Iterative Feature Transformation for Fast and Versatile Universal Style Transfer

Tai-Yin Chiu and Danna Gurari

University of Texas at Austin

This document supplements Sections 3 and 4 of the main paper. In particular, it includes the following:

- Derivation of the analytical gradient (supplements **Section 3.2**).
- Training details of the autoencoders (supplements **Section 4**).
- Stylized results for quantitative analysis of photo-realistic transfer (supplements **Section 4.2**).
- Formulation of NST and WCT for multi-style transfer and double-style transfer results from AdaIN and Avatar-net (supplements **Section 4.3**).

1 Derivation of the analytical gradient

For simplicity, we suppress the subscript N. Here we show that if

$$l_j(\mathbf{F}) = ||\mathbf{F} - \mathbf{F}^{(j)}||_F^2 + \lambda||\frac{1}{n}\mathbf{F}\mathbf{F}^{\mathrm{T}} - \frac{1}{m}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}}||_F^2, \tag{1}$$

then

$$\frac{\mathrm{d}l}{\mathrm{d}\mathbf{F}} = 2(\mathbf{F} - \mathbf{F}^{(j)}) + \frac{4\lambda}{n}(\frac{1}{n}\mathbf{F}\mathbf{F}^{\mathrm{T}} - \frac{1}{m}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}})\mathbf{F}. \tag{2}$$

Proof.

$$\begin{aligned}
&||\mathbf{F} - \mathbf{F}^{(j)}||_F^2 && (3)\\
=&\mathrm{tr}[(\mathbf{F} - \mathbf{F}^{(j)})(\mathbf{F} - \mathbf{F}^{(j)})^{\mathrm{T}}] && (4)\\
=&\mathrm{tr}[\mathbf{F}\mathbf{F}^{\mathrm{T}} - 2\mathbf{F}(\mathbf{F}^{(j)})^{\mathrm{T}} + \mathbf{F}^{(j)}(\mathbf{F}^{(j)})^{\mathrm{T}}], && (5)
\end{aligned}$$

and

$$\begin{aligned}
&||\frac{1}{n}\mathbf{F}\mathbf{F}^{\mathrm{T}} - \frac{1}{m}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}}||_F^2 && (6)\\
=&\mathrm{tr}[(\frac{1}{n}\mathbf{F}\mathbf{F}^{\mathrm{T}} - \frac{1}{m}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}})(\frac{1}{n}\mathbf{F}\mathbf{F}^{\mathrm{T}} - \frac{1}{m}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}})] && (7)\\
=&\mathrm{tr}[\frac{1}{n^2}\mathbf{F}\mathbf{F}^{\mathrm{T}}\mathbf{F}\mathbf{F}^{\mathrm{T}} - \frac{2}{nm}\mathbf{F}\mathbf{F}^{\mathrm{T}}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}} + \frac{1}{m^2}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}}]. && (8)
\end{aligned}$$

Let $\mathbf{F} = [f_1, f_2, \ldots, f_n]$, $\mathbf{F}^{(j)} = [f_1^{(j)}, f_2^{(j)}, \ldots, f_n^{(j)}]$, and $\mathbf{F}_s = [f_1^s, f_2^s, \ldots, f_m^s]$. We first find the partial derivatives with respect to f_i:

$$\frac{\partial\,||\mathbf{F} - \mathbf{F}^{(j)}||_F^2}{\partial\, f_i} = \frac{\partial\,\mathrm{tr}[\mathbf{F}\mathbf{F}^{\mathrm{T}}]}{\partial\, f_i} - 2\frac{\partial\,\mathrm{tr}[\mathbf{F}(\mathbf{F}^{(j)})^{\mathrm{T}}]}{\partial\, f_i}, \tag{9}$$

