# Revisiting Deep Intrinsic Image Decompositions (*Supplementary Material*)

Qingnan Fan*[1] Jiaolong Yang[2] Gang Hua[2] Baoquan Chen[3,4] David Wipf[2]
[1]Shandong University [2]Microsoft Research
[3]Shenzhen Research Institute, Shandong University [4]Peking University
fqnchina@gmail.com, {jiaoyan,ganghua,davidwip}@microsoft.com, baoquan@pku.edu.cn

## 1 Outline

This document contains several technical details and experiments that could not be included in the main paper due to space constraints. The remaining content is organized as follows.

- **Section 2:** More details of joint training with multiple datasets.
- **Section 3:** Details of the hinge loss used for pairwise comparison data in the IIW dataset.
- **Section 4:** Training details for the proposed framework.
- **Section 5:** More qualitative results on the IIW benchmark with intermediate outputs.
- **Section 6:** Test of generalization to outdoor images.
- **Section 7:** More visual results on the MIT intrinsic dataset.
- **Section 8:** Additional details regarding MPI-Sintel data.
- **Section 9:** More visual comparisons with state-of-the-art methods on MPI-Sintel dataset.

## 2 More Implementation Details about Joint Training of Multiple Data Sources

Joint training under this circumstances can be conducted with any combination of the three datasets. Since improving the performance on real images is a more practical and meaningful solution to intrinsic image decompositions, we base our discussion mainly on the IIW dataset and its specific network structure.

Note as the guidance network and domain filter is employed only to further improve the piecewise constant effects while the primary intrinsic results are predicted by the direct intrinsic network, our modifications are mainly based on the DI network, which are detailed below.

(**i**). The image split of the main MPI benchmark is incorporated for joint training, instead of MIT. The MPI dataset contains enough quantity of training data compared to IIW data (4370 v.s. 4184), while MIT only has 110 training images, which is far from comparable to the others. (**ii**). The last layer of DI network with 1-channel output is adaptively changed to 3 channels to be compatible with MPI data. Then the albedo intensity is later calculated as the mean of three output channels. (**iii**). The loss layer for MPI data contains only MSE criterion for the albedo image without gradient supervision and consideration of the shading image. (**iv**). The weight to balance the loss function between IIW and MPI is empirically determined as 1 and 0.5 due to dramatically different back propagated gradients. (**v**). We jointly train the direct intrinsic network with both datasets, while finetuning the whole framework with only the IIW data.

*This work was done when Qingnan Fan was an intern at MSR.

# 3 Hinge Loss for Pairwise Comparison Data

The classification of reflectance pairs for the definition of WHDR error metric is calculated as

$$\hat{J}_\delta(\overline{R}_{k_1}, \overline{R}_{k_2}) = \begin{cases} 1 & \text{if } \overline{R}_{k_2}/\overline{R}_{k_1} > 1+\delta, \\ 2 & \text{if } \overline{R}_{k_1}/\overline{R}_{k_2} > 1+\delta, \\ E & \text{otherwise.} \end{cases} \tag{1}$$

Therefore, the SVM hinge loss [6] we used for pairwise comparison data can be written as

$$\mu(J_k, \overline{R}_{k_1}, \overline{R}_{k_2}, \delta, \xi) = \begin{cases} max(0, \frac{\overline{R}_{k_1}}{\overline{R}_{k_2}} - \frac{1}{1+\delta+\xi}) & \text{if } J_k = 1, \\ max(0, 1+\delta+\xi - \frac{\overline{R}_{k_1}}{\overline{R}_{k_2}}) & \text{if } J_k = 2, \\ max(0, \begin{cases} \frac{1}{1+\delta-\xi} - \frac{\overline{R}_{k_1}}{\overline{R}_{k_2}}, \\ \frac{\overline{R}_{k_1}}{\overline{R}_{k_2}} - (1+\delta-\xi) \end{cases}) & \text{if } J_k = E. \end{cases} \tag{2}$$

# 4 Training Details

The proposed framework is implemented with Torch. We adopt the Adam solver [5] for optimizing the loss functions. The learning rate is initialized with 0.01, and decreased to 0.001 to generate further improvement. The network parameters are initialized with the method described in [3]. The training samples are randomly shuffled at the beginning of each epoch. We first train the direct intrinsic network and guidance network separately and in parallel, then jointly train (finetune) the entire network. The whole training procedure takes up to 12 hours on an NVIDIA Tesla P100 GPU.

