Bridging the Visual Gap: Wide-Range Image Blending
Supplementary Materials

Chia-Ni Lu Ya-Chu Chang Wei-Chen Chiu
National Chiao Tung University (NCTU), Taiwan
MediaTek-NCTU Research Center, Taiwan
julialu67.cs08g@nctu.edu.tw jenna.cs07g@nctu.edu.tw walon@cs.nctu.edu.tw

1. Details of Proposed Method

Here we provide more details of our proposed method, including the implementation of SHC, GRB, and the encoder-decoder architecture, as well as the overall objectives employed in our model training.

1.1. Detailed Architecture

Figure 1 illustrates our U-Net-alike model together with the Bidirectional Content Transfer (BCT) model and the revised contextual attention on skip connection. Moreover, we provide detailed illustration of Global Residual Block (GRB) and Skip Horizontal Connection (SHC) in Figure 2 and Figure 3 respectively. Our implementation of GRB is identical to the one in [4], while we adapt the original implementation of SHC in [4] to fit our scenario of wide-range image blending. To be specific, as shown in Figure 3(a), we take the feature maps of I_{left} and I_{right} extracted from a certain layer L in the encoder as well as the ones extracted from the corresponding layer of L in the decoder and concatenate them along the channel dimension, SHC is then exploited to fuse the information obtained from both encoder and decoder to generate more refined feature maps. Regarding the layer where we apply our revised contextual attention mechanism, besides using SHC to improve the results of the left and the right regions, we uses SHC to combine the reference information computed from the contextual attention mechanism (i.e. the reconstruction of the intermediate region based on the patches from the left and the right regions) and the feature map of the intermediate region extracted from the decoder, the illustration is shown in Figure 3(b). Please notice that for each layer, the weights of SHC used for both the left and the right regions are shared in our implementation, while the weights of SHC used for the intermediate region are not shared.

1.2. Overall Objectives

Our objective function for the self-reconstruction stage in our training procedure can be expressed as:

$$\mathcal{L}^{\text{SR}} = \mathcal{L}^{\text{SR}}_{pixel} + \mathcal{L}^{\text{SR}}_{feat\_rec} + \lambda_{mrf}\mathcal{L}^{\text{SR}}_{mrf} + \mathcal{L}_{feat\_con} + \lambda_{adv}\mathcal{L}_{adv\_G}, \tag{1}$$

where λ_{mrf} and λ_{adv} are used to weight the loss functions for controlling their balance, and λ_{mrf} and λ_{adv} are set to 0.01 and 0.0018 respectively in our experiments.

Subsequently, our objective function for the fine-tuning stage in our training procedure can be expressed as:

$$\mathcal{L}^{\text{FT}} = \mathcal{L}^{\text{SR}}_{pixel} + \mathcal{L}^{\text{FT}}_{pixel} + \mathcal{L}^{\text{SR}}_{feat\_rec} + \lambda_{mrf}\mathcal{L}^{\text{SR}}_{mrf} + \mathcal{L}_{feat\_con} + \lambda_{adv}\mathcal{L}_{adv\_G}. \tag{2}$$

It is worth noting that here we use both training examples of having I_{left} and I_{right} obtained from same image as well as the ones with I_{left} and I_{right} obtained from different images, and therefore the training objective of this stage contains loss functions for both types of training examples.

1.3. Training Details

Our implementation is based on Pytorch with Nvidia Tesla V100 SXM2 32GB GPU. The networks are trained with Adam optimizer [1], with the learning rate set to 10^{-3} for the self-reconstruction training stage, and 2×10^{-3} for the fine-tuning training stage with decaying step 50 and decaying rate 0.5. Both training stages are run for 200 epochs.

2. More Qualitative Results

2.1. Qualitative Comparisons

Here we provide more examples of qualitative comparisons with respect to several baselines (from image inpainting or outpainting) in Figure 4.

2.2. Qualitative Results

Here we provide more qualitative results in Figure 5, 6, and 7, as well as more full panoramic images in Figure 8, 9 and 10, produced by our proposed method.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[2] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 5

[3] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[4] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 5

[5] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[6] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[7] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[8] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 5

