Scene Reconstruction from High Spatio-Angular Resolution Light Fields Supplementary Material

Changil Kim[1,2] Henning Zimmer[1,2] Yael Pritch[1] Alexander Sorkine-Hornung[1] Markus Gross[1,2]

[1]Disney Research Zurich [2]ETH Zurich

1 Stereo Comparison on Classic Datasets

1.1 Middlebury Stereo Data

In Figure 1 we compare our method to two stereo methods [Zitnick et al. 2004; Szeliski and Scharstein 2002]. Both methods initially match image segments or patches and then refine these coarse estimates using a smoothing or propagation strategy. For evaluation we use Middlebury stereo data sets with the ground truth[1], enabling a quantitative comparison. Note that we use all input images (5 images for *Tsukuba* and 8 images for *Venus* and *Sawtooth*) whereas the other methods use two images only. In Table 1 we compare the estimation errors, for which we compute the percentage of bad estimates. We consider an estimate bad if its difference from the ground truth disparity is larger than a threshold T.

It takes about two seconds for our method to process these data sets as they are both angularly and spatially at low resolution (5 or 8 images with 0.1–0.2 MP each). See Table 1 for comparison. Szeliski and Scharstein [2002] report 4.7 seconds for *Tsukuba* data set, which was measured on a 750 MHz Pentium III CPU. Zitnick et al. [2004] do not provide run times. The quality of our results for these data sets is not optimal. This can be due to the low spatio-angular resolutions since our method is specifically designed to operate at the pixel level by leveraging the redundancy and coherence in high resolution light fields. For such data sets, methods based on image patch comparisons and global regularization perform better.

Table 1: *Quantitative comparison on Middlebury stereo data. We report errors as the percentage of bad pixels with T=1.*

	Tsukuba	Venus	Sawtooth
[Zitnick et al. 2004]	1.87	1.85	n/a
[Szeliski and Scharstein 2002]	4.9	n/a	n/a
Ours	8.42	10.59	6.25
run time	*1.4 s*	*2.4 s*	*2.6 s*

1.2 Zitnick et al.'s Multi-View Stereo Data

In Figure 2 we compare our method to a multi-view stereo method [Zitnick et al. 2004] using their data sets[2] consisting of videos captured by eight synchronized cameras at a resolution of 0.8 MP.

Although the data sets are at a higher resolution than Middlebury stereo data, we found them particularly challenging for our method. The main reasons are considerable noise and exposure changes between cameras. As in the previous comparison, the assumptions our method is based on do not hold with these data sets. While we showed in the paper that our depth score is robust to large (but sporadic) outliers, the amount of outliers in these data sets is too high. Handling such difficult scenarios thus seems an interesting direction for future research.

2 Depth from 4D Light Fields

As described in Section 5.4 of the paper, our method can be extended for 4D light fields. In Figure 3 we use synthetic light field data from the HCI lab[3] to quantitatively evaluate our results on 4D light fields using the available ground truth depth. The error measures are summarized in Table 2 where we use the same measurement as in Section 1. We report two error measures with different threshold values to count the bad estimates. Note that these synthetic data sets were originally published with a paper [Wanner and Goldlücke 2012], but the data currently available on the web page differs from those reported in their paper. Thus, we omit a direct comparison, but in the paper we compare to their method using a 4D light field from the Stanford database. See Figure 14 in the paper.

As can be observed, our method produces high quality depth estimates on this data. Processing 9×9 images at a resolution of 0.6 MP requires 28 seconds.

Table 2: *Quantitative comparison on 4D light fields. We report errors as the percentage of bad pixels using different threshold values T.*

	Buddha	Mona	Papillon	StillLife
T = 0.1	0.89	3.81	4.93	4.33
T = 0.5	0.21	0.27	0.34	0.49

3 3D Meshes

To further assess our reconstruction quality we show in Figure 4 3D meshes corresponding to the data sets presented in Figure 6 in the paper. The meshes were obtained from our reconstructions by triangulating individual depth maps and merging them into a single model. To enhance visualization we color coded vertices according to their depth (red for near vertices and blue for far).