# 5 More Visual Results on the IIW dataset

Figure 1 shows more visual results on the IIW dataset. The intermediate results are also presented for better understanding of our framework. The guidance network predicts more sparse and salient structures ($E'$) compared to the input ($E(I)$). Guided by the learned edge map ($E'$), the initially predicted reflectance ($R'$) is flattened by domain filter and appears more realistic and piecewise constant. From the final output albedo and shading images, it can be observed that lighting and surface reflectance are well separated.

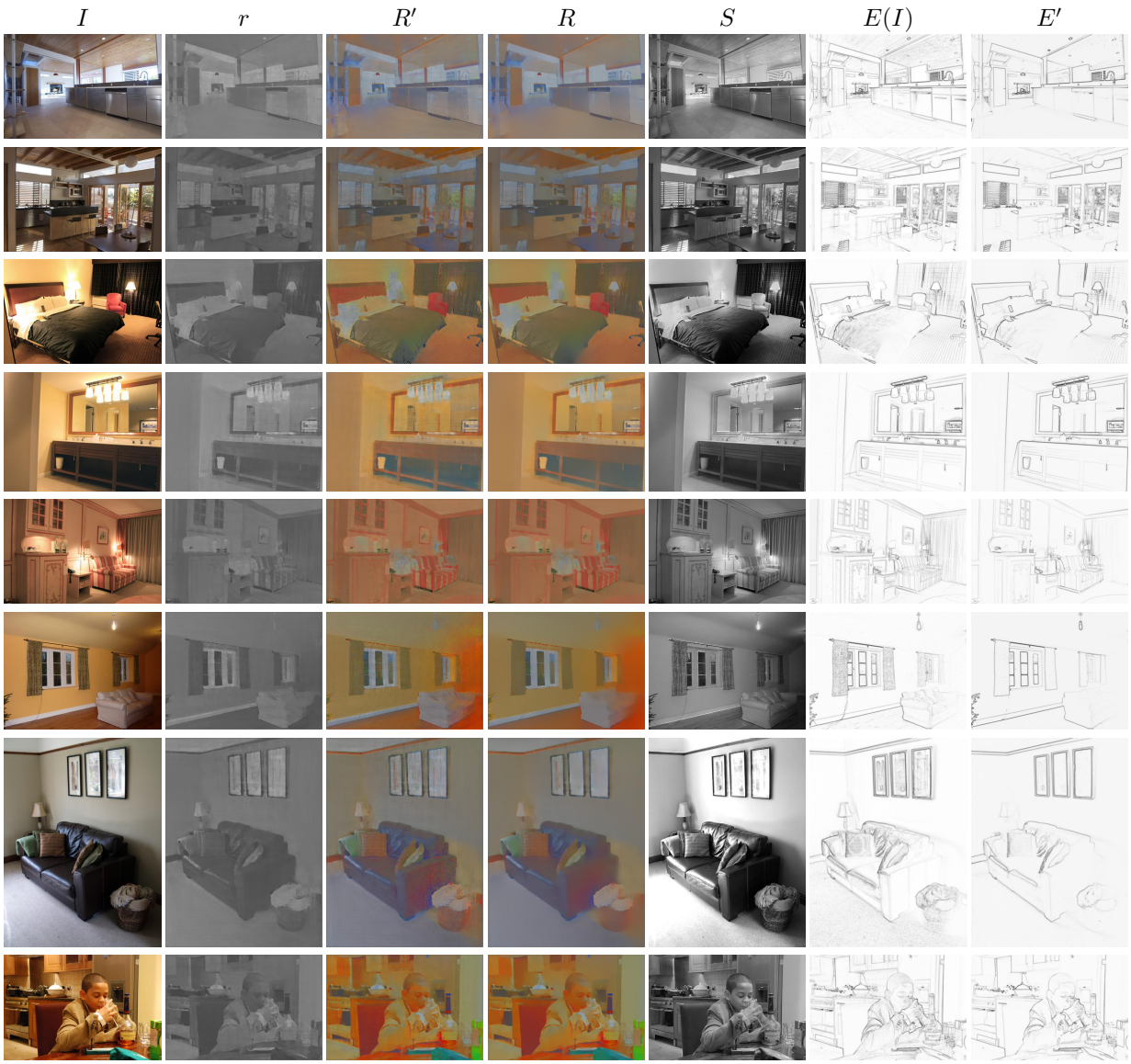


Figure 1: More visual results on the IIW dataset with intermediate outputs. Note the filtered albedo images ($R$) guided by predicted edge map ($E'$) are smoother and more realistic compared to the unfiltered albedo ($R'$). As can be seen, the lighting $S$ is mostly separated from the reflectance $R$.

