Supplementary Material
Shadow Harmonization for Realistic Compositing

LUCAS VALENÇA, Université Laval, Canada
JINSONG ZHANG, Université Laval, Canada
MICHAËL GHARBI, Adobe, USA
YANNICK HOLD-GEOFFROY, Adobe, USA
JEAN-FRANÇOIS LALONDE, Université Laval, Canada

ACM Reference Format:
Lucas Valença, Jinsong Zhang, Michaël Gharbi, Yannick Hold-Geoffroy, and Jean-François Lalonde. 2023. **Supplementary Material** Harmonization for Realistic Compositing. In *SIGGRAPH Asia 2023 Conference Papers (SA Conference Papers '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 39 pages. https://doi.org/10.1145/3610548.3618227

1 IMPLEMENTATION DETAILS

As stated in the main paper, for the generator network we use a UNet [Ronneberger et al. 2015] with fixed update initialization [Zhang et al. 2019] (implementation from [Griffiths et al. 2022]) and a simple patch discriminator [Isola et al. 2017]. Diagrams of our generator can be found in the supplemental of Griffiths et al. [2022].

In the PyTorch implementation, our generator has as input 4 channels (except for the baseline variation that received MTMT's detection as input, in which case it received 5). The output has 4 channels: 3 for the gain map and 1 for the detection. The output activation function is a *sigmoid*, with intermediary activation functions as *ReLU*. The UNet has 5 down layers, 3 identity layers, and 6 bottleneck layers, with a maximum amount of features of 256. All skip links are enabled. For our discriminator, a diagram can be found in fig. 1.

Our generator and discriminator were trained using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a learning rate of $1e-4$. Our generator losses had weights $5e-1, 1e2, 3e1$ for ℓ_{GAN}, ℓ_c, and ℓ_s respectively. Each model was trained for around 1000 epochs.

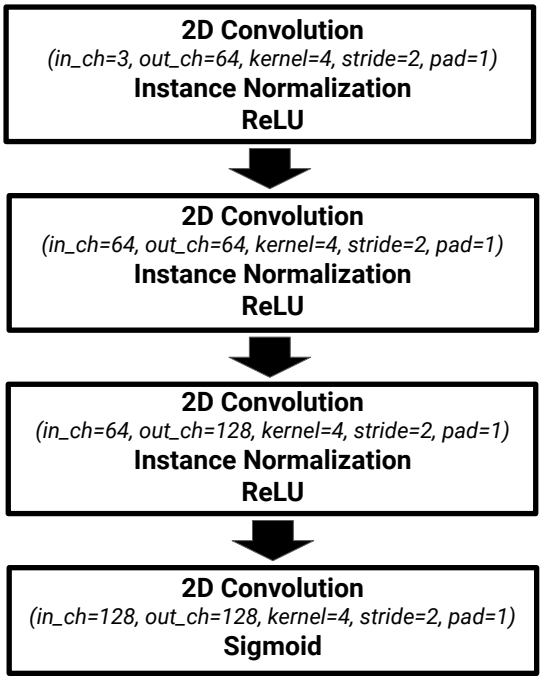

Fig. 1. **Discriminator architecture**. As implemented in PyTorch.

2 EXTRA EVALUATIONS

To complement the evaluations conducted in the main paper, we include the following tests in tab. 1 and figs. 2 and 3:

- **Ablation tests on loss functions:** in tab. 1, we evaluate how the shadow detection loss ℓ_s and the adversarial loss ℓ_{GAN} influence the performance of our model when combined with the L1 ground loss ℓ_c. More discussions in sec. 2.1.
- **Ablation tests on training datasets:** in tab. 1, we evaluate how our synthetic dataset bridges the real-synthetic domain gap and measure the impact of supplementation with shadow removal and shadow detection data. More discussions in sec. 2.2.
- **Cross-dataset validation:** in tab. 1, we evaluate how models perform on each test set when trained only on the other datasets (i.e., we train one model without ISTD, dubbed "no ISTD", and another without DESOBA, dubbed "no DESOBA"). More discussions in sec. 2.3.
- **Softer shadow samples:** in our work, we showcase seamless shadow blending by accurately detecting soft shadow edge intensities. To investigate further how soft the detected shadows can be, in figs. 2 and 3, we provide a qualitative comparison of shadow detection results from our proposed GAN and from MTMT on real images with softer shadows (as often found in cloudier weather, turbid atmospheres, or cast by detailed objects farther from the ground). All images were unseen during training. More discussions in sec. 2.4.

SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0315-7/23/12.
https://doi.org/10.1145/3610548.3618227

	ISTD [2018]						DESOBA [2022]						averages					
	SSIM$_\uparrow$	PSNR$_\uparrow$	L1$_\downarrow$	S$_\downarrow$	NS$_\downarrow$	BER$_\downarrow$	SSIM$_\uparrow$	PSNR$_\uparrow$	L1$_\downarrow$	S$_\downarrow$	NS$_\downarrow$	BER$_\downarrow$	SSIM$_\uparrow$	PSNR$_\uparrow$	L1$_\downarrow$	S$_\downarrow$	NS$_\downarrow$	BER$_\downarrow$
ℓ_c **only**	0.924	28.696	6.072	9.526	0.934	5.230	0.981	34.177	1.222	9.133	0.278	4.705	0.952	31.437	3.647	9.329	0.606	4.968
+ ℓ_s	0.929	29.952	5.222	3.197	0.698	1.948	**0.983**	35.689	0.844	3.481	0.211	1.846	**0.956**	32.821	3.033	3.339	0.454	1.897
+ ℓ_{GAN} **(ours)**	0.929	30.321	5.166	7.718	0.389	4.053	**0.983**	35.720	**0.804**	4.785	0.143	2.464	**0.956**	33.020	2.985	6.251	0.266	3.258
synthetic only	0.908	23.773	8.959	8.479	5.326	6.902	0.977	31.104	1.719	6.572	1.252	3.912	0.943	27.439	5.339	7.526	3.289	5.407
+ removal	**0.930**	**30.472**	**5.115**	5.839	0.420	3.130	**0.983**	**35.727**	0.846	4.870	0.193	2.532	**0.956**	**33.100**	**2.980**	5.355	0.307	2.831
+ detection **(ours)**	0.929	30.321	5.166	7.718	0.389	4.053	**0.983**	35.720	**0.804**	4.785	0.143	2.464	**0.956**	33.020	2.985	6.251	0.266	3.258
- synthetic	0.929	30.163	5.195	5.820	0.530	3.175	**0.983**	35.700	0.831	4.118	0.213	2.166	**0.956**	32.931	3.013	4.969	0.372	2.670
ours	0.929	30.321	5.166	7.718	0.389	4.053	**0.983**	35.720	**0.804**	4.785	0.143	2.464	**0.956**	33.020	2.985	6.251	0.266	3.258
no ISTD	0.925	29.305	5.569	7.398	0.724	4.061	**0.983**	35.637	0.835	4.043	0.227	2.135	0.954	32.471	3.202	5.720	0.476	3.098
no DESOBA	0.929	29.858	5.328	**3.057**	0.915	**1.986**	0.981	33.809	1.082	**1.992**	0.563	**1.277**	0.955	31.834	3.205	**2.525**	0.739	**1.632**
MTMT	0.925	30.120	5.173	13.360	**0.119**	6.740	0.981	34.402	0.923	21.136	**0.056**	10.596	0.953	32.261	3.048	17.248	**0.088**	8.668

Table 1. **Extra quantitative results for shadow compositing and detection.** We quantitatively evaluate our method on shadow renders on SSIM, PSNR (dB), and L1 error. We also report the precision on shadow (S) and non-shadow (NS) regions as well as the BER metric. Three evaluations are conducted (from top to bottom): loss ablations, dataset ablations, cross-dataset validation, see text for details. Colors indicate best, medium and worst performance for each method independently, and **bold-underlined** the best performance across all.

Here, all models are evaluated with "pbla" and RGB scaling (see main paper). Results for "ours" and "MTMT" are extracted from the main paper.

2.1 Loss functions

As can be seen in the top section of tab. 1, the L1 loss on ground regions ℓ_c by itself is insufficient for both detection and blending. Adding a shadow detection loss ("+ ℓ_s") improves detection significantly. Finally, our GAN loss ("ℓ_{GAN} (ours)") slightly worsens true positives for detection, but significantly improves compositing quality and level of detail (fewer false negatives).

2.2 Training data

It is clear from the results in the middle section of tab. 1 that training with synthetic data alone ("synthetic only") is unable to bridge the reality gap. Adding removal data ("+ removal") provides the best compositing scores on our test datasets. Adding detection data ("+ detection") causes a marginal loss of compositing quality, but further improves the level of detail of detected shadows by having fewer false negatives on pixels near shadow boundaries (i.e., avoiding missed shadow seams that create either shadow gaps or double shadows, due to bad soft edge estimation). This is further confirmed by its superior detection performance on the DESOBA dataset, which is more detailed and realistic (i.e., consisting of common photos instead of careful single-shadow photos like ISTD). For ISTD's shadows (which are large, connected, and coarse), downsampling the input proved to be the best option (see main paper).

Finally, removing just the synthetic data ("- synthetic") caused a loss of compositing quality and shadow detail, in exchange for slightly fewer false positives. For our intended task, this trade-off is undesirable. We hypothesize that this behavior arises from our synthetic soft shadows ground truths, resulting in sharper estimated shadow edges without synthetic data (undesirable for seamless shadow compositing).

2.3 Cross-dataset validation

The bottom section of tab. 1 shows cross-dataset validation results. For both datasets (ISTD and DESOBA) in tab. 1, it can be seen that our proposed model ("ours") surpasses the models trained with less data ("no ISTD" and "no DESOBA"), even when those were trained in-domain, with the training set of the same dataset being evaluated. This indicates that our model is generalizing as expected (i.e., not overfitting to individual distributions). Both our models trained with less data are outperformed by MTMT, which was also expected, since MTMT's publicly-available model was trained using a rather large number of image sources (including ISTD), while ours relies heavily on the two smaller shadow removal datasets, especially DESOBA's multiple instances per image (each separately-annotated shadow instance forms multiple augmented samples).

Interestingly, the model without DESOBA ("no DESOBA") was the most accurate for true positives in detailed shadows, even in DESOBA. This was achieved at the expense of a higher number of false negatives (i.e., missed details and hard-to-classify pixels, like non-binary edges). This indicates our model without DESOBA might be the best choice for users who wish to perform detailed shadow detection instead of compositing. It is also possible the model is not necessarily more accurate, but instead learned to match the binary intensities of shadow annotations better, as it was trained with all our detection data but less varied shadow removal data, which is crucial for learning shadow compositing in our pipeline.

2.4 Softer shadow samples

It can be seen in fig. 2 and fig. 3 that our model learned to generalize the soft shadow annotations from our synthetic dataset, as showcased in the third column of fig. 4, fig. 5, and fig. 6. This is a good indication of how to bridge the reality gap for the open problem of soft shadow estimation. However, as shadows get increasingly softer (fig. 2), especially with overlapping details (fig. 3, last two rows), we notice the synthetic training data alone might be insufficient for a completely precise estimation on real images. This is particularly challenging given unknown surface textures.

On the other hand, for less extreme cases of softness, our model can capture considerable detail (e.g., the stick shadows behind the right-side fence in fig. 3, second row).

Overall, we believe these results highlight the complexity of the task at hand. While our proposed model is more general, variations in training data can yield different strengths and weaknesses. The best pre-trained model to be used for our proposed pipeline, thus, will depend on the user's intended application, including required level of generalization, shadow detail and softness, size of connected bodies, and tolerance to false positives. Regardless, we hope to facilitate different specific application needs by making our multiple pre-trained models available. We also hope our black-box shadow detector variation (in this work, used with, but not limited to, MTMT) can help future-proof our proposed pipeline.

3 SYNTHETIC SAMPLES

In the following pages, we provide samples of our synthetic dataset generated using Blender and SceneCity [Couturier 2023]. For our model, we generated 3000 training samples. However, the variations are virtually limitless.

4 DATASET SAMPLES

In the following pages, we provide samples of our augmented real data, as follows:

- Samples from the test sets of DESOBA [Hong et al. 2022] and ISTD [Wang et al. 2018], augmented for our shadow matting quantitative evaluation.
- Samples of the shadow erosion live augmentation performed by our network on ISTD training samples.
- Samples of the SRD [Qu et al. 2017] dataset showing our median filter adjustment to the binary masks of Cun et al. [2020].

The same matting augmentation algorithm shown in the test set images was used in the training sets to teach our model shadow matting, however, that was done online. For the test set evaluation, we enforced the masks to be no less than 20% of the full shadow and no more than 80%, with a fixed number of 30 ellipsoids, to guarantee matting on every frame. During training, the odds of our model seeing such matting augmentation were roughly 25%, due to the other augmentations described in the main paper (not considering the distribution of datasets, out of which real shadow removal data was roughly 40%, the remainder being roughly 35% shadow detection and 25% synthetic).

To generate the ellipsoids for the masks in our shadow augmentation algorithm (see the main paper) we generated the ellipsoids with a random 2D rotation (from 0 to 360 degrees) at random scale (from 0 to half the image's resolution), centered anywhere within the mask. The Gaussian kernel used is 5×5 with $\sigma = 1$.

5 EXTRA MATTING RESULTS

In the following pages, we provide samples of matting results measured during our quantitative evaluation for the following models:

- **MTMT** (+crf, +value scaling, +rgb scaling)
- **compositing network** (+pbla, +value scaling, +rgb scaling)
- **gain network** (+pbla, +value scaling, +rgb scaling)
- **ours** (+pbla, +value scaling, +rgb scaling)

REFERENCES

Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. 2020. A Multi-Task Mean Teacher for Semi-Supervised Shadow Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE, Seattle, WA, USA, 5610–5619. https://doi.org/10.1109/CVPR42600.2020.00565

Arnaud Couturier. 2023. SceneCity blender plugin. https://www.cgchan.com/store/scenecity.