Although our reconstructions have a lower accuracy in terms of absolute distance compared to a laser scanner, we manage to faithfully reproduce fine details of the complex, cluttered scenes and obtain precise reconstruction of object contours.

References

SZELISKI, R., AND SCHARSTEIN, D. 2002. Symmetric sub-pixel stereo matching. In *ECCV*.

WANNER, S., AND GOLDLÜCKE, B. 2012. Globally consistent depth labeling of 4D light fields. In *CVPR*.

ZITNICK, C. L., KANG, S. B., UYTTENDAELE, M., WINDER, S., AND SZELISKI, R. 2004. High-quality video view interpolation using a layered representation. *ACM Trans. Graph. 23*, 3.

[1] http://vision.middlebury.edu/stereo/

[2] http://research.microsoft.com/en-us/um/people/larryz/videoviewinterpolation.htm

[3] http://hci.iwr.uni-heidelberg.de/HCI/Research/LightField

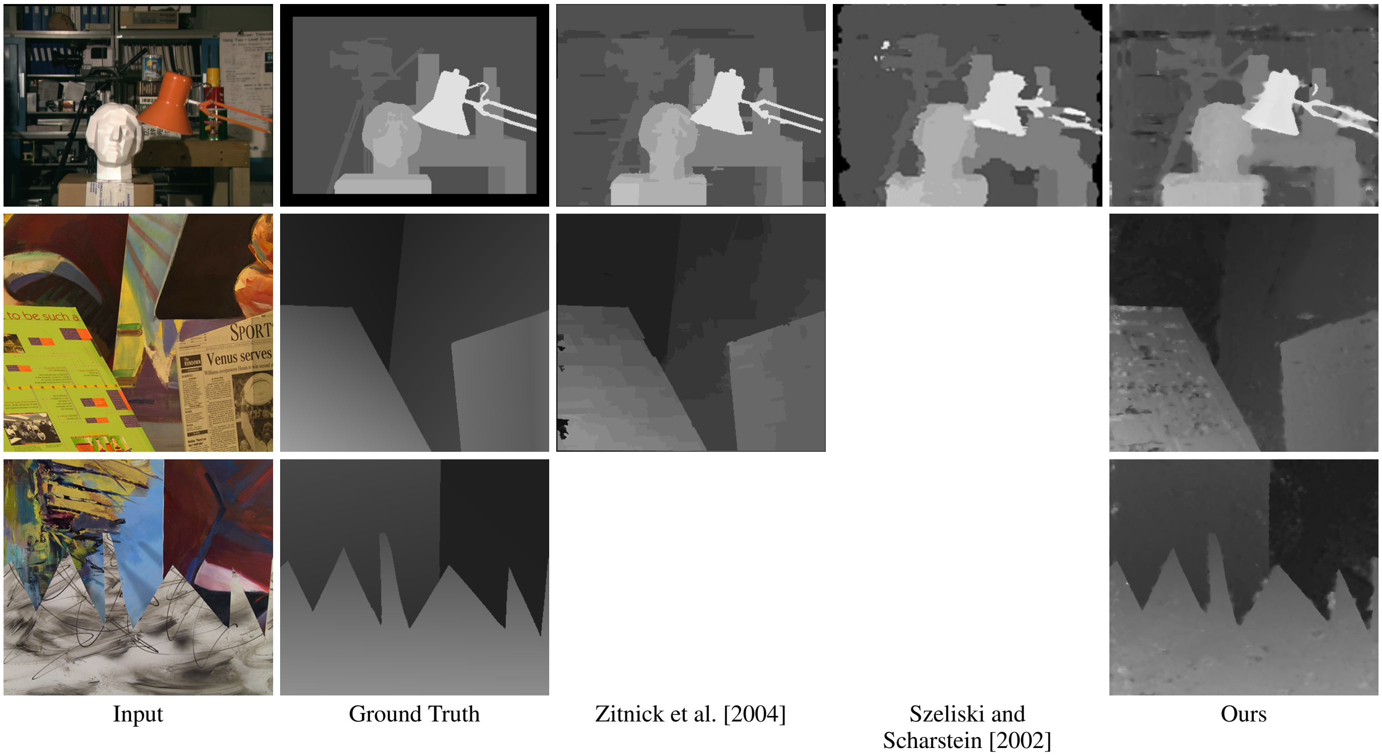

Figure 1: *Comparison to stereo methods on Middlebury data sets with ground truth.* ***From top to bottom:*** Tsukuba*,* Venus *and* Sawtooth*.*

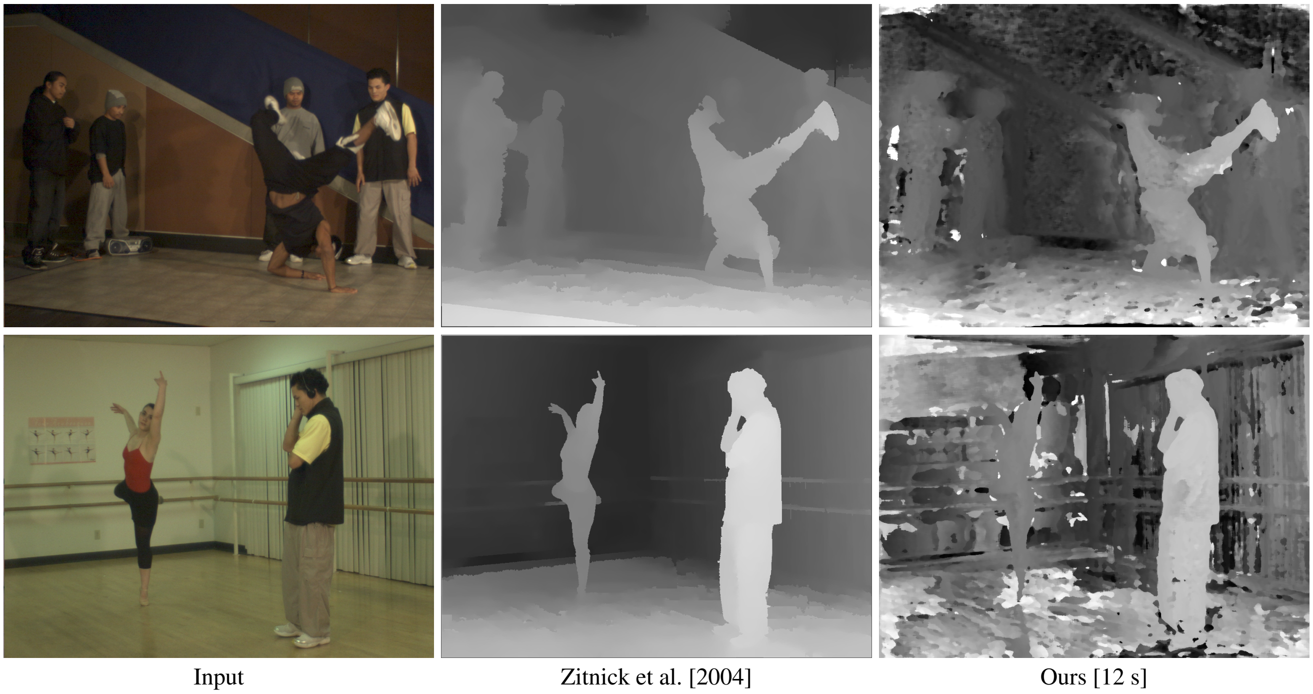

Figure 2: *Comparison to the multi-view stereo method of Zitnick et al.* ***From top to bottom:*** Breakdancing *and* Ballet*.*

