Supplementary Materials: Photorealistic Style Transfer via Wavelet Transforms

Jaejun Yoo* Youngjung Uh* Sanghyuk Chun* Byeongkyu Kang Jung-Woo Ha
Clova AI Research, NAVER Corp.
{jaejun.yoo,youngjung.uh,sanghyuk.c,bk.kang,jungwoo.ha}@navercorp.com

1. Frame-based signal reconstruction

Our proposed model WCT^2 is inspired by the recent theoretical advancement of frame-based signal reconstruction approaches [9, 10]. To make the paper self-contained, we provide a brief introduction to the frame theory (Section 1.1), tightness of Haar wavelets (Section 1.2) and our theoretical motivation (Section 1.3).

1.1. Perfect reconstruction condition

Consider an *analysis operator* $\Phi = \begin{bmatrix}\phi_1 & \cdots & \phi_m\end{bmatrix} \in \mathbb{R}^{n\times m}$, where $\{\phi_k\}_{k=1}^m$ is a family of functions in a Hilbert space H. Then, $\{\phi_k\}_{k=1}^m$ is called a *frame* if it satisfies the following inequality [2]:

$$\alpha\|f\|^2 \leq \|\Phi^\top f\|^2 \leq \beta\|f\|^2, \quad \forall f \in H, \tag{1}$$

where $f \in \mathbb{R}^n$ is an input signal and $\alpha, \beta > 0$ are called the frame bounds.

The original signal f can be exactly recovered from the frame coefficient $z = \Phi f$ when there is the *dual frame* $\tilde{\Phi}$ (*i.e.*, *synthesis operator*) satisfying *the perfect reconstruction (PR) condition*: $\tilde{\Phi}\Phi^\top = I$, since $f = \tilde{\Phi}z = \tilde{\Phi}\Phi^\top f = f$. Here, we call such frame *tight* (*i.e.*, $\alpha = \beta$ in (1)) which is equivalent to $\tilde{\Phi} = \Phi$ or $\Phi\Phi^\top = I$. Note that a tight frame does not amplify the power of the input and thus it has the minimum noise amplification factor. To achieve the best reconstruction performance, frame bases should satisfy another property, called energy compaction. This is particularly important to parametric models, which have to adaptively deal with varying amounts of information with a fixed number of parameters, *e.g.*, deep neural networks (DNNs). For example, singular value decomposition (SVD) provides both tight and energy compact bases given an arbitrary signal. However, SVD is data-dependent, which makes it hard to use for a large dataset.

1.2. Wavelet frames

Wavelets are known to compactly represent signals while maintaining important information such as edges, thus resulting in a good energy compaction [1]. Therefore, by using a tight wavelet filter-bank, we can improve the reconstruction performance of encoder-decoder type of networks with minimal noise amplification. Specifically, the non-local basis Φ^T is now composed of a filter bank:

$$\Phi = [T_1 \cdots T_L], \tag{2}$$

where T_k denotes the k-th subband operator and the filter bank is tight, *i.e.*

$$\Phi\Phi^T = \sum_{k=1}^{L} T_k T_k^T = I. \tag{3}$$

In this paper, we use Haar wavelets which is one of the simplest tight filter bank frames with low and high sub-band decomposition. Here, $T_1 \in \mathbb{R}^{\frac{n}{2}\times n}$ is the low-pass subband. This is equivalent to the average pooling:

$$T_1^\top = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 1 & \cdots & 0 & 0 \\ & \vdots & & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix}. \tag{4}$$

Then, T_2 is the high pass filtering given by

$$T_2^\top = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ & \vdots & & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \tag{5}$$

and we can easily see that

$$T_1 T_1^T + T_2 T_2^T = I, \tag{6}$$

so the Haar wavelet frame is tight.

1.3. Theoretical motivation

In the perspective of the frame-based signal reconstruction, the commonly used encoder-decoder convolution structure of deep neural networks (DNNs), such as U-net [8], can be interpreted as the data-driven way of learning the local bases Ψ (*e.g.*, convolution filters) with hand-crafted global bases Φ (*e.g.*, max-pooling) [9]. Recently, Ye *et al.* [9] interpreted training DNNs as finding a multi-layer realization of the convolution framelets [10]:

$$Z = \Phi^T (f \circledast \Psi) \tag{7}$$

$$f = \left(\tilde{\Phi} Z\right) \circledast \tilde{\Psi}, \tag{8}$$

where $\Phi = [\phi_1, \cdots, \phi_n]$ and $\tilde{\Phi} = [\tilde{\phi}_1, \cdots, \tilde{\phi}_n] \in \mathbb{R}^{n \times n}$ (resp. $\Psi = [\psi_1, \cdots, \psi_q]$ and $\tilde{\Psi} = [\tilde{\psi}_1, \cdots, \tilde{\psi}_q] \in \mathbb{R}^{d \times q}$) are frames and their duals. Here, $\circledast$ stands for the convolution operation.

