Supplementary Material: Deep Image Harmonization

Yi-Hsuan Tsai[1] Xiaohui Shen[2] Zhe Lin[2] Kalyan Sunkavalli[2] Xin Lu[2] Ming-Hsuan Yang[1]
[1]University of California, Merced [2]Adobe Research
[1]{ytsai2, mhyang}@ucmerced.edu [2]{xshen, zlin, sunkaval, xinl}@adobe.com

1. Comparison between Joint Training and Separate Model

To validate the effectiveness of our joint training scheme, we also try an alternative of incorporating an off-the-shelf state-of-the-art scene parsing model [3] into our single encoder-decoder harmonization framework to provide semantic information. This network architecture is shown in Figure 1. We show quantitative comparisons on our synthesized dataset in Table 1 and 2. The MSE and PSNR of the results generated from the framework with the separate scene parsing model is worse than our joint model, where the scene parsing decoder is learned from scratch as described in the main manuscript. It shows that our joint training scheme can achieve better harmonization results than the one based on separate training.

Table 1. Comparisons of different methods on three synthesized datasets using mean-squared errors (MSE) .

	MSCOCO	MIT-Adobe	Flickr
copy-and-paste	400.5	552.5	701.6
Lalonde [1]	667.0	1207.8	2371.0
Xue [2]	351.6	568.3	785.1
Zhu [4]	322.2	360.3	475.9
Ours (w/o semantics)	80.5	168.8	491.7
Ours (separate model)	86.0	227.5	511.2
Ours (joint training)	**76.1**	**142.8**	**406.8**

Table 2. Comparisons of different methods on three synthesized datasets using PSNR.

	MSCOCO	MIT-Adobe	Flickr
cut-and-paste	26.3	23.9	25.9
Lalonde [1]	22.7	21.1	18.9
Xue [2]	26.9	24.6	25.0
Zhu [4]	26.9	25.8	25.4
Ours (w/o semantics)	32.2	27.5	27.2
Ours (separate model)	32.3	27.2	26.7
Ours (joint training)	**32.9**	**28.7**	**27.4**

Table 3. PSNR improvement by adding semantics on MSCOCO.

	bear	giraffe	horse	zebra	boat	train	bird
PSNR diff	+1.26	+1.19	+0.94	+0.85	+0.40	+0.56	+0.24

Table 4. PSNR improvement by adding semantics with scene parsing IoUs on ADE20K.

	sky	tree	person	sofa	table
PSNR diff	+2.21	+2.81	+2.38	+1.16	+1.26
IoU	86.1	57.9	52.5	17.7	18.2

2. Results on Different Categories

Table 3 shows the results on different object categories of the MSCOCO dataset. For objects that have specific patterns (*bear, giraffe*, *zebra*), the PSNR values are improved significantly, while for categories that have diverse appearances, the improvement is marginal. Table 4 shows PSNR improvements using semantics and the corresponding scene parsing IoUs on the ADE20K test set. Similarly, more improvements are achieved for categories with stronger semantic patterns and better scene parsing results.

3. Results on Real Composite Images

We present all the results of real composite images used in our user study, including examples created by ourselves (Figure 2 to 8) and examples from [2] (Figure 9 to 16). We compare harmonization results generated by our joint model to other state-of-the-art algorithms, including Lalonde [1], Xue [2] and Zhu [4].

References

[1] J.-F. Lalonde and A. A. Efros. Using color compatibility for assessing image realism. In *ICCV*, 2007. 1, 2

[2] S. Xue, A. Agarwala, J. Dorsey, and H. Rushmeier. Understanding and improving the realism of image composites. *ACM Trans. Graph. (proc. SIGGRAPH)*, 31(4), 2012. 1, 2

[3] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ADE20K dataset. *CoRR*, abs/1608.05442, 2016. 1, 2

[4] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Learning a discriminative model for the perception of realism in composite images. In *ICCV*, 2015. 1, 2

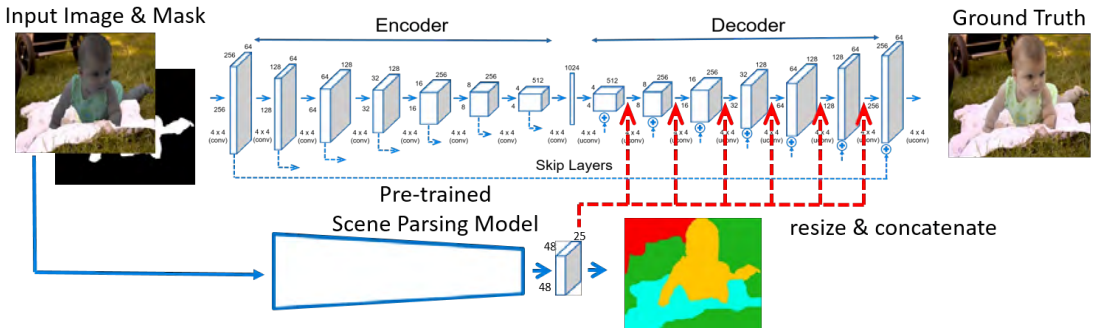
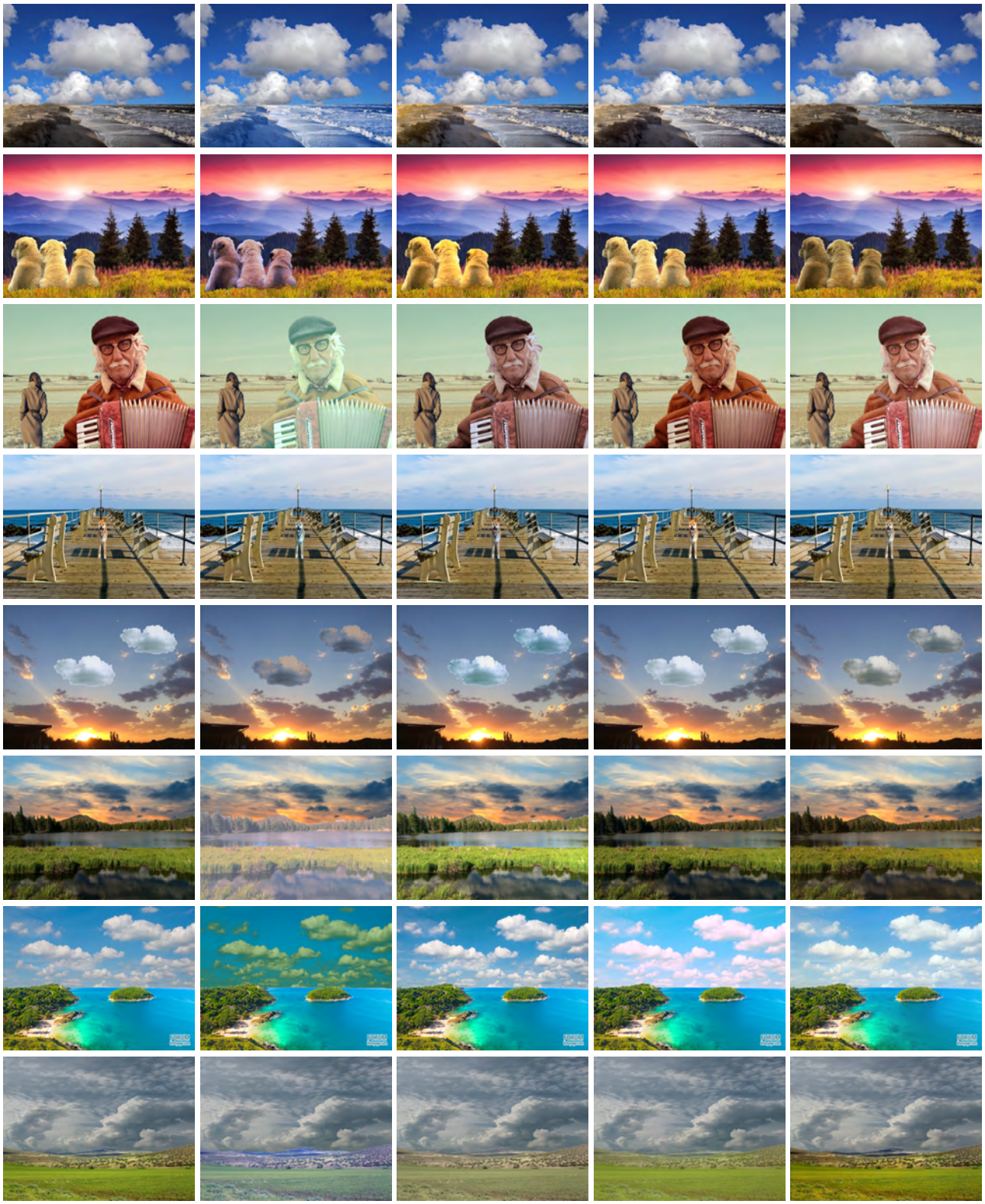