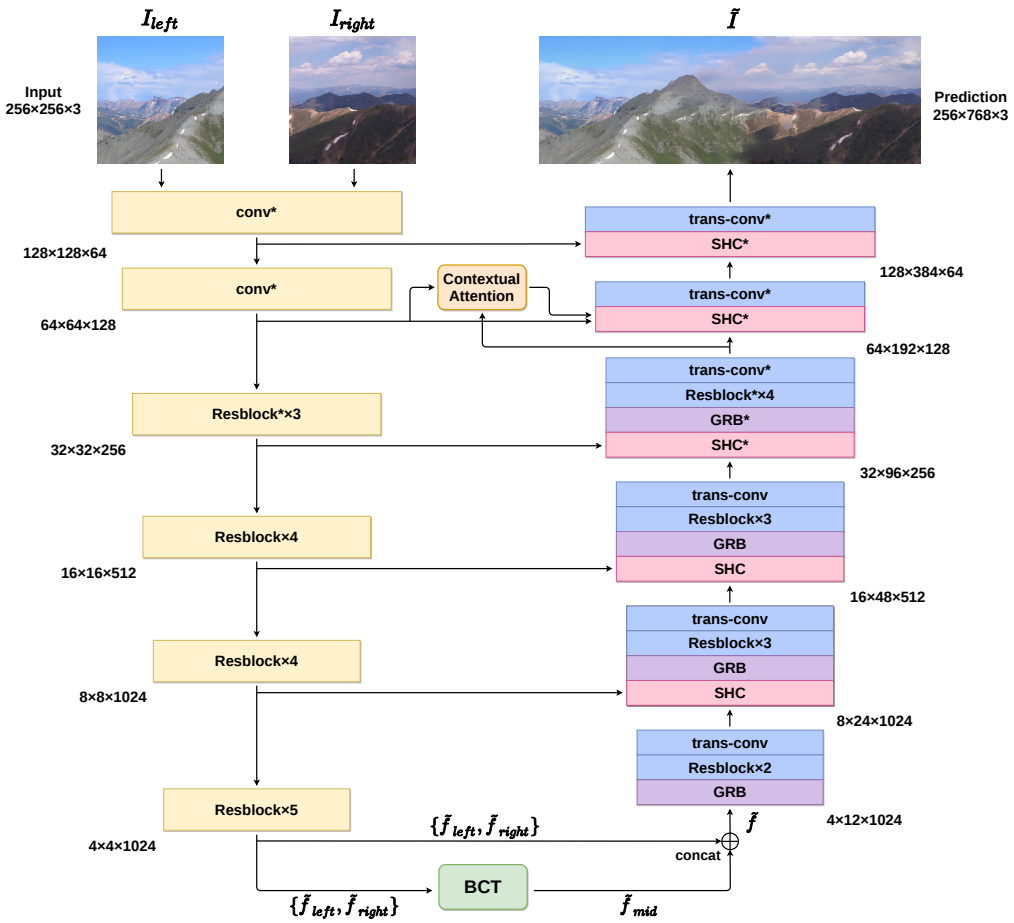


Figure 1: Illustration of our full model architecture. Our full model takes I_{left} and I_{right} as input, and compresses them into compact feature representations $\tilde{f}_{left}$ and $\tilde{f}_{right}$ individually via the encoder. Afterwards, our novel Bidirectional Content Transfer (BCT) module is used to predict $\tilde{f}_{mid}$ from $\tilde{f}_{left}$ and $\tilde{f}_{right}$. Lastly, based on the feature $\tilde{f}$, which is obtained by concatenating $\{\tilde{f}_{left}, \tilde{f}_{mid}, \tilde{f}_{right}\}$ along the horizontal direction, the decoder $\mathcal{D}$ generates our final result $\tilde{I}$. Noting that there is a contextual attention mechanism on the skip connection between the encoder and decoder, which helps to enrich the texture and details of our blending result. Please notice that all the instance normalization is removed in the layers marked with "*".

