PCA-Based Knowledge Distillation Towards Lightweight and Content-Style Balanced Photorealistic Style Transfer Models

Tai-Yin Chiu
University of Texas at Austin
chiu.taiyin@utexas.edu

Danna Gurari
University of Colorado Boulder
danna.gurari@colorado.edu

Supplementary Materials

This document supplements the main paper with the following.

1. Results demonstrating the insufficiency of LST for photorealistic style transfer (supplements Section 2 of the main paper).
2. Derivation of the equivalence of Equations 3 and 4 in the main paper.
3. Derivation of the loss objective in Equation 6 in the main paper.
4. Channel length selection for Ours-Mob model distilled from MobileNet (supplements Section 3.3 in the main paper).
5. Results demonstrating that high-frequency residuals improve the high-frequency detail construction of PhotoWCT and the original CKD-distilled model (supplements Section 4 of the main paper).
6. Inference time of different models on the CPU (supplements Section 4.2 of the main paper).
7. Demonstration that CKD does not improve performance when implemented using our PCA-derived channel lengths instead of the empirical ones (supplements Section 4.3 of the main paper).
8. Qualitative results to supplement those in Section 4.3 of the main paper.
9. Results demonstrating that global eigenbases reflect style better than local eigenbases (supplements Section 4.3 of the main paper).
10. Results of stylized images for 4K+ resolutions from our PCA-distilled models (supplements Section 4.3 of the main paper).

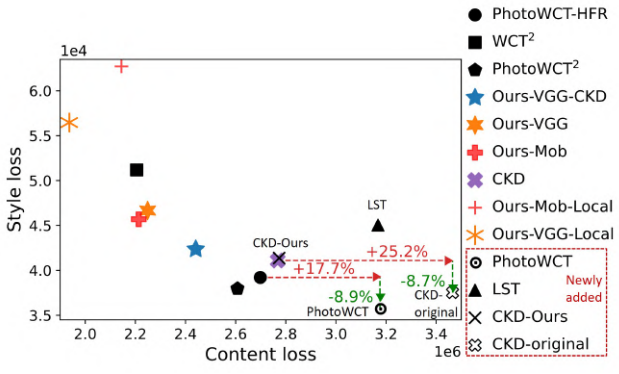

Figure 1. Content and style losses resulting from different models. This figure extends the Figure 5(a) in the main paper. Four newly added points are for LST [3], PhotoWCT [4], CKD-original [6], and CKD-Ours.

Insufficiency of LST for photorealistic style transfer

In the Section 2 of the main paper, we mention LST [3], an autoencoder-based model for artistic style transfer, is claimed to be capable of photorealistic style transfer. However, it does not provide any quantitative analysis for that assertion. Here we show LST is not sufficient for photorealistic style transfer by providing both quantitative and qualitative results.

First, we observe in Fig. 1 that LST results in a content loss and a style loss both worse than those of CKD, which has the worst performance of all methods considered in the main paper. Qualitative results in Fig. 2 also show that compared to the results from our models, the results from LST are prone to blurred boundaries (low sharpness) and dullness (low contrast).

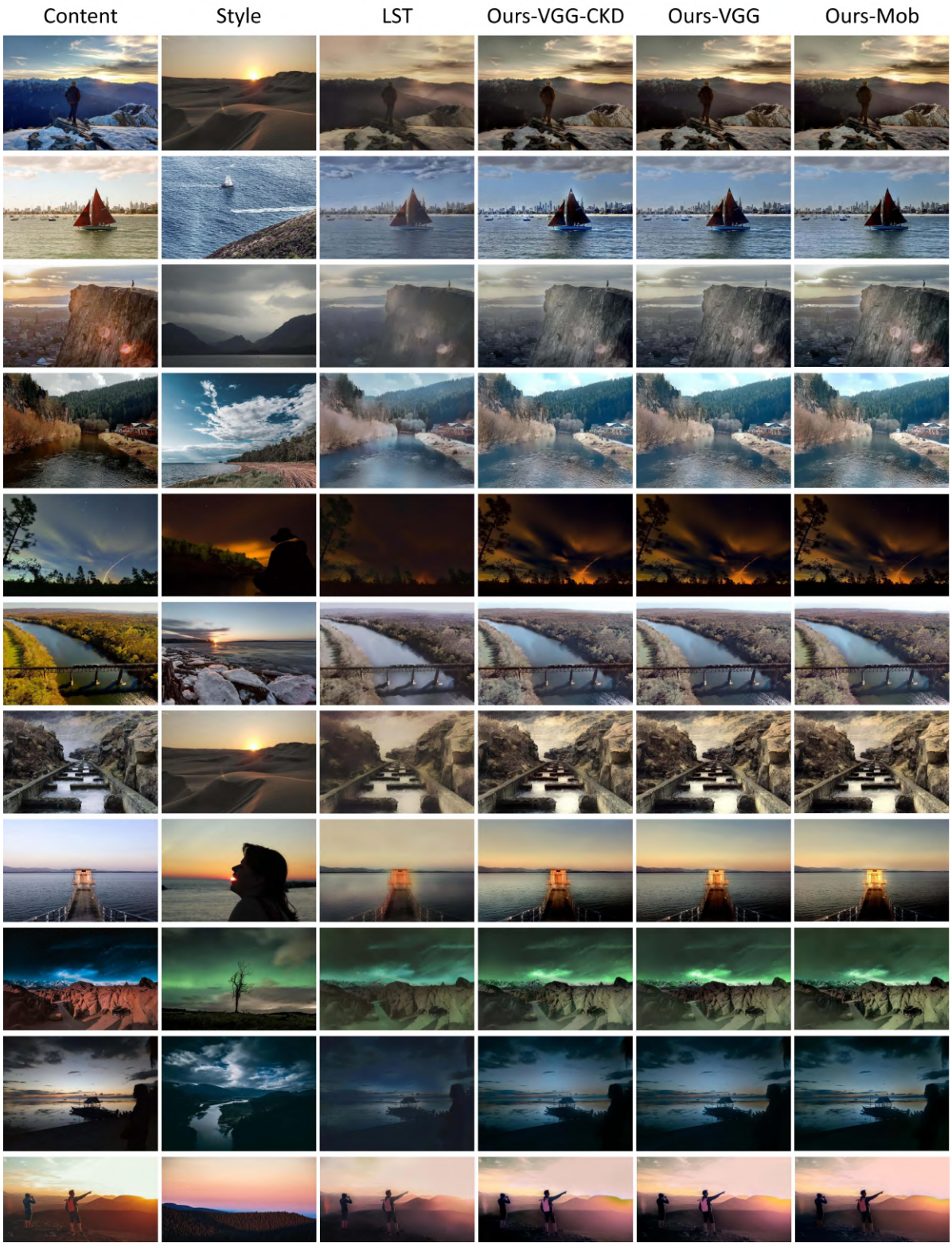

Figure 2. Comparison of the qualitative results from LST [3] and our PCA-distilled models. Compared to the results from our models, the results from LST are prone to blurred boundaries (low sharpness) and dullness (low contrast).