$$\frac{\partial\,||\frac{1}{n}\mathbf{F}\mathbf{F}^{\mathrm{T}}-\frac{1}{m}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}}||_F^2}{\partial\, f_i}=\frac{1}{n^2}\frac{\partial\,\mathrm{tr}[\mathbf{F}\mathbf{F}^{\mathrm{T}}\mathbf{F}\mathbf{F}^{\mathrm{T}}]}{\partial\, f_i}-\frac{2}{nm}\frac{\partial\,\mathrm{tr}[\mathbf{F}\mathbf{F}^{\mathrm{T}}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}}]}{\partial\, f_i}, \tag{10}$$

where $\mathrm{tr}[\mathbf{F}\mathbf{F}^{\mathrm{T}}]=\sum_{a=1}^{n} f_a^{\mathrm{T}} f_a$, $\mathrm{tr}[\mathbf{F}(\mathbf{F}^{j})^{\mathrm{T}}]=\sum_{a=1}^{n} f_a^{\mathrm{T}} f_a^{(j)}$,

$$\begin{aligned}
&\mathrm{tr}[\mathbf{F}\mathbf{F}^{\mathrm{T}}\mathbf{F}\mathbf{F}^{\mathrm{T}}] &(11)\\
=&\mathrm{tr}[\sum_{a=1}^{n} f_a f_a^{\mathrm{T}}\sum_{b=1}^{n} f_b f_b^{\mathrm{T}}] &(12)\\
=&\sum_{a=1}^{n}\sum_{b=1}^{n}\mathrm{tr}[f_a f_a^{\mathrm{T}} f_b f_b^{\mathrm{T}}] &(13)\\
=&\sum_{a=1}^{n}\sum_{b=1}^{n}(f_a^{\mathrm{T}} f_b)^2, &(14)
\end{aligned}$$

and similar to $\mathrm{tr}[\mathbf{F}\mathbf{F}^{\mathrm{T}}\mathbf{F}\mathbf{F}^{\mathrm{T}}]$, we have $\mathrm{tr}[\mathbf{F}\mathbf{F}^{\mathrm{T}}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}}]=\sum_{a=1}^{n}\sum_{b=1}^{m}(f_a^{\mathrm{T}} f_b^s)^2$.

For the partial derivatives with respect to f_i, we only have to focus on the terms associated with f_i. Therefore,

$$\frac{\partial\,\mathrm{tr}[\mathbf{F}\mathbf{F}^{\mathrm{T}}]}{\partial\, f_i}=\frac{\partial\, f_i^{\mathrm{T}} f_i}{\partial\, f_i}=2f_i, \tag{15}$$

$$\frac{\partial\,\mathrm{tr}[\mathbf{F}(\mathbf{F}^{(j)})^{\mathrm{T}}]}{\partial\, f_i}=\frac{\partial\, f_i^{\mathrm{T}} f_i^{(j)}}{\partial\, f_i}=f_i^{(j)}, \tag{16}$$

$$\begin{aligned}
&\frac{\partial\,\mathrm{tr}[\mathbf{F}\mathbf{F}^{\mathrm{T}}\mathbf{F}\mathbf{F}^{\mathrm{T}}]}{\partial\, f_i} &(17)\\
=&\frac{\partial}{\partial\, f_i}\Big(\sum_{a\neq i}(f_a^{\mathrm{T}} f_i)^2+\sum_{b\neq i}(f_i^{\mathrm{T}} f_b)^2+(f_i^{\mathrm{T}} f_i)^2\Big) &(18)\\
=&2\sum_{a\neq i}(f_a^{\mathrm{T}} f_i)f_a+2\sum_{b\neq i}(f_i^{\mathrm{T}} f_b)f_b+4(f_i^{\mathrm{T}} f_i)f_i &(19)\\
=&4\sum_{a\neq i}(f_a^{\mathrm{T}} f_i)f_a+4(f_i^{\mathrm{T}} f_i)f_i &(20)\\
=&4\sum_{a=1}^{n}(f_a^{\mathrm{T}} f_i)f_a &(21)\\
=&4\mathbf{F}\mathbf{F}^{\mathrm{T}} f_i, &(22)
\end{aligned}$$