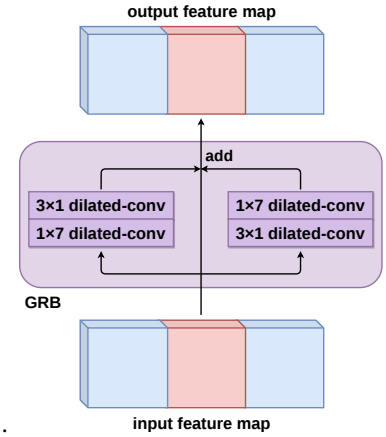
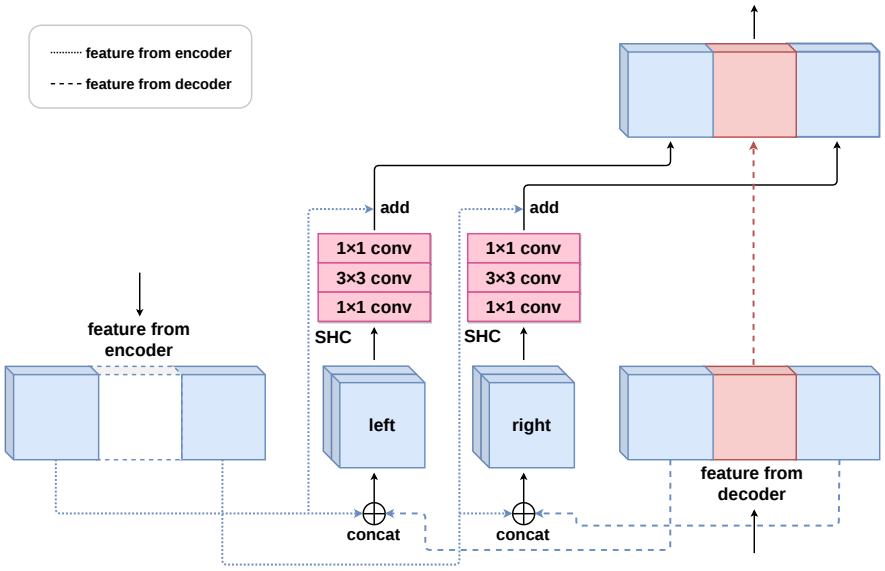
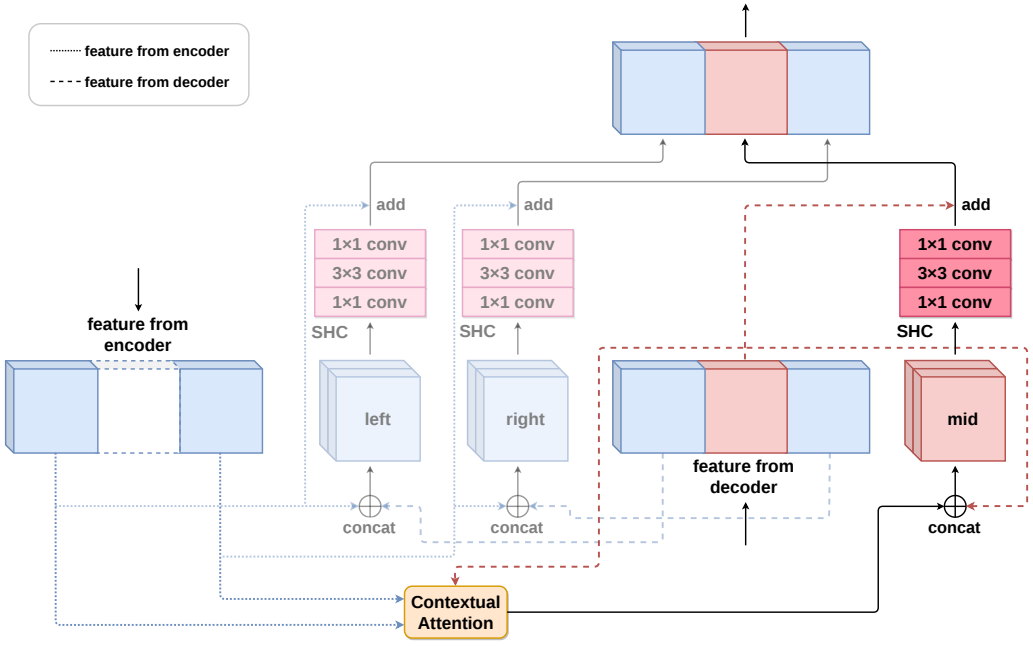


Figure 2: Illustration for the architecture of Global Residual Block (GRB).



(a) SHC w/o Contextual Attention.



(b) SHC w/ Contextual Attention.

Figure 3: Illustration for the architecture of Skip Horizontal Connection (SHC).