Figure 1. Overview of the network architecture that incorporates a pre-trained scene parsing model. Given a composite image and a provided foreground mask, we use the same encoder-decoder architecture for harmonization as described in the main manuscript. To propagate semantic information, we use a pre-trained scene parsing model (dilatedNet) [3] with the top 25 frequent labels in the ADE20K dataset. We first extract the response map before the output layer as semantic information. During the training process for harmonization, we then resize and concatenate this response to each deconvolutional layer in the harmonization decoder (denoted as red-dot lines). Note that, different from the proposed scene parsing decoder jointly trained with harmonization described in the main manuscript, this separate scene parsing model only provides semantic information without updating parameters through back propagation.

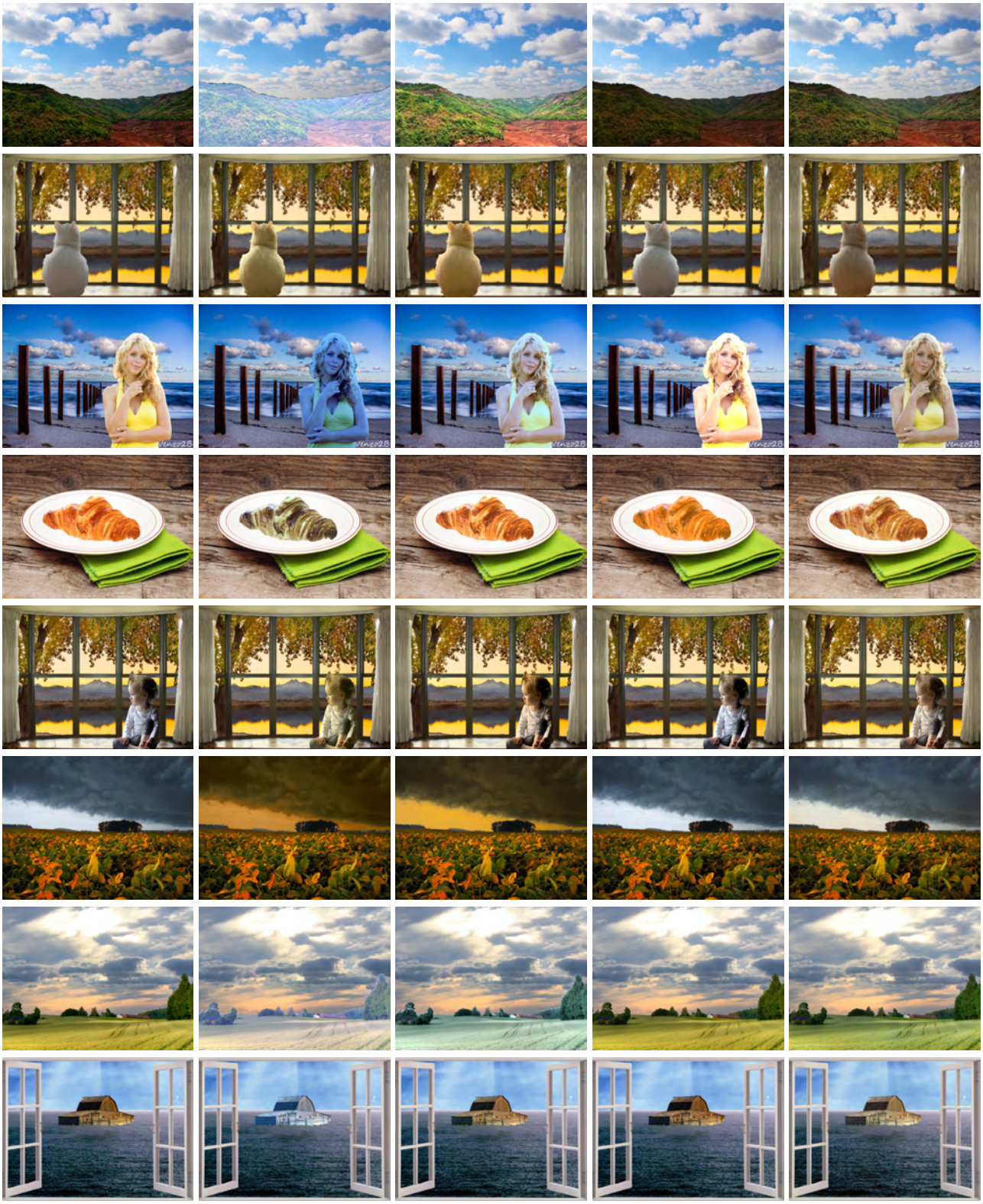
Input Lalonde [1] Xue [2] Zhu [4] Ours

Figure 2. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.



