Edge-aware Bidirectional Diffusion for Dense Depth Estimation from Light Fields

Numair Khan[1]
http://cs.brown.edu/~nkhan6/
Min H. Kim[2]
http://vclab.kaist.ac.kr/
James Tompkin[1]
https://www.jamestompkin.com

[1] Brown University
Providence, USA
[2] KAIST
Daejeon, South Korea

Abstract

We present an algorithm to estimate fast and accurate depth maps from light fields via a sparse set of depth edges and gradients. Our proposed approach is based around the idea that true depth edges are more sensitive than texture edges to local constraints, and so they can be reliably disambiguated through a bidirectional diffusion process. First, we use epipolar-plane images to estimate sub-pixel disparity at a sparse set of pixels. To find sparse points efficiently, we propose an entropy-based refinement approach to a line estimate from a limited set of oriented filter banks. Next, to estimate the diffusion direction away from sparse points, we optimize constraints at these points via our bidirectional diffusion method. This resolves the ambiguity of which surface the edge belongs to and reliably separates depth from texture edges, allowing us to diffuse the sparse set in a depth-edge and occlusion-aware manner to obtain accurate dense depth maps.

1 Introduction

Light fields record small view changes onto a scene. This allows them to store samples from both the spatial and angular distributions of light. The additional angular dimension allows imaging applications such as synthetic aperture photography and view interpolation [14, 33]. Most of these applications can be directly implemented in image space using image-based rendering (IBR) techniques [10, 21]. For applications such as light field editing and augmented reality, we require an explicit scene representation in the form of a point cloud, depth map, or derived 3D mesh, to allow occlusion-aware and view-consistent processing, editing, and rendering.

However, light field depth estimation is a difficult problem. Oftentimes, state-of-the-art methods strive for geometric accuracy without always considering occlusion edges, which are especially important for handling visibility in light field editing applications. Further, while the many views allow dense and accurate depth to be derived, the extra angular dimension carries large data costs that makes most depth estimation algorithms computationally inefficient [15, 36]. Recent methods have sought to overcome this barrier by learning data-driven priors with deep learning. While this can be effective, it requires additional training data, and may overfit to scenes or capture scenarios [22].

We present a first-principles method for estimating occlusion-accurate depth maps from light fields with no learned priors and demonstrate its application in light field editing tasks. This is achieved by estimating disparity at a sparse set of pixels identified as most important for the final result. These estimates are then propagated to all pixels using occlusion-aware diffusion. Traditionally,

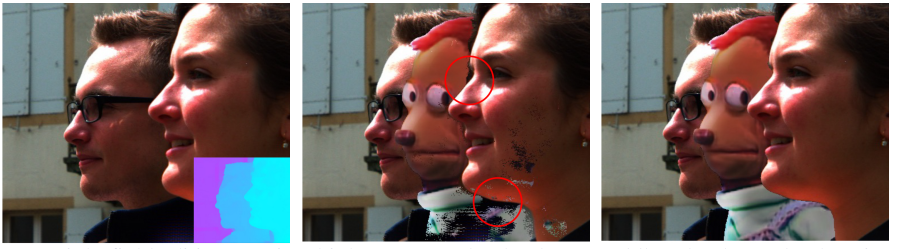


Figure 1: Light field editing requires high-accuracy depth edges. *Middle:* The dense depth estimated by Zhang et al. [36] illustrates the difficulty of extracting correct edges, with inserted content appearing at incorrect depths or with overhanging regions. *Right*: The accuracy of our depth edges allows effective occlusion handling when editing light fields. The inset shows our disparity map.

diffusion pipelines for depth completion attempt to recover a complete description of depth maps via a sparse set of depth edges and gradients [8]. Commonly, techniques follow three steps [11, 35]:

1. Obtain sparse depth labels accurately and efficiently,
2. Determine diffusion gradient at each labeled point, and
3. Perform dense depth diffusion.

Step 1 is critical yet difficult: finding sparse depth labels via edges in EPIs requires robustness to noise and occlussion-awareness. For this, we identify unwanted edges by observing gradients along and just next to the edge. Second, using large filter banks for subpixel depth precision is expensive. Thus, from an initial depth estimate from a moderately-sized filter bank, we propose a novel entropy-based depth refinement using efficient random search to obtain a subpixel estimate.