and

$$\frac{\partial\,\mathrm{tr}[\mathbf{F}\mathbf{F}^{\mathrm{T}}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}}]}{\partial\, f_i}=\frac{\partial}{\partial\, f_i}\sum_{b=1}^{m}(f_i^{\mathrm{T}} f_b^s)^2=2\sum_{b=1}^{m}(f_i^{\mathrm{T}} f_b^s)f_b^s=2\sum_{b=1}^{n}((f_b^s)^{\mathrm{T}} f_i)f_b^s=2\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}} f_i. \tag{23}$$

Putting everything together, we have

$$\frac{\partial\, l_j(\mathbf{F})}{\partial\, f_i} = 2f_i - 2f_i^{(j)} + \lambda(4\frac{1}{n^2}\mathbf{F}\mathbf{F}^{\mathrm{T}} f_i - 4\frac{1}{nm}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}} f_i) \tag{24}$$

$$= 2(f_i - f_i^{(j)}) + \frac{4\lambda}{n}(\frac{1}{n}\mathbf{F}\mathbf{F}^{\mathrm{T}} - \frac{1}{m}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}})f_i. \tag{25}$$

Finally,

$$\frac{\mathrm{d}l_j(\mathbf{F})}{\mathrm{d}\mathbf{F}} = \Big[\frac{\partial\, l_j}{\partial\, f_1}, \frac{\partial\, l_j}{\partial\, f_2}, \ldots, \frac{\partial\, l_j}{\partial\, f_n}\Big] \tag{26}$$

$$= 2(\mathbf{F} - \mathbf{F}^{(j)}) + \frac{4\lambda}{n}(\frac{1}{n}\mathbf{F}\mathbf{F}^{\mathrm{T}} - \frac{1}{m}\mathbf{F}_s\mathbf{F}_s^{\mathrm{T}})\mathbf{F}. \tag{27}$$

2 Multiple-style transfer

Starting from equation 8 in the main paper:

$$\min_{\mathbf{F}_N} ||\mathbf{F}_N - \mathbf{F}_N^{(j)}||_F^2 + \sum_{k=1}^{q} \lambda_N^k || \frac{1}{n_N}\mathbf{F}_N\mathbf{F}_N^{\mathrm{T}} - \frac{1}{m_N^k}\mathbf{F}_{N,s}^k(\mathbf{F}_{N,s}^k)^{\mathrm{T}}||_F^2, \tag{28}$$

since the following equivalence

$$\min_{\mathbf{X}} a||\mathbf{X} - \mathbf{Y}||_F^2 + b||\mathbf{X} - \mathbf{Z}||_F^2 \equiv \min_{\mathbf{X}} (a+b)||\mathbf{X} - \frac{a}{a+b}\mathbf{Y} - \frac{b}{a+b}\mathbf{Z}||_F^2 \tag{29}$$

holds, which can be shown by completing the square and removing the constant parts, we can rewrite equation 28 into an equivalent form with $\lambda_N \triangleq \sum_{k=1}^{q} \lambda_N^k$:

$$\min_{\mathbf{F}_N} ||\mathbf{F}_N - \mathbf{F}_N^{(j)}||_F^2 + \lambda_N || \frac{1}{n_N}\mathbf{F}_N\mathbf{F}_N^{\mathrm{T}} - \frac{1}{\lambda_N}\sum_{k=1}^{q} \frac{\lambda_N^k}{m_N^k}\mathbf{F}_{N,s}^k(\mathbf{F}_{N,s}^k)^{\mathrm{T}}||_F^2. \tag{30}$$

The gradient of objective is then given by

$$\frac{\mathrm{d}l_j}{\mathrm{d}\mathbf{F}_N} = 2(\mathbf{F}_N - \mathbf{F}_N^{(j)}) + \frac{4\lambda_N}{n_N}(\frac{1}{n_N}\mathbf{F}_N\mathbf{F}_N^{\mathrm{T}} - \frac{1}{\lambda_N}\sum_{k=1}^{q} \frac{\lambda_N^k}{m_N^k}\mathbf{F}_{N,s}^k(\mathbf{F}_{N,s}^k)^{\mathrm{T}})\mathbf{F}_N. \tag{31}$$

Note that when $q = 1$, equation 31 reduces to equation 2.