Derivation of the equivalence of Equations 3 and 4 in the main paper

In order to solve Equation 3 in the main paper with mini-batch gradient descent, we avoid the unstable minimization process due to the unbounded trace function by rewriting Equation 3 as Equation 4 where the objective is lower-bounded. Here we prove the equivalence of them (the following two equations).

$$\max_{\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}=\mathbb{1}} \frac{1}{M}\sum_{k=1}^{M}\mathrm{tr}(\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}), \quad (1)$$

and

$$\min_{\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}=\mathbb{1}} \frac{1}{M}\sum_{k=1}^{M}||\mathbf{W}_{N,g}^{\mathrm{T}}\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}-\bar{\mathbf{F}}_{N,k}||_2^2. \quad (2)$$

Proof.

$$\min_{\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}=\mathbb{1}} \frac{1}{M}\sum_{k=1}^{M}||\mathbf{W}_{N,g}^{\mathrm{T}}\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}-\bar{\mathbf{F}}_{N,k}||_2^2 \quad (3)$$

$$= \min_{\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}=\mathbb{1}} \frac{1}{M}\sum_{k=1}^{M}\mathrm{tr}\big[(\mathbf{W}_{N,g}^{\mathrm{T}}\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}-\bar{\mathbf{F}}_{N,k})\cdot(\mathbf{W}_{N,g}^{\mathrm{T}}\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}-\bar{\mathbf{F}}_{N,k})^{\mathrm{T}}\big] \quad (4)$$

$$= \min_{\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}=\mathbb{1}} \frac{1}{M}\sum_{k=1}^{M}\mathrm{tr}\begin{bmatrix}\mathbf{W}_{N,g}^{\mathrm{T}}\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}\mathbf{W}_{N,g}\\ -\mathbf{W}_{N,g}^{\mathrm{T}}\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\\ -\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}\mathbf{W}_{N,g}\\ +\cancel{\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}}\end{bmatrix} \quad (5)$$

The last term $\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}$ is crossed out without affecting the optimization result. By using the identity that $\mathrm{tr}[\mathbf{AB}] = \mathrm{tr}[\mathbf{BA}]$ for any two multiplicable matrices $\mathbf{A}$ and $\mathbf{B}$, the objective can be further simplified as follows:

$$\min_{\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}=\mathbb{1}} \frac{1}{M}\sum_{k=1}^{M}\mathrm{tr}\begin{bmatrix}\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}\\ -\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}\\ -\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}\end{bmatrix} \quad (6)$$

$$= \min_{\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}=\mathbb{1}} \frac{1}{M}\sum_{k=1}^{M}\mathrm{tr}\begin{bmatrix}\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}\mathbb{1}\\ -2\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}\end{bmatrix} \quad (7)$$

$$= \min_{\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}=\mathbb{1}} \frac{1}{M}\sum_{k=1}^{M}\mathrm{tr}[-\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}] \quad (8)$$

$$= \max_{\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}=\mathbb{1}} \frac{1}{M}\sum_{k=1}^{M}\mathrm{tr}[\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}]. \quad (9)$$

■

Derivation of the loss objective in Equation 6 in the main paper

In the encoder distillation introduced in Section 3.2 in the main paper, we find using the feature reconstruction loss in Eq. (10) for encoder distillation results in a better convergence than directly taking $||\bar{\mathbf{F}}_{N,k}^{e}-\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}||_2^2$ as the loss function.

$$\mathcal{L}_{enc}^{N}(I_k) = ||\mathbf{W}_{N,g}^{\mathrm{T}}\bar{\mathbf{F}}_{N,k}^{e}-\bar{\mathbf{F}}_{N,k}||_2^2. \quad (10)$$

Here we show the equivalence of Eq. (10) and $||\bar{\mathbf{F}}_{N,k}^{e}-\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}||_2^2$ as the loss function.

Proof. First,

$$\min_{\bar{\mathbf{F}}_{N,k}^{e}} ||\bar{\mathbf{F}}_{N,k}^{e}-\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}||_2^2 \quad (11)$$

$$= \min_{\bar{\mathbf{F}}_{N,k}^{e}} \mathrm{tr}\big[(\bar{\mathbf{F}}_{N,k}^{e}-\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k})(\bar{\mathbf{F}}_{N,k}^{e}-\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k})^{\mathrm{T}}\big] \quad (12)$$

$$= \min_{\bar{\mathbf{F}}_{N,k}^{e}} \mathrm{tr}\begin{bmatrix}\bar{\mathbf{F}}_{N,k}^{e}(\bar{\mathbf{F}}_{N,k}^{e})^{\mathrm{T}}+\cancel{\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}}\\ -\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}(\bar{\mathbf{F}}_{N,k}^{e})^{\mathrm{T}}-\bar{\mathbf{F}}_{N,k}^{e}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}\end{bmatrix}, \quad (13)$$

where the term $\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}$ can be crossed out since it does not contain the variable $\bar{\mathbf{F}}_{N,k}^{e}$ we optimize for. Second,

$$\min_{\bar{\mathbf{F}}_{N,k}^{e}} ||\mathbf{W}_{N,g}^{\mathrm{T}}\bar{\mathbf{F}}_{N,k}^{e}-\bar{\mathbf{F}}_{N,k}||_2^2 \quad (14)$$

$$= \min_{\bar{\mathbf{F}}_{N,k}^{e}} \mathrm{tr}\big[(\mathbf{W}_{N,g}^{\mathrm{T}}\bar{\mathbf{F}}_{N,k}^{e}-\bar{\mathbf{F}}_{N,k})(\mathbf{W}_{N,g}^{\mathrm{T}}\bar{\mathbf{F}}_{N,k}^{e}-\bar{\mathbf{F}}_{N,k})^{\mathrm{T}}\big] \quad (15)$$

$$= \min_{\bar{\mathbf{F}}_{N,k}^{e}} \mathrm{tr}\begin{bmatrix}\mathbf{W}_{N,g}^{\mathrm{T}}\bar{\mathbf{F}}_{N,k}^{e}(\bar{\mathbf{F}}_{N,k}^{e})^{\mathrm{T}}\mathbf{W}_{N,g}+\cancel{\bar{\mathbf{F}}_{N,k}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}}\\ -\mathbf{W}_{N,g}^{\mathrm{T}}\bar{\mathbf{F}}_{N,k}^{e}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}-\bar{\mathbf{F}}_{N,k}(\bar{\mathbf{F}}_{N,k}^{e})^{\mathrm{T}}\mathbf{W}_{N,g}\end{bmatrix} \quad (16)$$