Therefore, the convolutional layers of the encoder learns the signal representation with a global pooling operation. We refer to Φ as global bases because it observes the entire image dimension n while Ψ learns local features from the data by $d \times d$ convolution kernels of q channels. When these frames satisfy the PR condition:

$$\tilde{\Phi} \Phi^\top = I_{n \times n}, \quad \Psi \tilde{\Psi}^\top = I_{d \times d}, \tag{9}$$

the input signal f can be exactly recovered from the learned representations. Note that the encoder-decoder architectures of WCT [5] and PhotoWCT [6] cannot satisfy the perfect reconstruction condition because of the max-pooling, which does not have its exact inverse (*i.e.*, not a frame). On the other hand, our model WCT^2 can fully exploit the information from the encoder due to the favorable property of the wavelet decomposition and reconstruction, *i.e.*, Haar wavelet pooling and unpooling.

1.4. Proposed network architecture

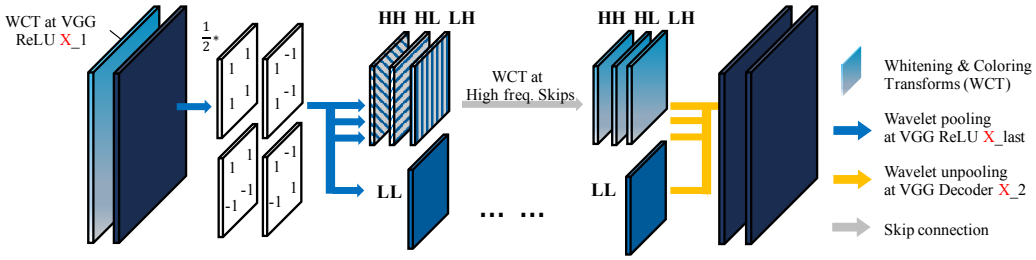


Figure 1: The proposed module using Haar wavelet pooling and unpooling. A pair of encoder and decoder at same scale are shown. WCT is performed on the output of VGG `convX_1` layer followed by subsequent VGG layers and wavelet pooling. Only the low component passes to the next layer and the high frequency components are directly skipped to the corresponding decoding layer. At the decoder, the components are aggregated by the wavelet unpooling.

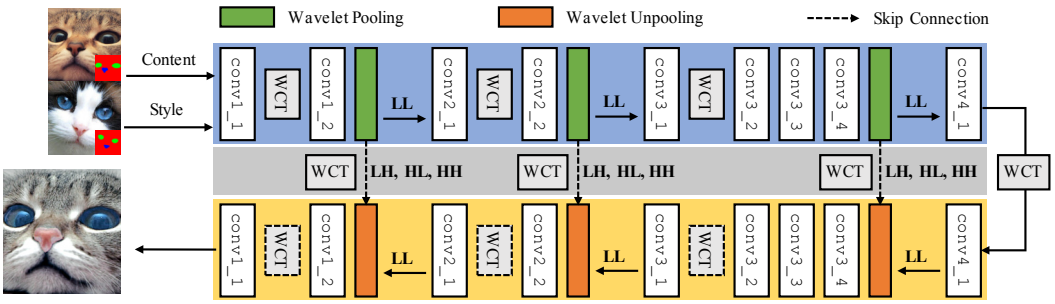
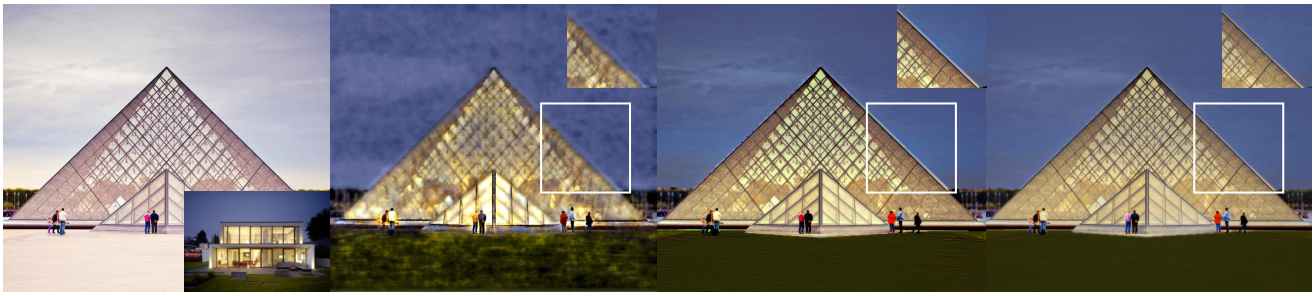