3 Training details of the autoencoders

The four autoencoders are trained by minimizing an image reconstruction loss and a perceptual loss. In particular, if the functions of the *encoder*$_N$ and *decoder*$_N$ are denoted $\phi_N(\cdot)$ and $\psi_N(\cdot)$, respectively, the *decoder*$_N$ is trained by minimizing the loss $\mathcal{L}_{AE}$:

$$\mathcal{L}_{AE} = ||\psi_N(\phi_N(I)) - I||_F^2 + ||\phi_N(\psi_N(\phi_N(I))) - \phi_N(I)||_F^2, \tag{32}$$

where I is an input image. We train the autoencoders on the MS-COCO dataset. To support batch training, each image from the dataset is resized to 512×512 and randomly cropped to 256×256 as a training example in a batch. For the autoencoders associated with *relu4_1* and *relu3_1* layers, they are trained with a batch size of 8 for 5 epochs, while for *relu2_1* and *relu1_1* cases, the autoencoders are trained for 3 epochs, due to their smaller sizes. We use Adam optimizer with the learning rate 1×10^{-4} and without weight decay. Moreover, we use up-sampling layers with bilinear interpolation in the decoders as the symmetric part of the max-pooling layers in the encoders.

4 Stylized results for quantitative analysis of photo-realistic transfer

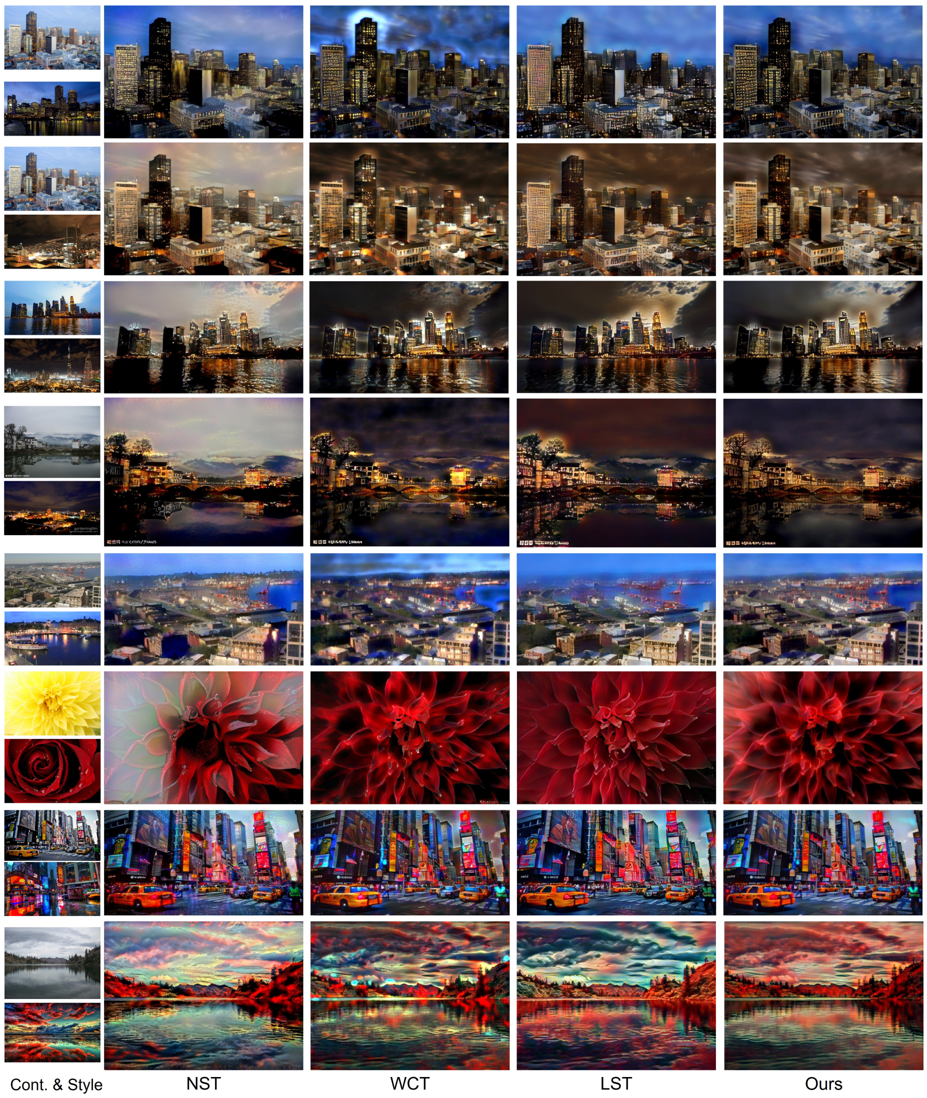

Fig. 1: Photo-realistically stylized images from 30 pairs of a content and a style images for quantitative analysis. No spatial control and no post-processing are applied (Part 1/4).

Table 1: Speed performance of our method under $n_{upd} = 15$ and $n_{iter} = 1$ for generating the results in figures 1, 2, 3, and 4. **Unit**: Second.

