Supplemental Material for Differentiable Diffusion for Dense Depth Estimation from Multi-view Images

Numair Khan
Brown University

Min H. Kim
KAIST

James Tompkin
Brown University

Table 1: **(Best viewed in color)** Quantitative comparison of our method against five baseline methods on real-world Stanford (top) and EPFL (bottom) datasets. As a metric, we use reprojection error $\times 10^{-2}$ as a proxy for depth accuracy as no ground truth exists. The top three results are highlighted in gold, silver, and bronze.

Light Fields	Reprojection Error					
	[6]	[4]	[1]	[5]	[2]	Ours
Bulldozer	2.06	1.90	3.57	1.87	2.44	1.67
Bunny	1.18	1.08	1.57	1.08	1.10	0.96
Chess	1.72	3.19	1.63	1.54	1.66	2.22
Eucalyptus	1.19	1.03	1.55	1.11	1.14	1.01
Jelly Beans	1.36	1.78	1.71	1.40	1.68	1.27
Lego	3.13	3.50	4.98	2.57	3.19	2.38
Tarot	15.1	12.1	17.1	10.5	17.2	9.98
Treasure	2.51	1.89	2.35	1.96	2.18	1.87
Truck	1.49	1.39	1.80	1.33	1.37	1.24
Average	3.30	3.10	4.03	2.60	3.55	2.51
Bikes	2.26	4.85	2.24	5.13	2.39	2.24
Grid	3.33	6.46	3.05	6.84	3.20	2.87
Silos	1.77	3.34	1.85	3.41	2.01	1.69
Sphynx	2.37	4.80	2.31	5.21	2.50	2.30
Average	2.43	4.86	2.36	5.14	2.53	2.28

A. Expanded Results

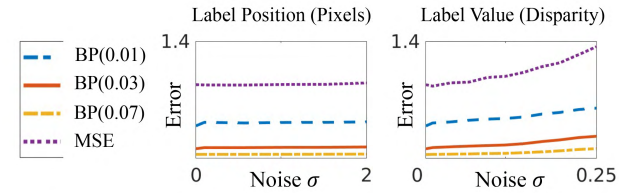
In Table 1, we compare performance on the real-world Stanford and EPFL light fields, with the methods of Zhang at al. [6], Li et al [4], Jiang et al. [1], Shi et al. [5], and the central-view results [2] of Khan et al. [3]. No ground truth depth data exists for these scenes. As a proxy for depth accuracy we use reprojection error $\times 10^{-2}$ in RGB color space induced by warping the central view onto the corner views using the estimated disparity map.

Our approach is competitive or better than other methods except on the *Chess* light field, which exhibits strong specular effects due to polished metal materials. However, featureless backgrounds cause our edges to be diffuse (Figure 2).

We also include error maps for all light field datasets, along with depth maps for all example light fields listed in the main and supplemental documents. Please see Figures 3 and 4 for synthetic scenes, Figures 5 and 6 for the Stanford scenes, and Figures 7 and 8 for the EPFL scenes. For the additional results presented here, we only run a single pass over each parameter.

B. Noise Handling

We evaluate the robustness of our method to noise by randomly adding Gaussian noise of increasing variance to 50% of the original points. We find our method robust to position noise and still capable under significant disparity noise. Figure 1 shows results for the *Dino* scene



Visualization of Gaussian noise ($\sigma = 0.25$) in label value (disparity):

Figure 1: Robustness of our method to noise (*Dino* light field). The ground truth scene has mean disparity -0.1 with SD 0.73.

C. Supervised Loss

In the main paper, we describe a validation of our differentiable rendering and diffusion approach using a supervised loss against ground truth data. In Figure 9, we present MSE and bad pixel metrics over iterations of the optimization for both the HCI dataset and our new *living room* and *piano* realistic scenes. For comparison, we also mark the performance of five existing methods. In Figure 10, this experiment shows that our approach

can produce errors close to zero, and validates the potential of such an approach in the best case.

Still, why do the errors not reduce to zero? As Figure 10 shows, the presence of outliers in the original point set and the lack of labels in regions with fine detail prevents the diffusion from entirely eliminating errors. Such sources of error occur in all six light fields to varying degrees.

D. Point Sparsity

In Figure 12, we evaluate the performance of the optimization routine by varying the point set size. We observe that for the bad pixels metrics, the routine converges at the same minima even when only half the original point set is used. For MSE, having more points is helpful. This is expected as a larger point set is more helpful in filtering the effect of outliers.

E. Ray Attenuation Formula

The attenuation factor T at distance s from the image plane is defined as

$$\begin{aligned}
\mathrm{T}(x,y,s) &= \exp\left(-\int_0^s \rho \sum_{\mathbf{x}\in\mathcal{P}} \exp\left(-\left(\frac{(x-x_\mathbf{x})^2}{2\sigma_\mathrm{S}^2}+\frac{(y-y_\mathbf{x})^2}{2\sigma_\mathrm{S}^2}+\frac{(z-Z_\mathbf{x})^2}{2\sigma_Z^2}\right)\right)dz\right) \\
&= \exp\left(-\int_0^s \rho \sum_{\mathbf{x}\in\mathcal{P}} \exp\left(-\frac{(x-x_\mathbf{x})^2}{2\sigma_\mathrm{S}^2}-\frac{(y-y_\mathbf{x})^2}{2\sigma_\mathrm{S}^2}-\frac{(z-Z_\mathbf{x})^2}{2\sigma_Z^2}\right)dz\right) \\
&= \exp\left(-\int_0^s \rho \sum_{\mathbf{x}\in\mathcal{P}} \exp\left(-\frac{(x-x_\mathbf{x})^2}{2\sigma_\mathrm{S}^2}-\frac{(y-y_\mathbf{x})^2}{2\sigma_\mathrm{S}^2}\right)\exp\left(-\frac{(z-Z_\mathbf{x})^2}{2\sigma_Z^2}\right)dz\right) \\
&= \exp\left(-\int_0^s \rho \sum_{\mathbf{x}\in\mathcal{P}} \frac{\mathrm{S}^\mathbb{Z}_\mathbf{x}(x,y)}{Z_\mathbf{x}}\exp\left(-\frac{(z-Z_\mathbf{x})^2}{2\sigma_Z^2}\right)dz\right) \quad \text{(From Equation 4, main paper.)} \\
&= \exp\left(-\sum_{\mathbf{x}\in\mathcal{P}}\int_0^s \rho\, \frac{\mathrm{S}^\mathbb{Z}_\mathbf{x}(x,y)}{Z_\mathbf{x}}\exp\left(-\frac{(z-Z_\mathbf{x})^2}{2\sigma_Z^2}\right)dz\right)
\end{aligned}$$