By using the identity that $\mathrm{tr}[\mathbf{AB}] = \mathrm{tr}[\mathbf{BA}]$ for any two multiplicable matrices $\mathbf{A}$ and $\mathbf{B}$, the objective can be further simplified as follows:

$$\min_{\bar{\mathbf{F}}_{N,k}^{e}} \mathrm{tr}\begin{bmatrix}\bar{\mathbf{F}}_{N,k}^{e}(\bar{\mathbf{F}}_{N,k}^{e})^{\mathrm{T}}\cancel{\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}}\\ -\bar{\mathbf{F}}_{N,k}^{e}\bar{\mathbf{F}}_{N,k}^{\mathrm{T}}\mathbf{W}_{N,g}^{\mathrm{T}}-\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}(\bar{\mathbf{F}}_{N,k}^{e})^{\mathrm{T}}\end{bmatrix}, \quad (17)$$

where $\mathbf{W}_{N,g}\mathbf{W}_{N,g}^{\mathrm{T}}$ is crossed out since it is an identity matrix. Since the equality of Eq. (13) and Eq. (17), we prove the equivalence of these two optimization problems $\min_{\bar{\mathbf{F}}_{N,k}^{e}} ||\bar{\mathbf{F}}_{N,k}^{e}-\mathbf{W}_{N,g}\bar{\mathbf{F}}_{N,k}||_2^2$ and $\min_{\bar{\mathbf{F}}_{N,k}^{e}} ||\mathbf{W}_{N,g}^{\mathrm{T}}\bar{\mathbf{F}}_{N,k}^{e}-\bar{\mathbf{F}}_{N,k}||_2^2$. ■

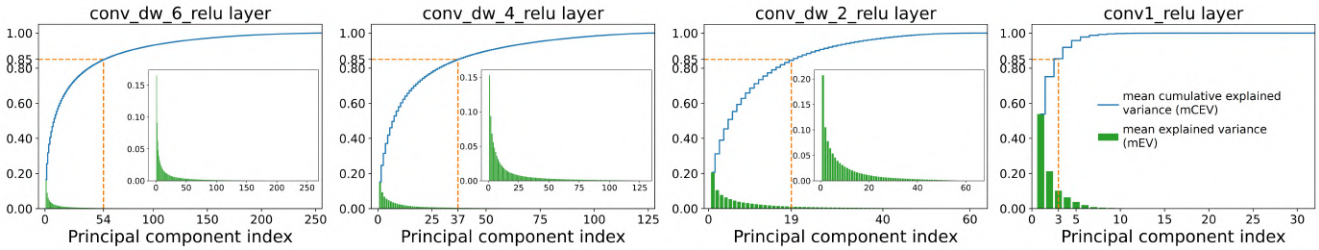

Figure 3. Mean explained variance (green histogram) and mean cumulative explained variance (blue curve) of the *conv_dw_relu* features and the *conv_1_relu* feature of MS-COCO images from MobileNet [2]. It is observed that on average 85% of the variance of a *conv_dw_6_relu*, *conv_dw_4_relu*, *conv_dw_2_relu*, or *conv_1_relu* feature can be explained by 54, 37, 19, or 3 eigenvectors of the feauture covariance, respectively.

Channel length selection for Ours-Mob model distilled from MobileNet

To demonstrate the generalizability of our PCA knowledge distillation, we apply it to distill style information from MobileNet [2]. By following how we select the layers from VGG-19 [5] for style representation: we select the layer right after each downsampling layer for four layers in total, we select the *conv_dw_6_relu*, *conv_dw_4_relu*, *conv_dw_2_relu*, and *conv_1_relu* layers from MobileNet for style representation.

We follow the same procedure as described in Section 4.1 in the main paper to distill style information from the selected layers to a smaller model which we call Ours-Mob in the main paper. We plot mCEV and mEV for each selected layer in Fig. 3. It is observed that on average 85% of the variance of a *conv_dw_6_relu*, *conv_dw_4_relu*, *conv_dw_2_relu*, or *conv_1_relu* feature can be explained by 54, 37, 19, or 3 eigenvectors of the feauture covariance, respectively. However, we find that if C_1^e is set to 3, it prevents the distillation from the *conv_dw_2_relu* layer if C_2^e is set to 19. We fix this issue by following the value of C_1^e we use in Ours-VGG. In the end, we set four channel lengths (C_1^e, C_2^e, C_3^e, C_4^e) to be (10, 19, 37, 54), resulting in Ours-Mob model.

High-frequency residuals improve the high-frequency detail construction of PhotoWCT and the original CKD-distilled model

Recall in the Section 2 in the paper that we mentioned PhotoWCT [4] is poor at preserving content due to two reasons: too strong stylization strength that introduces artifacts and the lossy architecture that does not hold the high-frequency detail well. To consider stylization strength as the main factor that affect the content preservation, we fix the lossy architecture by introducing high-frequency residuals [1] (HFR) to PhotoWCT. The resulting model which we call PhotoWCT-HFR reduces the content loss of PhotoWCT by 17.7% (Fig. 1). The better content preservation of PhotoWCT-HFR is due to its better high-frequency construction as exemplified in Fig. 4.

Similarly, the original model distilled with CKD [6] (which we call CKD-original) is for artistic style transfer and also poor at constructing high-frequency details. To have a fair comparison, we again introduce HFR to our CKD-distilled model. The resulting model which we call CKD in the main paper reduces the content loss of CKD-original by 25.2% (Fig. 1). The better content preservation of CKD is due to its better high-frequency construction as exemplified in Fig. 4.

Demonstration that CKD does not improve performance when implemented using our PCA-derived channel lengths instead of the empirical ones

Unlike our PCA distillation, which has clear guidelines for channel length selection, the previous method CKD [6] empirically sets the channel lengths to be ($C_1^e = 16$, $C_2^e = 32$, $C_3^e = 64$, $C_4^e = 128$) when distilling from VGG-19 and results in the CKD model. We show here that the model distilled with CKD using our smaller channel lengths ($C_1^e = 10$, $C_2^e = 20$, $C_3^e = 58$, $C_4^e = 64$), which we call CKD-Ours, does not change the performance of CKD as shown in Fig. 1. Qualitatively, as shown in Fig. 5, we observe that both CKD and CKD-Ours produce lots of artifacts in the synthesized images, and our models consistently result in more photorealistic images than both CKD-distlled models.