	256 × 256	**512 × 512**	**768 × 768**	**1024 × 1024**
time	0.13	0.31	0.62	0.92

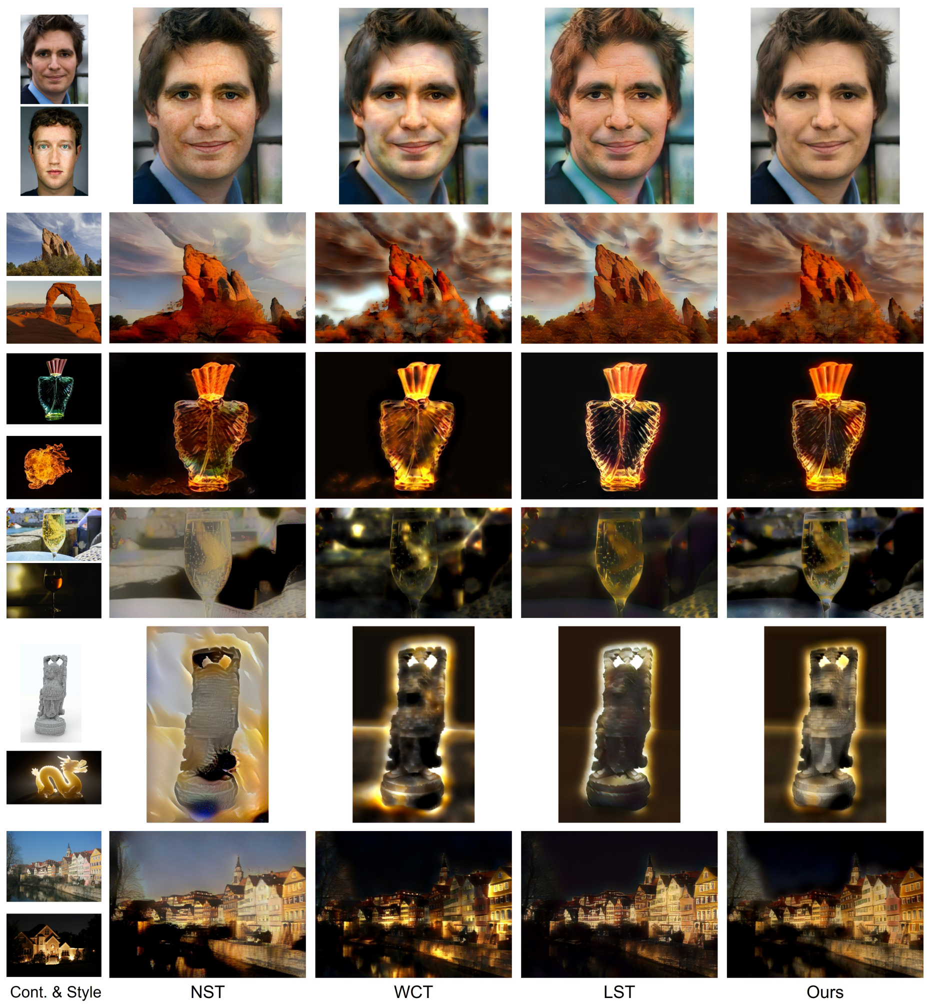

Fig. 2: Photo-realistically stylized images from 30 pairs of a content and a style images for quantitative analysis. No spatial control and no post-processing are applied (Part 2/4).