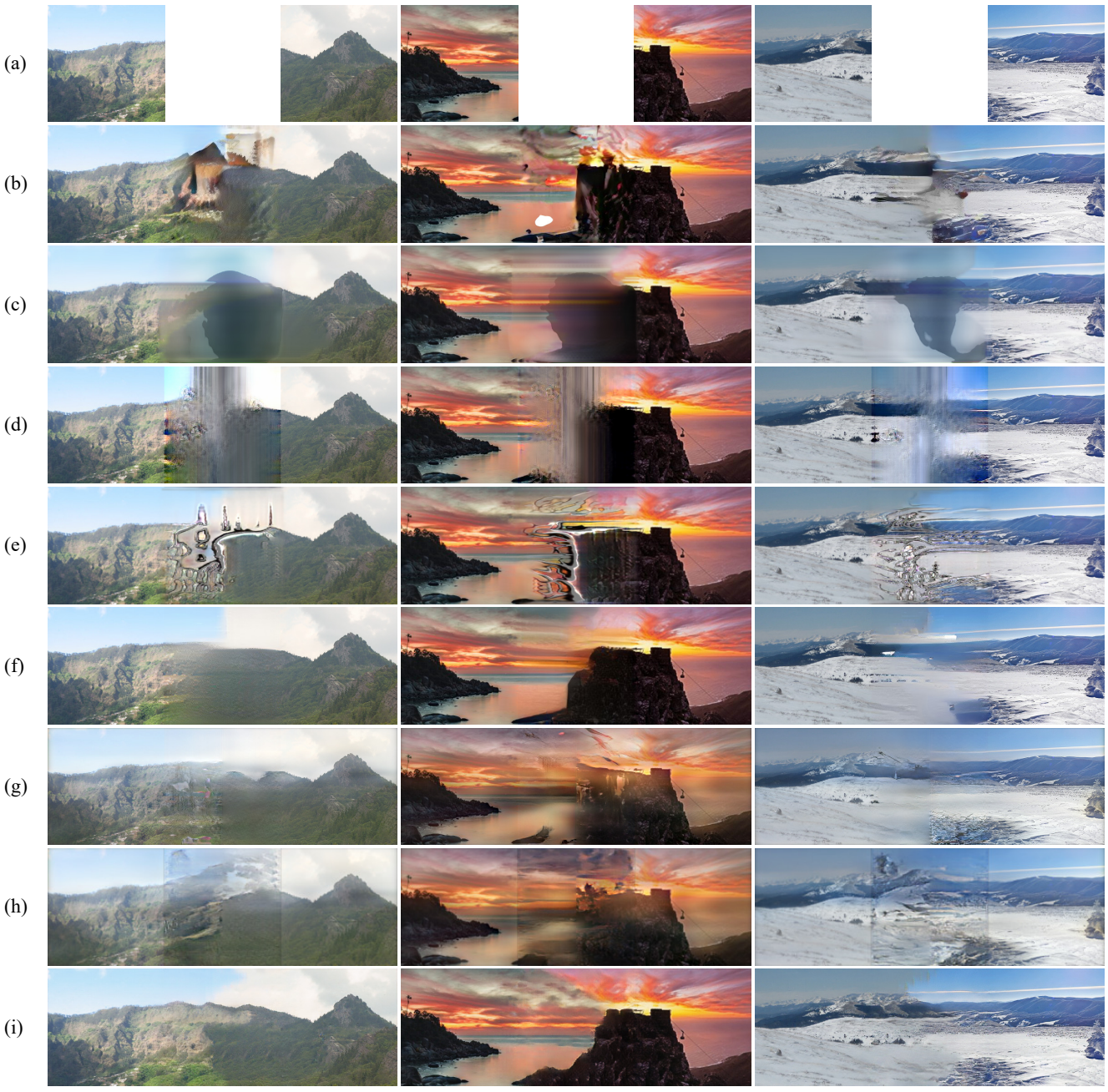


Figure 4: Qualitative comparison with respect to several baselines of image inpainting and image outpainting: (a) input images, (b) CA [6], (c) PEN-Net [7], (d) StructureFlow [2], (e) HiFill [5], (f) ProFill [8], (g) SRN [3], (h) Yang *et al*. [4], and (i) Ours.

Figure 5: Qualitative examples from our proposed method (noting that for each image here, its leftmost one-third and the rightmost one-third are the inputs I_{left} and I_{right} respectively).