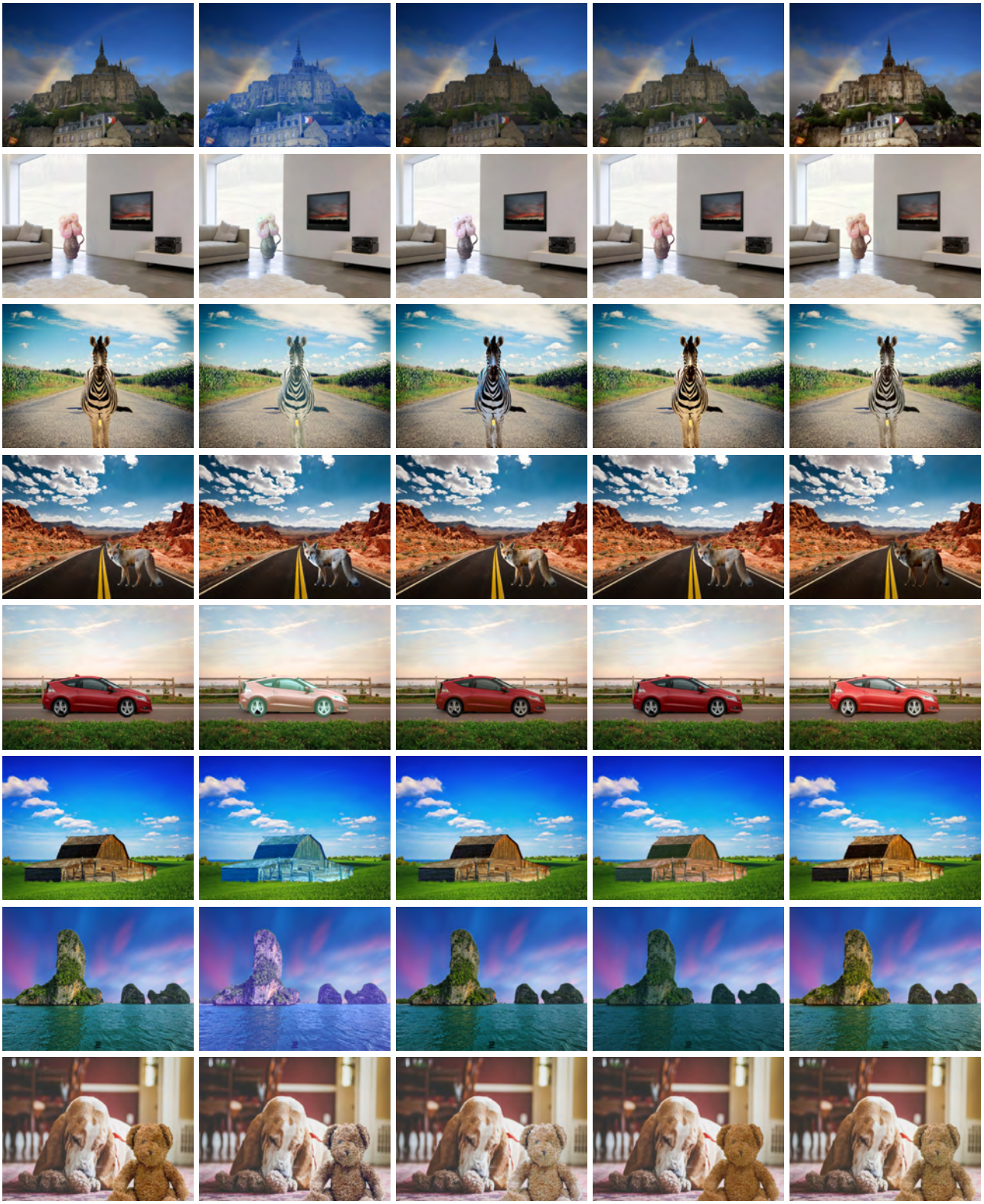
Input Lalonde [1] Xue [2] Zhu [4] Ours

Figure 3. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.



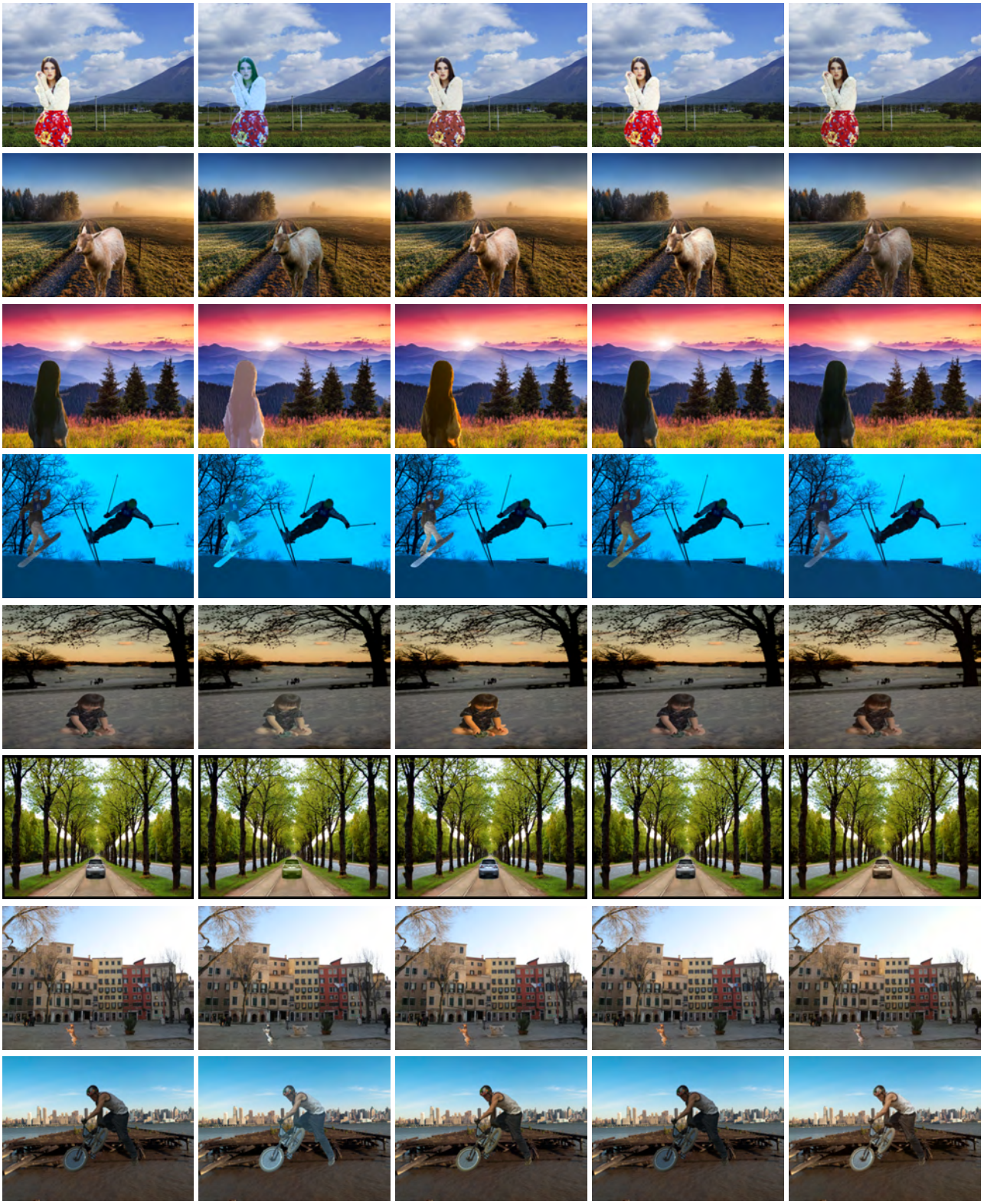
Input Lalonde [1] Xue [2] Zhu [4] Ours

Figure 4. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.



