling or pooling operations on the feature maps to reduce their spatial sizes. In Figure 7, we show results of high resolution style transfer of our method using average pooling on feature maps of R41

Figure 2. Visualization of feature distributions. In the first and second columns are the content images and style images respectively. The third column shows our stylized results. The last two columns shows the distribution of the original content and style features and the distribution of the style and the aligned features using t-SNE, where 'C', 'S' and 'CS' denote content, style and aligned features respectively. From the last two columns, the proposed method successfully aligns the content and style features.

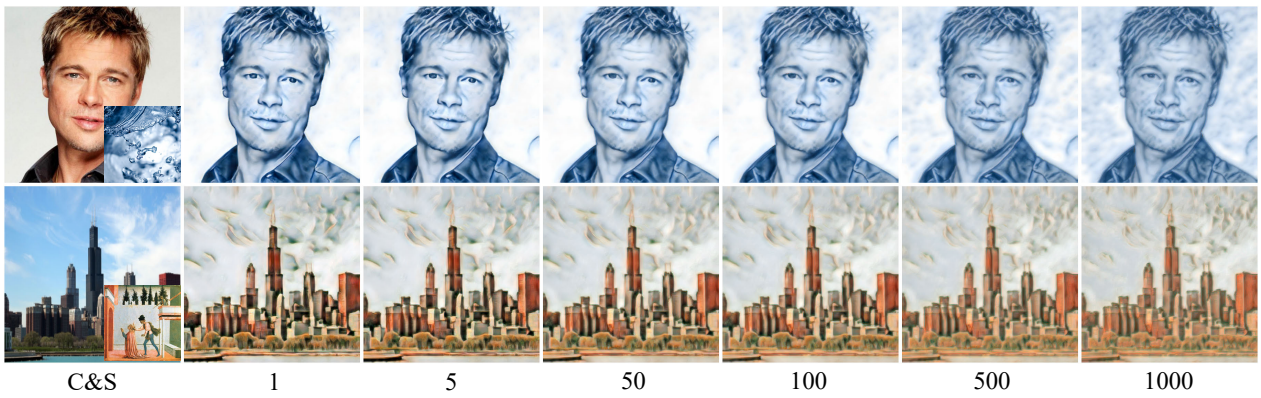

Figure 3. The influence of k nearest neighbors. With the increase of k, the results tend to become blurry.

(from 480×270 to 240×135) and R31 (from 960×540 to 240×135), under the multi-level style transfer framework.

8. Visualization of matched points

To show the proposed method can correctly find semantically aligned features, we have visualized the correspondence of content and style images in Figure 8. As can be seen, most corresponding points are correctly matched using the VGG encoder and normalized similarity metric.

9. Results under other style transfer frameworks

To show the proposed manifold alignment module can be plugged into other style transfer frameworks. We adopted the distilled auto-encoder style transfer framework of Wang *et al*. [3] as an example. The corresponding results are given in Figure 9. As seen, the proposed manifold alignment module can successfully stylize images under this new framework.

10. Results when the content and style images are semantically aligned

In Figure 10, we present examples of content and style images sharing the same semantic regions (including face painting to photo face). Compared with WCT and AdaIN, the proposed manifold alignment method can successfully find semantically aligned regions and transfer styles among these regions. The sky, face cheek and hair look are more like those of style images. Structure preservation is also better.

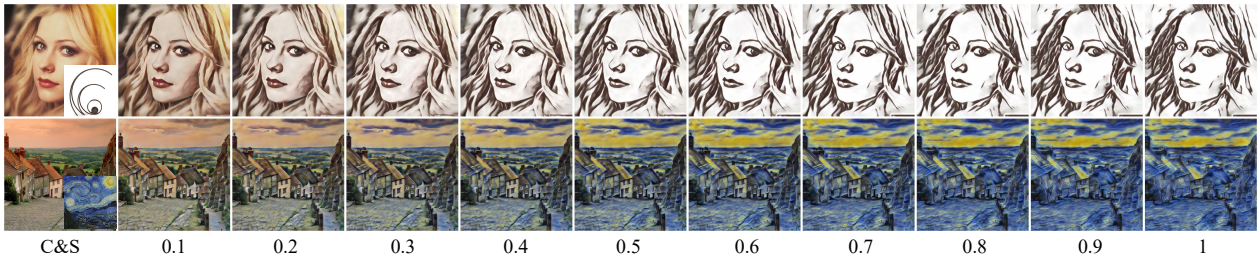

Figure 4. The influence of stylization weight α.