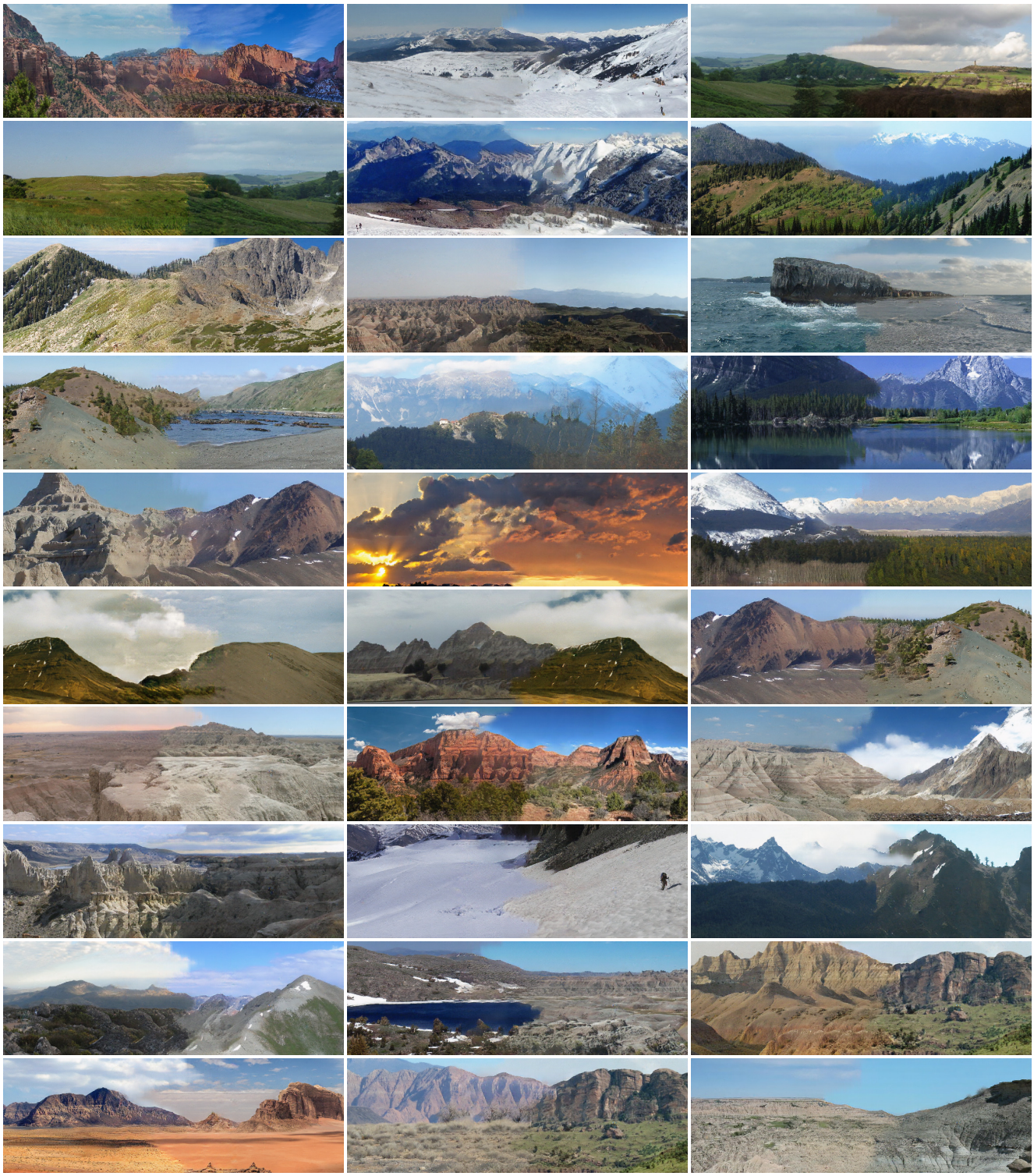


Figure 6: Qualitative examples from our proposed method (noting that for each image here, its leftmost one-third and the rightmost one-third are the inputs I_{left} and I_{right} respectively).