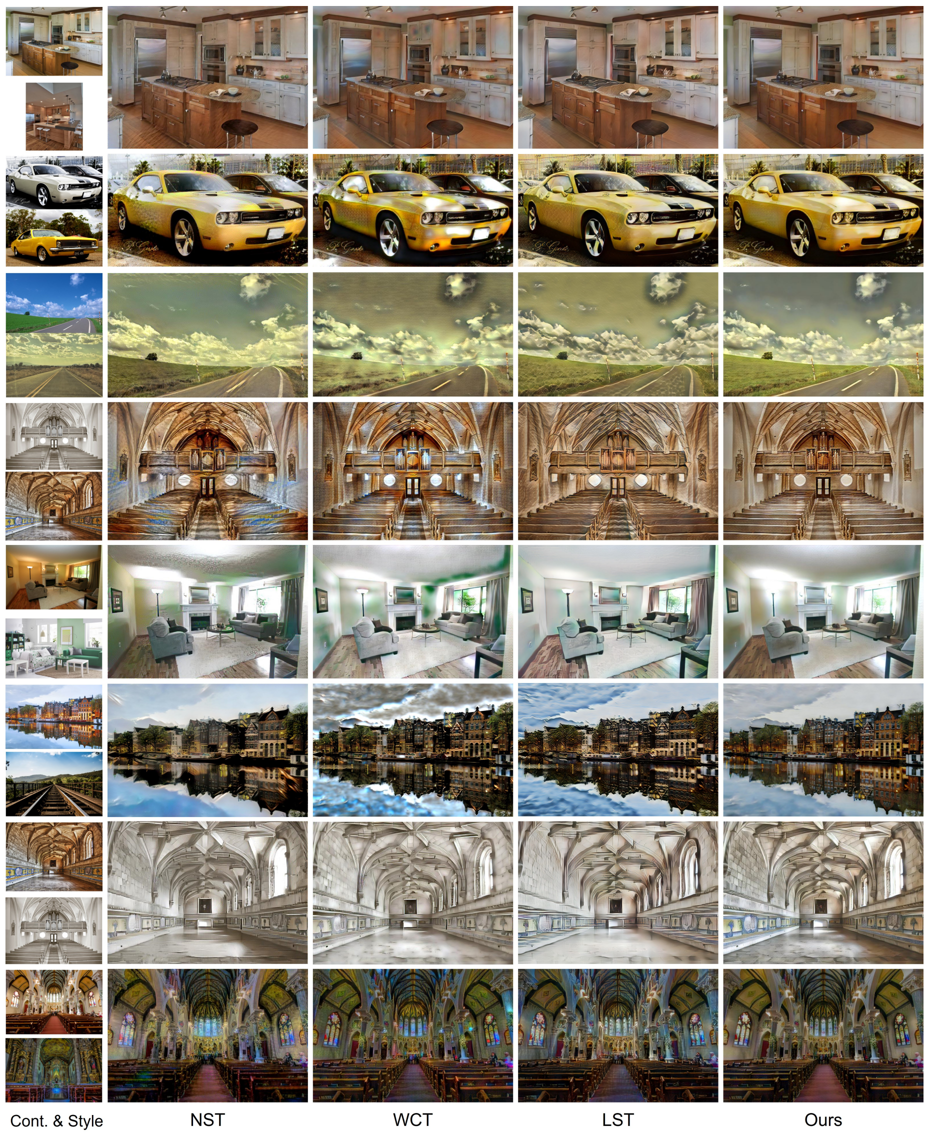

Fig. 3: Photo-realistically stylized images from 30 pairs of a content and a style images for quantitative analysis. No spatial control and no post-processing are applied (Part 3/4).

