Supplemental Material: Unifying Guided and Unguided Outdoor Image Synthesis

Muhammad Usman Rafique, Yu Zhang, Benjamin Brodie, Nathan Jacobs
University of Kentucky, Lexington, KY
{usman.rafique, y.zhang, benjamin.brodie, nathan.jacobs}@uky.edu

1. Network Architectures

We describe our network architectures in this section. We will release code for reproducibility and to support future research. Our overall approach resembles a U-Net [6]. Our encoders and docoders are mostly derived from the recent papers using adaptive instance normalization [1, 3, 4]. We use a convolutional network as a style encoder, as shown in Table 1. In the style encoder, we use kernel size 4×4 and a stride of 2×2 to reduce feature maps size. All layers use *LeakyReLU* activation and global average pooling (GAP)is used to extract global features. Table 1 shows architecture design and feature sizes for the input size 256×256 image and latent dimension 32.

We use two separate multi-layer perceptrons (MLPs) to predict unguided p and guided q distributions. Designs of these MLPs are shown in Tables 2 and 3. The unguided distribution captures the diverse plausible conditions of the source image; hence this is predicted on the basis of style and content of the source image, as shown in Table 2. The unguided MLP predicts parameters of distribution: μ_p and σ_p^2 for latent dimension of 32. The guided MLP predicts the parameters μ_q and σ_q^2 based on the style encoding of the guidance image, as shown in Table 3.

Our content encoder and image decoder are shown in Figure 1. The style vector (Figure 1 top) comes from the unguided or guided distribution, based on the mode: for unguided synthesis, this is a sample from the unguided distribution and for guided synthesis, this is a sample from the guided distribution (please see Figures 1 and 2 of the paper for an overview). An MLP predicts the parameters of adaptive instance normalization (AdaIN) [3]. In the content encoder, we use larger stride to downsample spatial size of features. For image synthesis, the decoder uses nearest-neighbor upsampling followed by convolutional layers, as this removes checkerboard artifacts [5].

2. More Visualizations

We show more qualitative results here.

Type (name)	Inputs	Output Channels	Spatial Size
conv2d (conv1)	Guidance Image	48	256×256
conv2d (conv2)	conv1	96	128×128
conv2d (conv3)	conv2	192	64×64
conv2d (conv4)	conv3	192	32×32
conv2d (conv5	conv4	192	16×16
GAP (gap1)	conv5	192	1×1
conv2d (conv6)	gap1	32	1×1

Table 1. Style encoder architecture for the latent size of 32 and image size 256×256.

Type (name)	Inputs	Output Size
linear (lin1)	conv6 style, Res4 content	288
batchnorm (bn1)	lin1	288
linear (lin2)	bn1	64
batchnorm (bn2)	lin2	32(μ_p), 32(σ_p^2)

Table 2. MLP for unguided distribution prediction. Note that *conv6 style* means style encoding of the source image and *Res4 content* means output of Residual4 block in the content encoder for the source image.

Type (name)	Inputs	Output Size
linear (lin1)	conv6 style	32
batchnorm (bn1)	lin1	32
linear (lin2)	bn1	64
batchnorm (bn2)	lin2	32(μ_q), 32(σ_q^2)

Table 3. MLP for guided distribution prediction. *Conv6 style* is the style encoding of the guidance image.

2.1. Unguided Synthesis Results

Results of unguided synthesis are shown in Figure 2. For every example, the input image is shown at the left and four images generated by sampling from the unguided distribution are displayed. We can see that our method can generate realistic renderings of the source image under diverse conditions while preserving the scene content.