As $\sigma_Z \to 0$, the density contribution at any point s along the ray will come from only a single Gaussian. Thus, we can write

$$\begin{aligned}
\mathrm{T}(x,y,s) &= \exp\left(-\int_0^{s_1} \rho\, \frac{\mathrm{S}^\mathbb{Z}_{\mathbf{x}_1}(x,y)}{Z_{\mathbf{x}_1}}\exp\left(-\frac{(z-Z_{\mathbf{x}_1})^2}{2\sigma_Z^2}\right)dz - \int_0^{s_2} \rho\, \frac{\mathrm{S}^\mathbb{Z}_{\mathbf{x}_2}(x,y)}{Z_{\mathbf{x}_2}}\exp\left(-\frac{(z-Z_{\mathbf{x}_2})^2}{2\sigma_Z^2}\right)dz - \ldots \right. \\
&\quad \left. - \int_0^{s_n} \rho\, \frac{\mathrm{S}^\mathbb{Z}_{\mathbf{x}_n}(x,y)}{Z_{\mathbf{x}_n}}\exp\left(-\frac{(z-Z_{\mathbf{x}_n})^2}{2\sigma_Z^2}\right)dz\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{T}(x,y,s) &= \exp\left(-\int_0^{t} \rho\, \frac{\mathrm{S}^\mathbb{Z}_{\mathbf{x}_1}(x,y)}{Z_{\mathbf{x}_1}}\exp\left(-\frac{(z-\mu_z)^2}{2\sigma_Z^2}\right)dz - \int_0^{t} \rho\, \frac{\mathrm{S}^\mathbb{Z}_{\mathbf{x}_2}(x,y)}{Z_{\mathbf{x}_2}}\exp\left(-\frac{(z-\mu_z)^2}{2\sigma_Z^2}\right)dz - \ldots \right. \\
&\quad \left. - \int_0^{t} \rho\, \frac{\mathrm{S}^\mathbb{Z}_{\mathbf{x}_n}(x,y)}{Z_{\mathbf{x}_n}}\exp\left(-\frac{(z-\mu_z)^2}{2\sigma_Z^2}\right)dz\right) \quad \text{(Figure 5, main paper describes the bounds } [0,t].\text{)} \\
&= \prod_\mathbf{x} \exp\left(-\int_0^t \rho \frac{\mathrm{S}^\mathbb{Z}_\mathbf{x}(x,y)}{Z_\mathbf{x}}\exp\left(-\frac{(z-\mu_z)^2}{2\sigma_Z^2}\right)dz\right) \\
&= \prod_\mathbf{x} \mathrm{T}_\mathbf{x}(x,y)
\end{aligned}$$