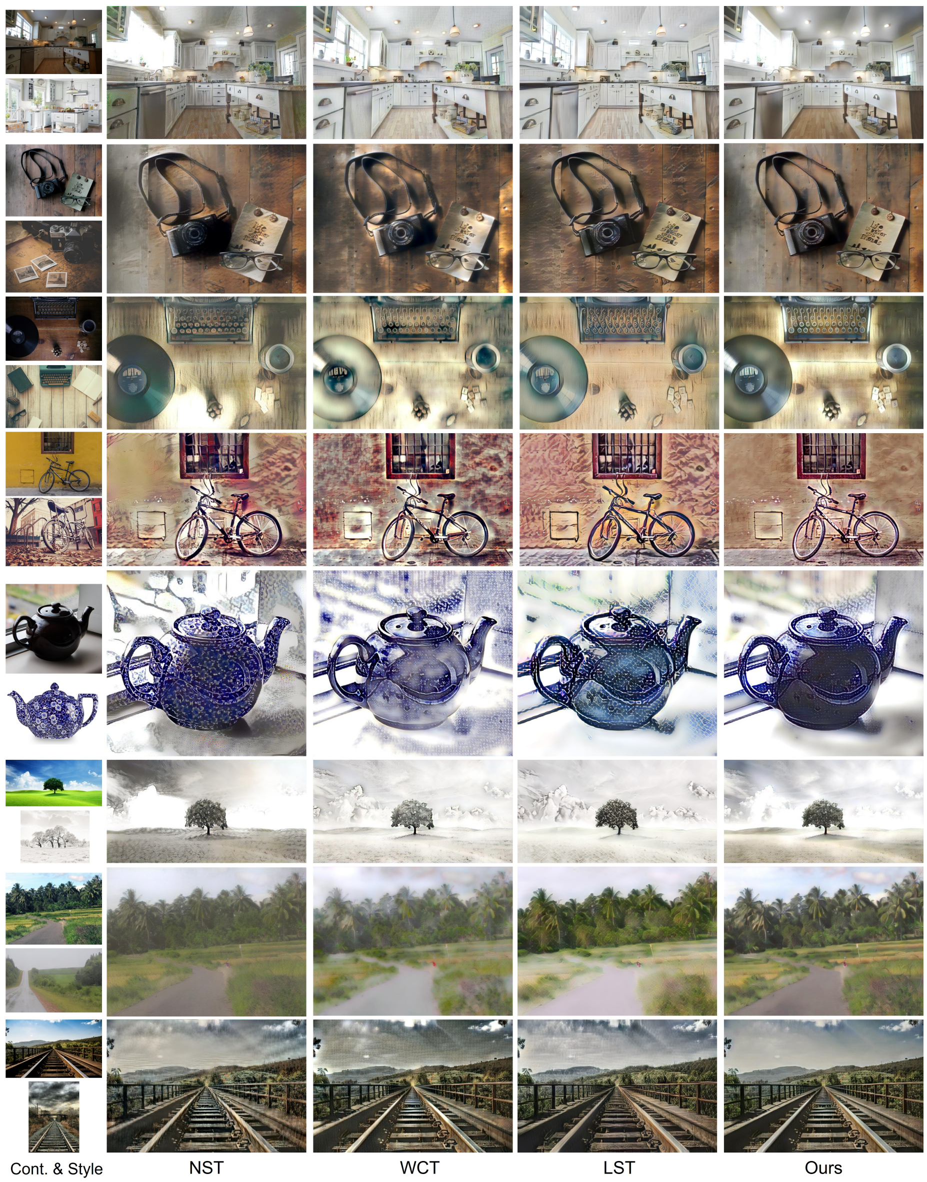

Fig. 4: Photo-realistically stylized images from 30 pairs of a content and a style images for quantitative analysis. No spatial control and no post-processing are applied (Part 4/4).

5 Formulation of NST and WCT for multi-style transfer

The objective of NST for multi-style transfer is as follows:

$$\min_I ||\mathbf{F}_4(I) - \mathbf{F}_{4,c}||_F^2 + \sum_{N=1}^{4}\sum_{k=1}^{q} \lambda_N^k || \frac{1}{n_N}\mathbf{F}_N(I)\mathbf{F}_N(I)^{\mathrm{T}} - \frac{1}{m_N^k}\mathbf{F}_{N,s}^k(\mathbf{F}_{N,s}^k)^{\mathrm{T}}||_F^2, \tag{33}$$

where $\mathbf{F}_{N,s}^k$'s are the feature maps of q style images extracted from $encoder_N$. The stylized image is then derived by solving equation 33 using gradient descent by back-propagation. How different style features are included in equation 33 is non-linear.