Xiaodong Cun, Chi-Man Pun, and Cheng Shi. 2020. Towards Ghost-Free Shadow Removal via Dual Hierarchical Aggregation Network and Shadow Matting GAN. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, New York, NY, USA, 10680–10687. https://doi.org/10.1609/aaai.v34i07.6695

David Griffiths, Tobias Ritschel, and Julien Philip. 2022. OutCast: Outdoor Single-image Relighting with Cast Shadows. *Comput. Graph. Forum* 41, 2 (2022), 179–193. https://doi.org/10.1111/cgf.14467

Yan Hong, Li Niu, and Jianfu Zhang. 2022. Shadow Generation for Composite Image in Real-World Scenes. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press, Virtual Event, 914–922. https://doi.org/10.1609/aaai.v36i1.19974

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, Honolulu, HI, USA, 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. 2017. DeshadowNet: A Multi-context Embedding Deep Network for Shadow Removal. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, Honolulu, HI, USA, 2308–2316. https://doi.org/10.1109/CVPR.2017.248

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *18th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI Proceedings, Part III (Lecture Notes in Computer Science, Vol. 9351)*, Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi (Eds.). Springer, Munich, Germany, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Tomás F. Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. 2016. Large-Scale Training of Shadow Detectors with Noisily-Annotated Shadow Examples. In *14th European Conference on Computer Vision, ECCV Proceedings, Part VI (Lecture Notes in Computer Science, Vol. 9910)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer, Amsterdam, The Netherlands, 816–832. https://doi.org/10.1007/978-3-319-46466-4_49

Jifeng Wang, Xiang Li, and Jian Yang. 2018. Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE Computer Society, Salt Lake City, UT, USA, 1788–1797. https://doi.org/10.1109/CVPR.2018.00192

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. 2019. *Fixup Initialization: Residual Learning Without Normalization*. arXiv. arXiv:1901.09321 http://arxiv.org/abs/1901.09321

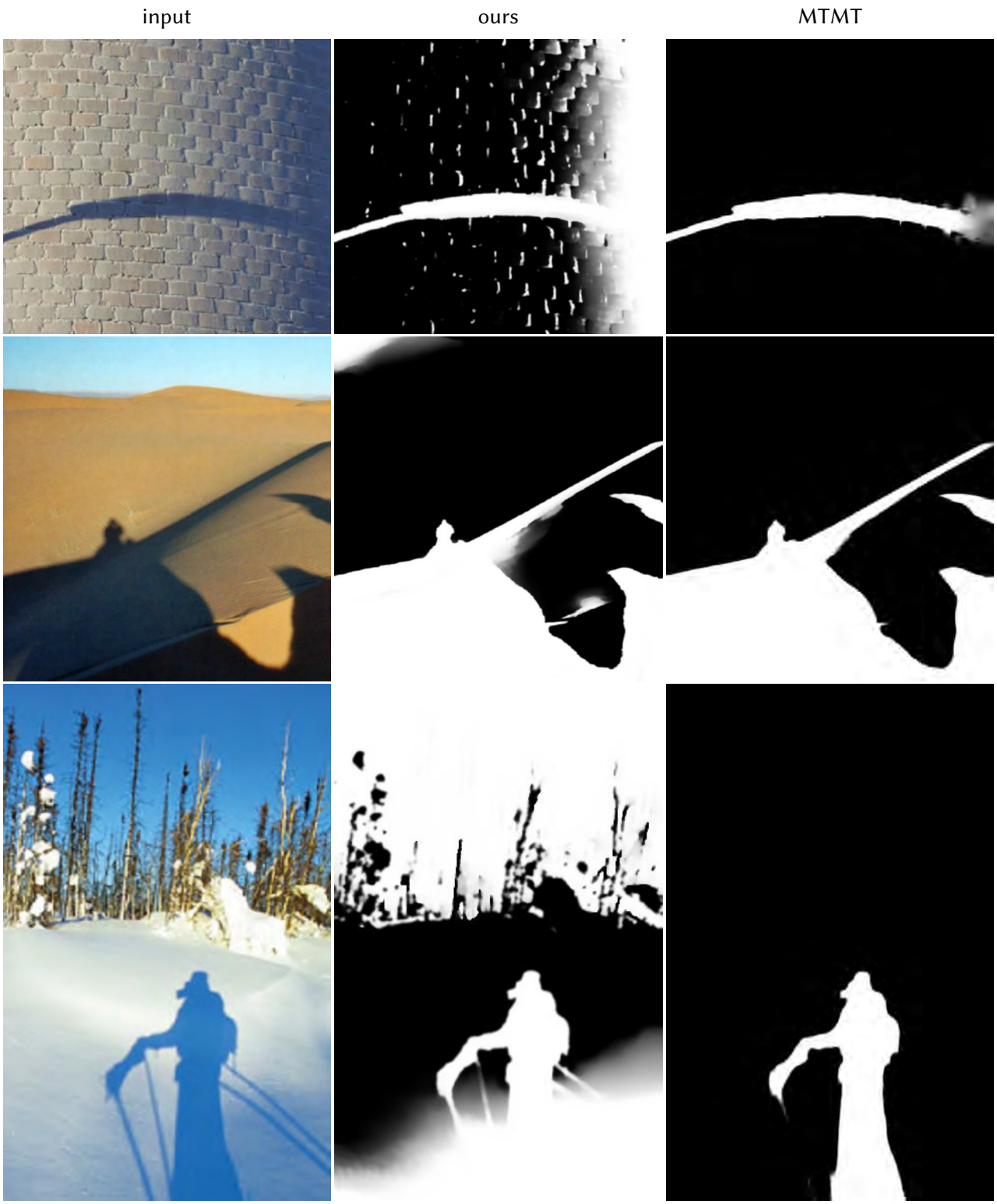

Fig. 2. **Soft shadow detection.** Our model (2nd column) produces a more detailed soft shadow map than the current state-of-the-art shadow detection method MTMT[Chen et al. 2020]. Our algorithm with stride 1 was used (see the main paper). Images from SBU [Vicente et al. 2016].

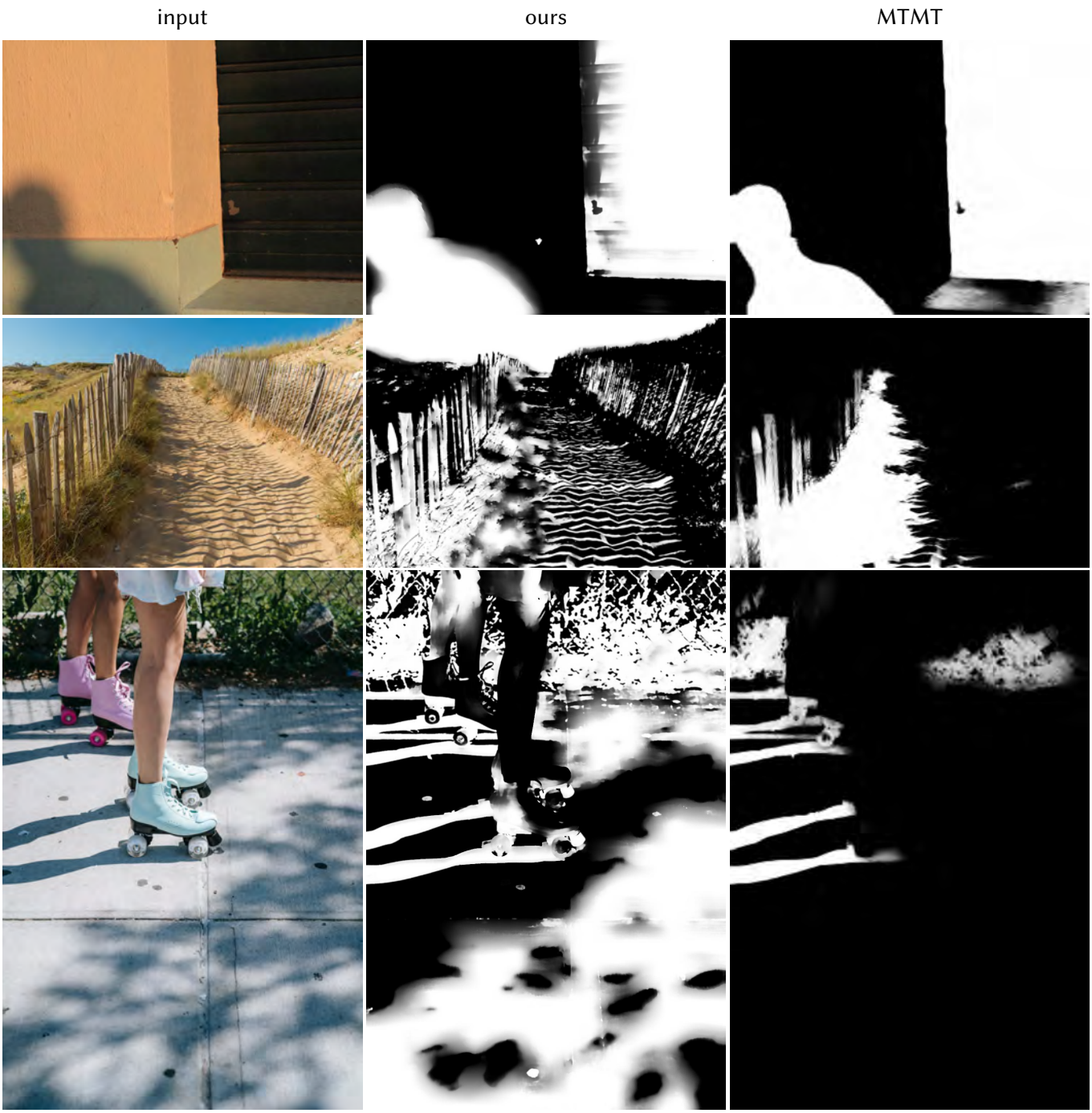

Fig. 3. **Soft shadow detection.** Our model (2nd column) produces a more detailed soft shadow map than the current state-of-the-art shadow detection method MTMT[Chen et al. 2020]. Our algorithm with stride 1 was used (see the main paper). First row from SRD [Qu et al. 2017], last 2 rows from our "in the wild" set of images found online under free Pexels license. Image credits (top to bottom): Pixabay, and Katya Wolf.

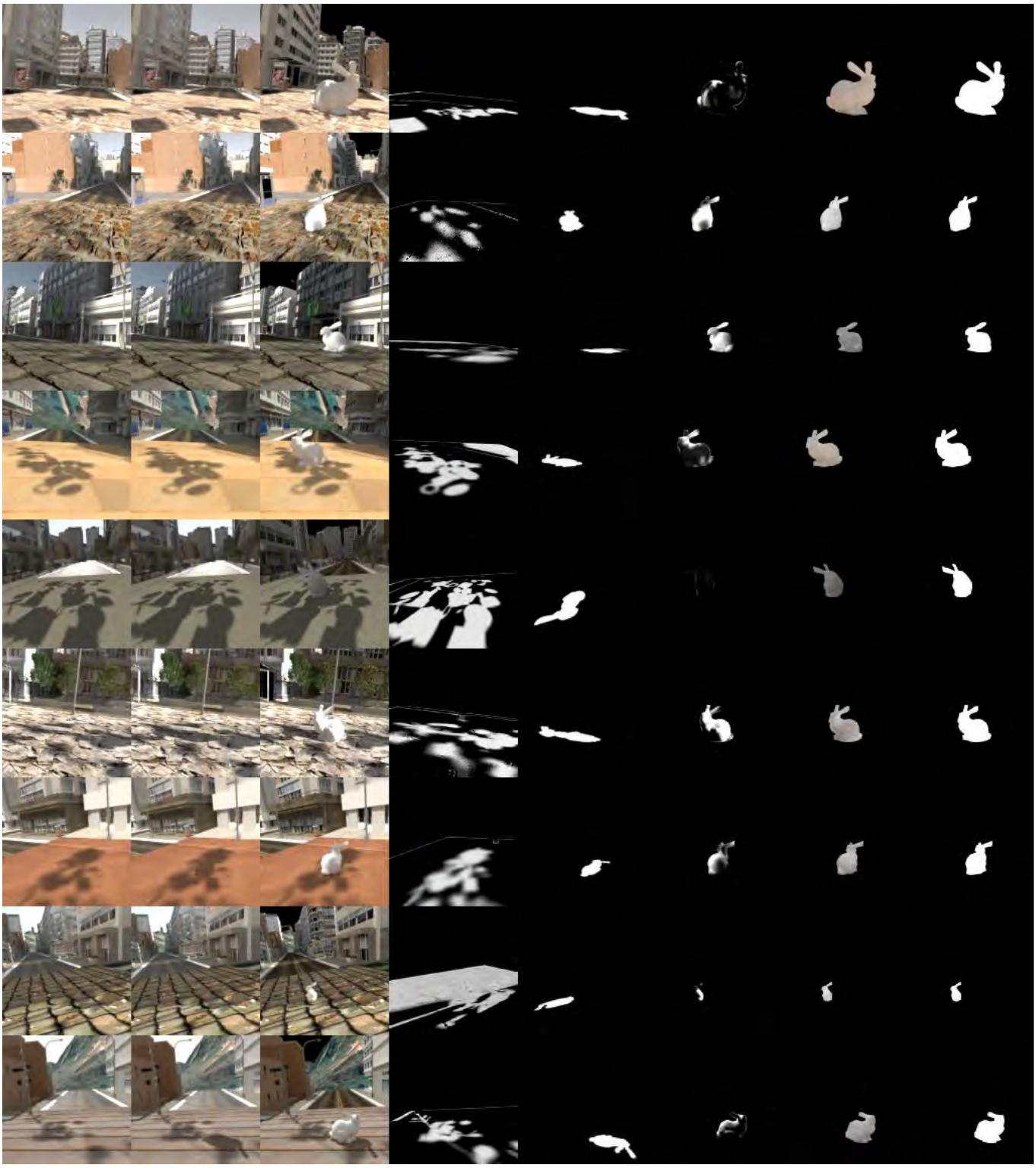

Fig. 4. **Blender dataset samples**. From left to right: input image, target image with shadows, virtual object shaded (as reference), input soft shadow mask, virtual shadow mask to insert, direct light render of the shaded object, indirect light render of the shaded object, object mask.