# 6 Test of Generalization to Outdoor Images

In this section, we evaluate the network trained on IIW dataset, which is composed of mostly indoor images, on outdoor scenes from a completely different, unrelated data source. As shown in Figure 2, the lighting and reflectance are well separated from each other for these outdoor images. For example, the dark shadow under the truck is mostly classified as part of shading layer. These results demonstrate the effectiveness of our framework on various different scenes.

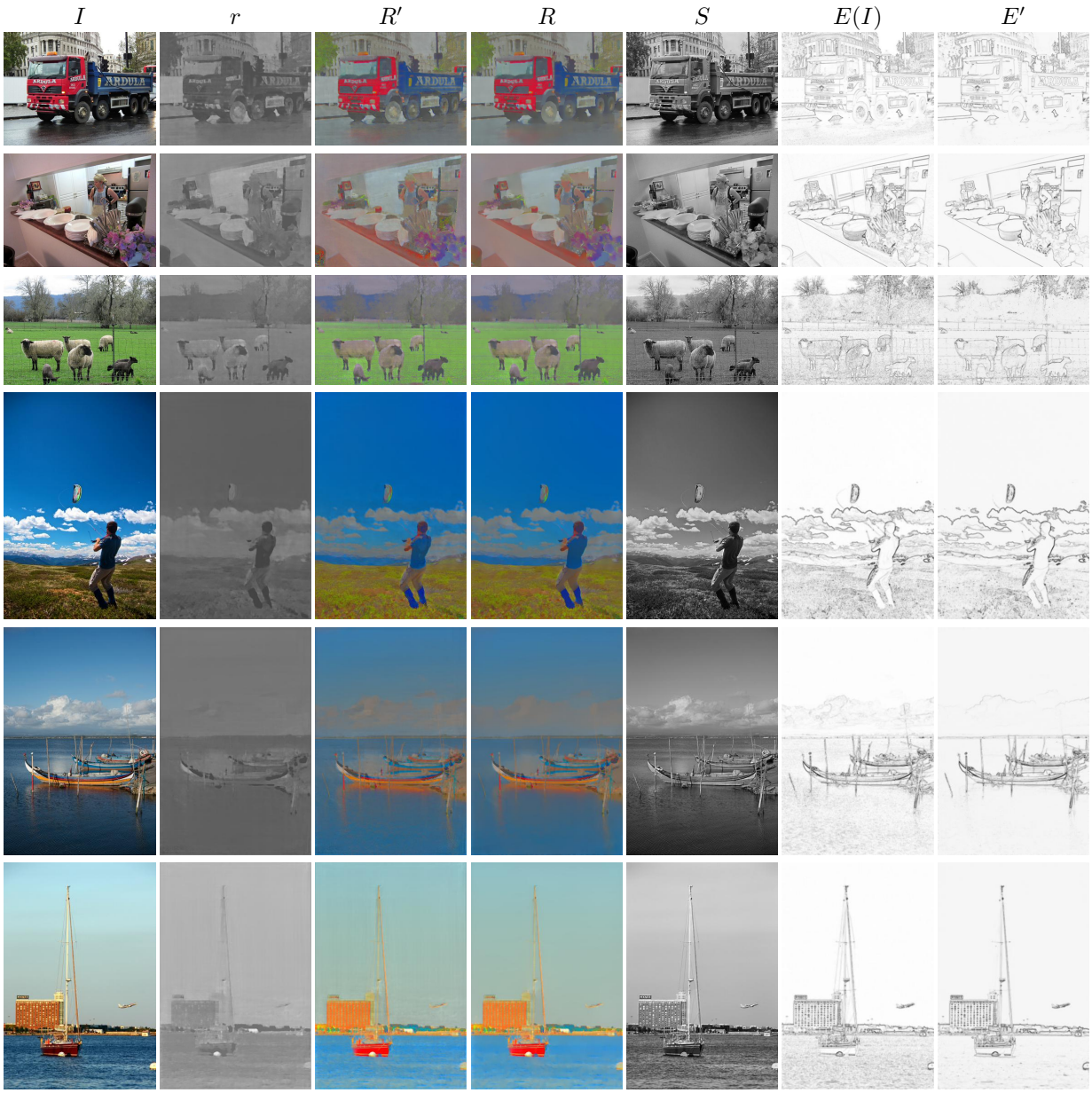


Figure 2: Evaluation of the generalization ability to outdoor scenes of the network trained on the IIW dataset which contains mostly indoor images.