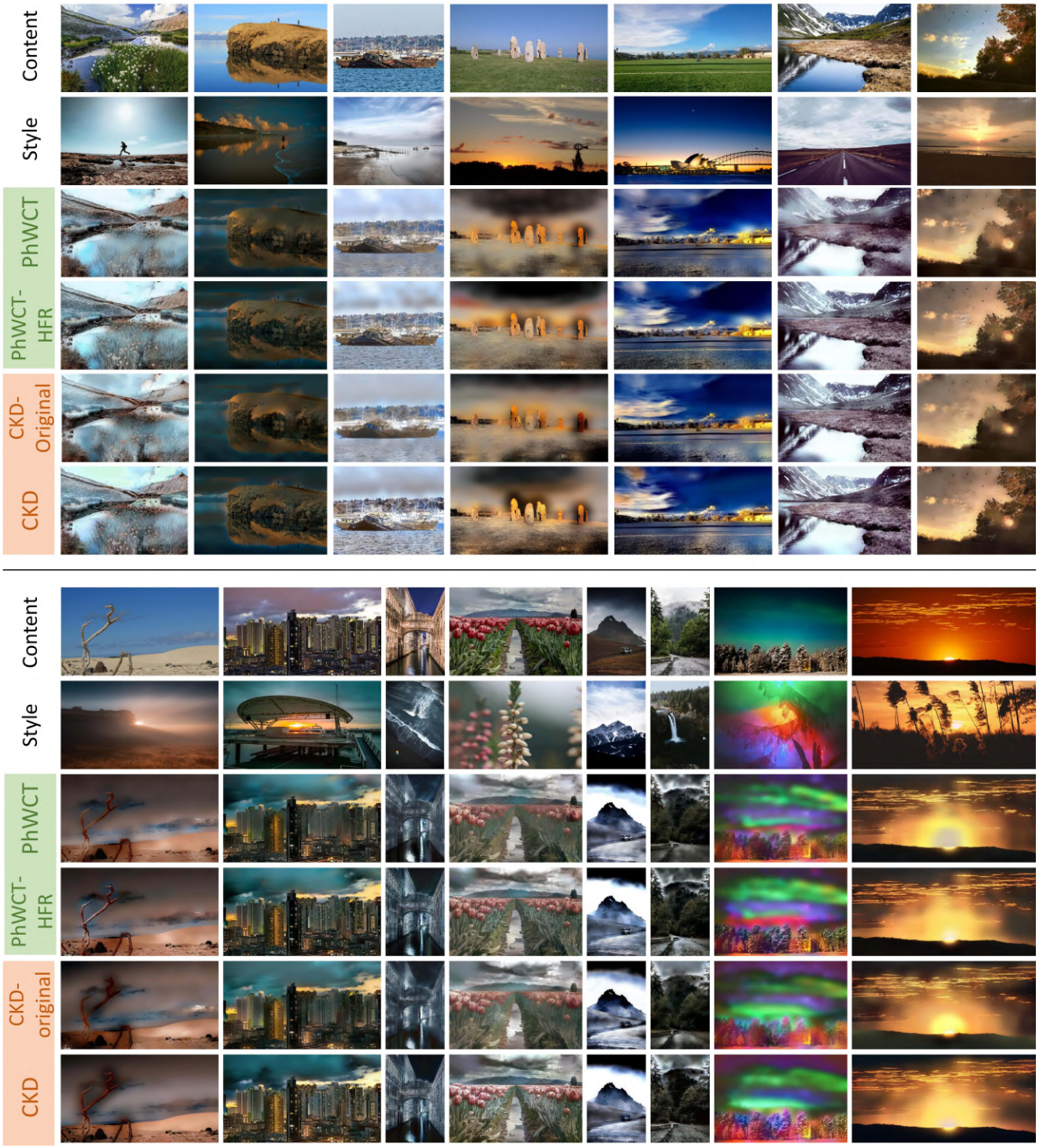

Figure 4. High-frequency residuals (HFR) [1] improve the content preservation of PhotoWCT [4] and CKD-original [6] by reinforcing the high-frequency detail construction.

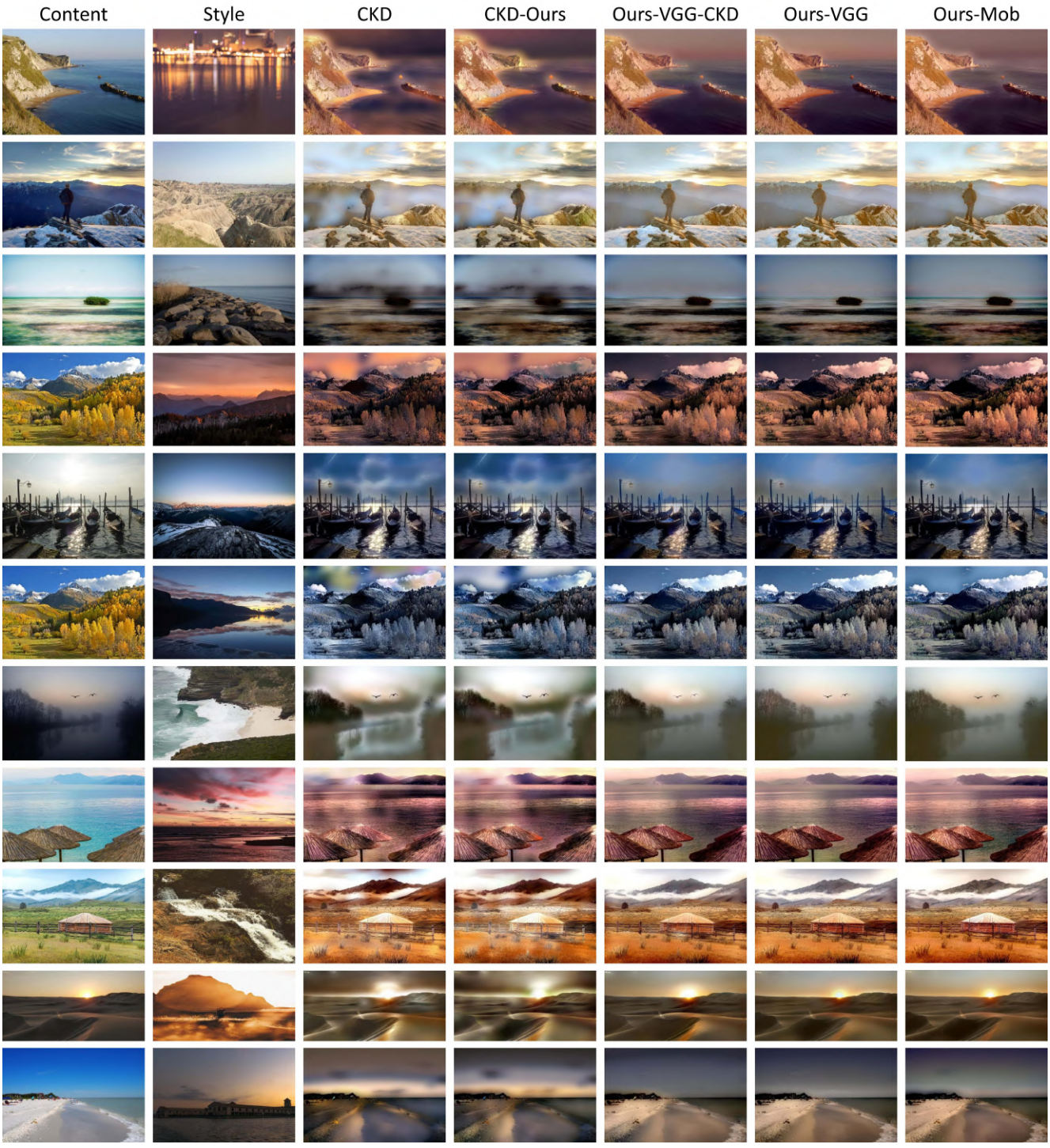

Figure 5. Qualitative comparison between our models, CKD and CKD-Ours. We observe that CKD and CKD-Ours produce very similar results with lots of artifacts, and our models consistently result in more photorealistic images than both CKD-distlled models.