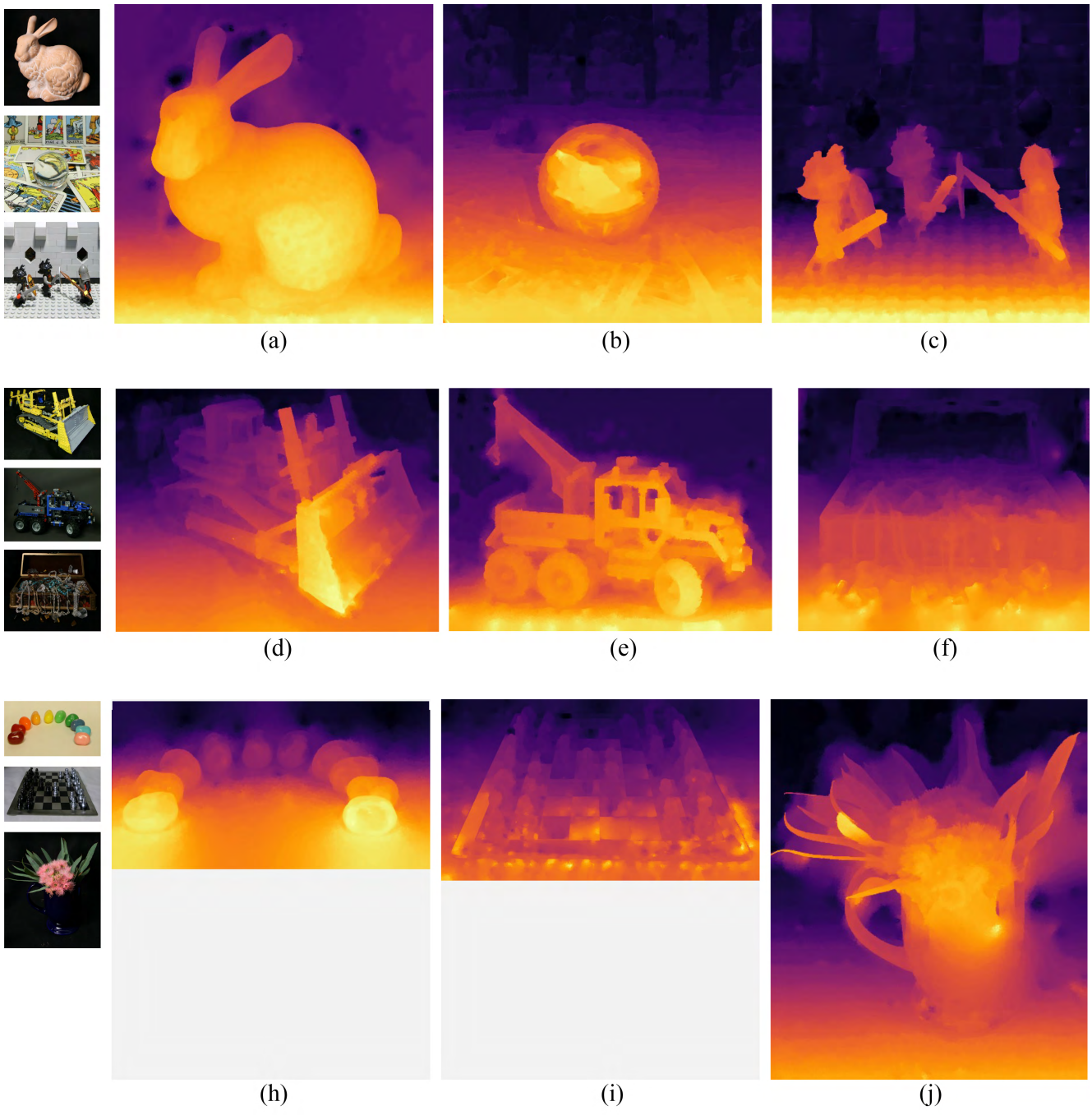


Figure 2: Our results on the real world light fields of the Stanford dataset.

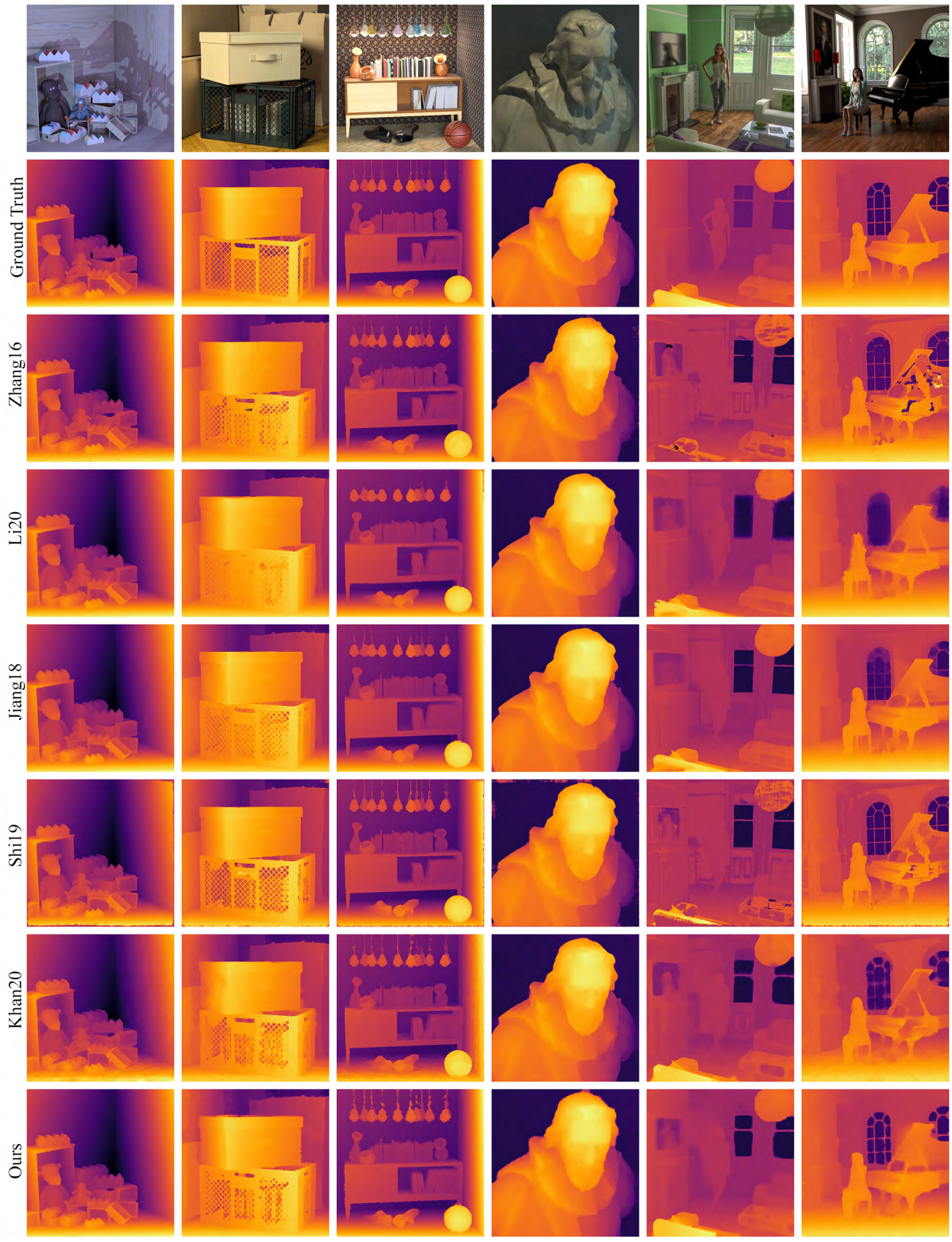


Figure 3: Qualitative comparison of our method with all baselines on the HCI dataset, and the light fields *Living Room* and *Piano*. From top to bottom: Zhang et al. [6], Li et al. [4], Jiang et al. [1], Shi et al. [5], Khan et al. [2], and our approach.