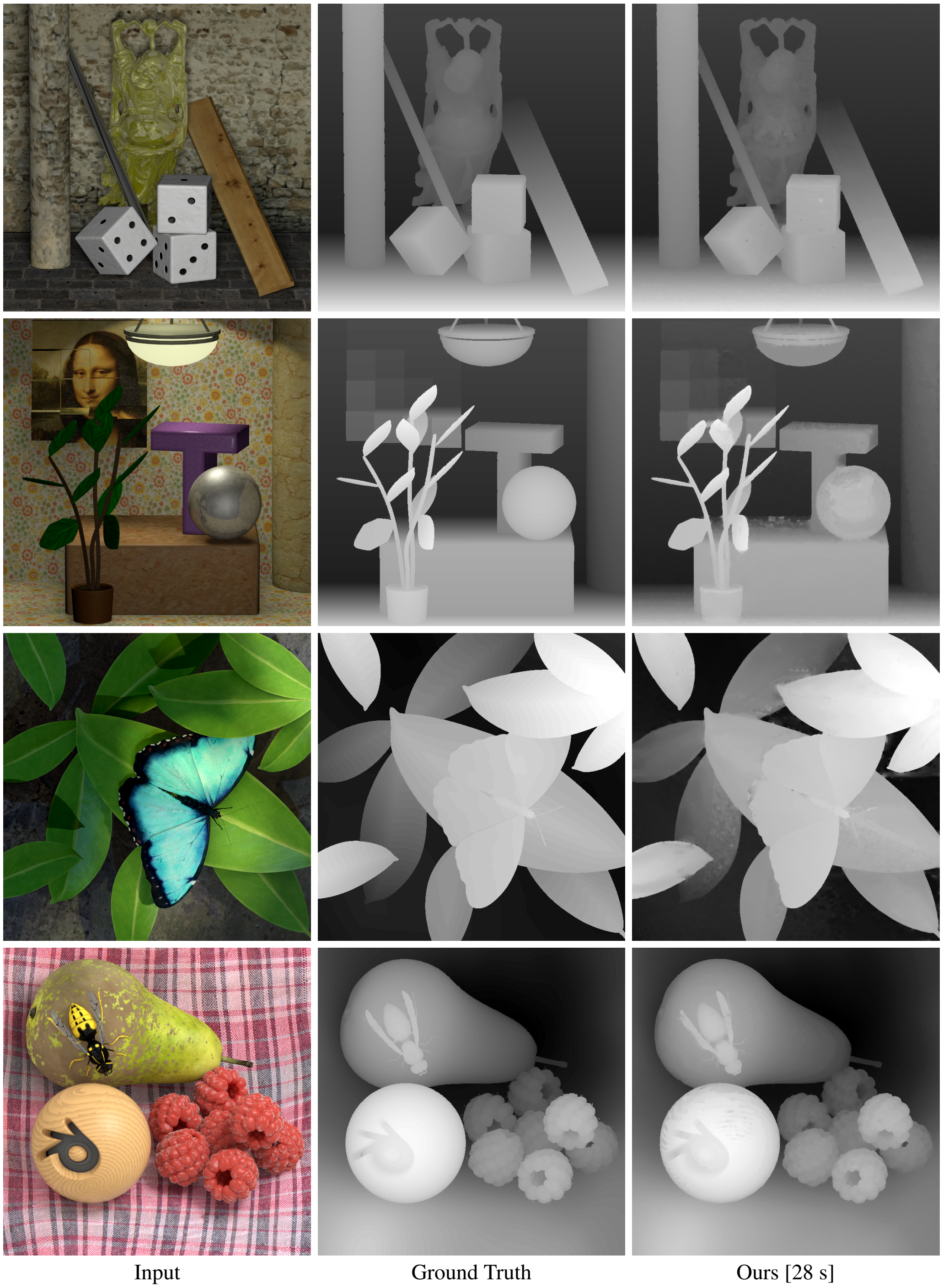

Figure 3: *Results on 4D HCI light field data with ground truth.* ***From top to bottom:*** Buddha, Mona, Papillon *and* StillLife.