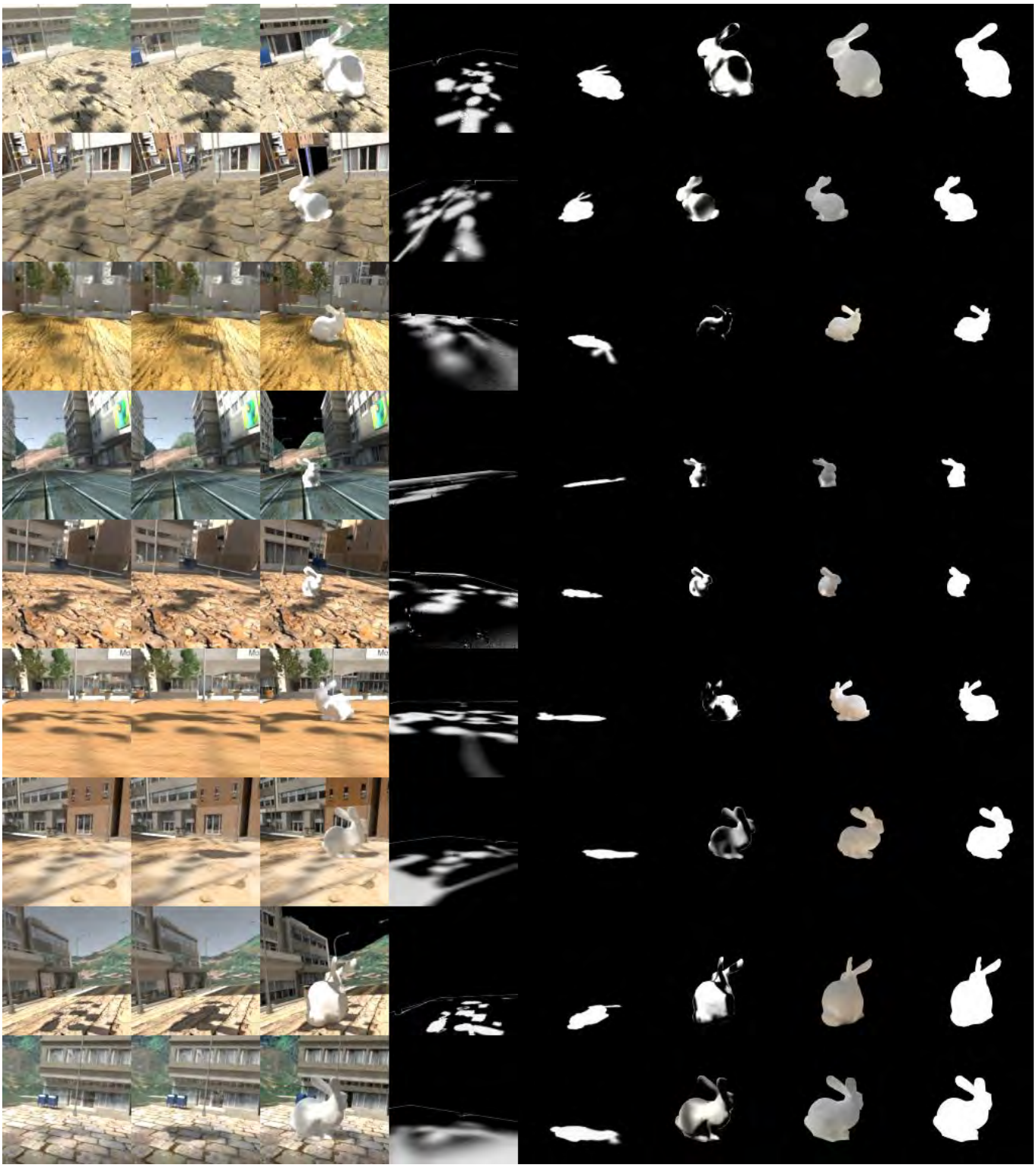

Fig. 5. **Blender dataset samples**. From left to right: input image, target image with shadows, virtual object shaded (as reference), input soft shadow mask, virtual shadow mask to insert, direct light render of the shaded object, indirect light render of the shaded object, object mask.

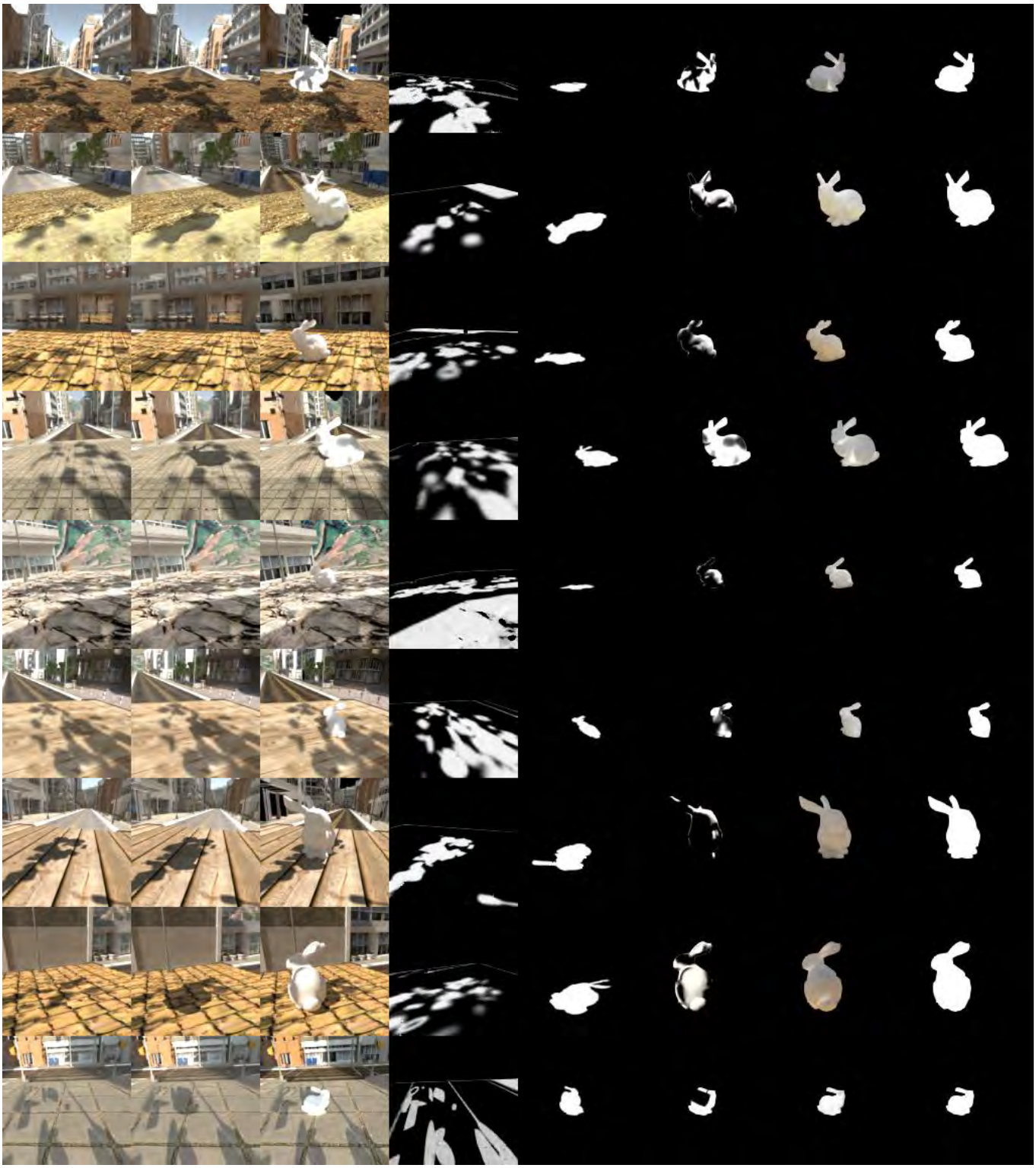

Fig. 6. **Blender dataset samples**. From left to right: input image, target image with shadows, virtual object shaded (as reference), input soft shadow mask, virtual shadow mask to insert, direct light render of the shaded object, indirect light render of the shaded object, object mask.

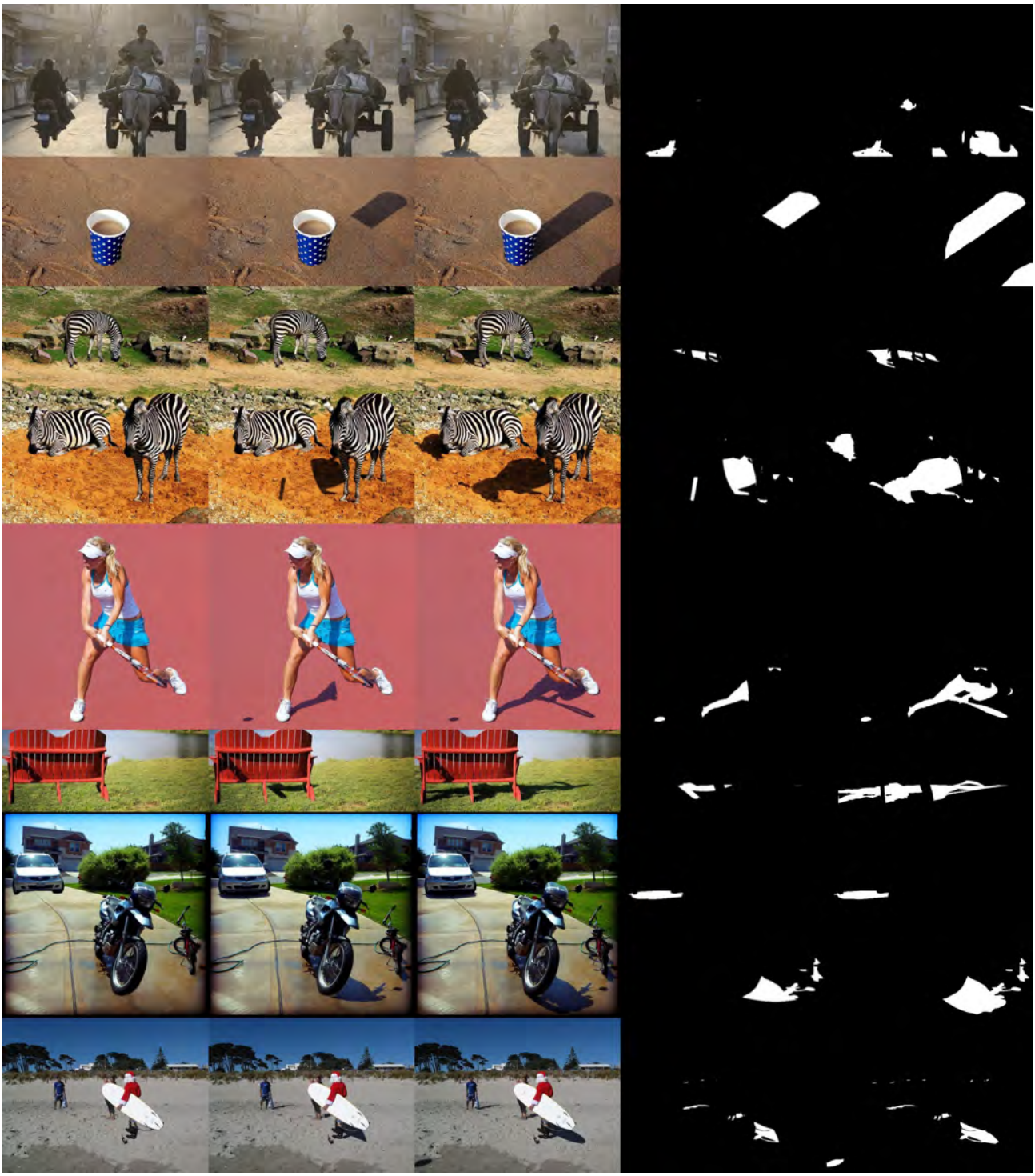

Fig. 7. **Samples from our DESOBA augmentation for shadow matting**. From left to right: original shadow-less image, augmented image with partial shadows (new input), fully-shaded image (ground truth), new input shadow mask (target for shadow detection), full shadow mask (used to ask the models to insert the remaining shadows and compare to the ground truth).

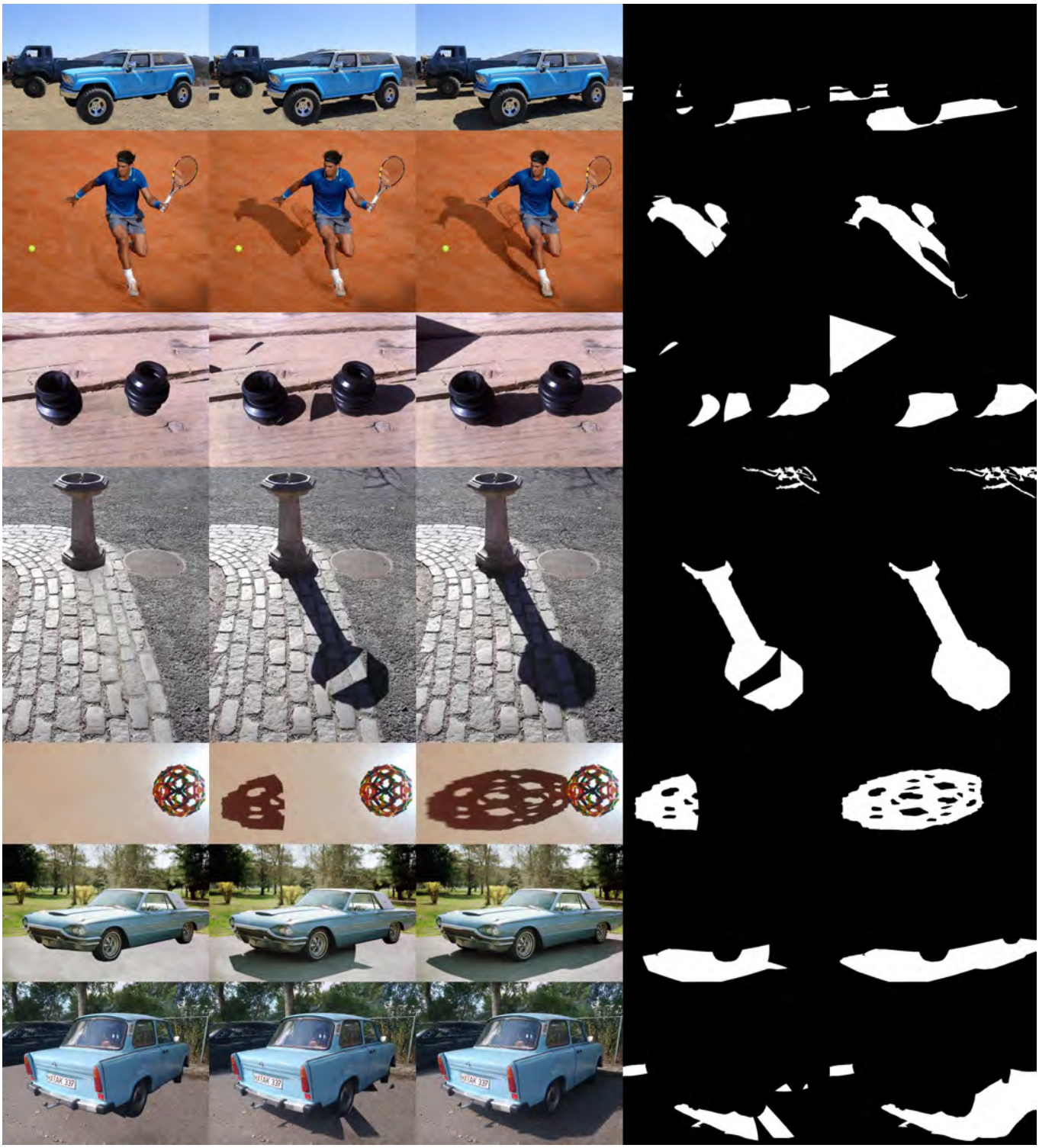

Fig. 8. **Samples from our DESOBA augmentation for shadow matting**. From left to right: original shadow-less image, augmented image with partial shadows (new input), fully-shaded image (ground truth), new input shadow mask (target for shadow detection), full shadow mask (used to ask the models to insert the remaining shadows and compare to the ground truth).