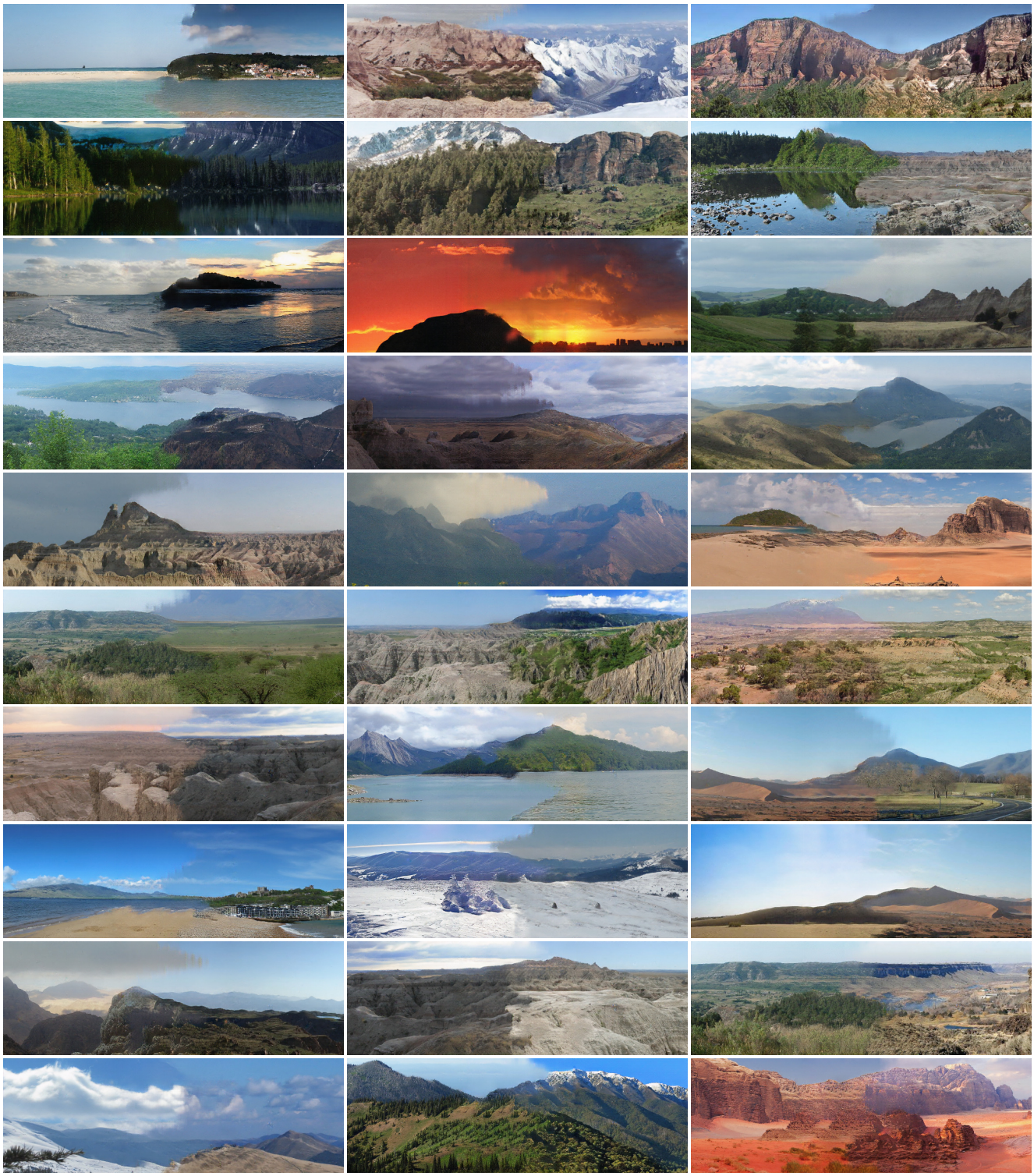


Figure 7: Qualitative examples from our proposed method (noting that for each image here, its leftmost one-third and the rightmost one-third are the inputs I_{left} and I_{right} respectively).