Figure 2: Overview of the proposed progressive stylization. For the encoder, we perform WCT on the output of `convX_1` layer and skip connections. For the decoder, we apply WCT on the output of `convX_2` layer, which is optional.

1.5. Differences to PhotoWCT and Wavelet corrected transfer based on AdaIN (WCT-AdaIN)

Our method shares the motivation with PhotoWCT but the way we posit the problem and reach to its solution is fundamentally different from PhotoWCT: *i*) We showed that the reason why PhotoWCT fails in preserving spatial information is because **pooling and unpooling operations cannot satisfy the frame condition** *ii*) Based on this theoretical analysis, our model architecture is **specifically designed to perfectly preserve spatial structure**, which is **proved effective in theory and practice.** This removes the necessity of post-processing, thus making our model far more practical and powerful than the previous methods. *iii*) The wavelet corrected model is **by no means limited to a specific stylization method**. It can serve as a **general architecture for photorealistic style transfer**, which is compatible with various methods, *e.g.*, AdaIN (Figure 2 (c)). Currently, our method (WCT^2) can process 1k resolution image in 4.7 seconds and this can be accelerated further ($\sim$1 second) by employing adaptive instance normalization (AdaIN) instead of time-consuming SVD procedure.



(a) Inputs **(b) PhotoWCT** **(c) Ours (AdaIN)** **(d) Ours (WCT)**

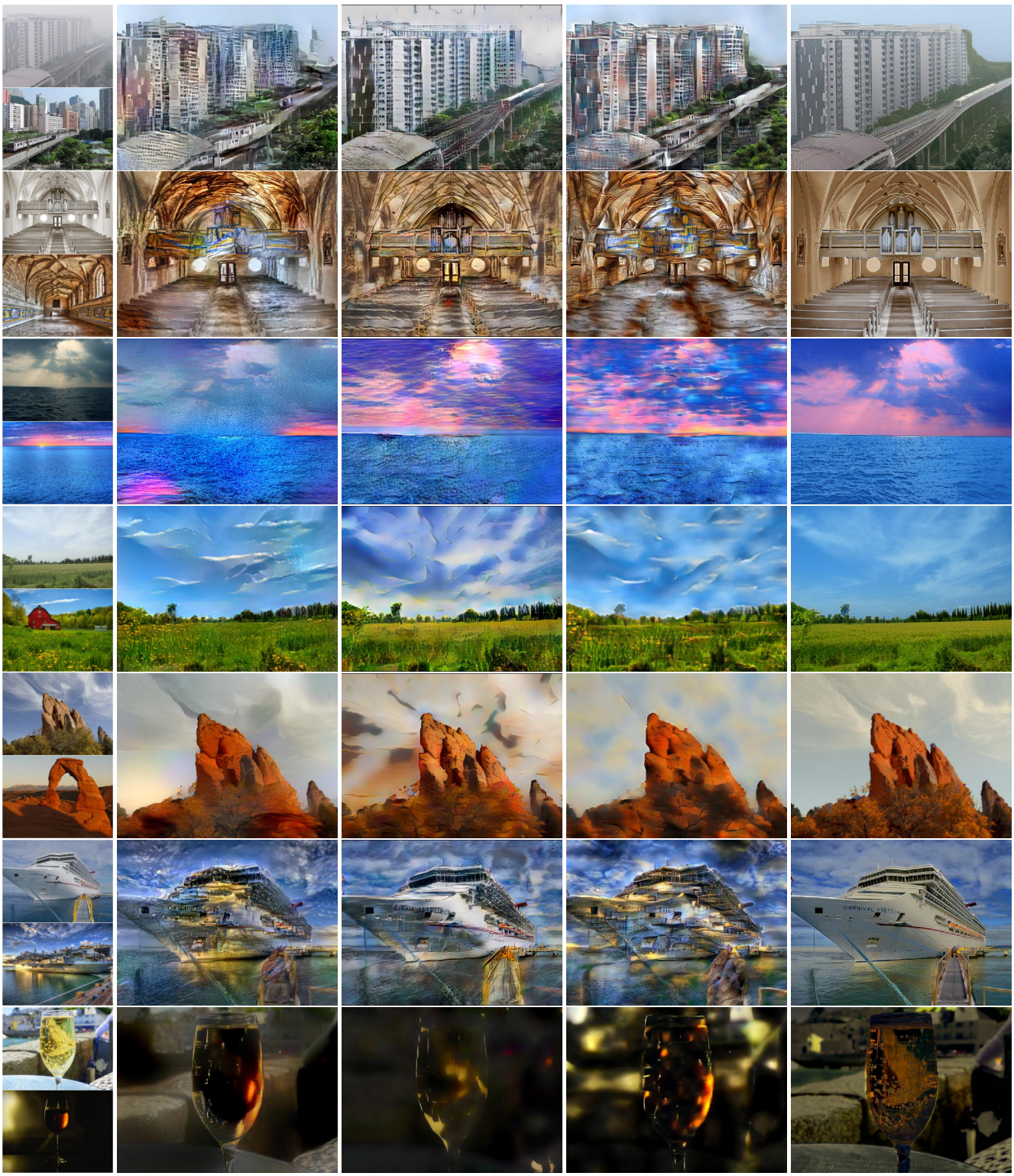
Figure 3: Photorealistic style transfer results of (a) input pairs using (b) PhotoWCT, (c) Ours (AdaIN) and (d) Ours (WCT). (c) is the results using our model architecture combined with AdaIN as the stylization method, and (d) is WCT^2 (proposed).