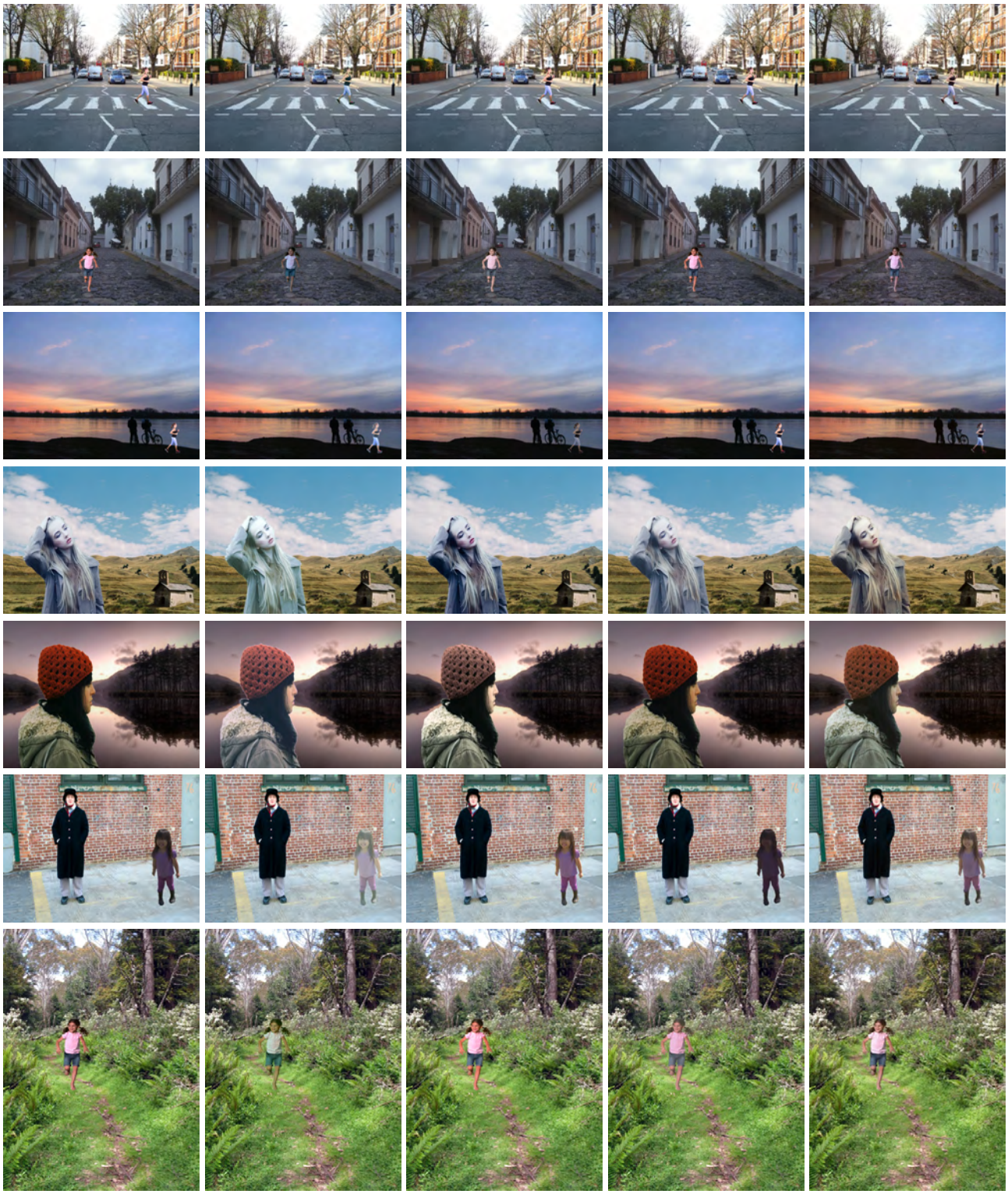
Input Lalonde [1] Xue [2] Zhu [4] Ours

Figure 5. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.



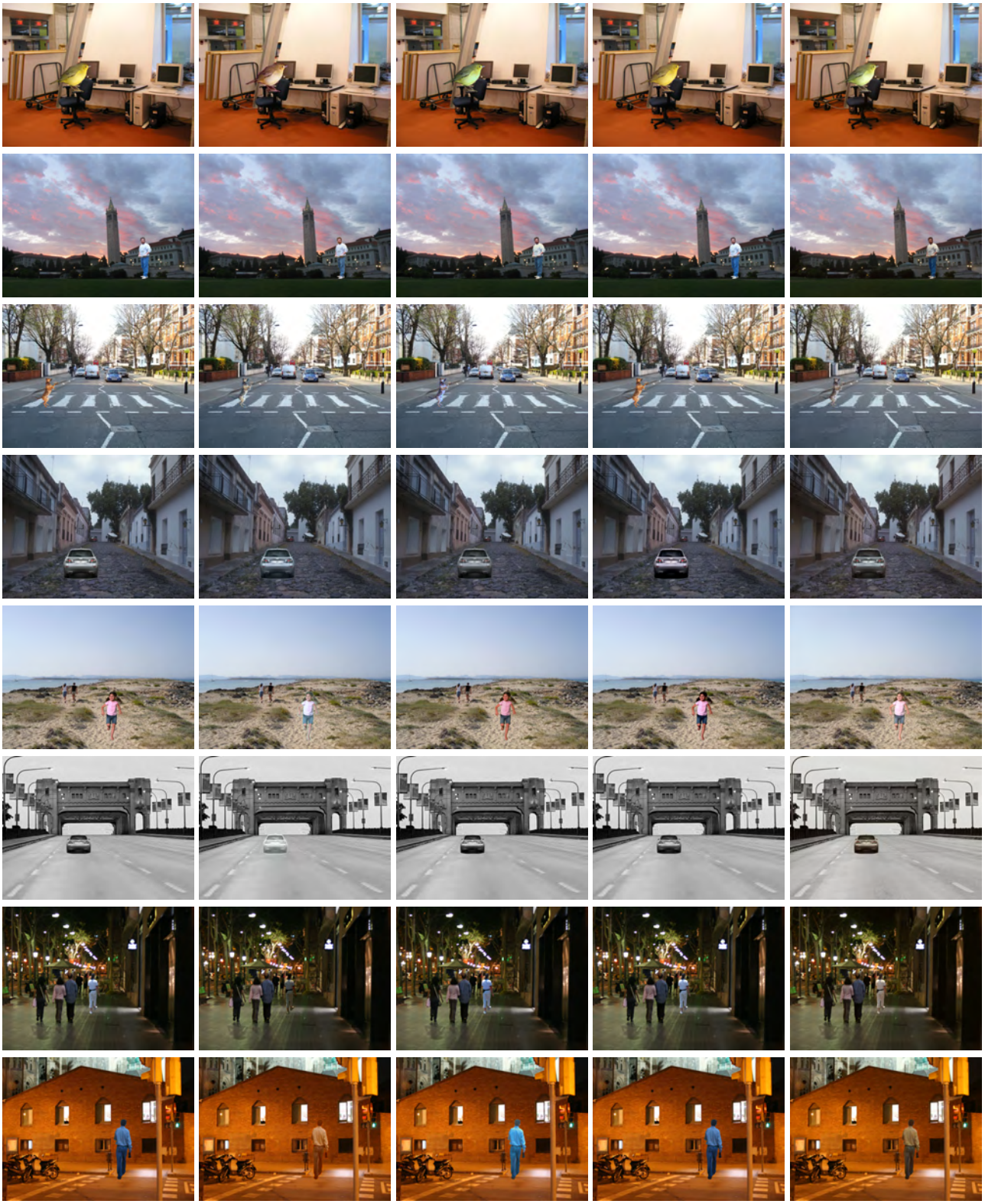
Input Lalonde [1] Xue [2] Zhu [4] Ours

Figure 6. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.