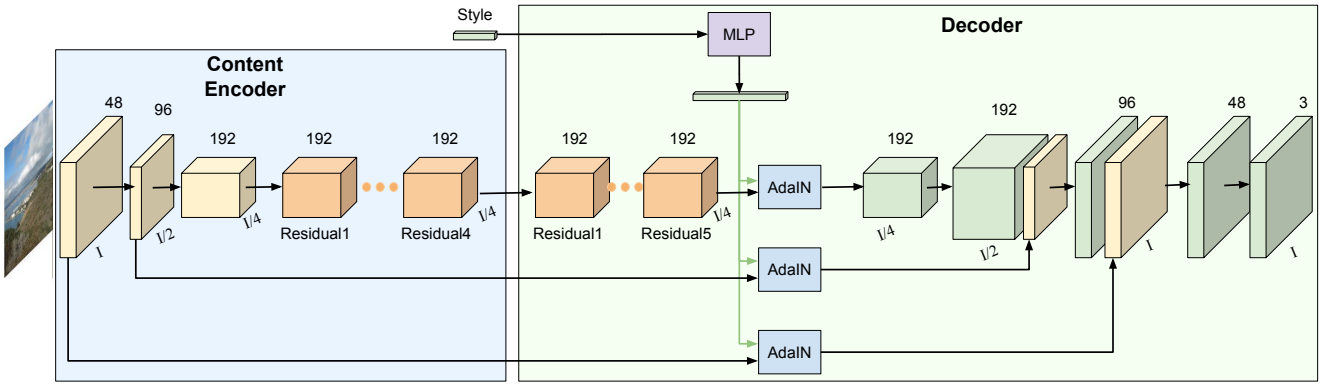


Figure 1. Architectures of content encoder (left) and image decoder (right). Feature size are shown with respect to the image size I.

2.2. Guided Synthesis Results

Qualitative results of guided synthesis are shown in Figure 3. A benefit of our probabilistic sampling is that once we predict a guided distribution based on the guidance image, we can draw multiple samples from this distribution to generate plausible images. We show two synthesized images for every example in Figure 3. We can see different sky colors and slightly different illumination settings of the two predicted images in several examples.

2.3. Time-Lapse Generation

We show more qualitative results of time-lapse generation on test set sequences from the TLVDB dataset [7]. We compare our results with a modern time-lapse generation method by Cheng *et al*. [2] that also requires true segmentation labels of source and guidance images during training and inference. Our method does not need segmentation for training or inference. We show some qualitative results in Figures 4 and 5; it can be seen that our method generates more realistic sequences with natural colors of the sky.

2.4. Style Interpolation

We now show some visualizations about style interpolation. In these examples, we linearly interpolate the style vector from source image style to the style of guidance image. The results are shown in Figure 6. Even though we have sparse training data and we do not impose any continuity constraint on the latent style representation, we can see that every interpolated style leads to a plausible appearance of that scene. We can see in Figure 6 that the appearance of synthesized images gradually changes with realistic appearance for every latent representation. For example, in rows 2-5, as we interpolate style from daytime to night, we observe realistic sunset renderings (columns (c)-(d)) before we get to the final night synthesis (f).

References

[1] Ivan Anokhin, Pavel Solovev, Denis Korzhenkov, Alexey Kharlamov, Taras Khakhulin, Aleksei Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin. High-resolution daytime translation without domain labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[2] Chia-Chi Cheng, Hung-Yu Chen, and Wei-Chen Chiu. Time flies: Animating a still image with time-lapse video as reference. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5, 6