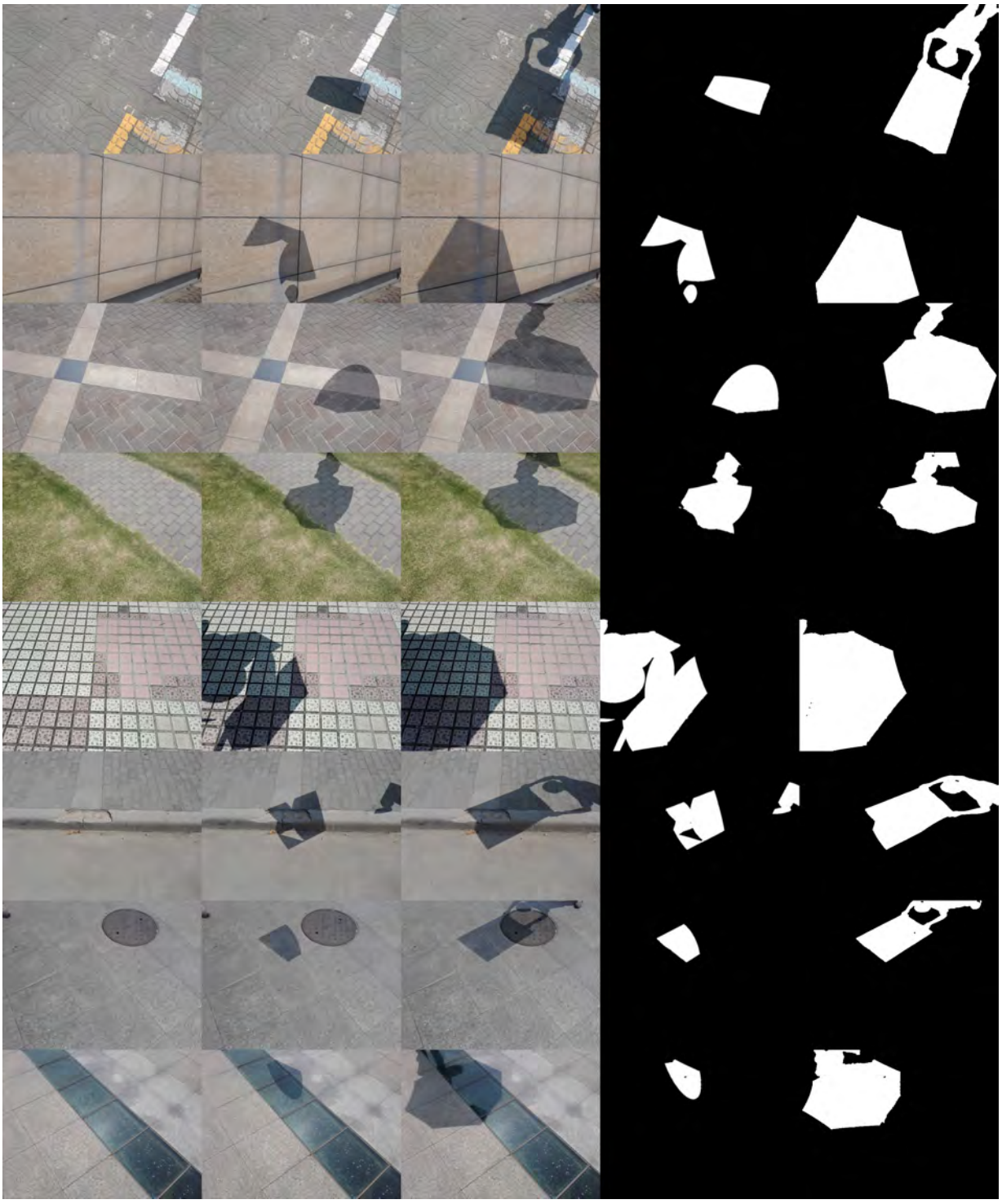

Fig. 9. **Samples from our ISTD augmentation for shadow matting**. From left to right: original shadow-less image, augmented image with partial shadows (new input), fully-shaded image (ground truth), new input shadow mask (target for shadow detection), full shadow mask (used to ask the models to insert the remaining shadows and compare to the ground truth).

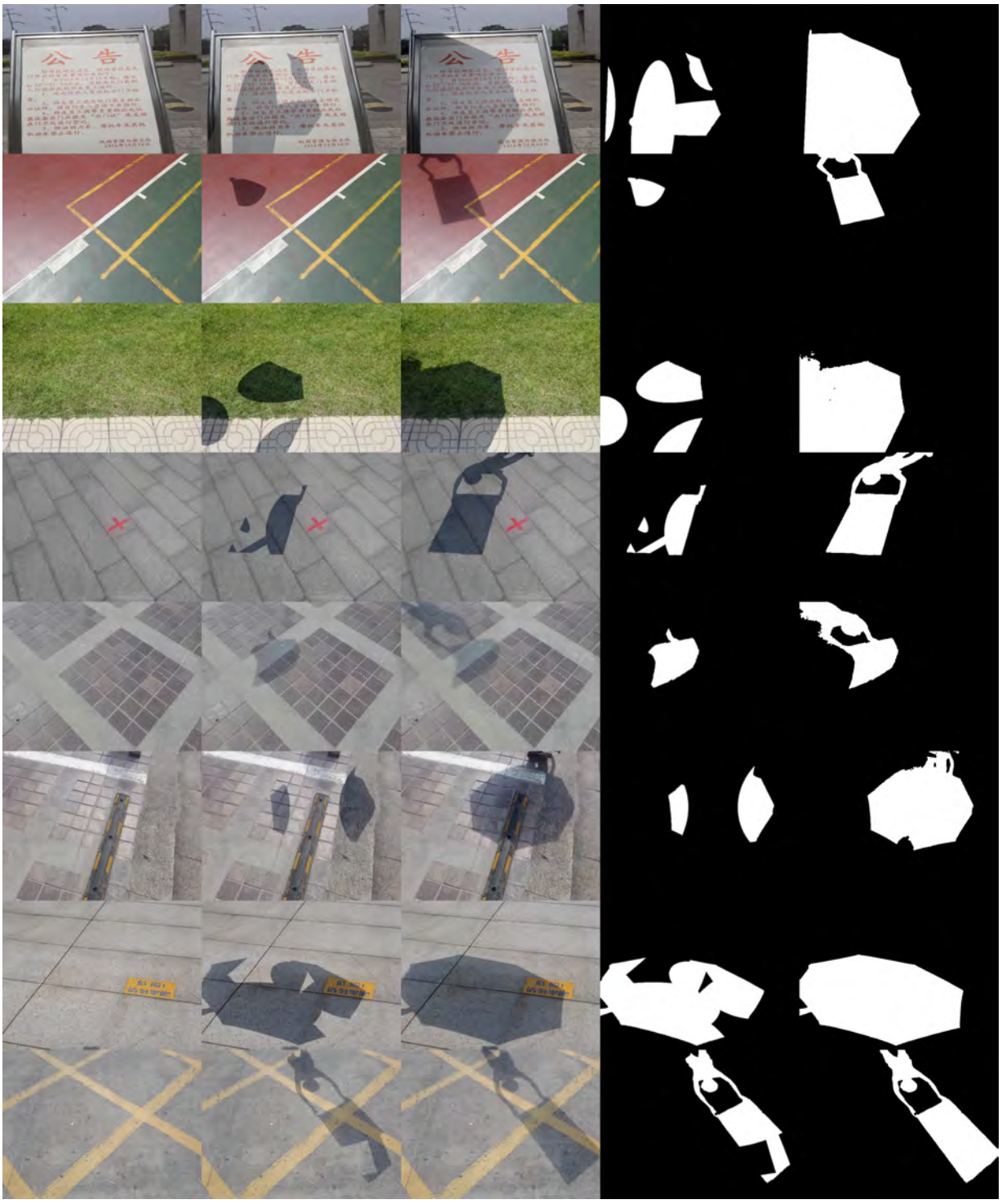

Fig. 10. **Samples from our ISTD augmentation for shadow matting**. From left to right: original shadow-less image, augmented image with partial shadows (new input), fully-shaded image (ground truth), new input shadow mask (target for shadow detection), full shadow mask (used to ask the models to insert the remaining shadows and compare to the ground truth).

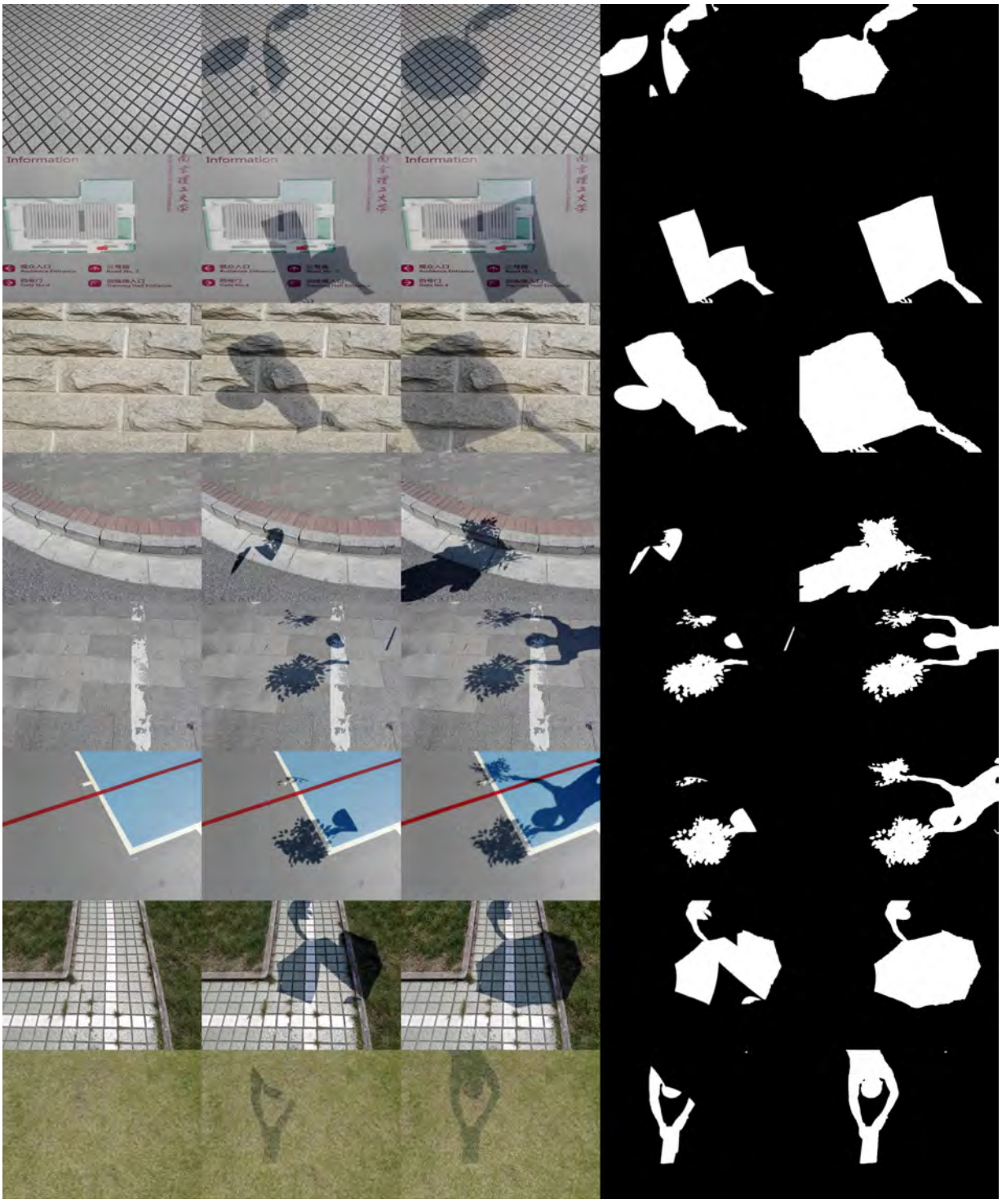

Fig. 11. **Samples from our ISTD augmentation for shadow matting**. From left to right: original shadow-less image, augmented image with partial shadows (new input), fully-shaded image (ground truth), new input shadow mask (target for shadow detection), full shadow mask (used to ask the models to insert the remaining shadows and compare to the ground truth).