Figure 4: A visualization per-pixel L1 error for all methods (lighter regions have higher error). From top to bottom: Zhang et al. [6], Li et al. [4], Jiang et al. [1], Shi et al. [5], Khan et al. [2], and our approach.

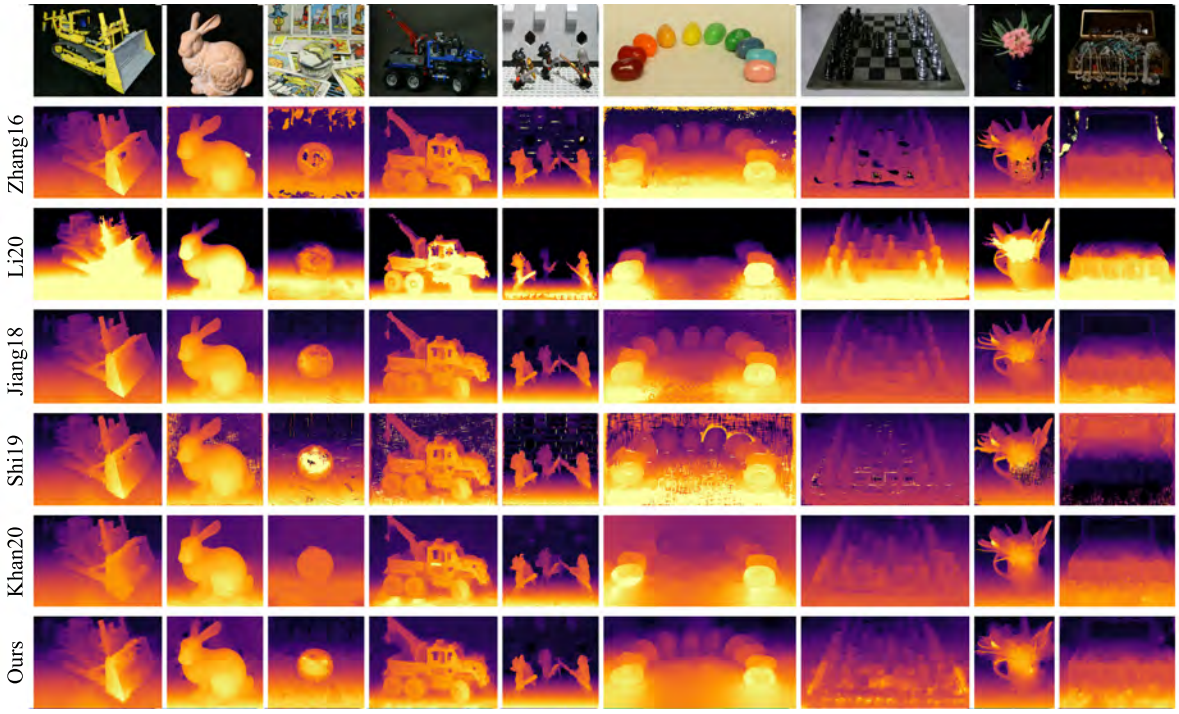


Figure 5: Qualitative comparison of our method with all baselines on the Stanford dataset. From top to bottom: Zhang et al. [6], Li et al. [4], Jiang et al. [1], Shi et al. [5], Khan et al. [2], and our approach.

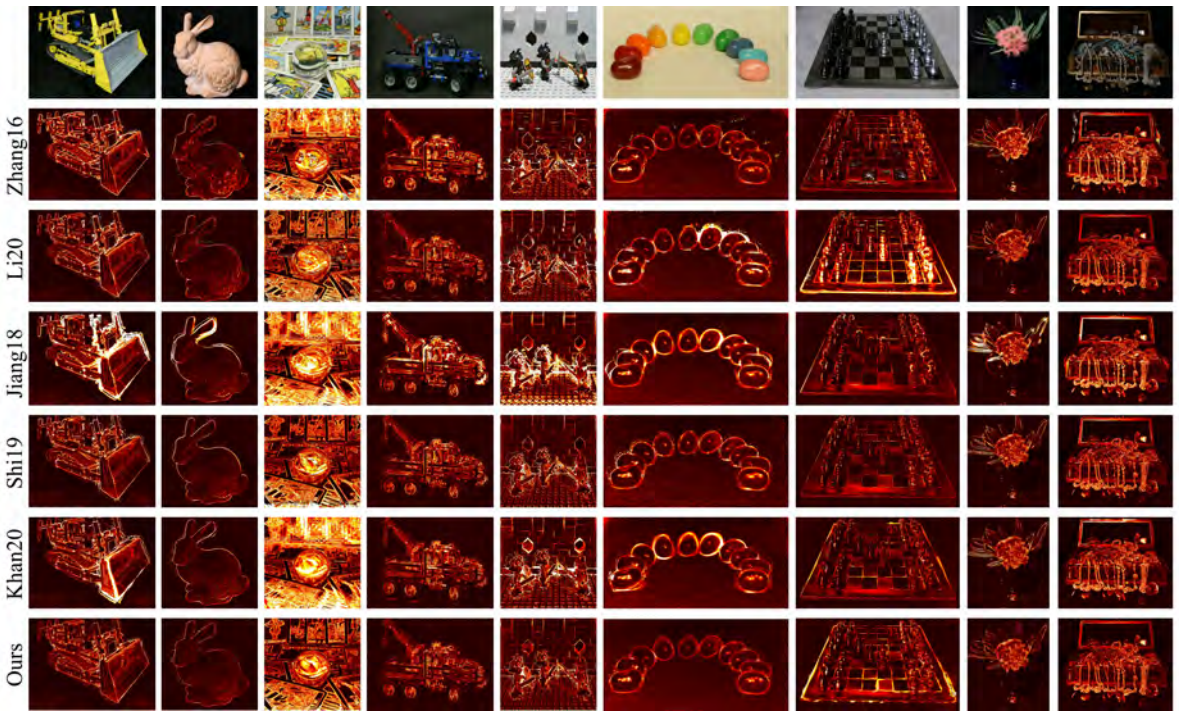


Figure 6: A visualization of the reprojection error for all methods on the Stanford dataset (lighter regions have higher error). From top to bottom: Zhang et al. [6], Li et al. [4], Jiang et al. [1], Shi et al. [5], Khan et al. [2], and our approach.