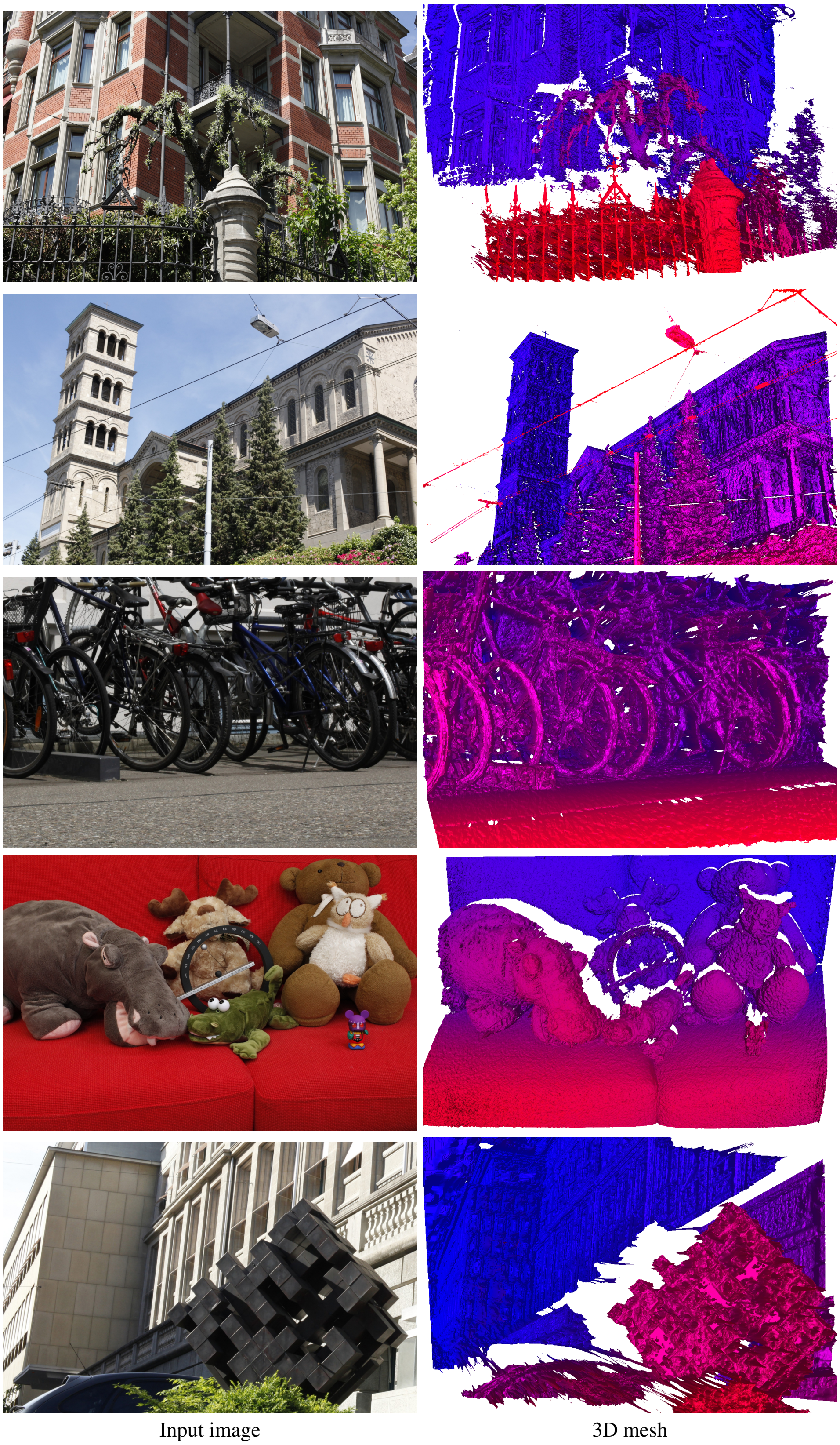

Figure 4: *Visualization of 3D meshes (color coded: red near, blue far).* ***From top to bottom:*** Mansion, Church, Bikes, Couch, *and* Statue.