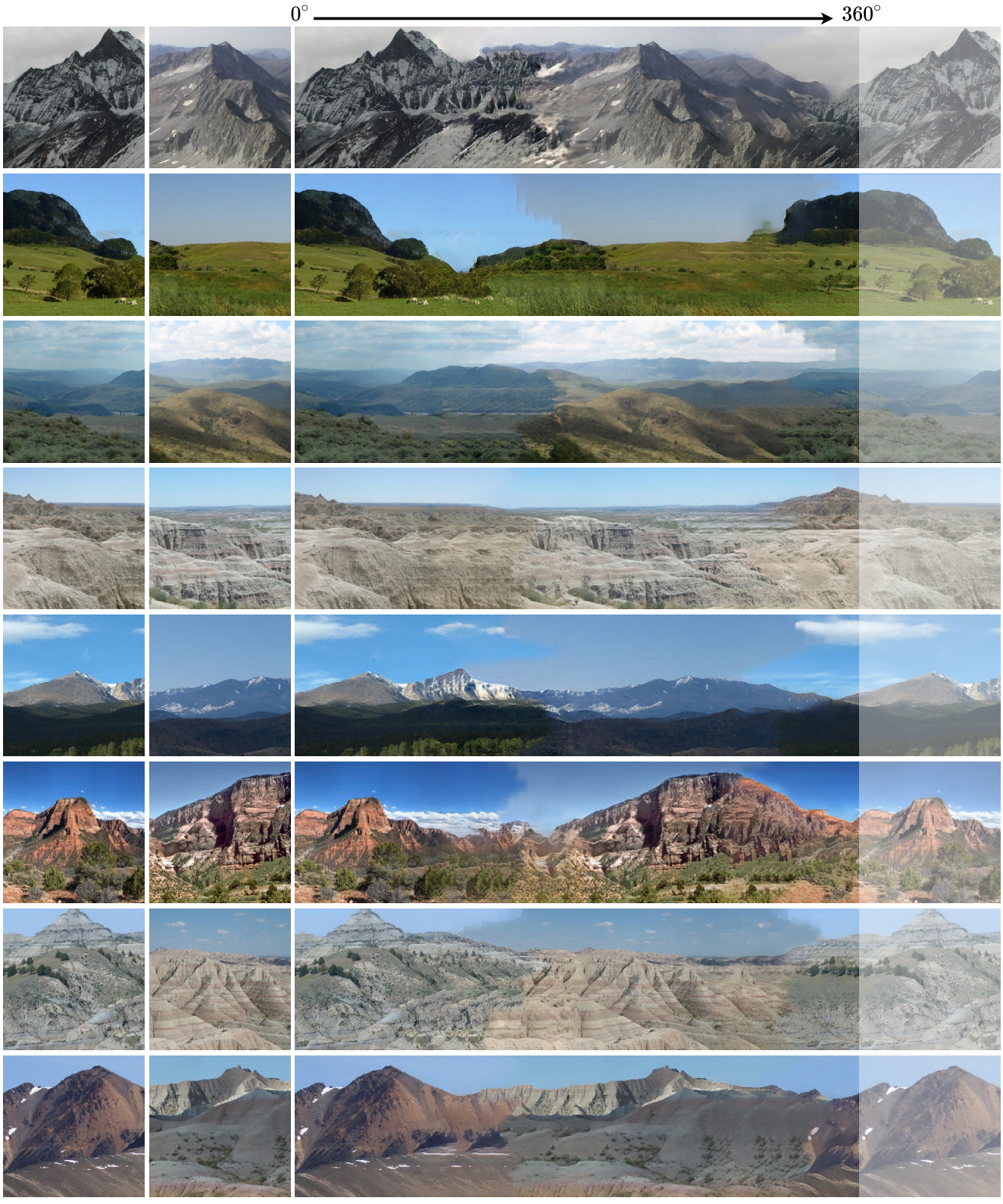


Figure 8: Example results of full panoramic images. Given two different input image (the first and the second columns), our method can construct a full panoramic image (the third column) that provides cyclic view by stitching the two blending results generated from two opposite spatial arrangements (i.e. first $\rightarrow$ second; and second $\rightarrow$ first).

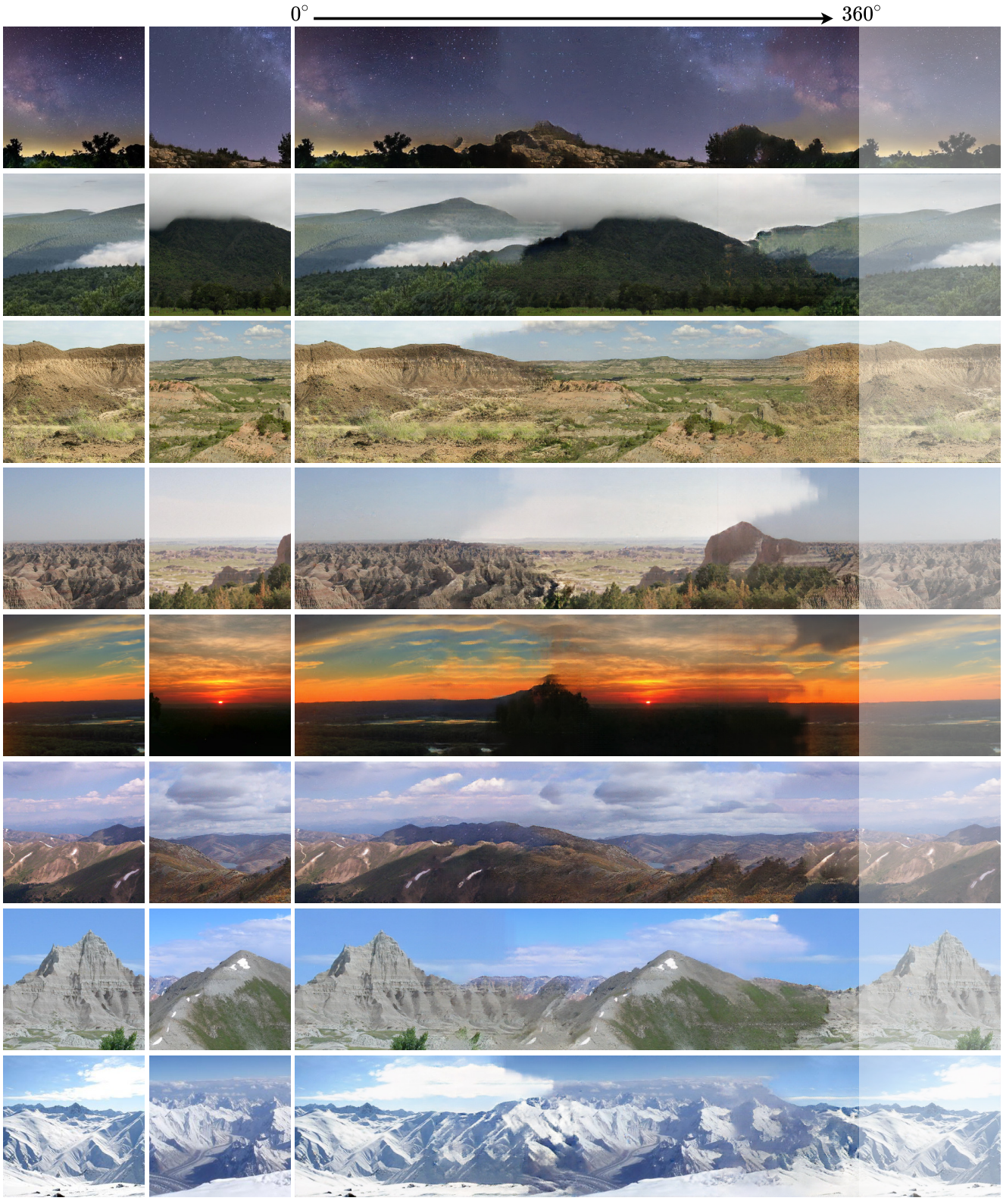


Figure 9: Example results of full panoramic images. Given two different input image (the first and the second columns), our method can construct a full panoramic image (the third column) that provides cyclic view by stitching the two blending results generated from two opposite spatial arrangements (i.e. first $\rightarrow$ second; and second $\rightarrow$ first).