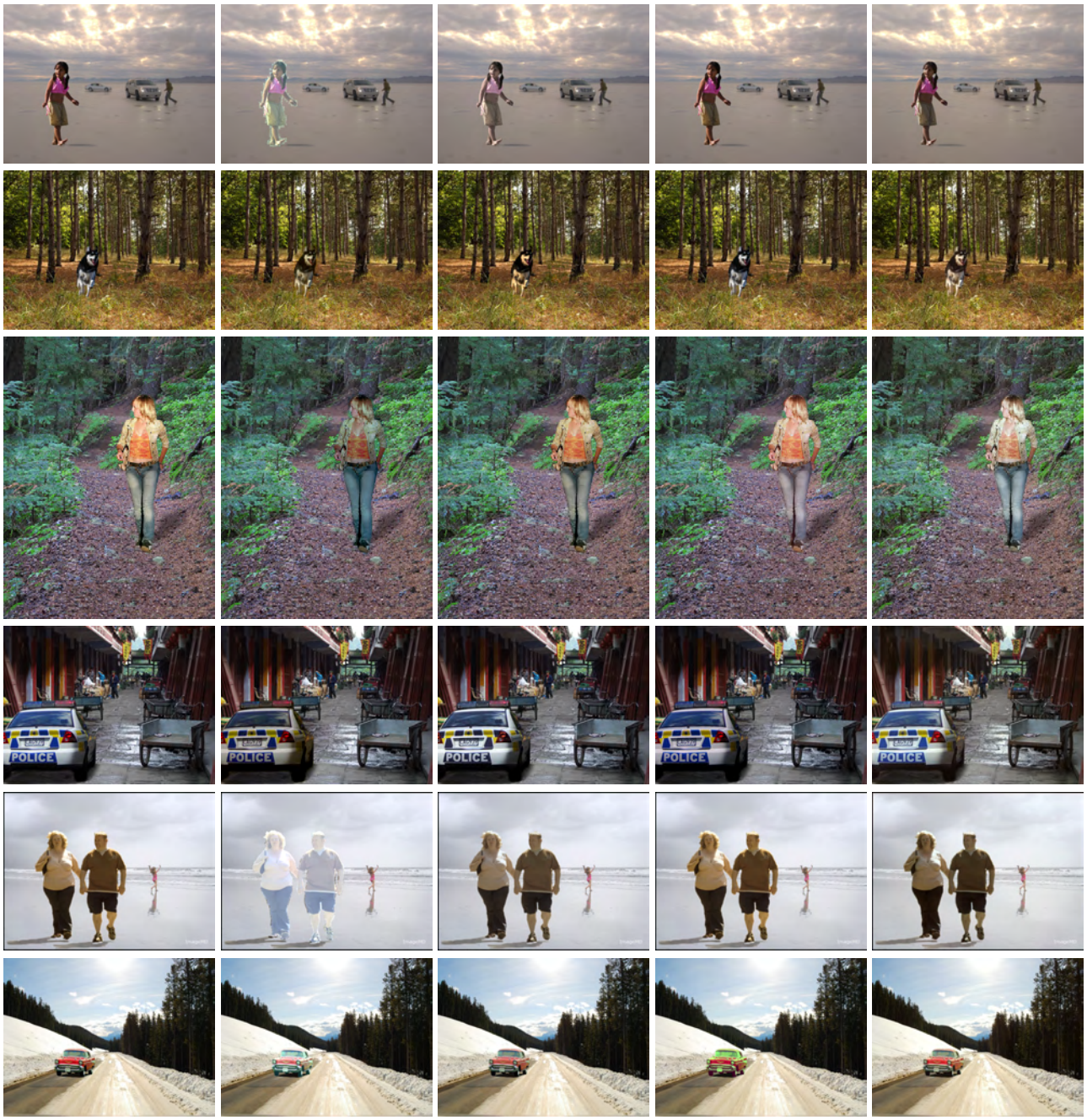
Input Lalonde [1] Xue [2] Zhu [4] Ours

Figure 7. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.



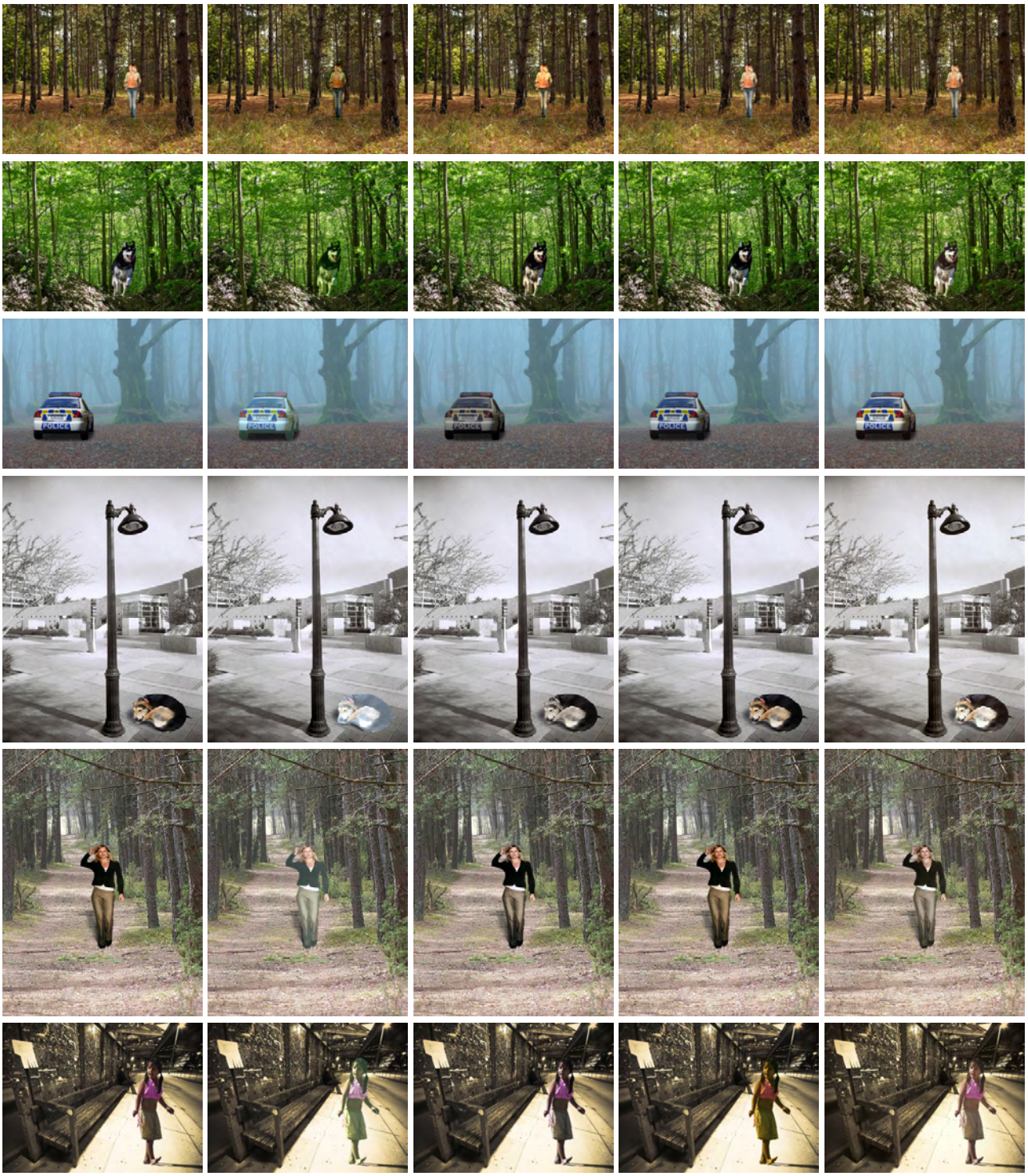
Input | Lalonde [1] | Xue [2] | Zhu [4] | Ours

Figure 8. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.