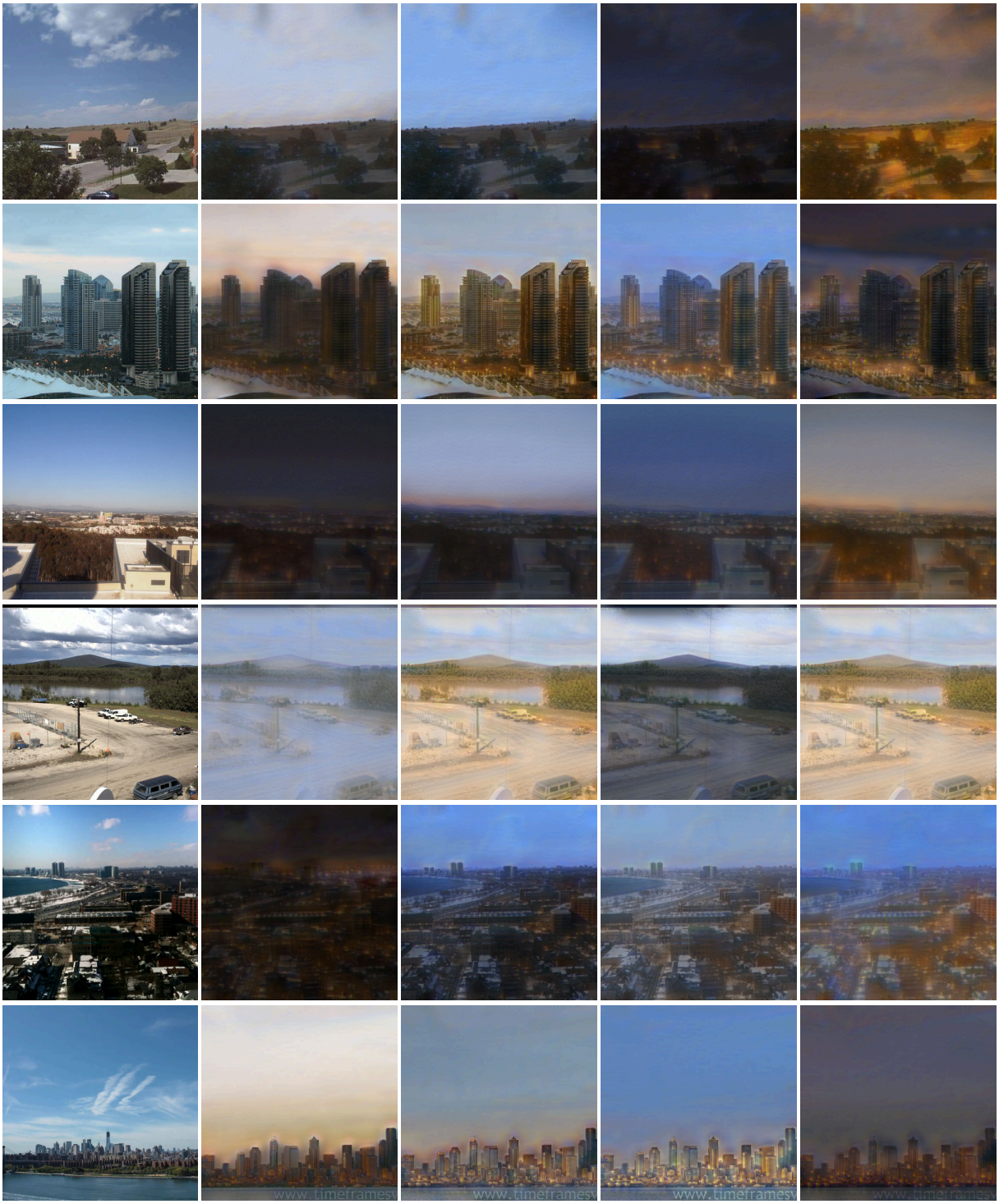
[3] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[4] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[5] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. 1