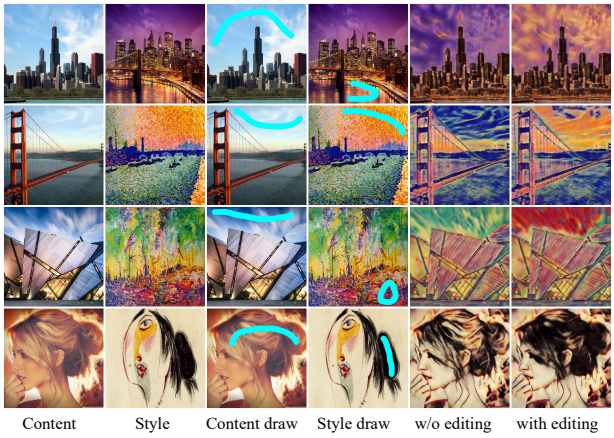

Figure 5. More user drawn editing results.

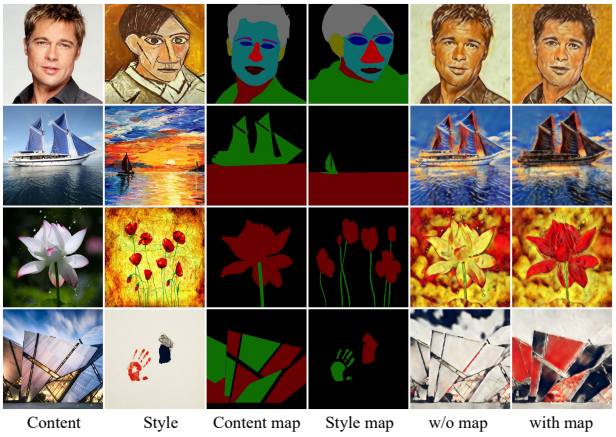

Figure 6. More style transfer results with segmentation mask as guidance.

11. More visual comparison of artistic and photorealistic style transfer

In addition to the visual results provided in the main paper, we provide more visual comparison results of artistic and photorealistic style transfer. In Figure 11, more comparison of artistic style transfer results is given. As can be seen, the proposed method is very good at preserving the semantic structure of the content image. In the meanwhile, the produced images highly resemble the input style images. In Figure 12, more comparison of photorealistic style transfer results is given. From the results, the proposed method is very competitive compared with existing photorealistic style transfer methods.

12. Discussions about limitations

Regarding limitations, one is that the unsupervised matching procedure of the proposed method may cause mismatched nearest neighbors. For example, the third results in Figure 6, where the result without guidance fails to match the flowers in the content and style images. However, this can be improved by training the encoder to encode semantically related features and by incorporating more effective similarity metrics. This is left to our future work.

References

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 2

[2] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 2

[3] Huan Wang, Yijun Li, Yuehai Wang, Haoji Hu, and Ming-Hsuan Yang. Collaborative distillation for ultra-resolution universal style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1860–1869, 2020. 3, 5

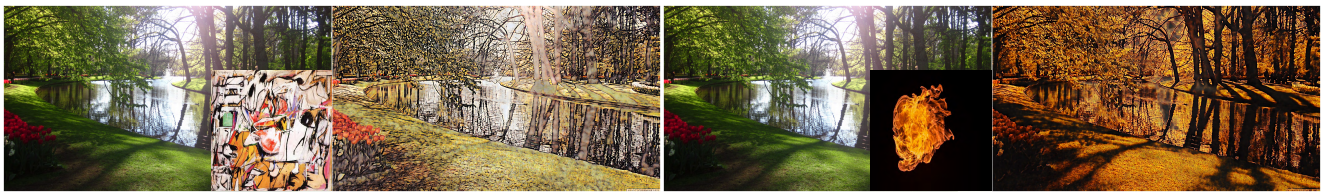

Figure 7. Example of high resolution style transfer. The content image is of size 3840×2160.

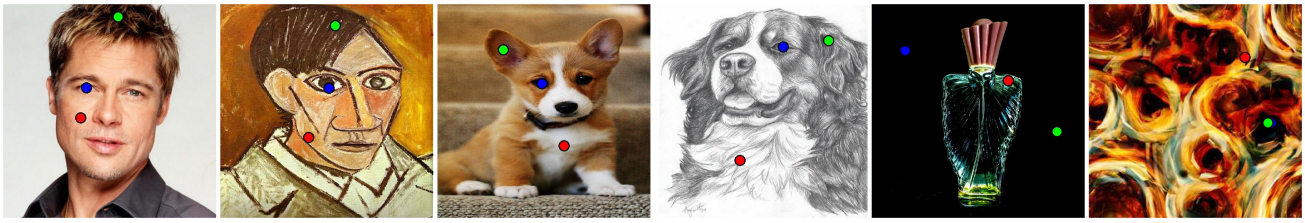

Figure 8. Matching points of three pairs of images.