Inference time of different models on the CPU

we report in Tab. 1 models' inference times on an Intel Xeon W-2195 CPU @ 2.30GHz with workstation memory of 256GB. The results have a similar trend to those in Table 1 in the main paper: our distilled models achieve the fastest inference time. Moreover, while the non-distilled models spend 1-2+ minutes rendering an 8K image with CPUs, our models require considerably less time; i.e., 10 seconds/image.

Model	HD	FHD	QHD	4K	5K	8K
WCT^2	5.92	18.04	31.61	71.28	109.13	X
PhotoWCT-HFR	4.91	10.31	18.06	38.44	69.22	149.22
PhotoWCT^2	2.33	5.06	9.08	20.39	36.15	79.74
CKD	1.04	2.20	3.80	8.03	14.66	34.16
Ours-VGG-CKD	0.47	0.96	1.68	3.60	6.51	14.21
Ours-VGG	**0.40**	**0.77**	**1.44**	**2.88**	**4.84**	**10.78**
Ours-Mob	**0.21**	**0.37**	**0.76**	**1.62**	**2.76**	**6.02**

Table 1. Inference time (s/img) of different models on CPU. The naming follows the main paper. X: segmentation fault (We found this is not due to the common reason of insufficient memory or system stack size. The reason remains unknown).

More stylized images from the PST dataset

We show several qualitative results from the PST dataset [7] in Figure 5 in the main paper. Here we show more of them in Fig. 7.

Global eigenbases reflect style better than local eigenbases

In Section 3 of the main paper, we explain the limitation of local eigenbases to faithfully reflect the style of style images. To overcome this, we propose using global eigenbases. The quantitative result (Fig. 1) justifies this limitation and our strategy. We show in Fig. 6 some qualitative results comparing the impact of global and local eigenbases.

Results of stylized images for 4K+ resolutions from our PCA-distilled models

Since images of large resolutions (e.g. 4K and beyond) contain more high-frequency details, which are hotbeds for artifacts to form in stylization, the ability to reduce artifacts of a photorealistic style transfer model can be better manifested in the stylization of images of 4K+ resolutions. We show such results from our PCA-distilled models in Fig. 8, Fig. 9, Fig. 10, and Fig. 11. We notice that Ours-VGG, which uses the channel lengths derived with PCA, constantly preserve better content than Ours-VGG-CKD, which uses the channel lengths empirically selected in the CKD paper [6], by trimming off the slight artifacts. We also notice that while Ours-Mob results in a slightly lower content loss than Ours-VGG in the stylization of images of lower resolutions (images in PST dataset [7]) as shown in Fig. 1, Ours-VGG produces fewer artifacts than Ours-Mob in the stylization of our 4K+ images.

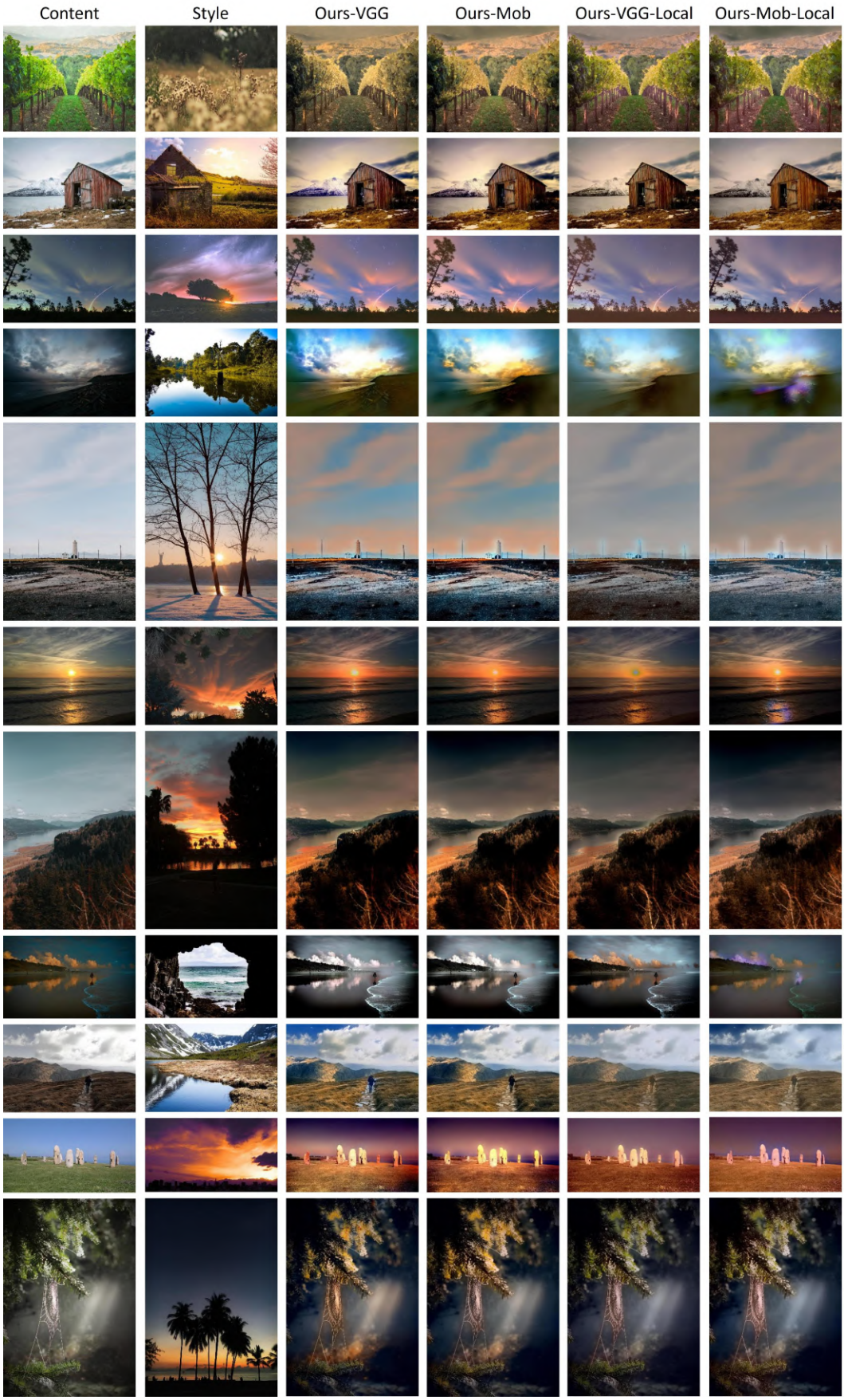

Figure 6. PCA distillation with global eigenbases reflects better style than local eigenbases.