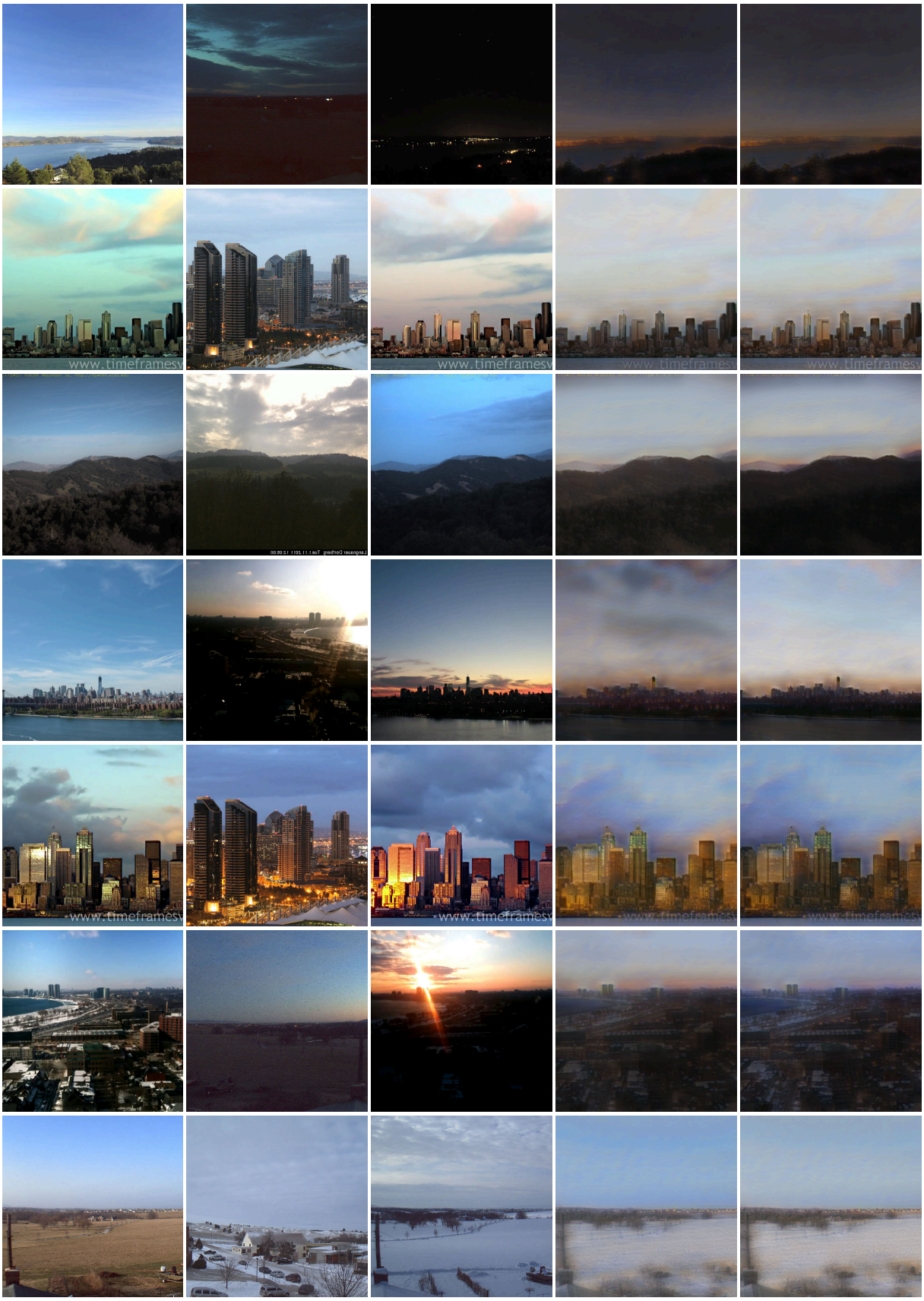
[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2015. 1

[7] Yichang Shih, Sylvain Paris, Frédo Durand, and William T Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *ACM Transactions on Graphics (TOG)*, 2013. 2



(a) Source (b) Synthesis1 (c) Synthesis2 (d) Synthesis3 (e) Synthesis4

Figure 2. Qualitative results: unguided synthesis. These results are from the unseen test set.



(a) Source (b) Guidance (c) Target (d) Synthesis 1 (e) Synthesis 2

Figure 3. Qualitative results: cross-scene guided synthesis on the test set. We show two different synthesized images, (d) and (e), which are sampled from the guided distribution q. for the given guidance image (b).