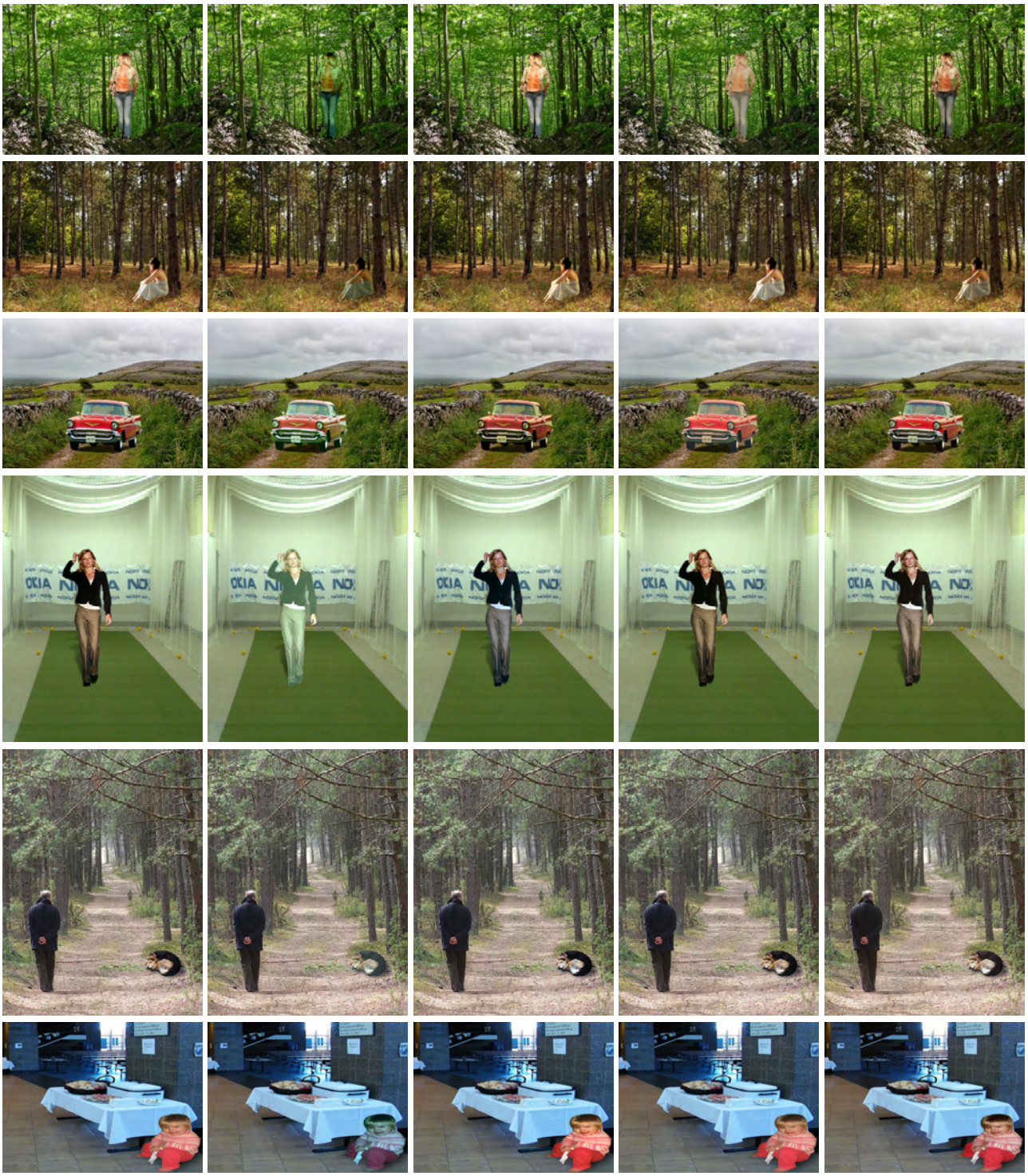
Input Lalonde [1] Xue [2] Zhu [4] Ours

Figure 9. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.



Input Lalonde [1] Xue [2] Zhu [4] Ours

Figure 10. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.



Input Lalonde [1] Xue [2] Zhu [4] Ours

Figure 11. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.

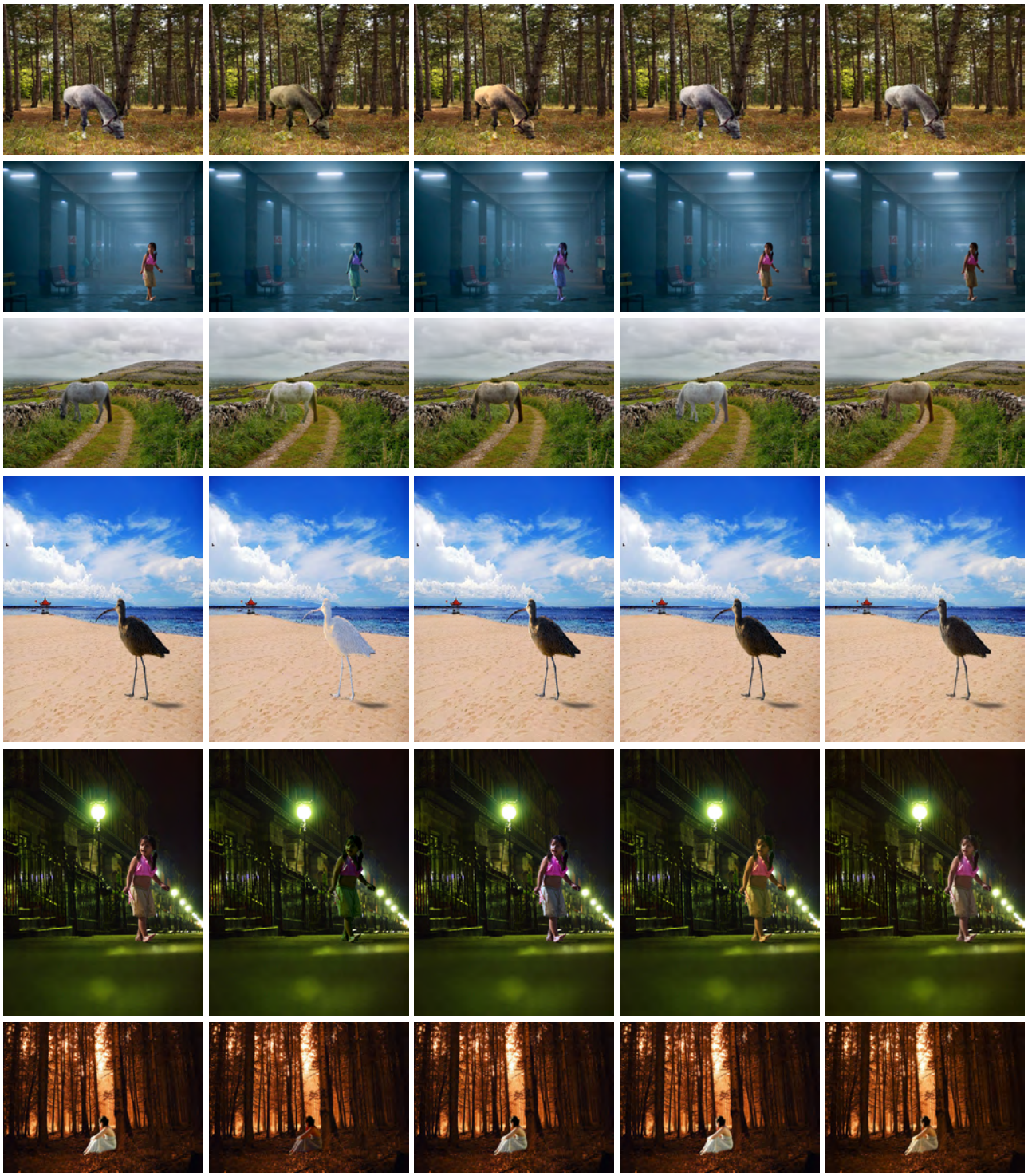


Figure 12. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.