1.6. Qualitative comparison with artistic style transfer results

We compare our proposed WCT^2 with popular artistic style transfer methods including NeuralStyle [3], AdaIN [4] and WCT [5] in Figure 4. To apply semantic segmentation map to the artistic style transfer methods, we followed the spatial control techniques proposed by the authors [7, 4, 5] respectively. In the figure, artistic style transfer methods generate undesired distortions and artifacts and often fail to maintain the structural information despite the spatial control with segmentation maps. In comparison, because of the proposed wavelet corrected transfer, our proposed WCT^2 prevents unrealistic artifacts and preserve the structure information such as edges.

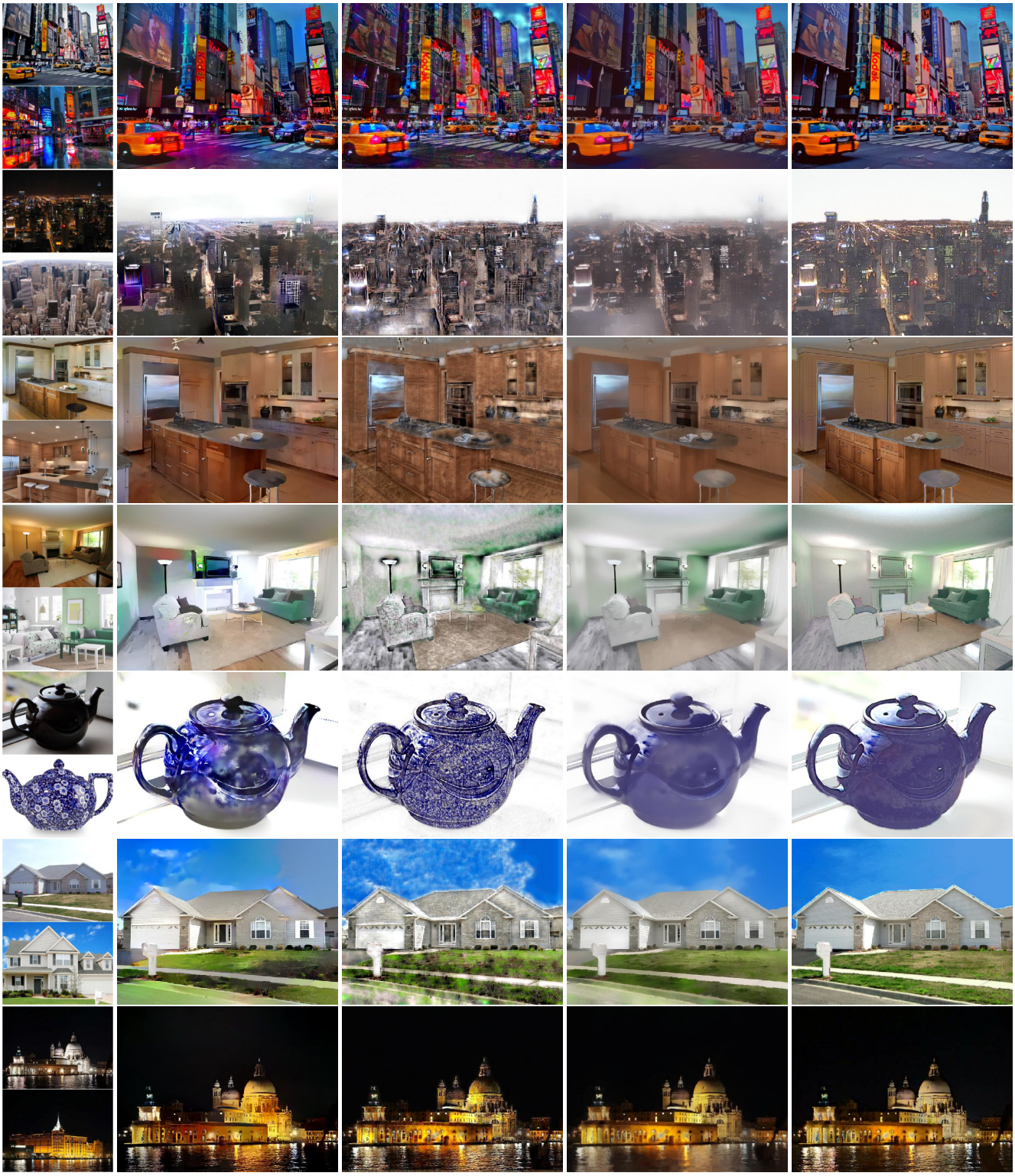
1.7. Additional Qualitative comparison with photorealistic style transfer

Additional qualitative results using WCT^2 and its variants are shown in Figure 5, Figure 6 and Figure 7. The video stylization results can be found in one of the other supplementary materials.



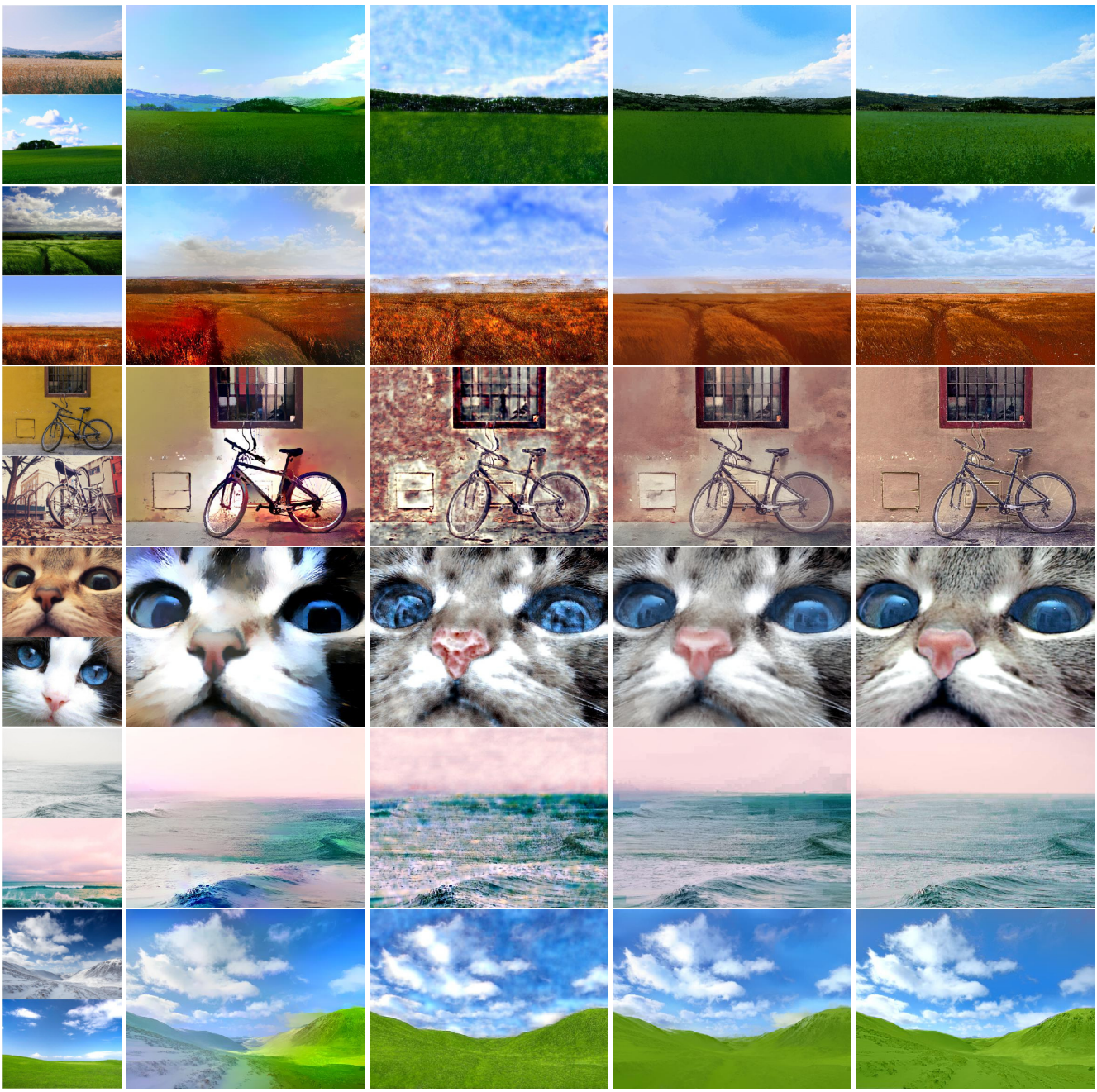
(a) Input (b) NeuralStyle [3] (c) AdaIN [4] (d) WCT [5] (e) Ours (WCT^2)