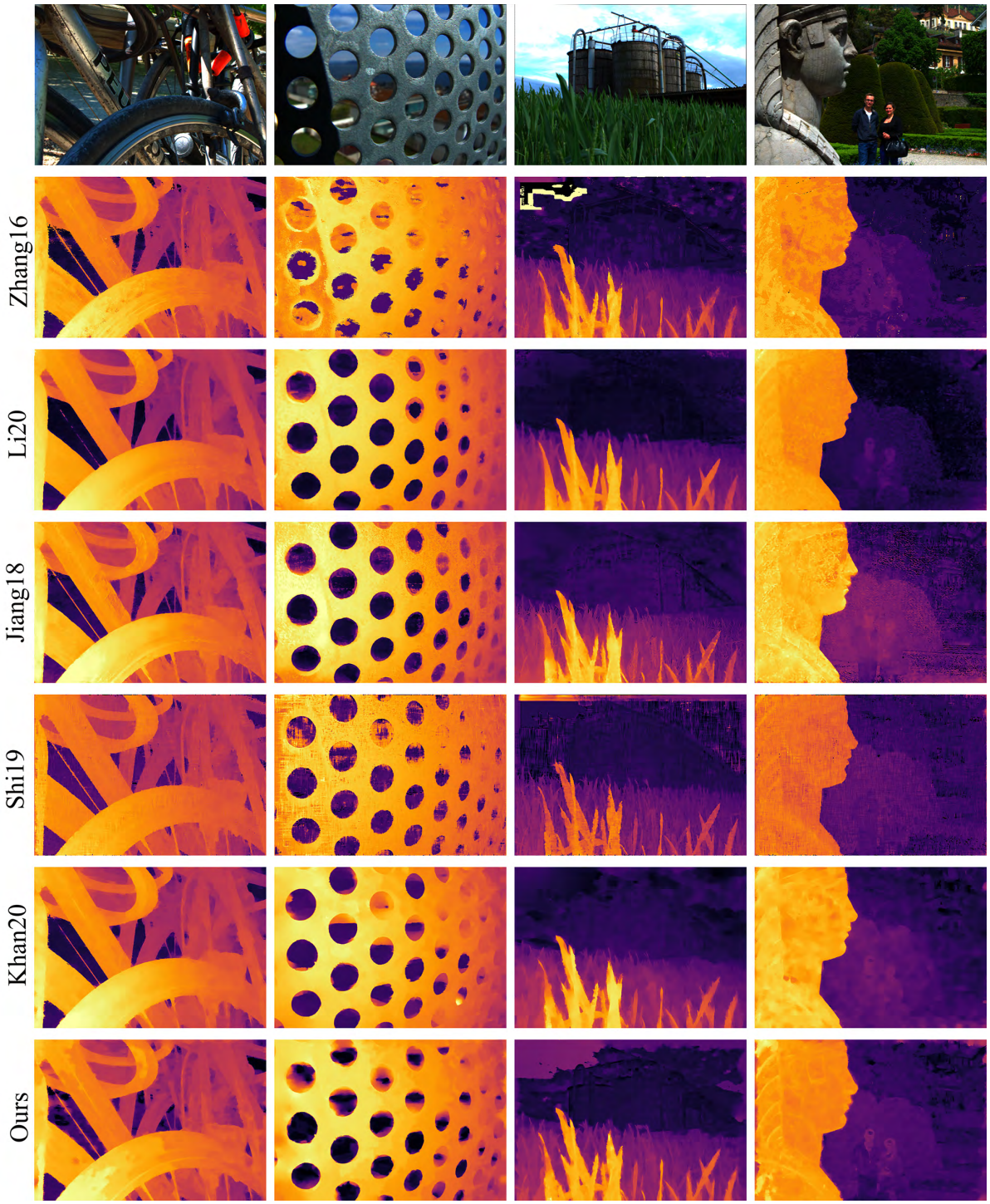


Figure 7: Qualitative comparison of our method with all baselines on the EPFL dataset. From top to bottom: Zhang et al. [6], Li et al. [4], Jiang et al. [1], Shi et al. [5], Khan et al. [2], and our approach.