Figure 9. Results of using the distilled style transfer framework [3]. In the first row are content and style images. The second row shows corresponding stylized images.

Figure 10. Results of style transfer when content and style images having same semantic regions.

Figure 11. More visual comparison of artistic style transfer results. Best view in large size.

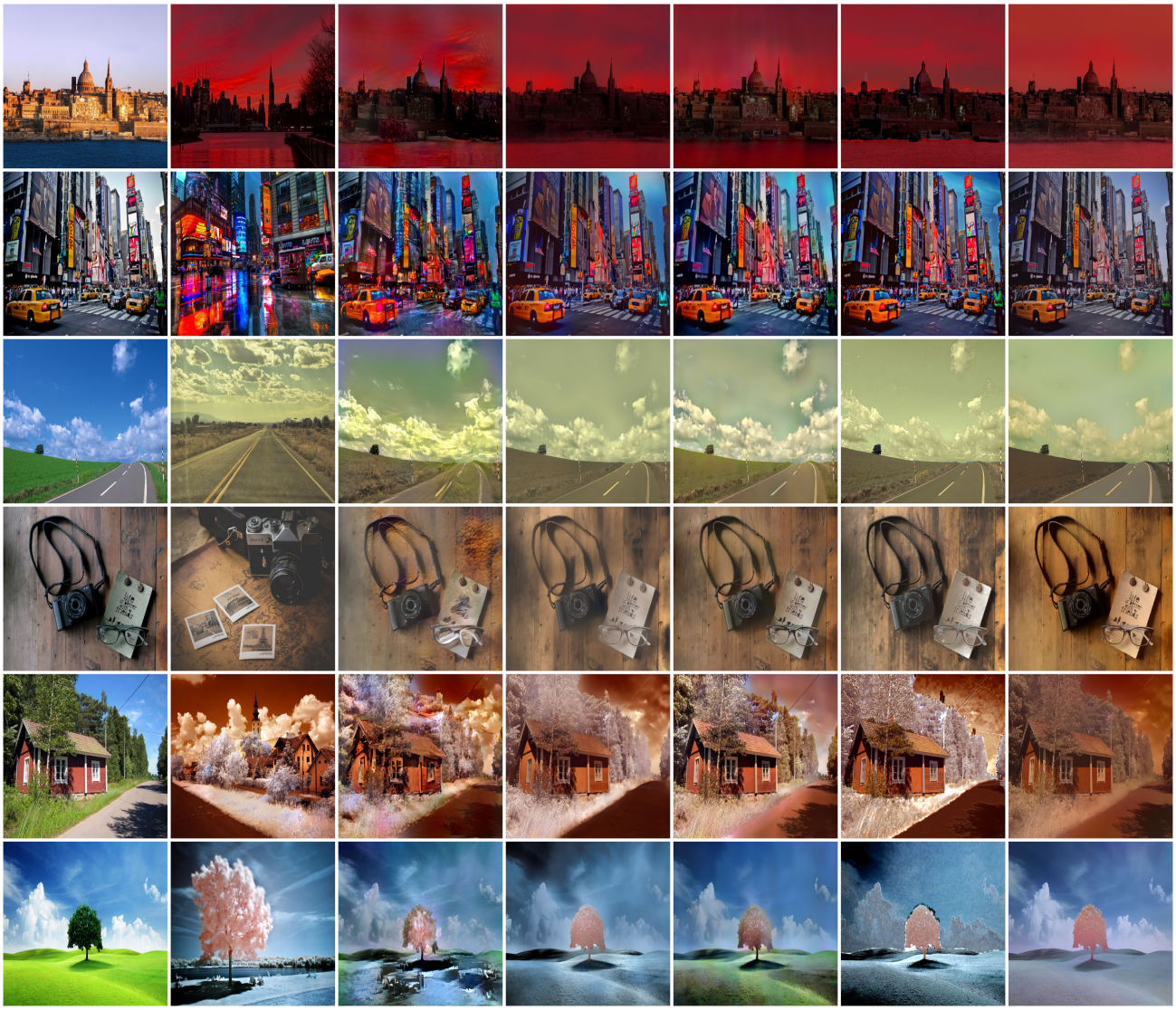

Figure 12. More visual comparison of photorealistic style transfer results. Best view in large size.