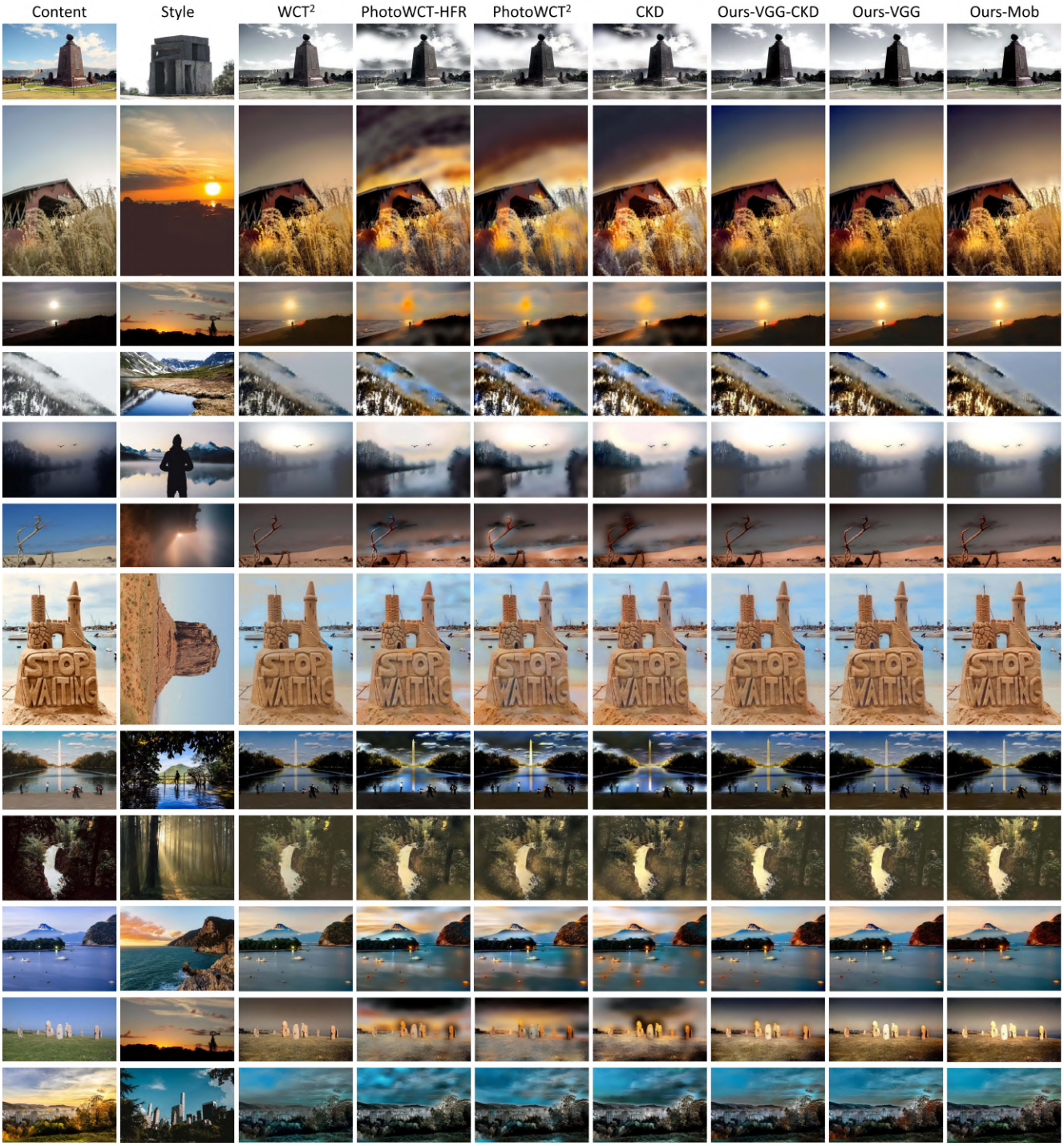

Figure 7. More stylized images from the PST dataset [7]. This figure expands the Figure 5 in the main paper.