On the other hand, WCT realizes multiple-style transfer by linear interpolation of transformed features. By applying WCT to each style feature $\mathbf{F}_{N,s}^k$ and the content feature $\mathbf{F}_{N,c}$, we can derive a transformed feature $\mathbf{F}_{N,wct}^k$. The final feature $\mathbf{F}_{N,wct}$ to be decoded is an affine combination:

$$\mathbf{F}_{N,wct} = \sum_{k=1}^{q} w_k \mathbf{F}_{N,wct}^k, \text{ with } \sum_{k=1}^{q} w_k = 1. \tag{34}$$

As such, each style is weakened due to $w_k < 1$ in the stylized image and could even not be observed.

6 Double-style transfer results from AdaIN and Avatar-net

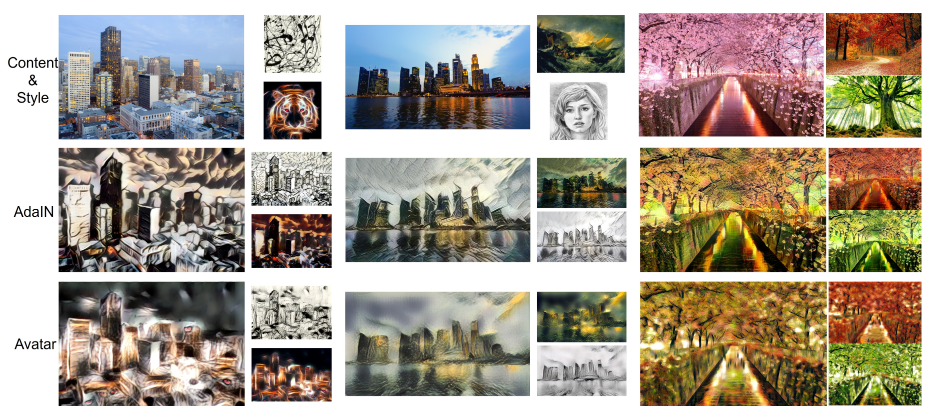

Fig. 5: Double-style transfer results from AdaIN and Avatar-net. Unlike our method that preserves the integrity of each style, styles in doubly stylized images from AdaIN and Avatar-net might be weakened due to the linear interpolation of feature maps.