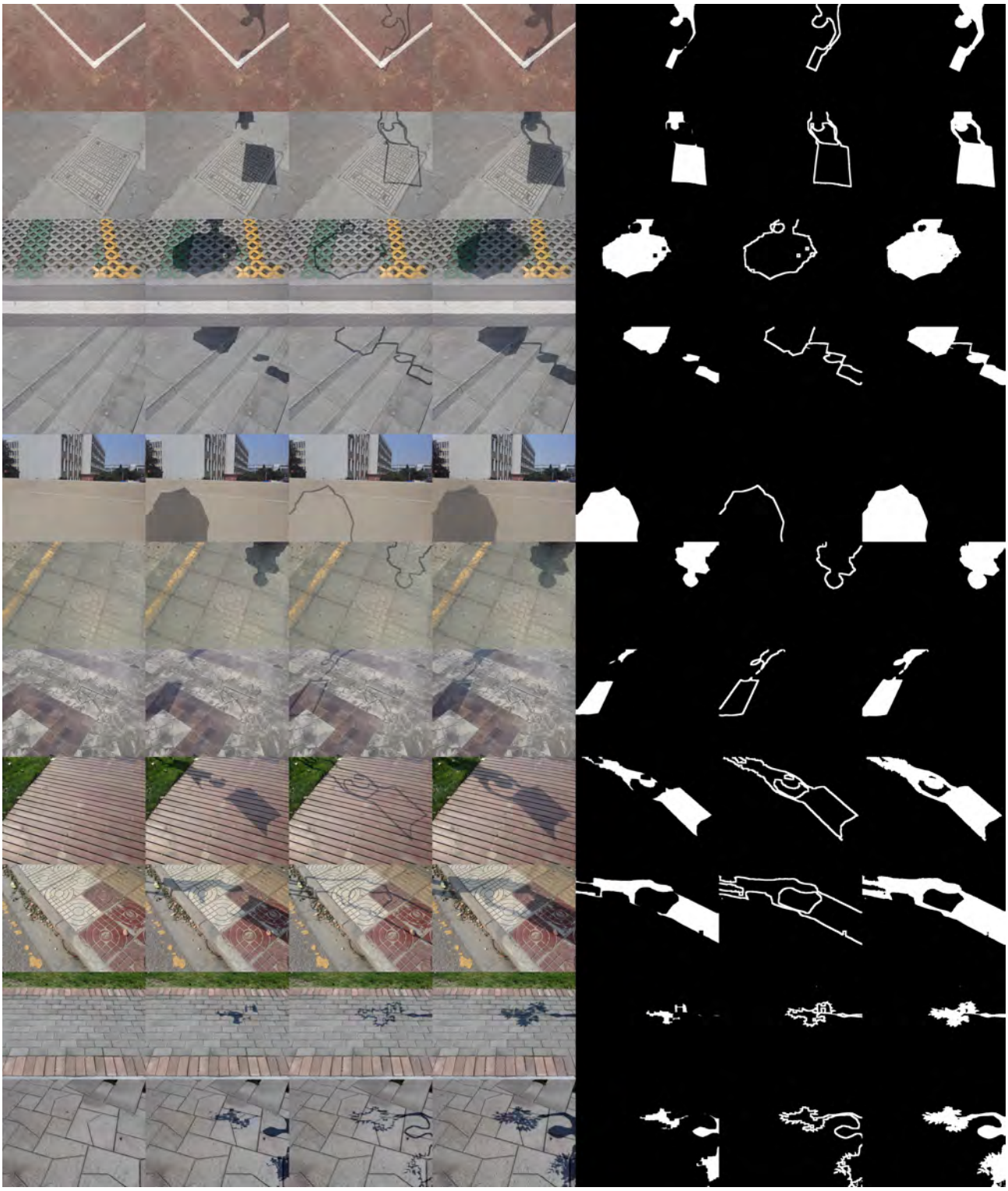

Fig. 12. **Samples of our erosion augmentation on ISTD**. From left to right: original shadow-less image, border-less augmented input, shadow edge augmented input, fully-shaded image (ground truth), border-less shadow mask, shadow edge mask, full shadow mask.

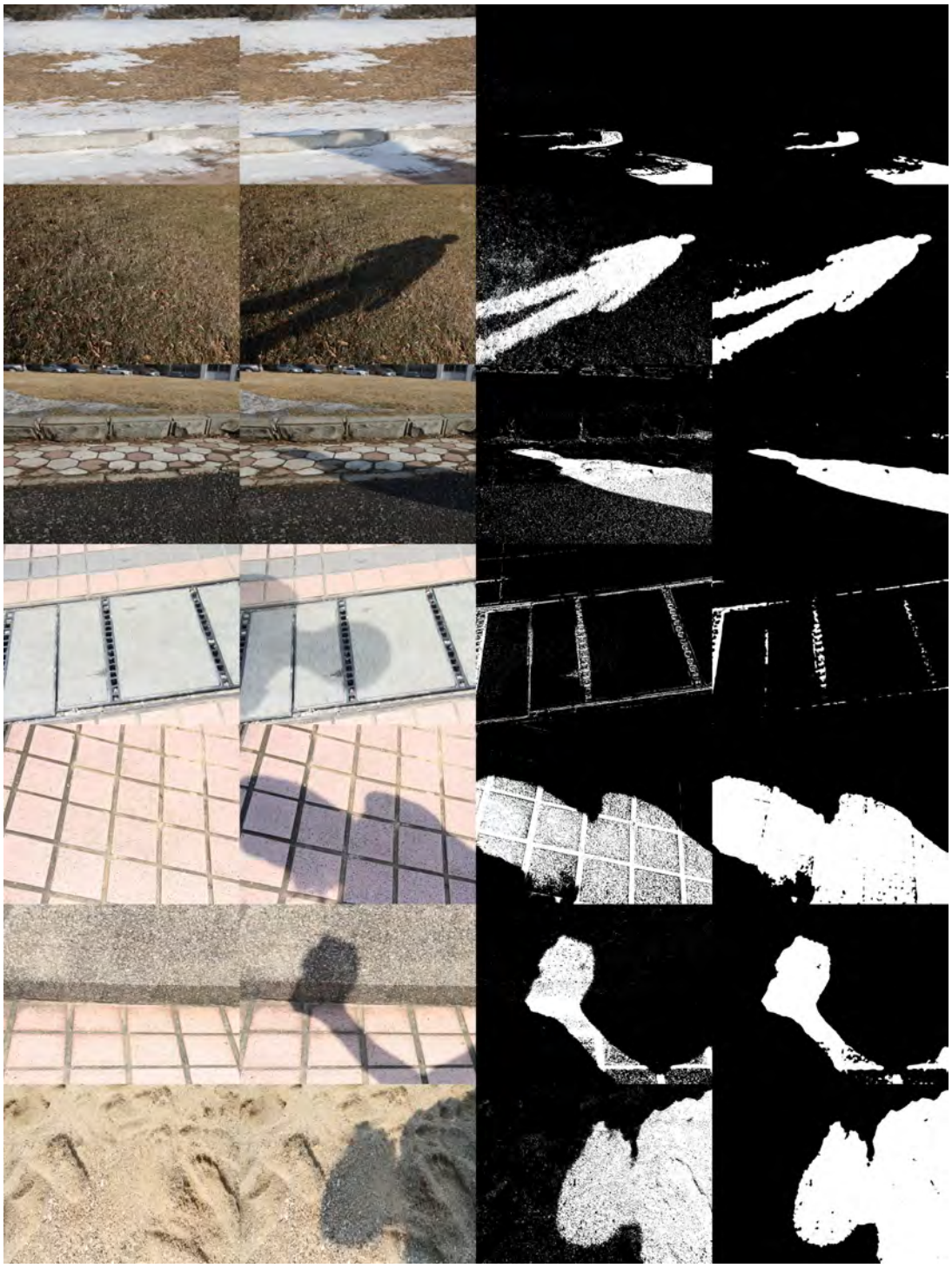

Fig. 13. **Samples of our adjustment of SRD masks**. From left to right: original shadow-less image, fully-shaded image, provided shadow masks, adjusted shadow masks. Our adjustment was sufficient for training, since there were over 8000 other accurate masks, but not meaningful for evaluations, since it would punish multiple correct detections and insert noise in the averages.

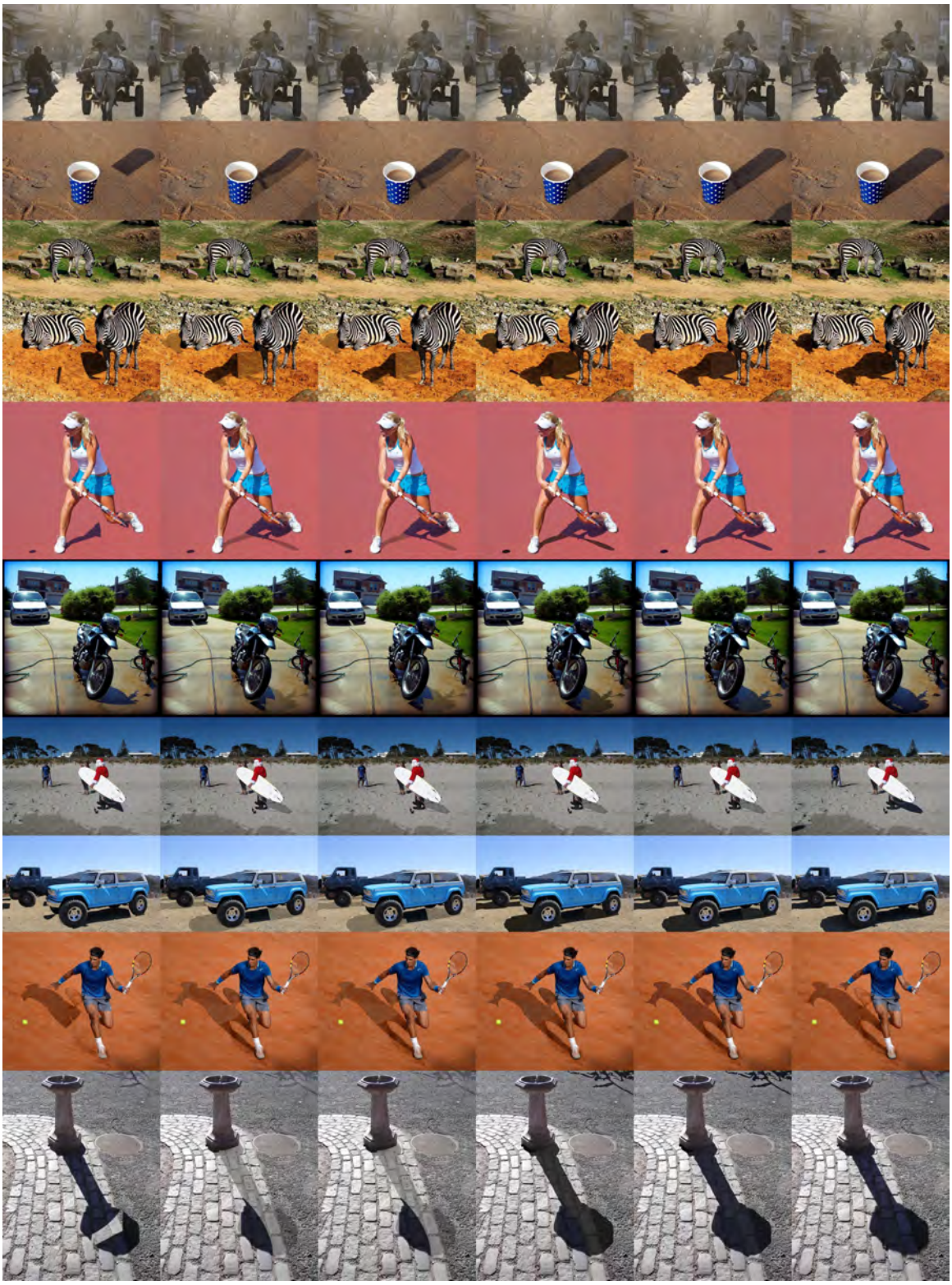

Fig. 14. **Matting results for "MTMT baseline" on DESOBA**. From left to right: input, matting, +crf, +value scaling, +rgb scaling, ground truth.

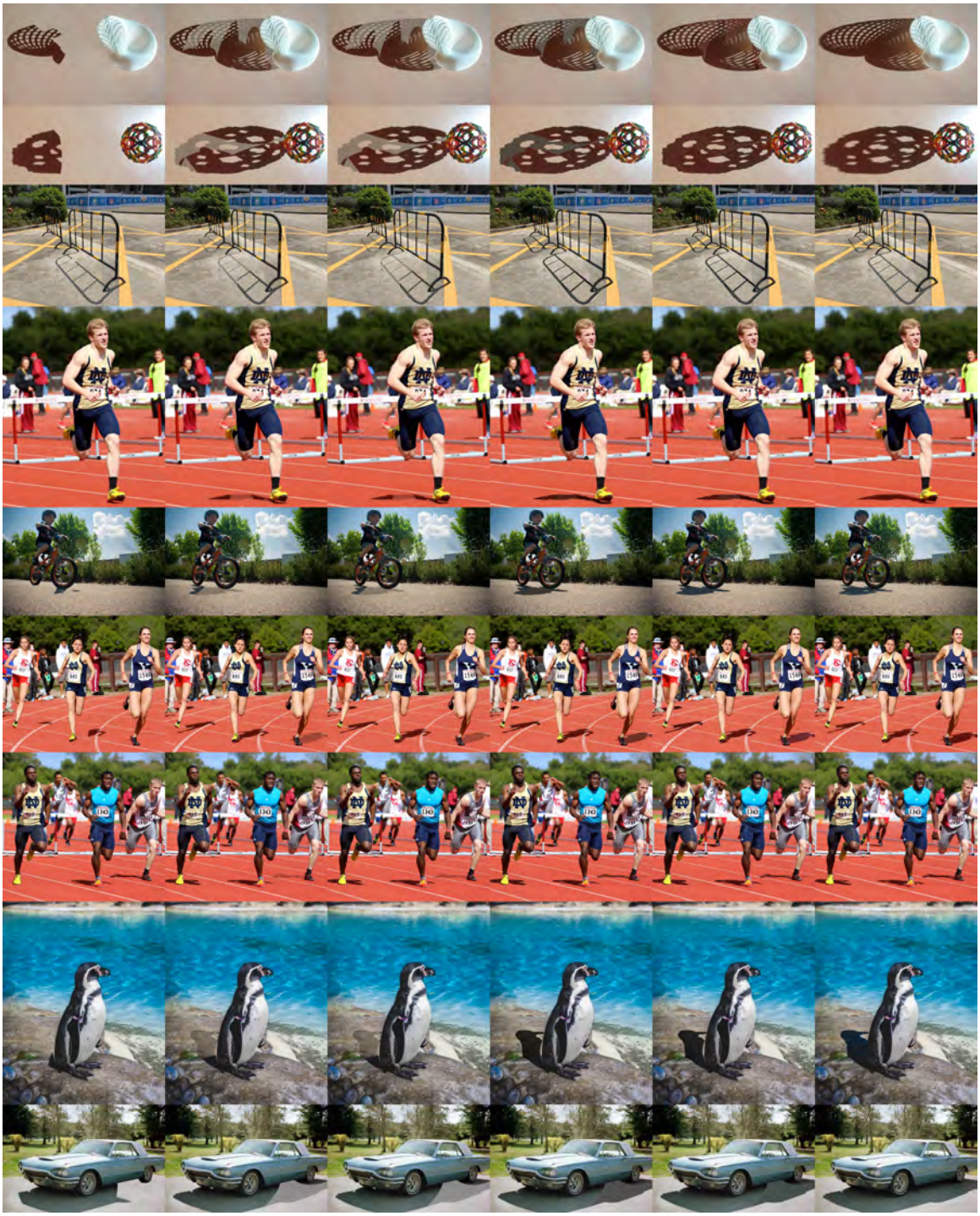

Fig. 15. **Matting results for "MTMT baseline" on DESOBA**. From left to right: input, matting, +crf, +value scaling, +rgb scaling, ground truth.