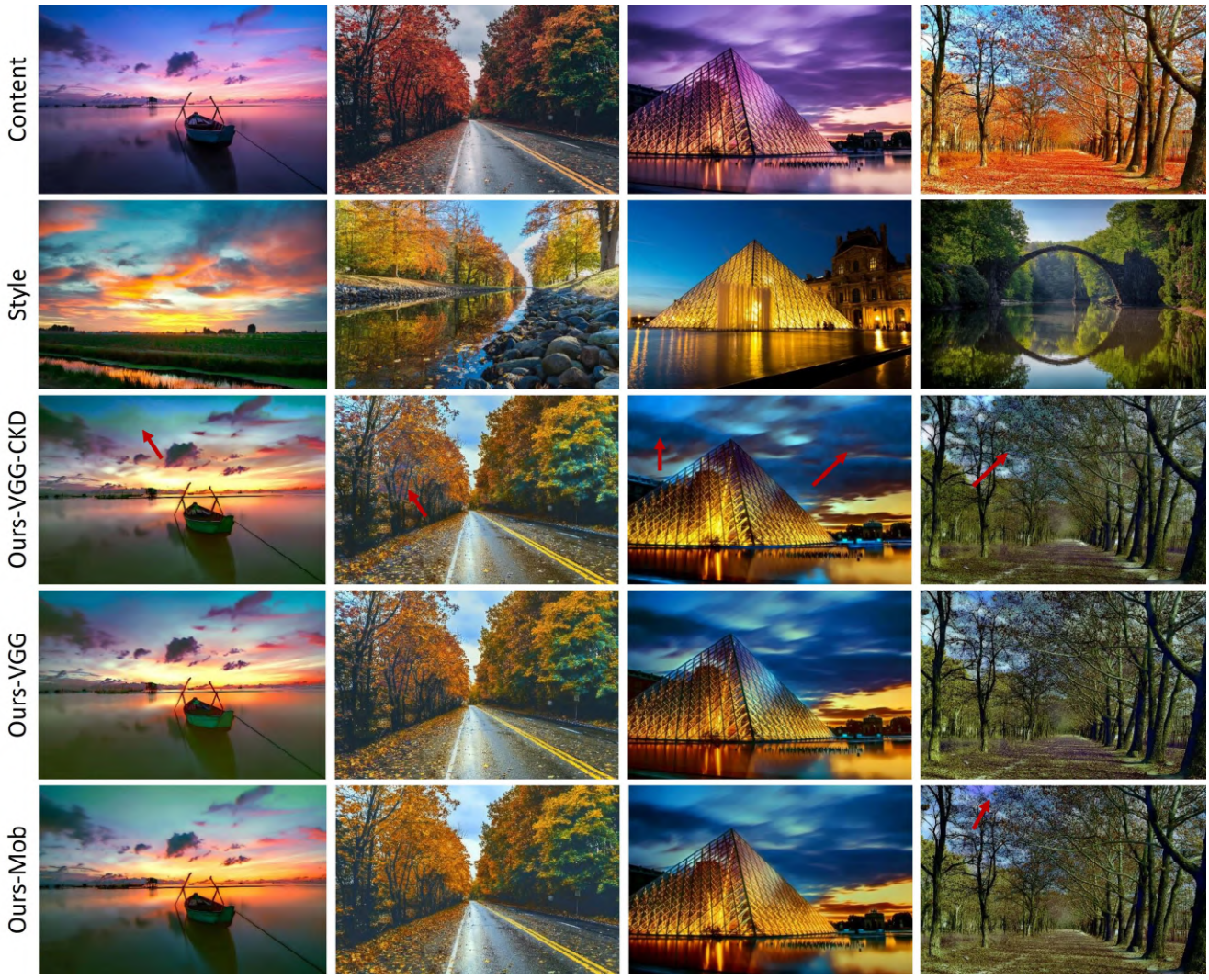

Figure 8. Stylized images of 4K+ resolutions resulting from our models. Several artifacts are pointed out with the red arrows. (Part 1/4)

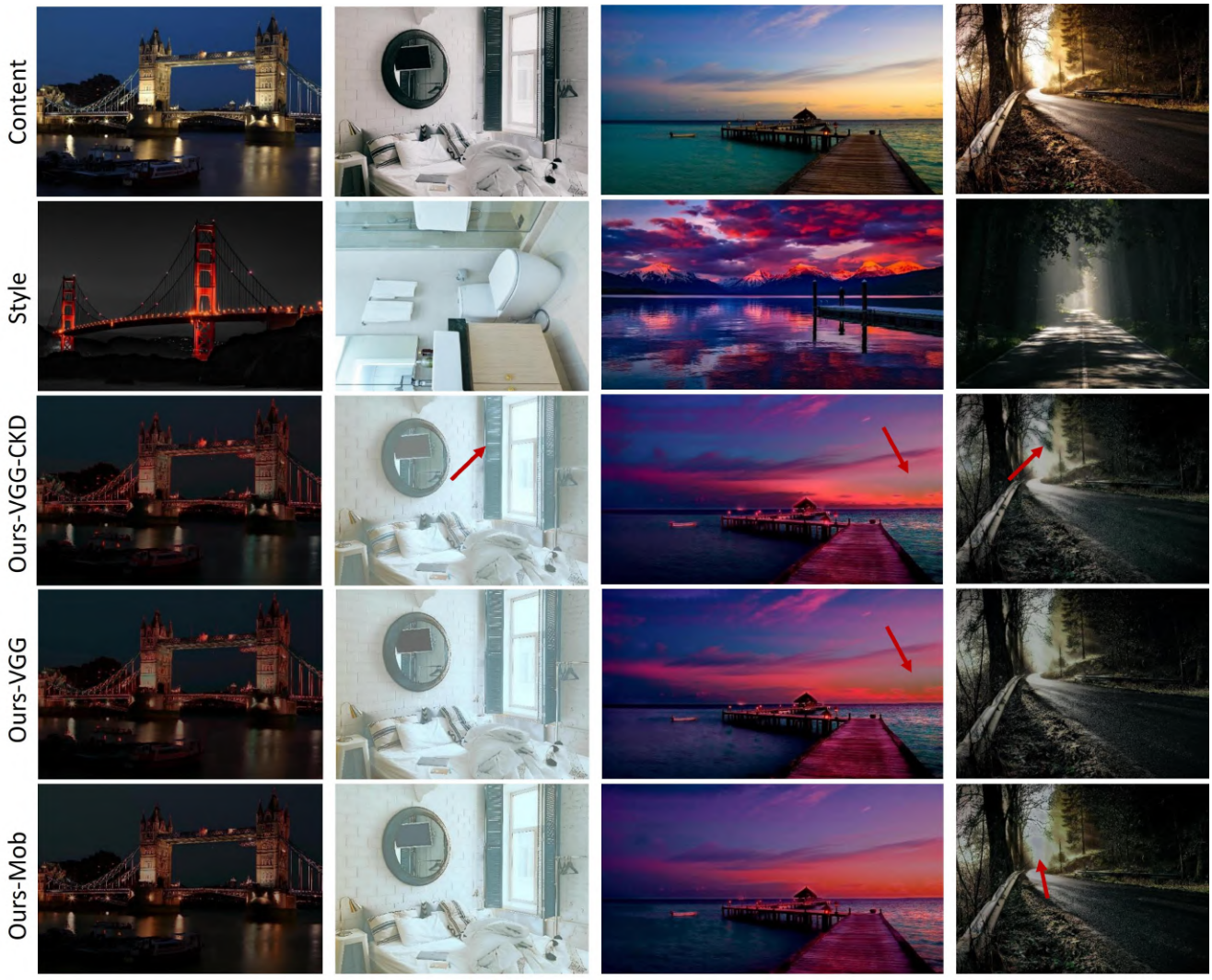

Figure 9. Stylized images of 4K+ resolutions resulting from our models. Several artifacts are pointed out with the red arrows. (Part 2/4)