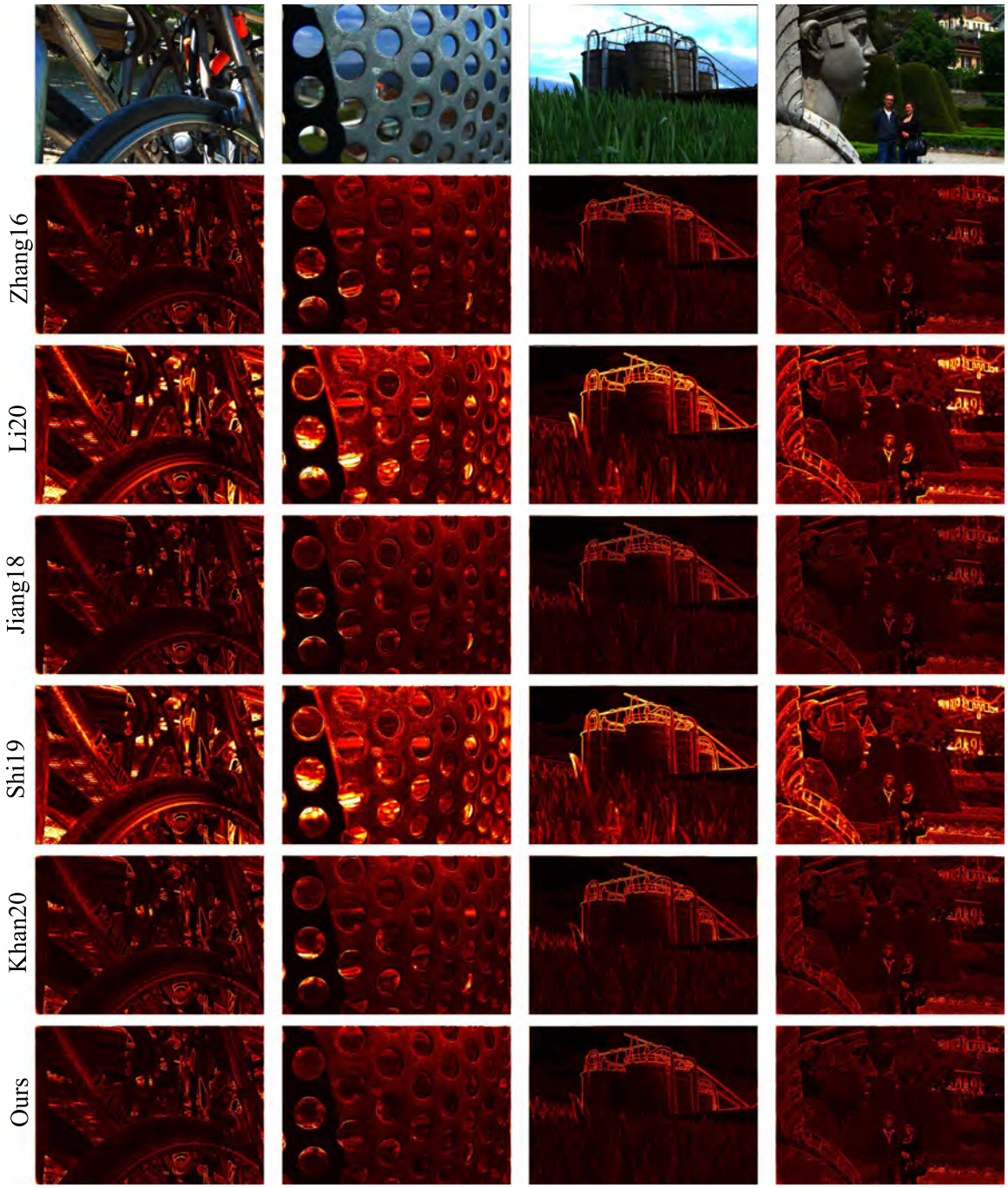
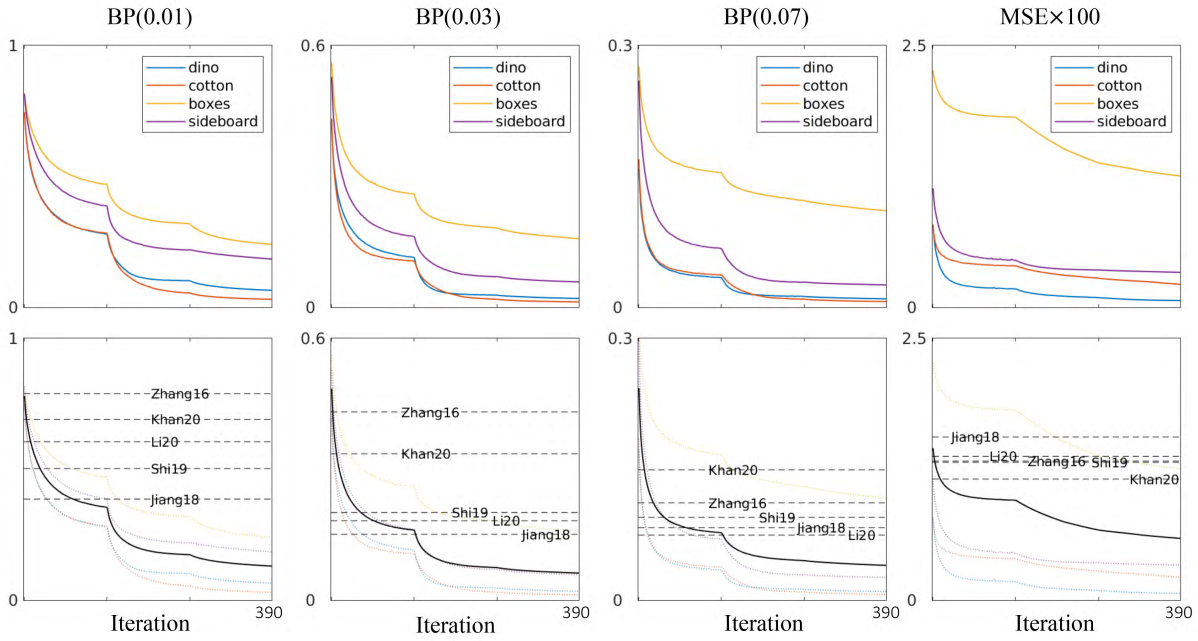
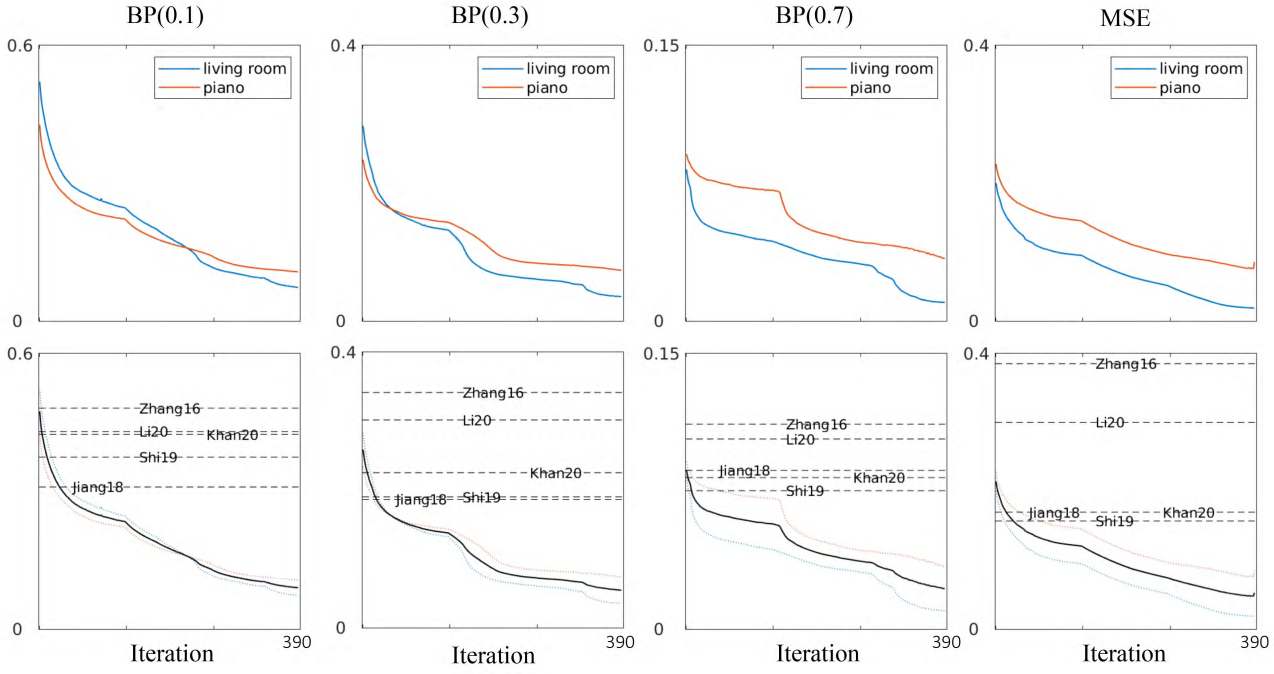