Figure 4: Qualitative comparison with artistic style transfer results. Given (a) an input pair (top: content, bottom: style), we compare the results of (b) NeuralStyle [3], (c) AdaIN [4] (d) WCT [5] and (e) ours (WCT^2).



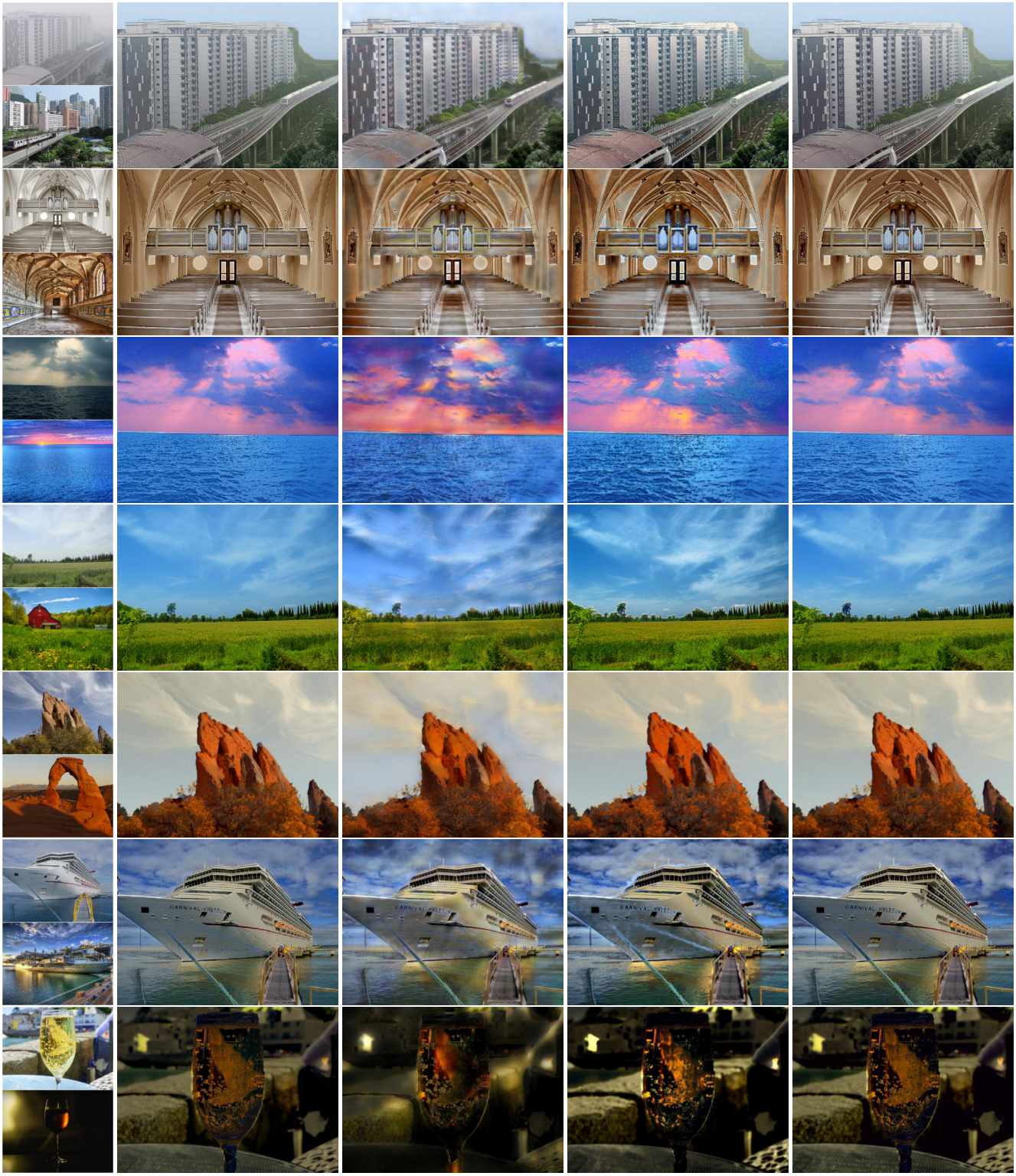
(a) Input (b) DPST [7] (c) PhotoWCT [6] (d) PhotoWCT (full) [6] (e) Ours (WCT^2)