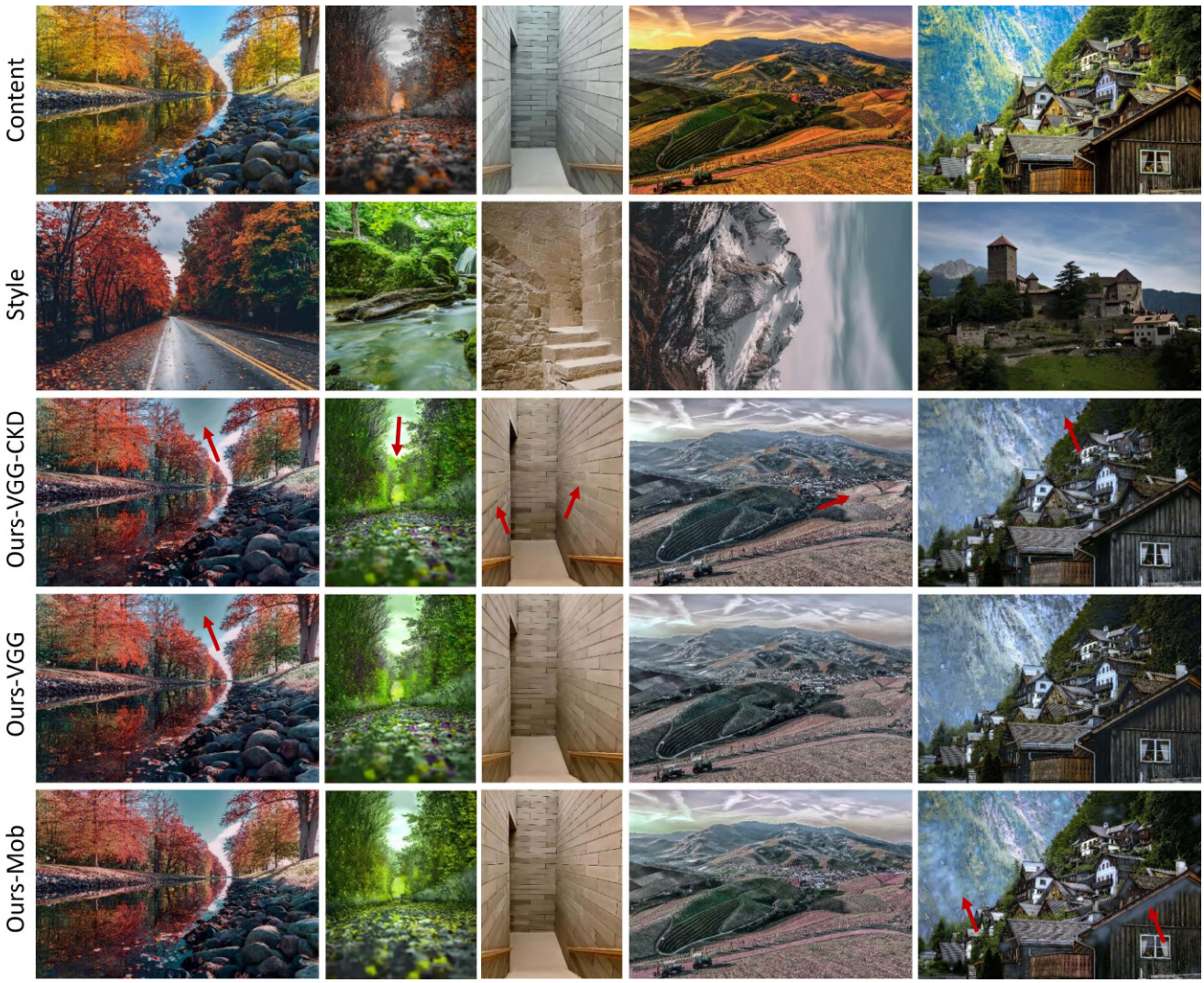

Figure 10. Stylized images of 4K+ resolutions resulting from our models. Several artifacts are pointed out with the red arrows. (Part 3/4)

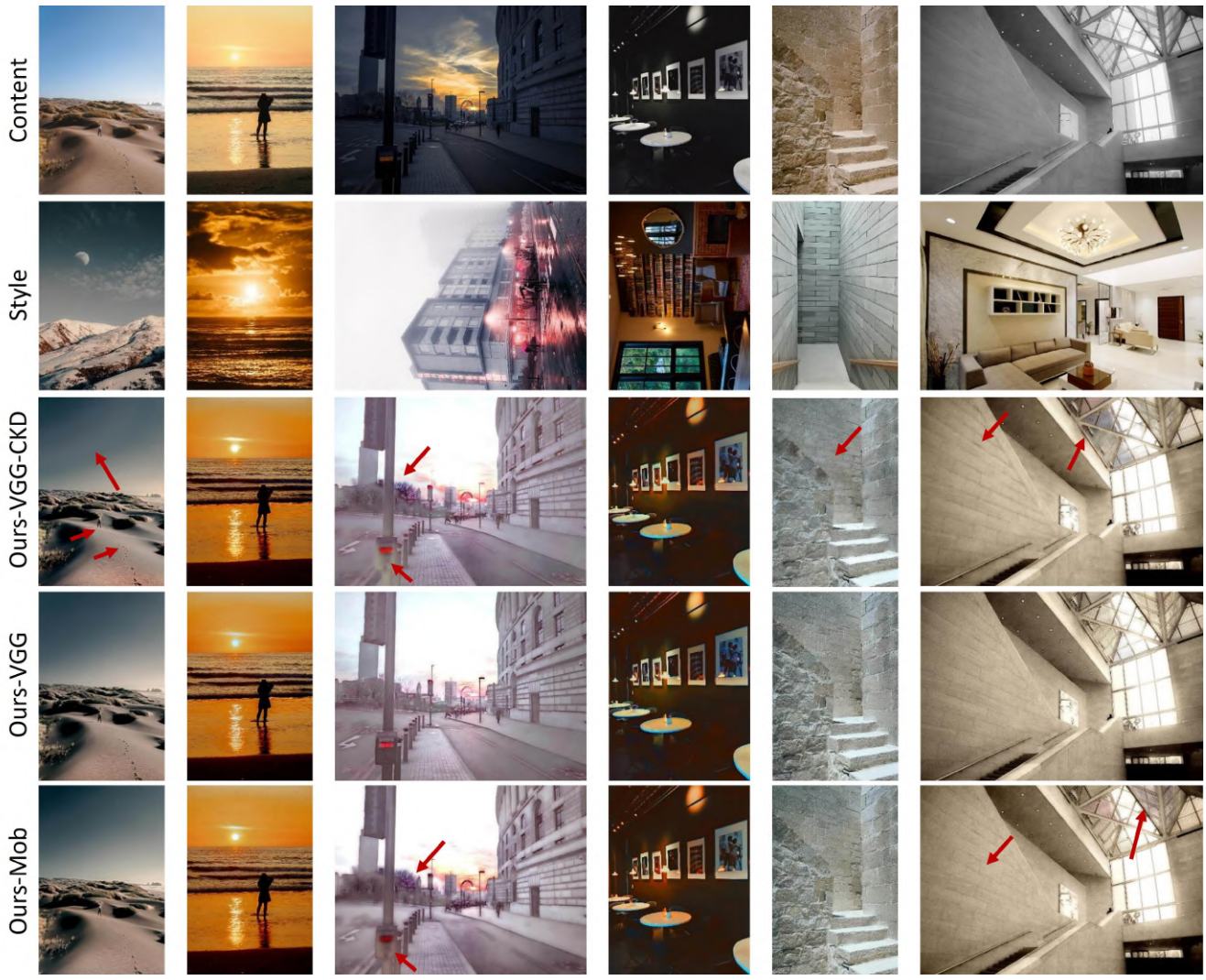

Figure 11. Stylized images of 4K+ resolutions resulting from our models. Several artifacts are pointed out with the red arrows. (Part 4/4)

References

[1] Tai-Yin Chiu and Danna Gurari. Photowct2: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals. *arXiv preprint arXiv:2110.11995*, 2021. 4, 5

[2] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 4

[3] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3809–3817, 2019. 1, 2

[4] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468, 2018. 1, 4, 5

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[6] 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. 1, 4, 5, 7

[7] Xide Xia, Meng Zhang, Tianfan Xue, Zheng Sun, Hui Fang, Brian Kulis, and Jiawen Chen. Joint bilateral learning for real-time universal photorealistic style transfer. *arXiv preprint arXiv:2004.10955*, 2020. 7, 9