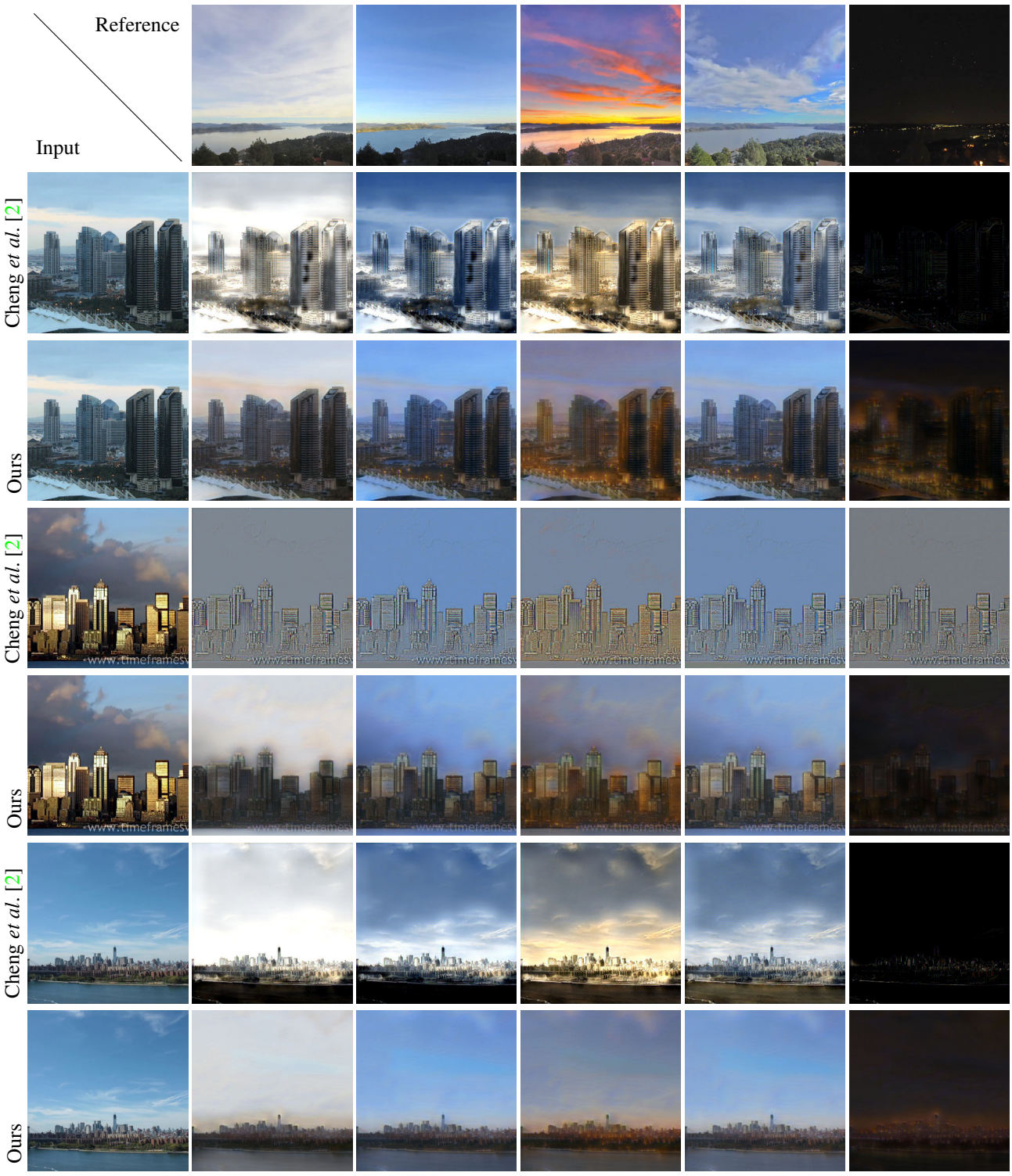


Figure 4. Time-lapse generation results. The reference time-lapse is shown on the top row and input images are shown in the left column. The method of Cheng *et al*. [2] also requires segmentation masks as input.