Fig. 16. **Matting results for "MTMT baseline" on DESOBA**. From left to right: input, matting, +crf, +value scaling, +rgb scaling, ground truth.

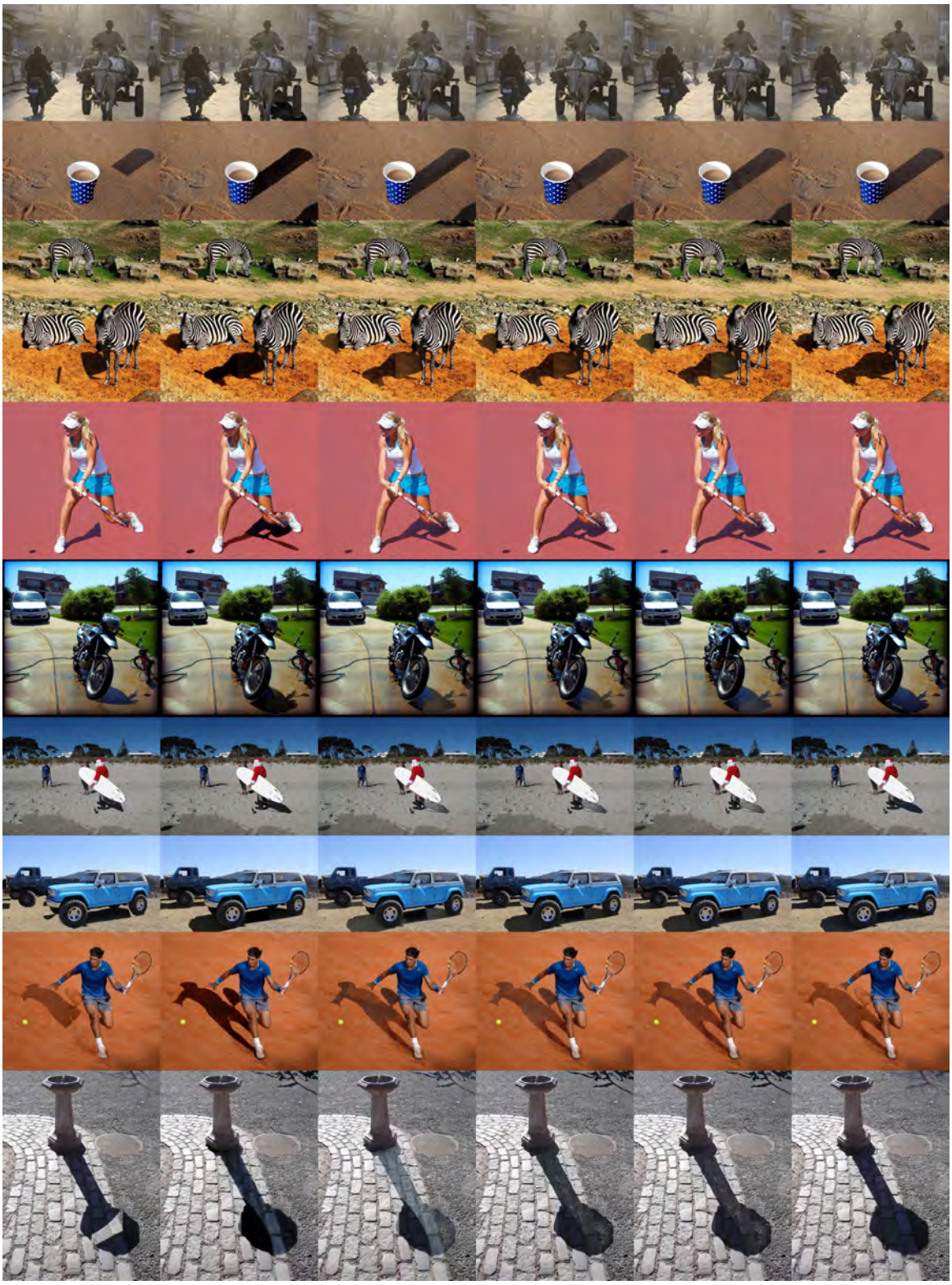

Fig. 17. **Matting results for "compositing network" on DESOBA**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

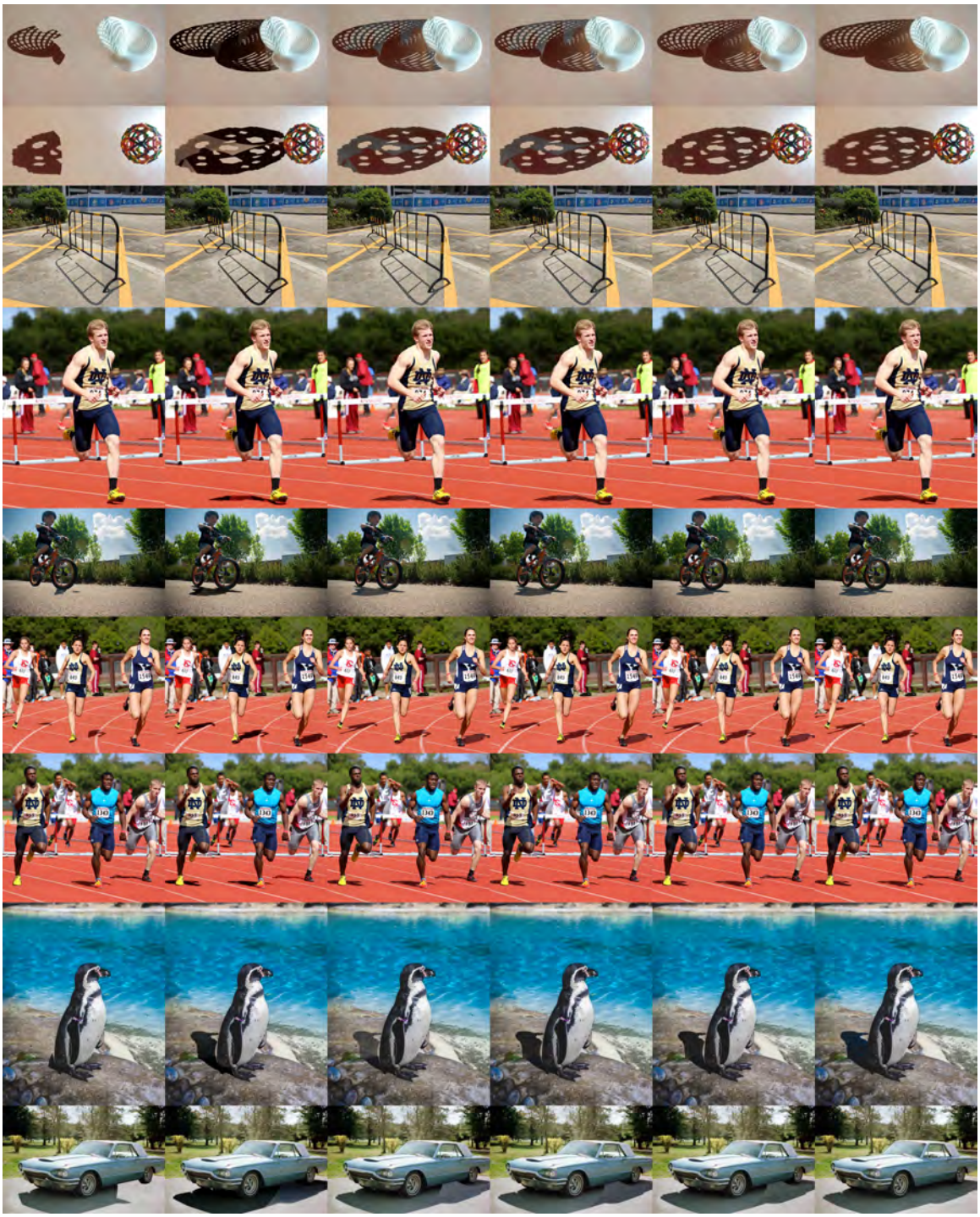

Fig. 18. **Matting results for "compositing network" on DESOBA**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

Fig. 19. **Matting results for "compositing network" on DESOBA**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

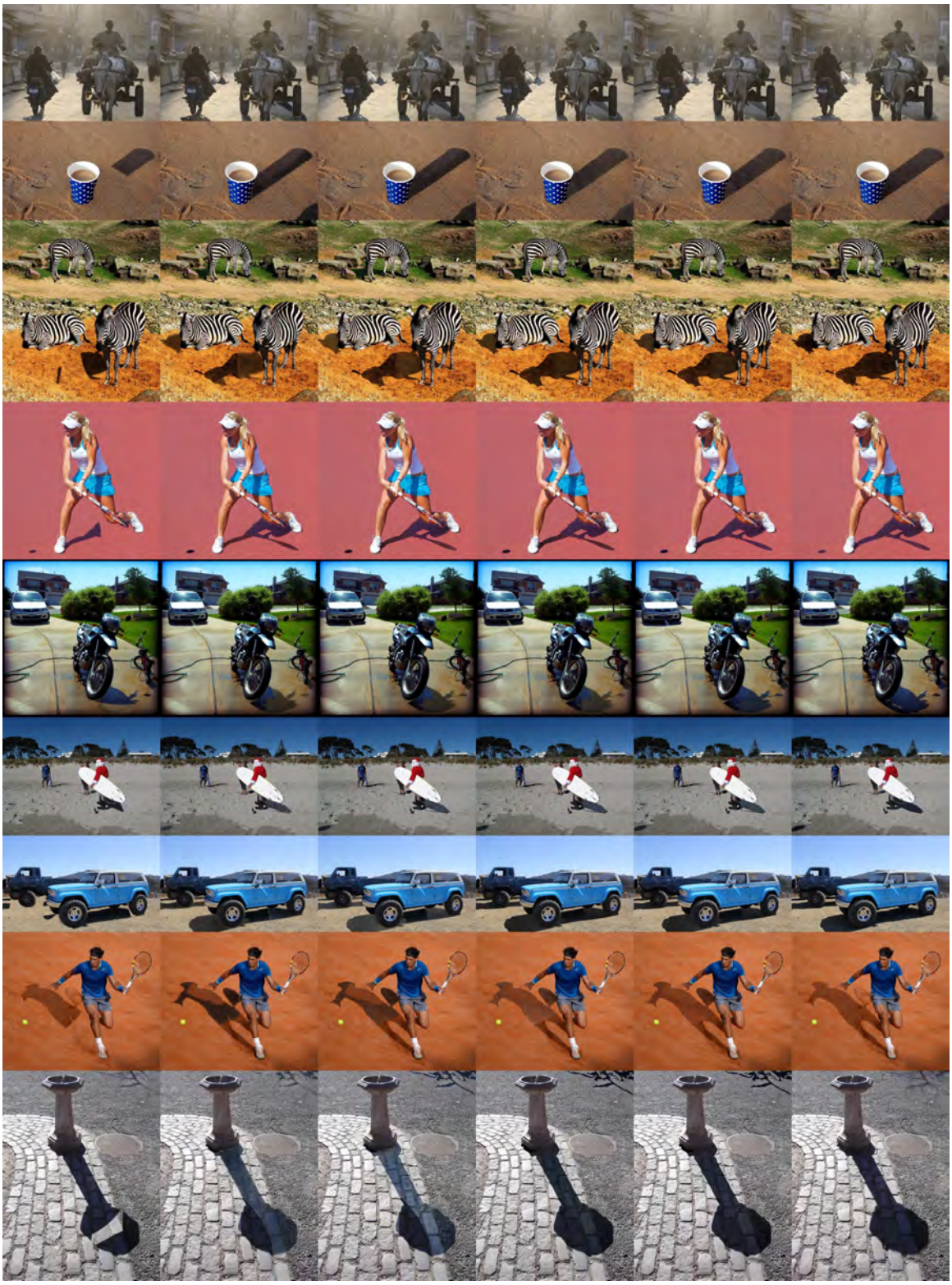

Fig. 20. **Matting results for "gain network" on DESOBA**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

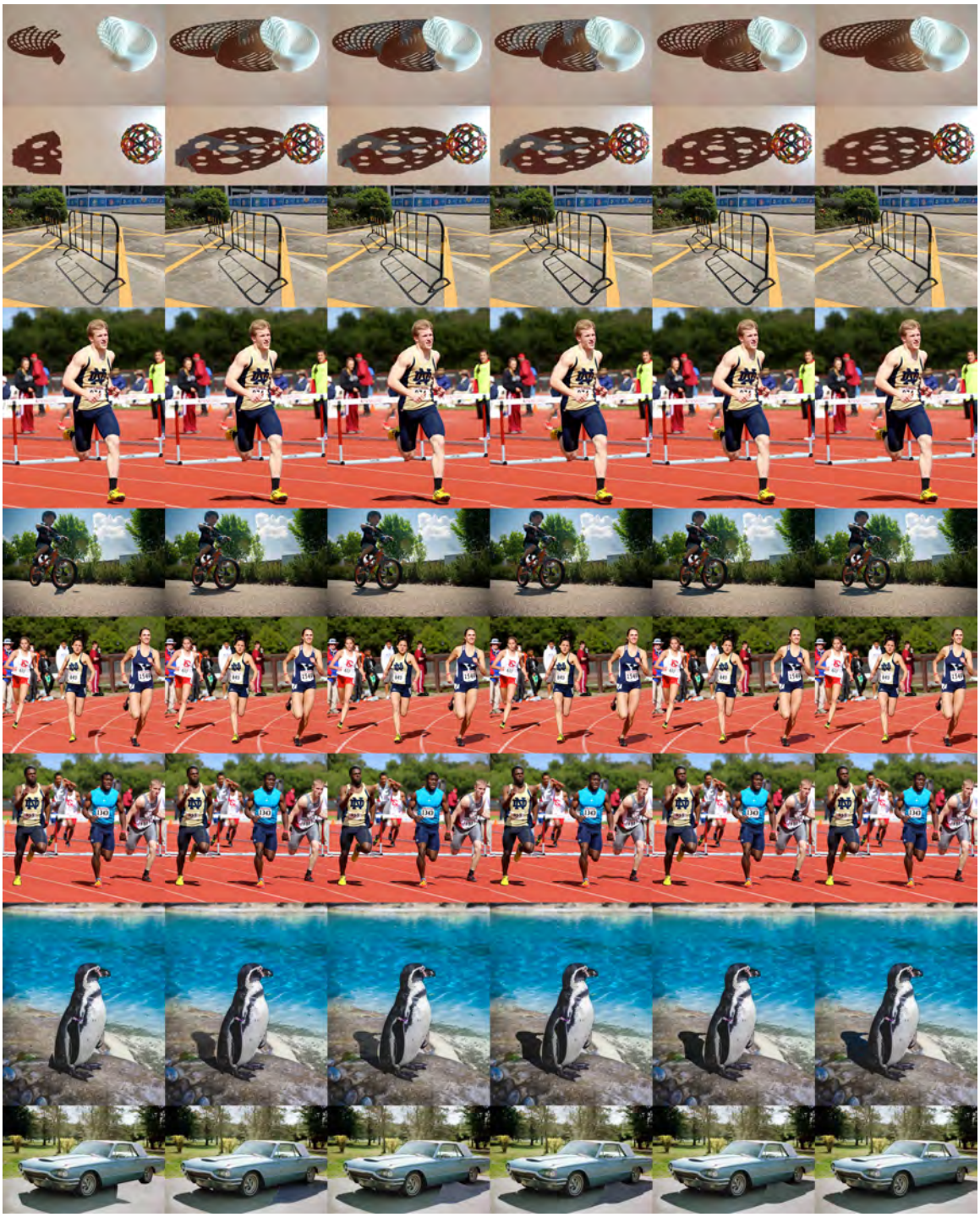

Fig. 21. **Matting results for "gain network" on DESOBA**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