Figure 5: Photorealistic style transfer results. Given (a) an input pair (top: content, bottom: style), we compare the results of (b) deep photo style transfer (DPST) [7], (c) and (d) PhotoWCT [6] and (e) ours (WCT^2). (c) is the results of PhotoWCT without any post-processing and (d) shows the results after applying two post-processing steps proposed by the authors [6].



(a) Input (b) DPST [7] (c) PhotoWCT [6] (d) PhotoWCT (full) [6] (e) Ours (WCT^2)

Figure 6: Photorealistic style transfer results. Given (a) an input pair (top: content, bottom: style), we compare the results of (b) deep photo style transfer (DPST) [7], (c) and (d) PhotoWCT [6] and (e) ours (WCT^2). (c) is the results of PhotoWCT without any post-processing and (d) shows the results after applying two post-processing steps proposed by the authors [6].



(a) Input (b) WCT^2 (c) WCT^2 (sum) [6] (d) WCT^2 (+multi-level) [6] (e) WCT^2 (+decoder)

Figure 7: Photorealistic style transfer results. Given (a) an input pair (top: content, bottom: style), we compare the results of WCT^2 and its variants, *i.e.*, (b) WCT^2, (c) WCT^2 (sum) (d) WCT^2 (+multi-level) and (e) WCT^2 (+decoder).

References

[1] Tony F Chan and Jianhong Jackie Shen. *Image processing and analysis: variational, PDE, wavelet, and stochastic methods*, volume 94. Siam, 2005. 1

[2] Richard J Duffin and Albert C Schaeffer. A class of nonharmonic Fourier series. *Transactions of the American Mathematical Society*, 72(2):341–366, 1952. 1

[3] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423. IEEE, 2016. 3, 4

[4] Xun Huang and Serge J Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519, 2017. 3, 4

[5] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, pages 386–396, 2017. 2, 3, 4

[6] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. *arXiv preprint arXiv:1802.06474*, 2018. 2, 5, 6, 7

[7] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. 3, 5, 6

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[9] Jong Chul Ye, Yoseob Han, and Eunju Cha. Deep convolutional framelets: A general deep learning framework for inverse problems. *SIAM Journal on Imaging Sciences*, 11(2):991–1048, 2018. 1, 2

[10] Rujie Yin, Tingran Gao, Yue M Lu, and Ingrid Daubechies. A tale of two bases: Local-nonlocal regularization on image patches with convolution framelets. *SIAM Journal on Imaging Sciences*, 10(2):711–750, 2017. 1, 2