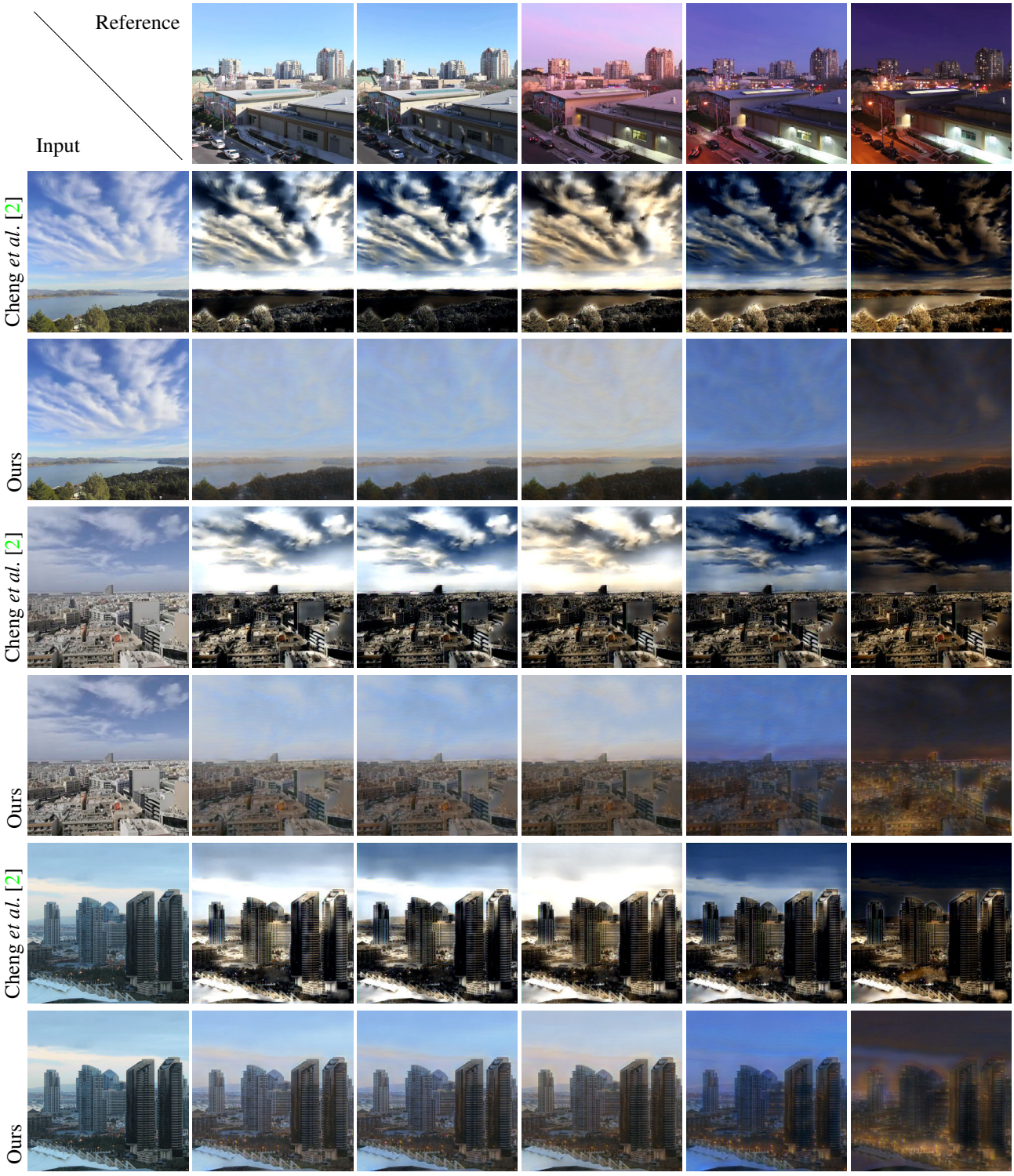
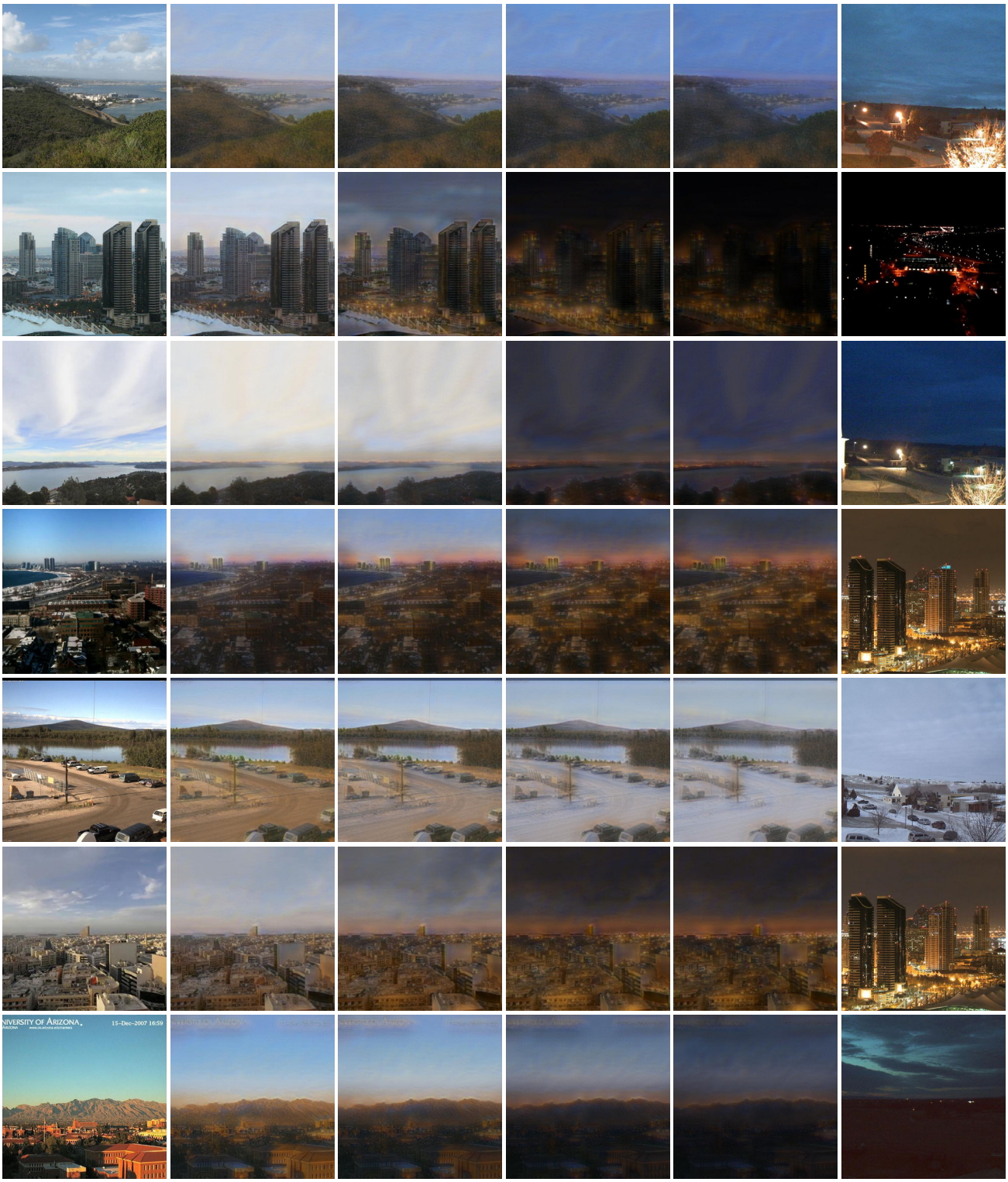


Figure 5. More Time-lapse results. The reference time-lapse is shown in the top row.



(a) Source (b) Synthesis (0.25) (c) Synthesis (0.5) (d) Synthesis (0.75) (e) Synthesis (1) (f) Guidance

Figure 6. Visualization of style interpolation. The style vector is gradually interpolated from source image (a) style to guidance image (f) style. These results are from the test set and the guidance image is from a different scene.