# 7 More Visual Results on the MIT intrinsic dataset

In this section, we show more qualitative results on the MIT intrinsic benchmark. Note our proposed network is only trained on 110 images, and the training and testing splits have no overlap. It can be seen in Figure 3 that the shading is also recovered well in some dark shadow regions such as on the top right of turtle case.

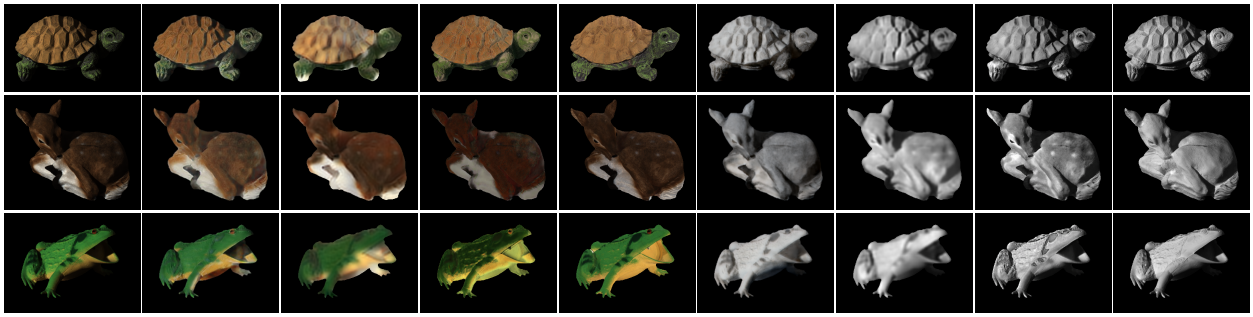


Figure 3: **Qualitative Comparison on MIT Intrinsic Benchmark**. Compared with Shi *et al.* [7] and MSCR [8] on Barron *et al.*'s test split of MIT intrinsic dataset, our algorithm achieves the best/sharpest results.

# 8 Additional Details Regarding MPI-Sintel dataset

Figure 4 compares the images in the *main* and *auxiliary* versions (discussed in the main paper) of the MPI-Sintel benchmark. Note also that the main MPI-Sintel benchmark used by most previous methods contains many defective pixels. We remove the defective images for evaluation on the *image split* following [8], and remove two scenes, "bandage_1" and "shaman_3", for evaluation on the *scene split*. Additionally, to calculate the error on the MIT intrinsic dataset, we use the publicly-available evaluation code of [1] for all methods.

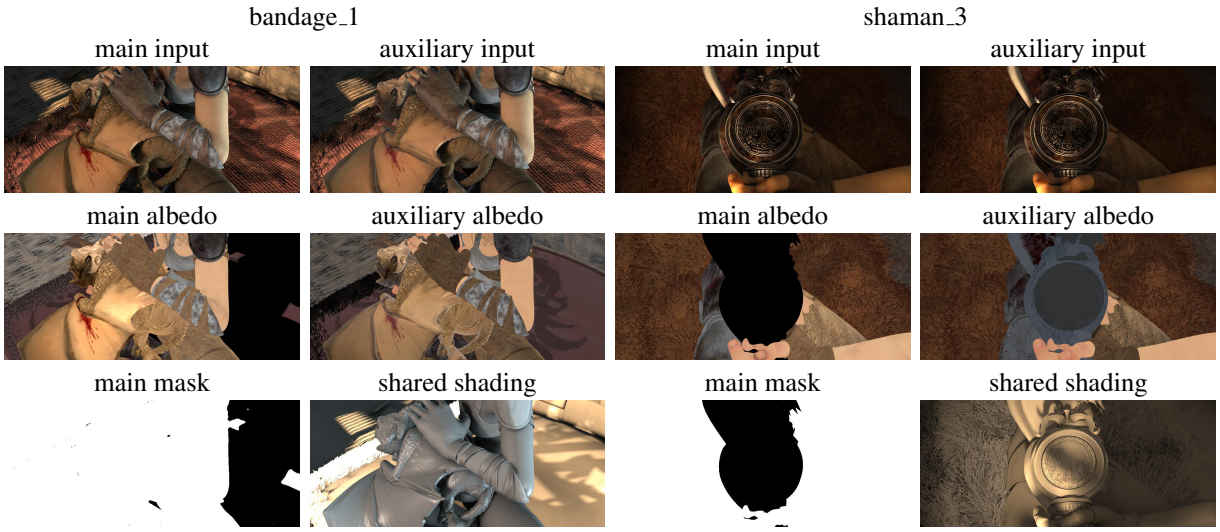


Figure 4: Illustration of differences between the *main* and *auxiliary* versions of the MPI-Sintel benchmark. Note that in the main version, the above albedo images contain many defective pixels due to rendering issues, which is not the case in the auxiliary version. The shading image is shared by the main and auxiliary datasets.