Fig. 22. **Matting results for "gain network" on DESOBA**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

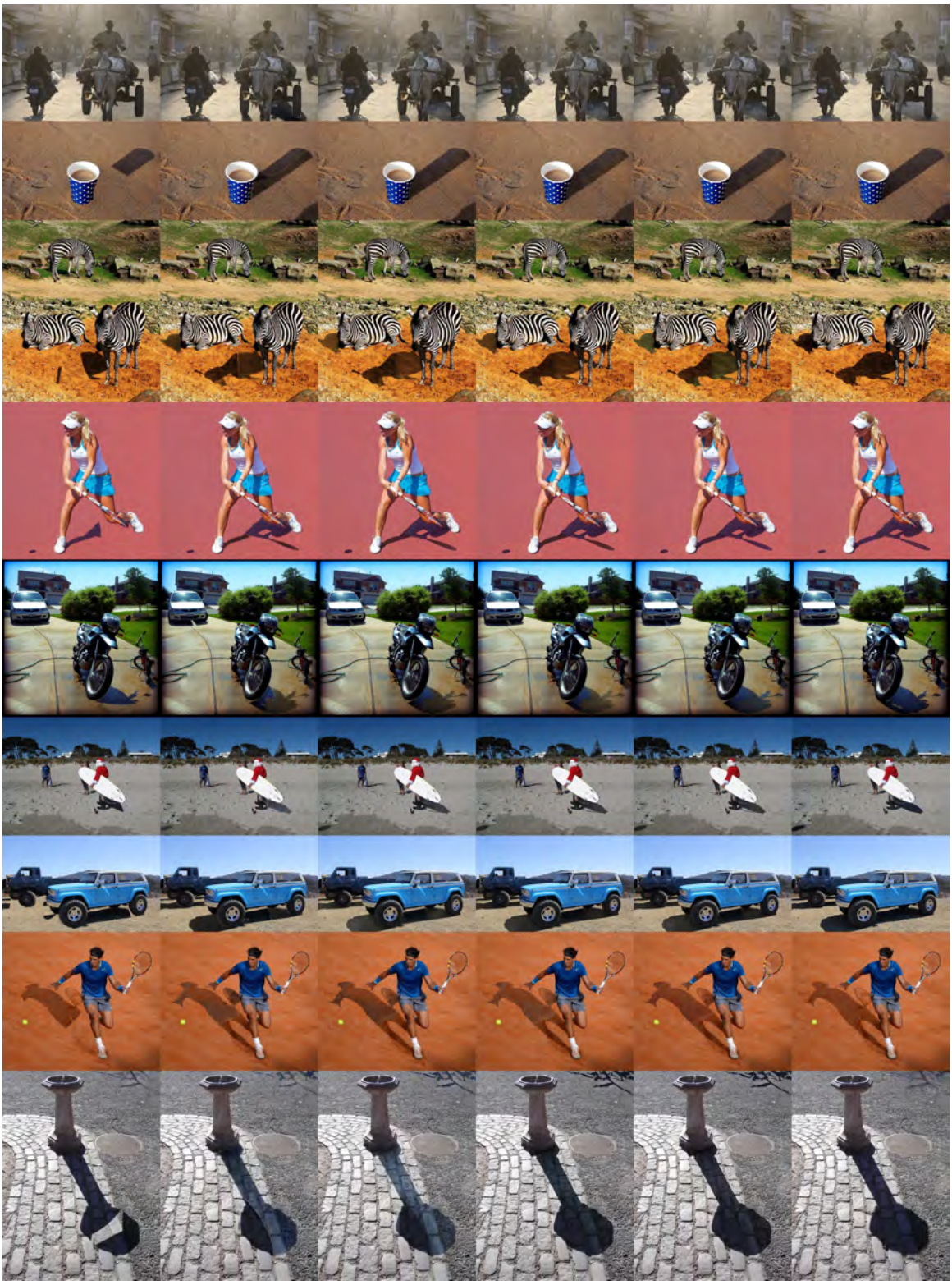

Fig. 23. **Matting results for "ours" on DESOBA**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

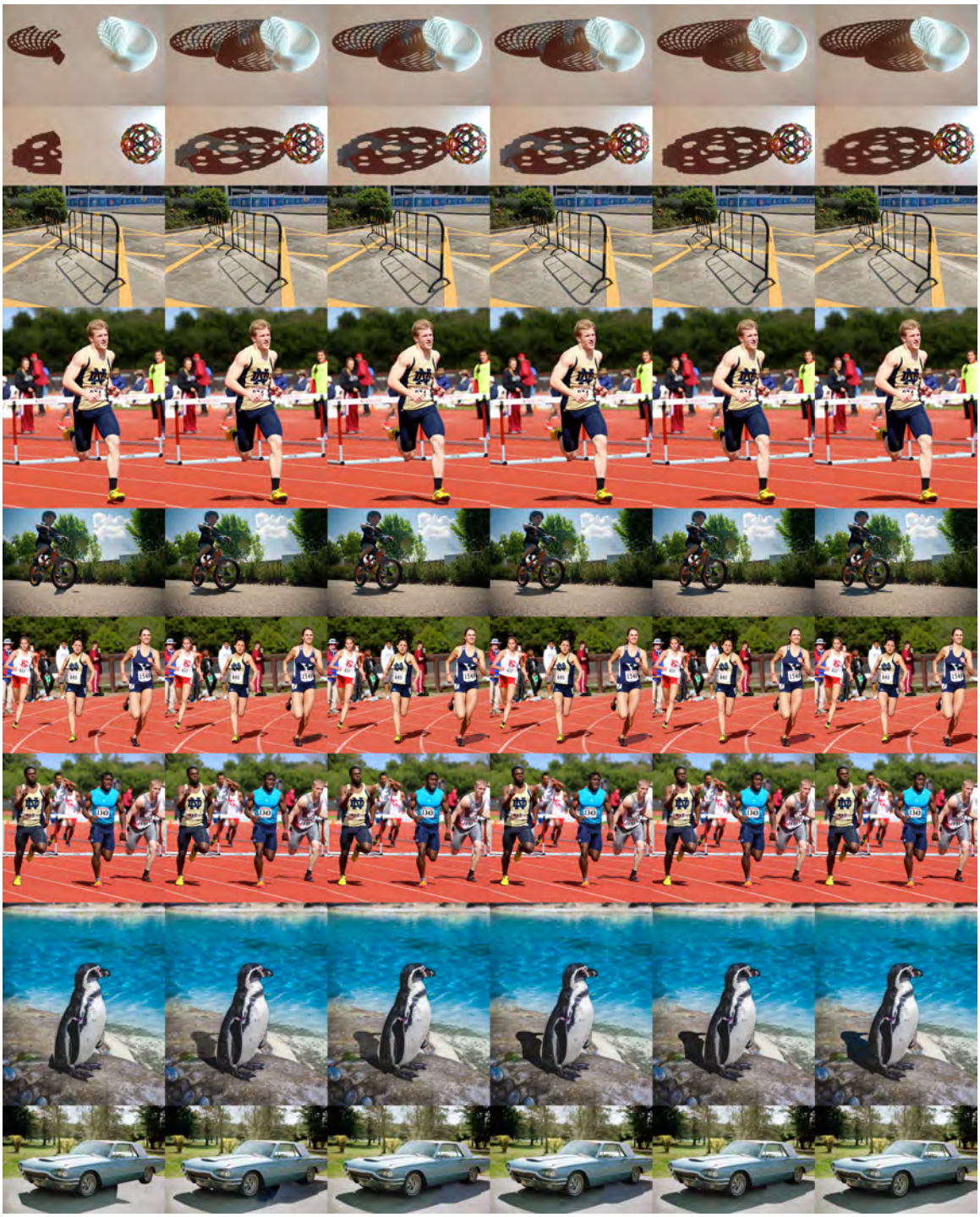

Fig. 24. **Matting results for "ours" on DESOBA**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

Fig. 25. **Matting results for "ours" on DESOBA**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

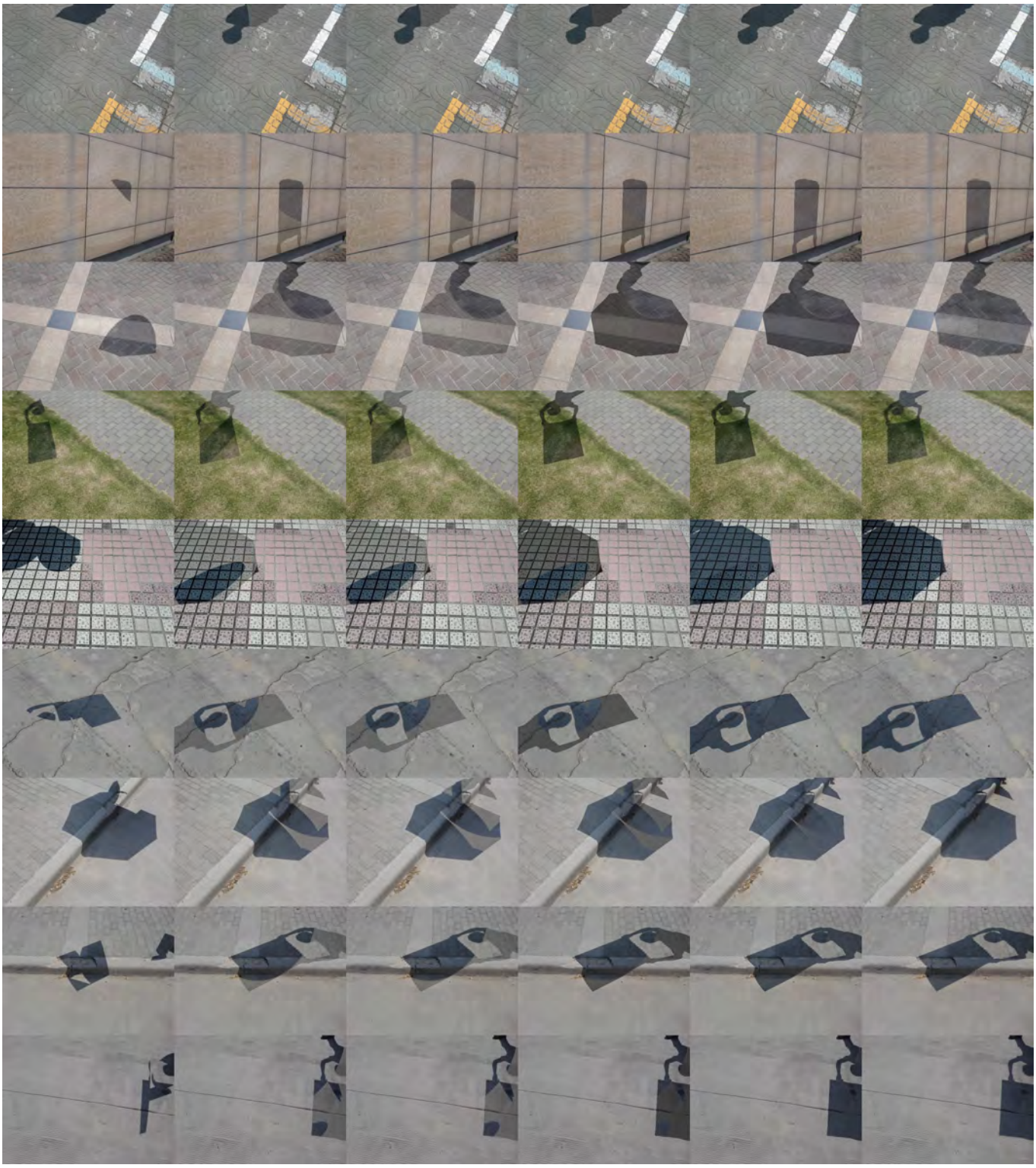

Fig. 26. **Matting results for "MTMT baseline" on ISTD**. From left to right: input, matting, +crf, +value scaling, +rgb scaling, ground truth.