Figure 8: A visualization of the reprojection error for all methods on the EPFL dataset (lighter regions have higher error). From top to bottom: Zhang et al. [6], Li et al. [4], Jiang et al. [1], Shi et al. [5], Khan et al. [2], and our approach.



(a) Evaluation metrics plotted over all iterations of the optimization with supervised loss (*HCI Dataset*). The top row shows results on individual light fields in the dataset. The bottom row compares average performance over the dataset with the baseline methods of Zhang et al. [6], Khan et al. [2], Li et al. [4], Shi et al. [5] and Jiang et al. [1]. The bumps in the curve occur where we switch optimization parameters.



(b) Evaluation metrics for the *Piano* and *Living Room* light fields with supervised loss.

Figure 9: We validate our image-space representation and optimization using ground truth depth to supervise the optimization using the same routine as Section 3.5. This generates high-quality results on all four evaluation metrics (MSE is plotted on a logarithmic scale). This confirms the potential of our differentiable sparse point optimization and diffusion method.

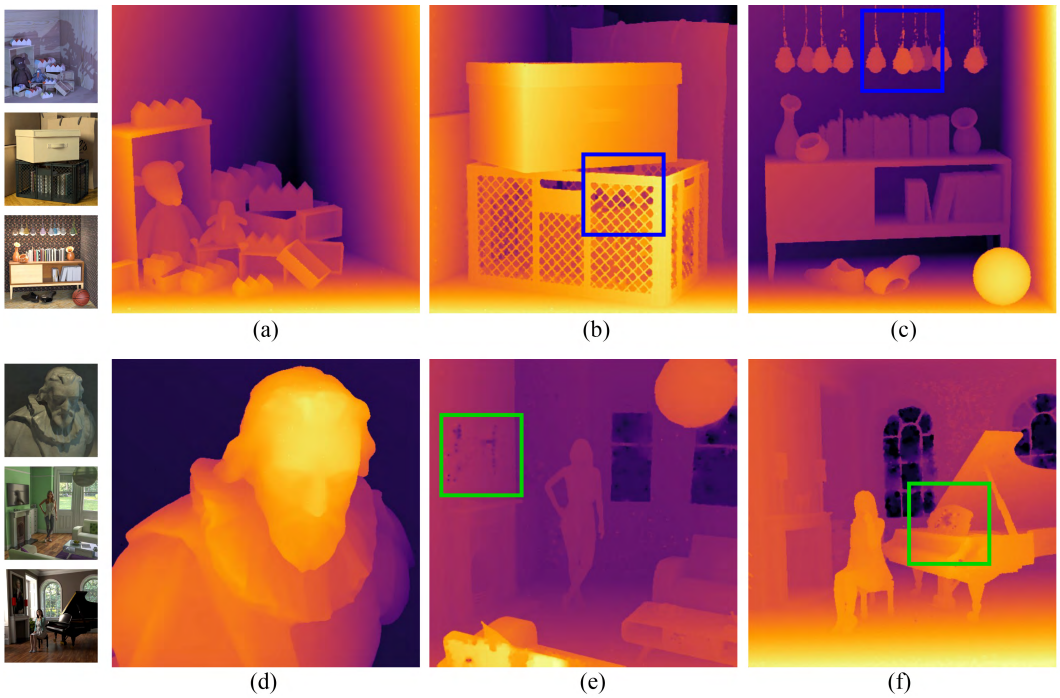


Figure 10: Disparity maps generated by our method using supervised loss. **(a)–(f)**: *Boxes*, *Dino*, *Sideboard*, *Cotton*, *Living Room* and *Piano*. The presence of outliers in the original point set (green boxes), and the lack of labels in regions with fine detail (blue boxes) prevents the diffusion from entirely eliminating errors. Such sources of error occur in all six light fields to varying degrees.