# 9 More Visual Results on MPI-Sintel dataset

This section presents more visual results of our approach compared against existing intrinsic image decomposition methods. Specifically, using the *image split* of the MPI-Sintel benchmark, we provide visual comparisons with several traditional and deep learning based algorithms on the main version of the dataset in Figure 5 and 6. Additionally, we compare our algorithm with [4] on the auxiliary version of the dataset in Figure 7, since [4] is the only previous method trained and tested on this version. On the *scene split* which is much more difficult, we show some qualitative results in Figure 8 and 9.

Our method does not require any additional training data or information such as depth. Compared to the state-of-the-art methods, our results are much sharper, clearer and more piecewise constant on both the *image split* and *scene split* of the main and auxiliary versions of the MPI Sintel benchmark.

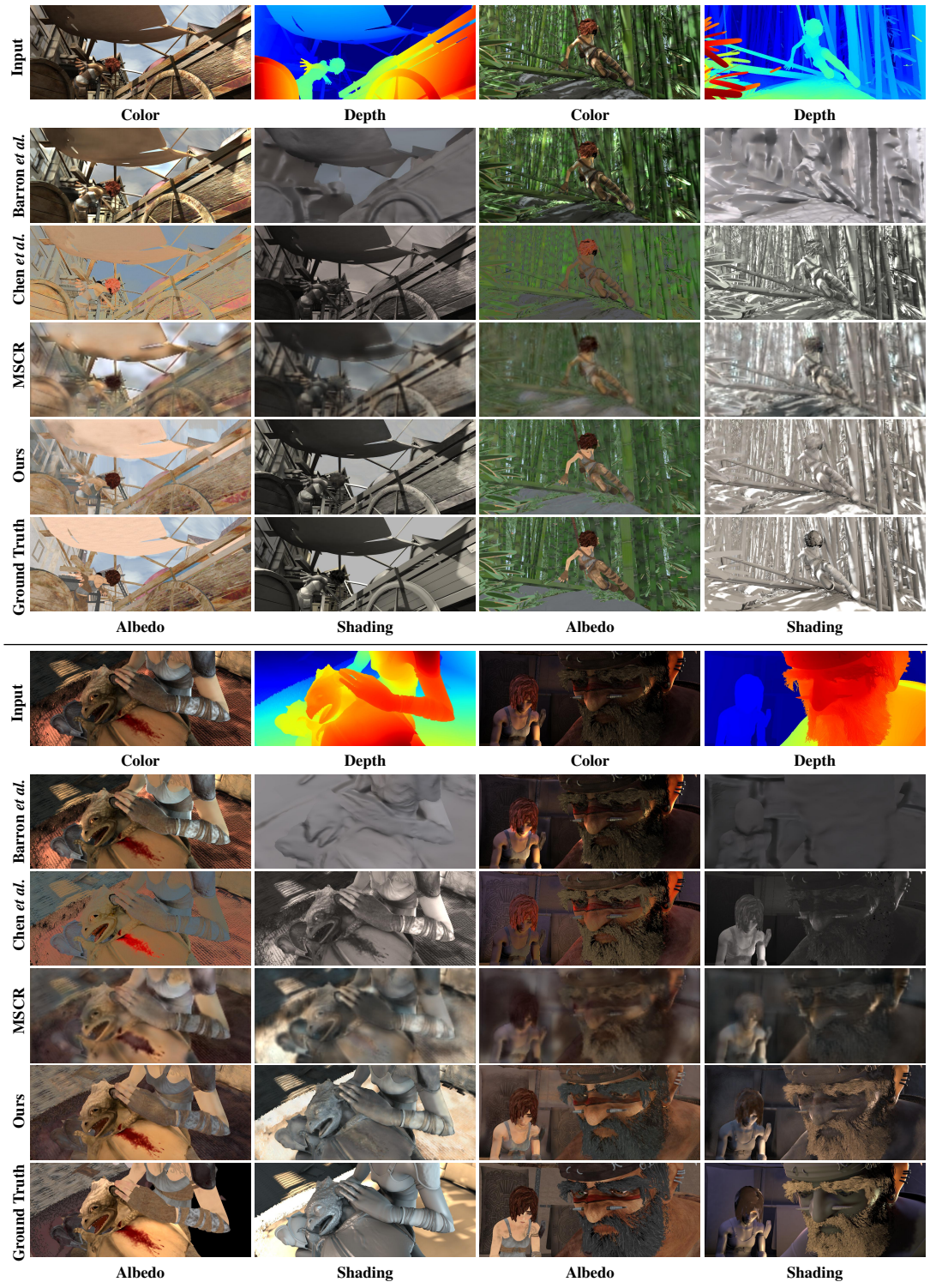


Figure 5: Qualitative comparisons with Barron *et al.* [1], Chen *et al.* [2] and MSCR [8] on the *image split* of the main MPI-Sintel Benchmark .