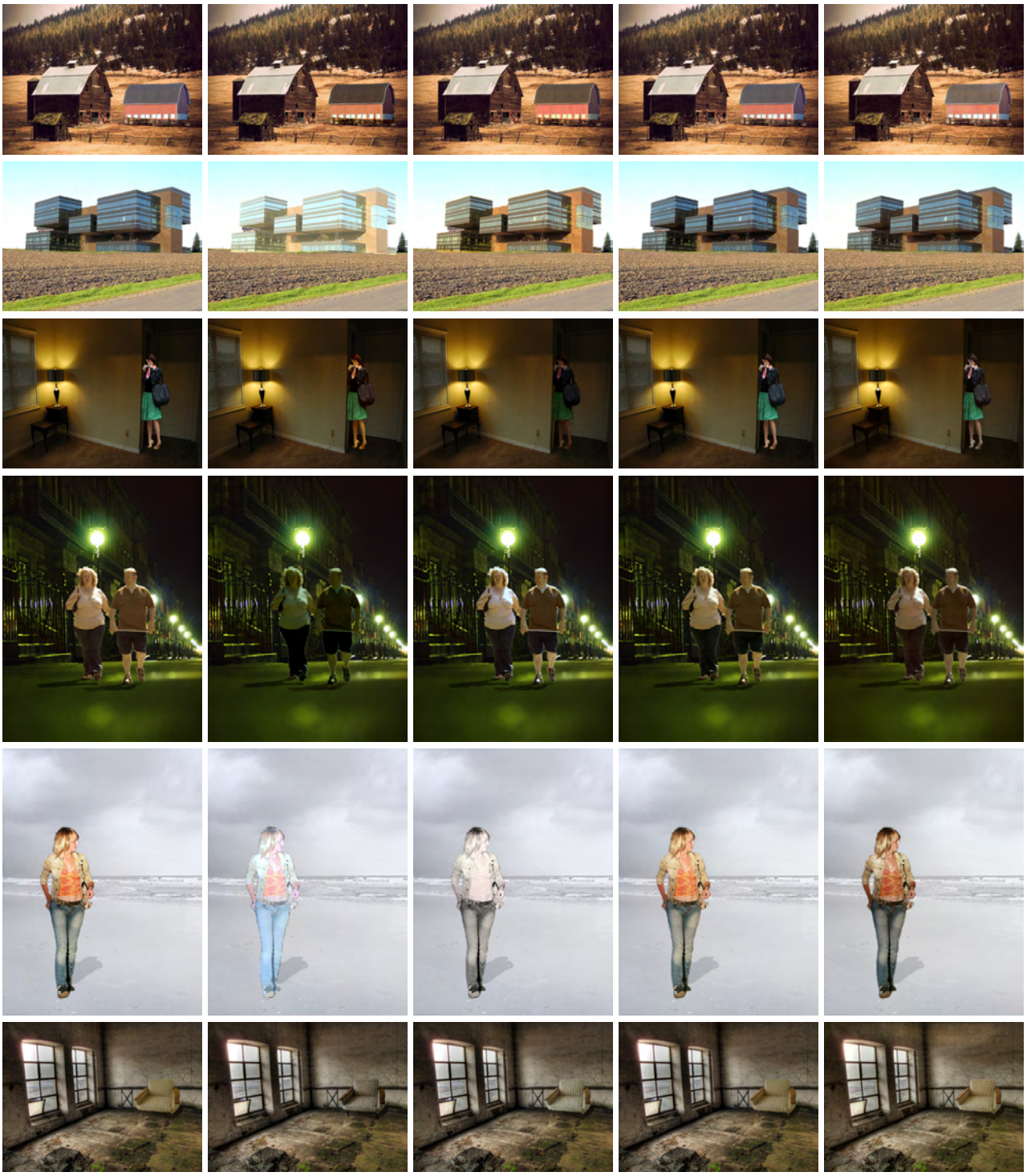


Figure 13. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.