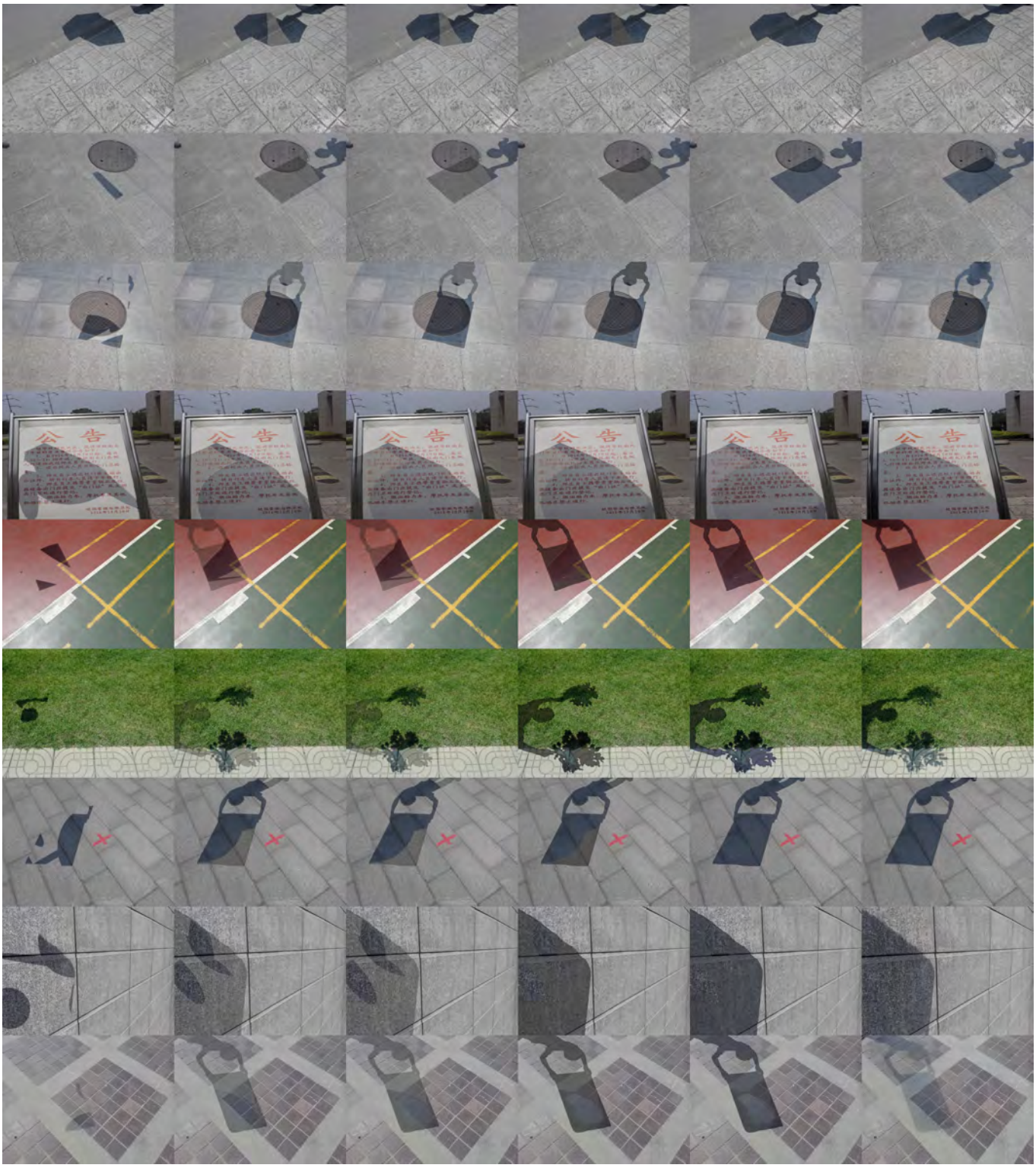

Fig. 27. **Matting results for "MTMT baseline" on ISTD**. From left to right: input, matting, +crf, +value scaling, +rgb scaling, ground truth.

Fig. 28. **Matting results for "MTMT baseline" on ISTD**. From left to right: input, matting, +crf, +value scaling, +rgb scaling, ground truth.

Fig. 29. **Matting results for "compositing network" on ISTD**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

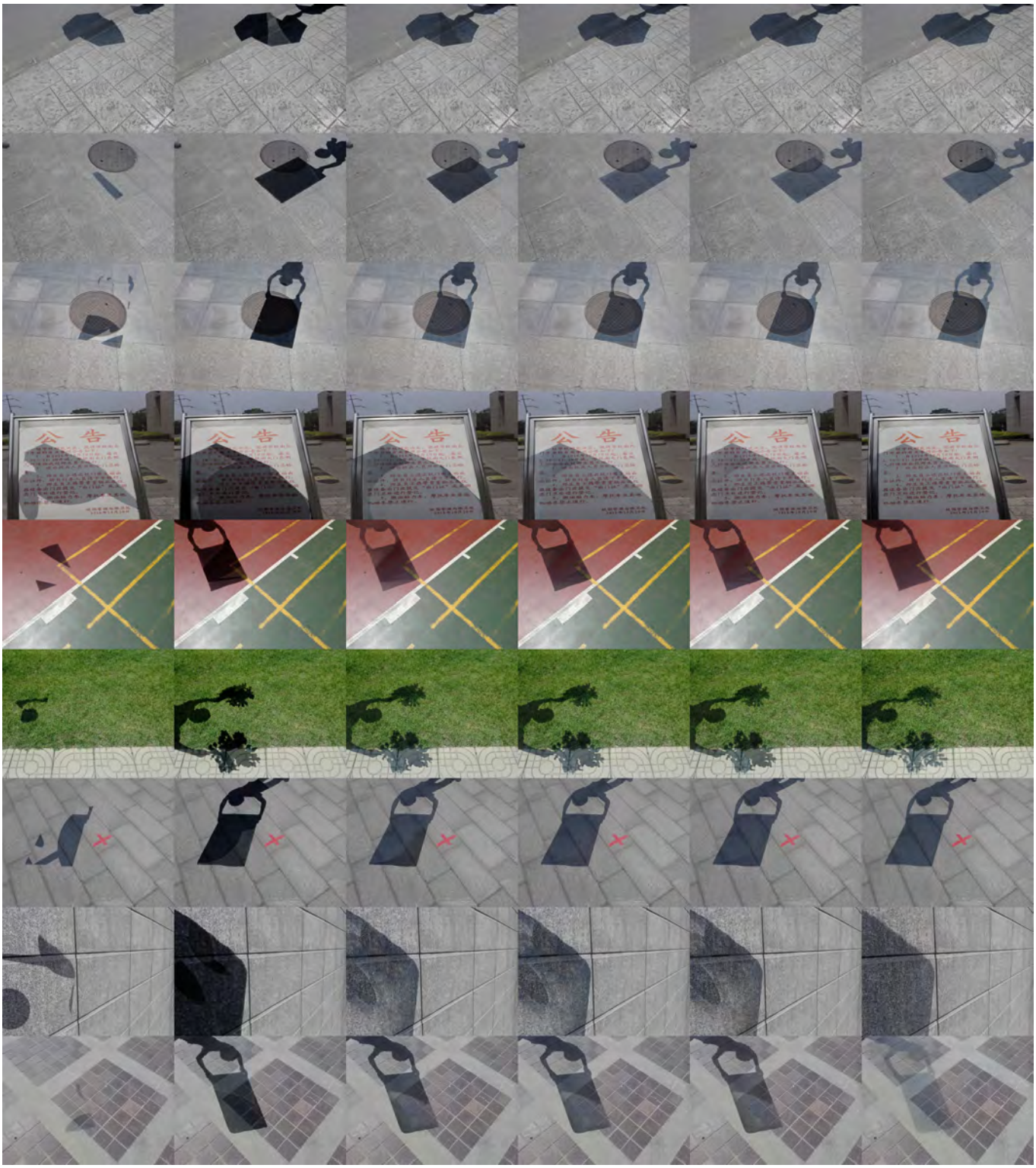

Fig. 30. **Matting results for "compositing network" on ISTD**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

Fig. 31. **Matting results for "compositing network" on ISTD**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

Fig. 32. **Matting results for "gain network" on ISTD**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

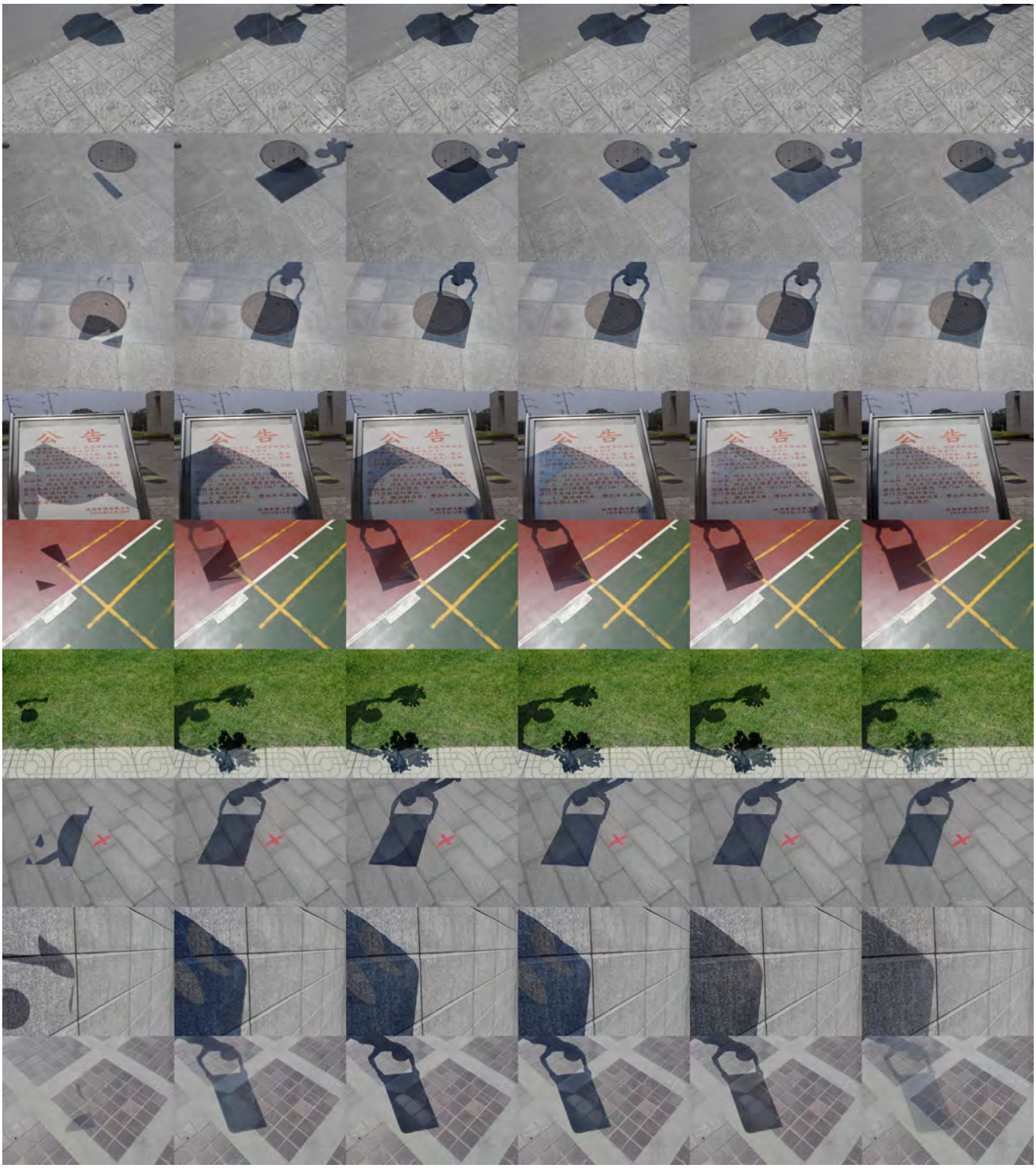

Fig. 33. **Matting results for "gain network" on ISTD**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

Fig. 34. **Matting results for "gain network" on ISTD**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

Fig. 35. **Matting results for "ours" on ISTD**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

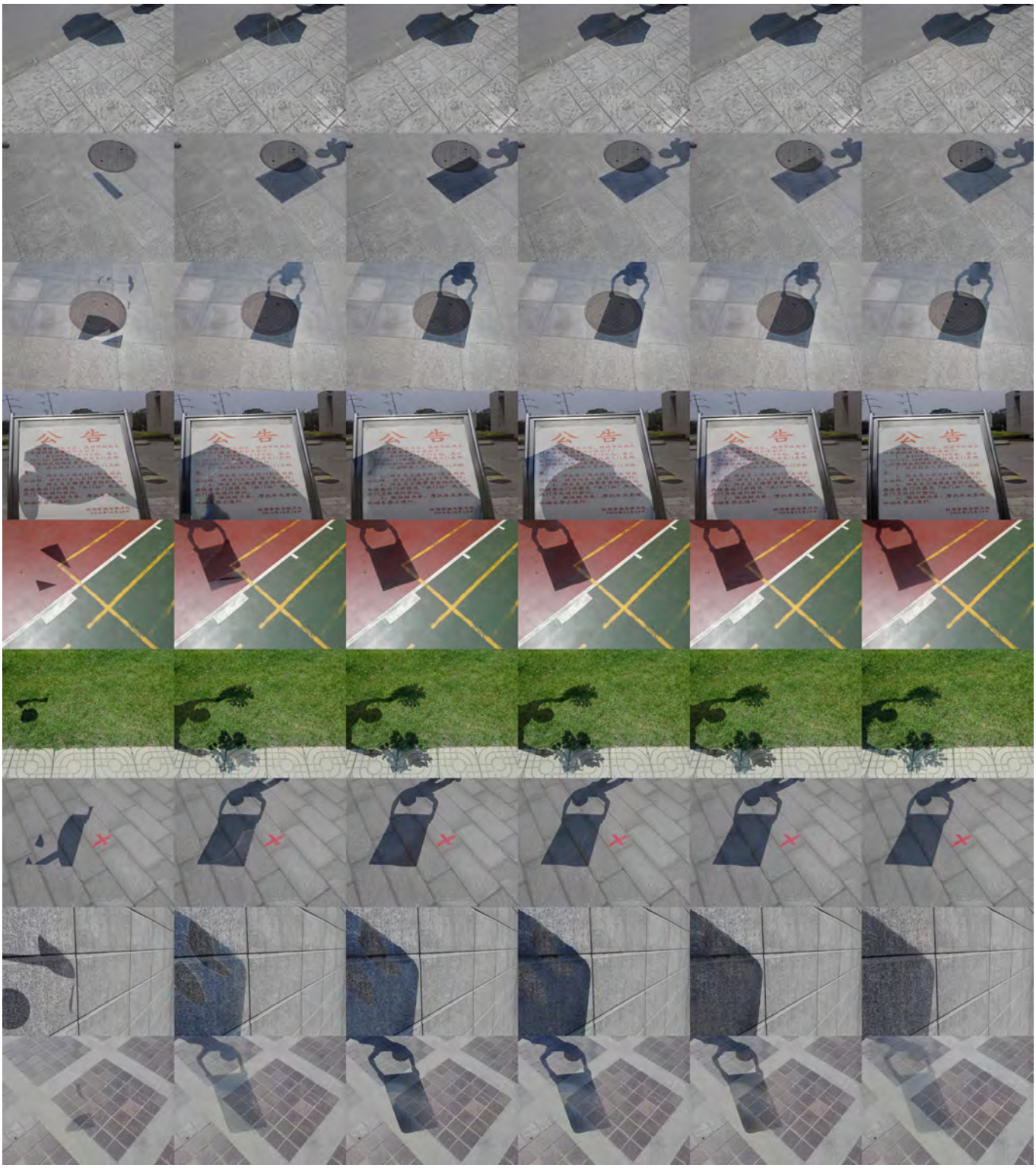

Fig. 36. **Matting results for "ours" on ISTD**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.

Fig. 37. **Matting results for "ours" on ISTD**. From left to right: input, matting, +pbla, +value scaling, +rgb scaling, ground truth.