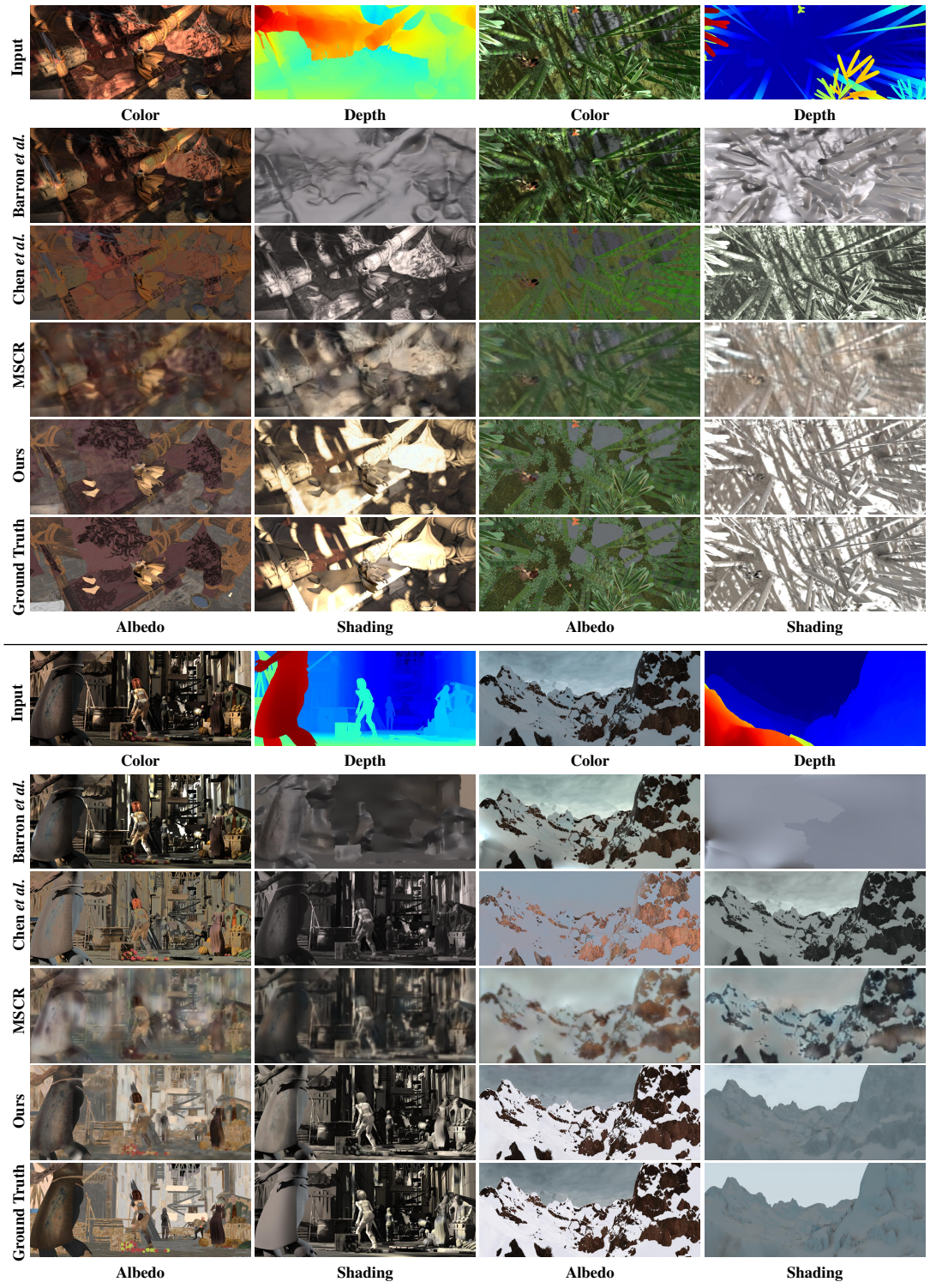


Figure 6: Qualitative comparisons with Barron *et al.* [1], Chen *et al.* [2] and MSCR [8] on the *image split* of the main MPI-Sintel Benchmark .