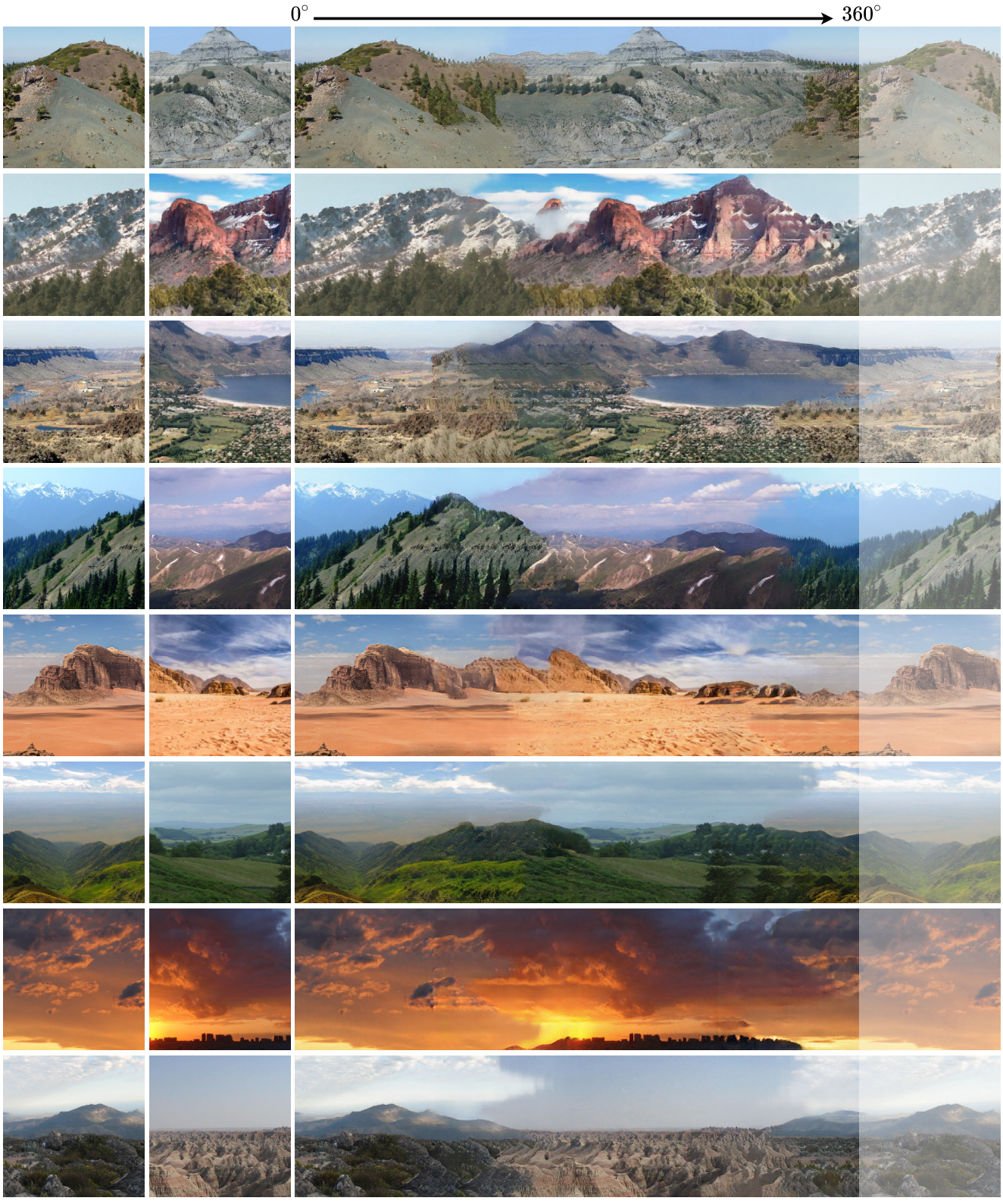


Figure 10: Example results of full panoramic images. Given two different input image (the first and the second columns), our method can construct a full panoramic image (the third column) that provides cyclic view by stitching the two blending results generated from two opposite spatial arrangements (i.e. first $\rightarrow$ second; and second $\rightarrow$ first).