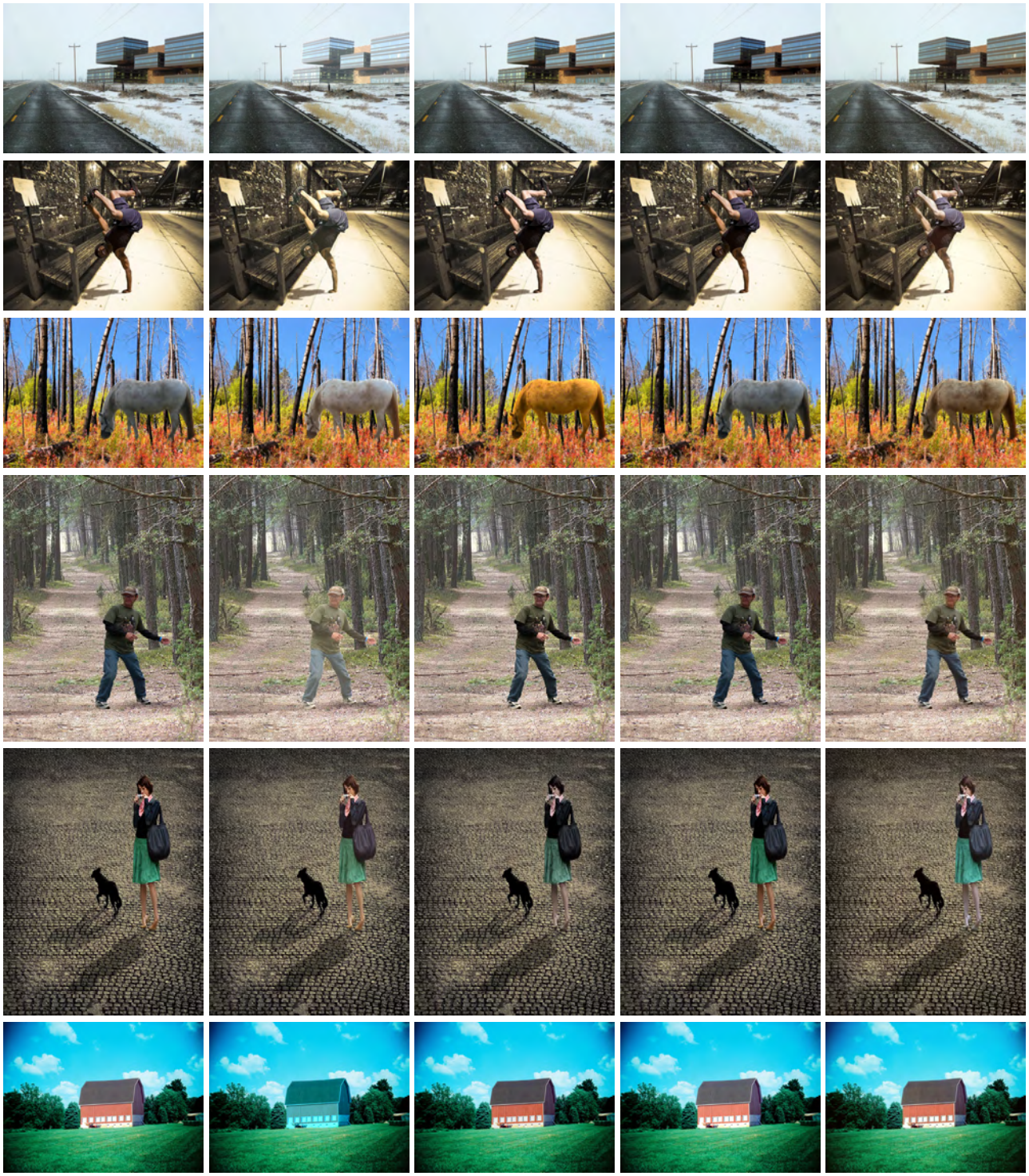
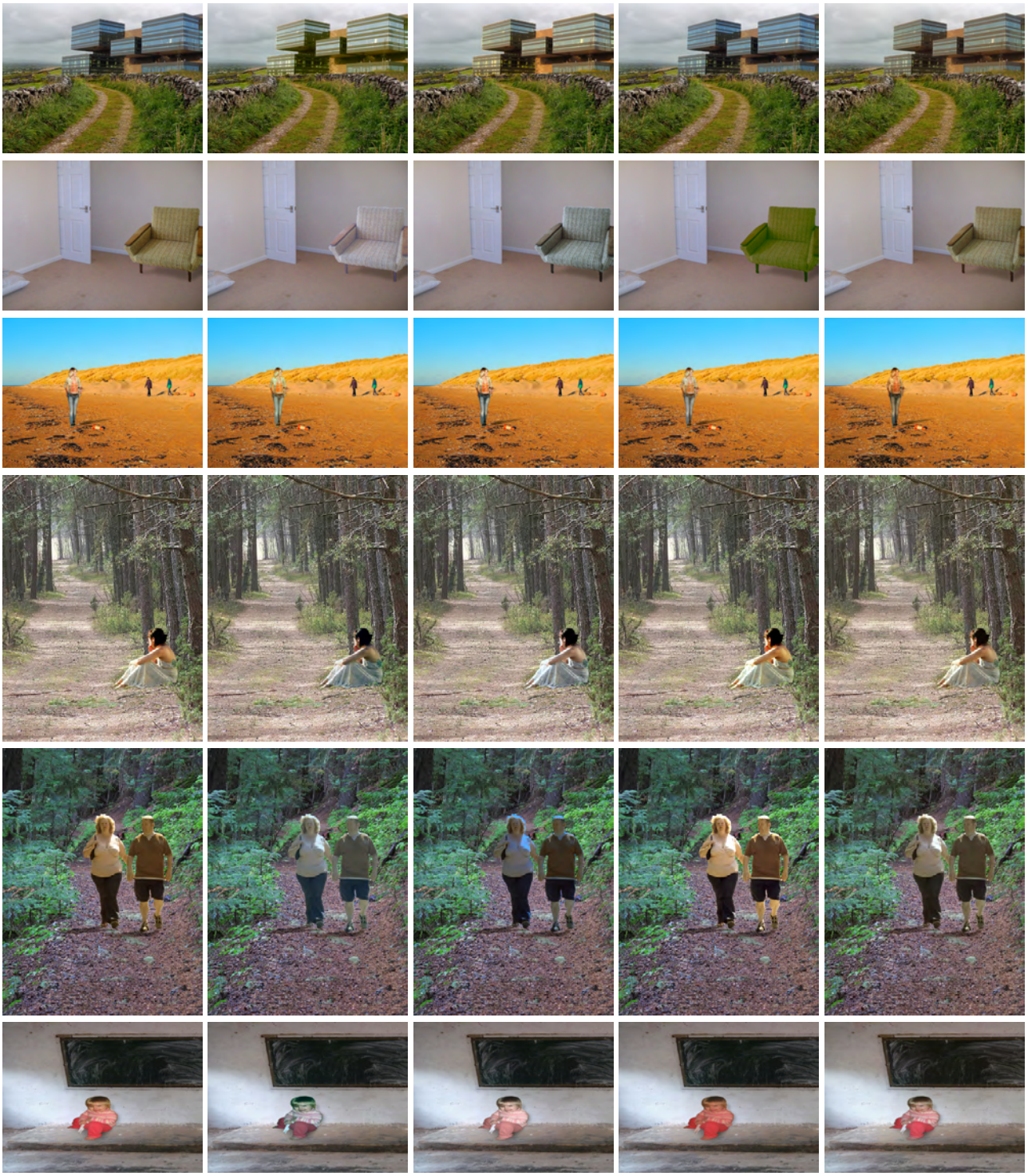
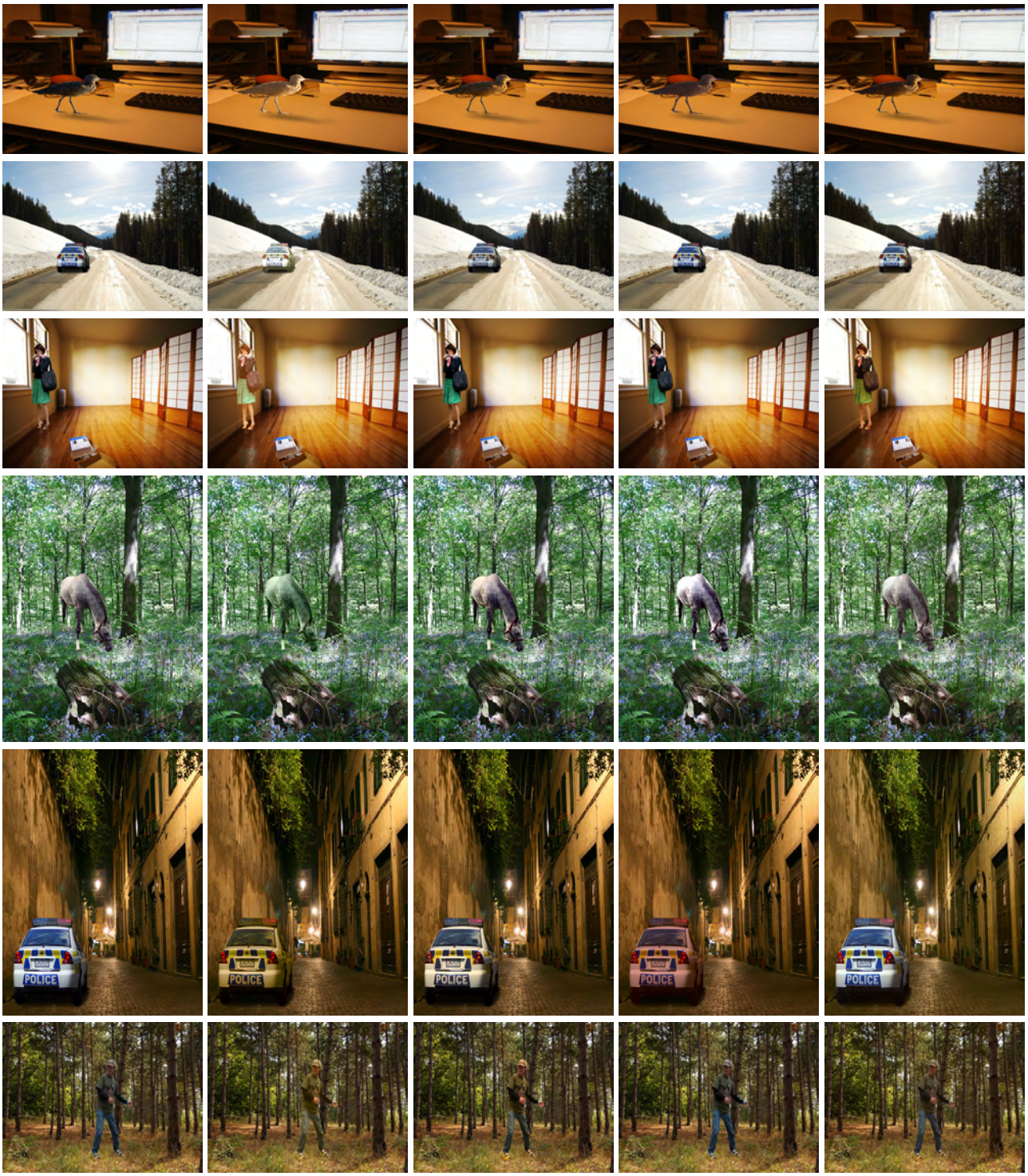


Figure 14. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.



Input | Lalonde [1] | Xue [2] | Zhu [4] | Ours

Figure 15. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.



Input Lalonde [1] Xue [2] Zhu [4] Ours

Figure 16. Sample results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.