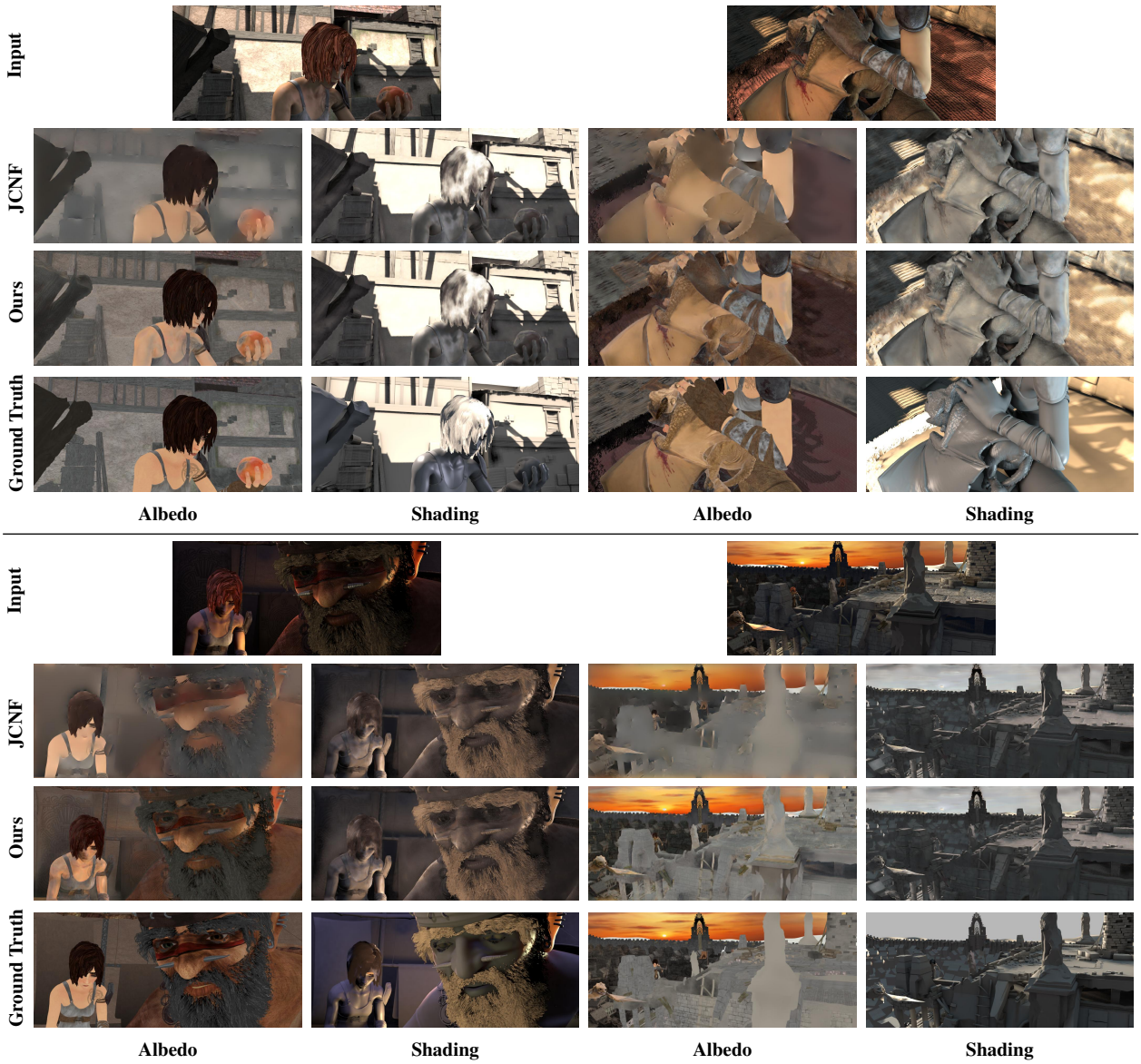


Figure 7: Qualitative comparisons with JCNF [4] on the *image split* of the auxiliary MPI-Sintel Benchmark