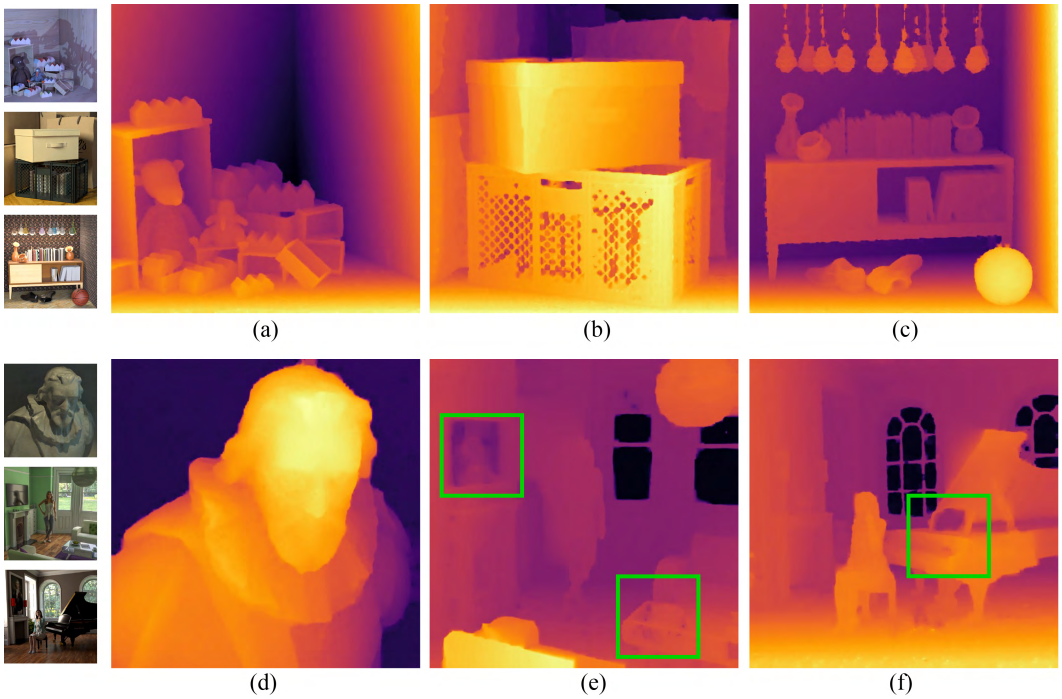


Figure 11: Our disparity maps with self-supervised loss. As expected, specular surfaces (green boxes) are especially difficult to label correctly with a self-supervised reprojection loss; Figure 3 shows that all methods suffer in these areas.

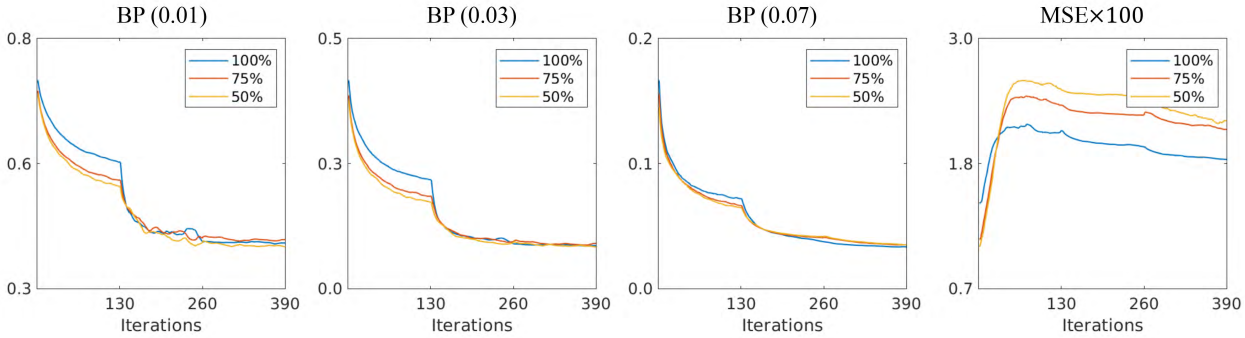


Figure 12: We evaluate optimization performance with varying point set size. We report average performance over all iterations on *cotton* and *dino* using 100%, 75% and 50% of the original point set.

References

[1] Xiaoran Jiang, Mikaël Le Pendu, and Christine Guillemot. Depth estimation with occlusion handling from a sparse set of light field views. In *25th IEEE International Conference on Image Processing (ICIP)*, pages 634–638. IEEE, 2018. 1, 4, 5, 6, 7, 8, 9

[2] Numair Khan, Min H. Kim, and James Tompkin. Fast and accurate 4D light field depth estimation. Technical Report CS-20-01, Brown University, 2020. 1, 4, 5, 6, 7, 8, 9

[3] Numair Khan, Min H. Kim, and James Tompkin. View-consistent 4d light field depth estimation. *British Machine Vision Conference*, 2020. 1

[4] Kunyuan Li, Jun Zhang, Rui Sun, Xudong Zhang, and Jun Gao. Epi-based oriented relation networks for light field depth estimation. *British Machine Vision Conference*, 2020. 1, 4, 5, 6, 7, 8, 9

[5] Jinglei Shi, Xiaoran Jiang, and Christine Guillemot. A framework for learning depth from a flexible subset of dense and sparse light field views. *IEEE Transactions on Image Processing (TIP)*, 28(12):5867–5880, 2019. 1, 4, 5, 6, 7, 8, 9

[6] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148–159, 2016. 1, 4, 5, 6, 7, 8, 9