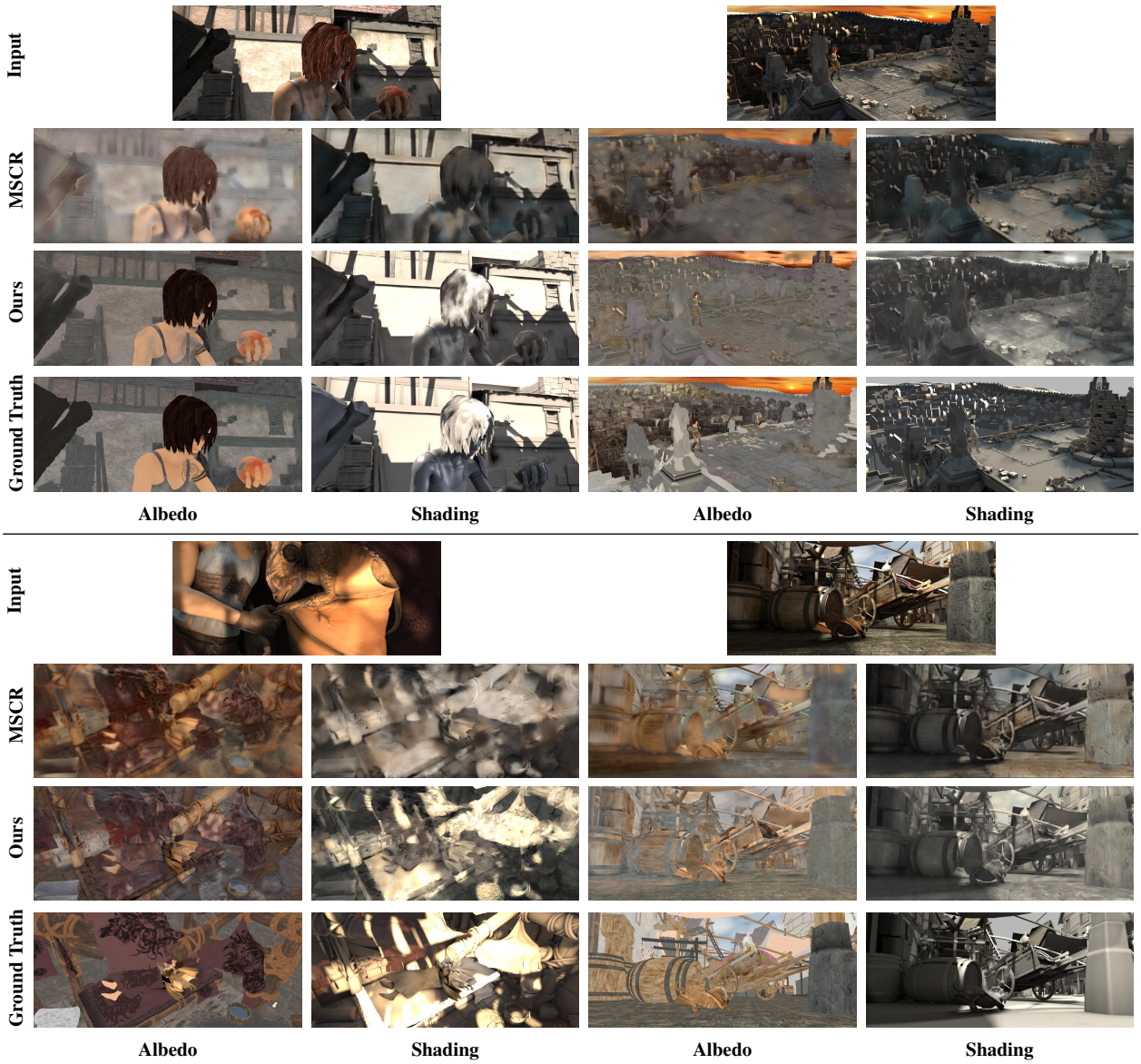


Figure 8: Qualitative comparisons with MSCR [8] on the challenging *scene split* of the MPI-Sintel Benchmark.


Figure 9: Qualitative comparisons with MSCR [8] on the challenging *scene split* of the MPI-Sintel Benchmark.

# References


[1] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *PAMI*, 37(8):1670–1687, 2015. 5, 7, 8

[2] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *ICCV*, 2013. 7, 8

[3] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 2

[4] S. Kim, K. Park, K. Sohn, and S. Lin. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *ECCV*, 2016. 6, 9

[5] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 2

[6] T. Nestmeyer and P. V. Gehler. Reflectance adaptive filtering improves intrinsic image estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[7] J. Shi, Y. Dong, H. Su, and S. X. Yu. Learning non-lambertian object intrinsics across shapenet categories. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[8] M. M. Takuya Narihira and S. X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015. 5, 7, 8, 10, 11