Step 2 is *also* critical yet difficult: determining the diffusion direction requires us to know the depth at pixels around each label, but for efficiency we only have a sparse set of labeled points. Holynski and Kopf [11] deal with this by assuming that sparse labels do not lie on depth edges so that neighboring pixels have a similar label. Yucer et al. [35] handle labels on depth edges, but their method is designed for light fields with a large number (≈3000+) of views. Our novel contribution here is that we determine diffusion direction from other sparse labels within context via a bidirectional 'backward-forward' diffusion process. Together, improvements in these steps allow fast and accurate occlusion estimation for light fields. https://visual.cs.brown.edu/lightfielddepth/

2 Related Work

The information implicit within an EPI (Epipolar-Plane Image) is useful for depth or disparity estimation algorithms, and the regular structure of an EPI obviates the need for extensive angular regularization. Thus, many light field operations seek to exploit it [23]. Wanner et al.'s [32] was among the earliest widely-applicable method to use EPI lines for local depth estimates. These were then optimized in a global framework with visibility constraints. Their results, while accurate, are computationally expensive to compute. Many subsequent methods have adopted a similar approach by posing depth estimation as an energy-minimization problem in EPI space. Zhang et al. [36] replace the structure tensor of Wanner et al. with a spinning parallelogram operator. Wang et al. [29, 30] propose a photo-consistency based energy term to address occlusion. Tao et al.'s [28] work considers higher dimensional representations of EPIs which allows them to use both correspondence and defocus to get depth. The latter two works rely upon the earlier work of Kolmogorov and Zabih [20] to minimize the NP-hard energy function using graph-cuts. The relation between defocus and depth is also exploited by the sub-pixel cost volume of Jeon et al. [15], who also present a method for dealing with the distortion induced by micro-lens arrays. The variational methods used

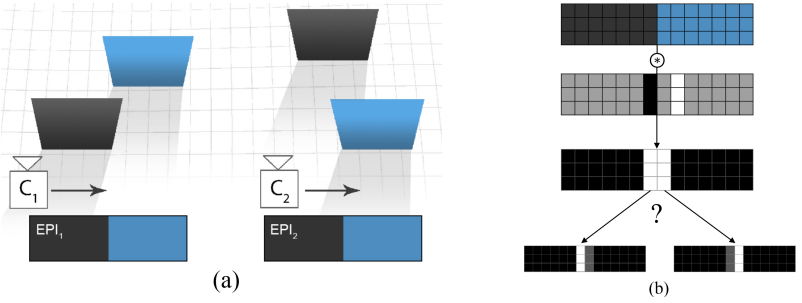


Figure 2: Sparse labels at edges are difficult to propagate because the edge is weakly localized at the boundary of two projected surfaces. As a result, labels may be assigned to the incorrect side of a depth boundary. **(a)** Two different scene configurations captured with cameras C_1 and C_2 may generate similar EPIs. The EPI edge represents the boundary of the occluding surface. For C_1 this is the surface on the left (black); for C_2 it is on the right (blue). **(b)** The direction from which occlusion happens cannot be disambiguated from edge activations alone, leading to incorrect label placement.

by the above-mentioned works lead to high computational costs and—in the case of Tao et al. [28] and Wang et al. [29, 30]—large outliers and quantization artifacts resulting from the graph-cut.

An efficient and accurate method for wide-baseline light fields was proposed by Chuchwara et al. [7]. They use an oversegmentation of each view to get initial depth proposals, which are iteratively improved using PatchMatch [3]. Closely related to our method is the work of Holynski and Kopf [11], who present an efficient method for depth densification from a sparse set of points for augmented reality applications. However, they assume that the set of sparse points and their depth values are known beforehand. Our method does not make this assumption, and seeks to identify both the points and their depth as well as performing dense diffusion. Similarly, Khan et al. [18] use a set of large Prewitt filters to reliably detect oriented lines in EPIs, then diffuses these across all light field views using occlusion-aware edges to guide a depth inpainting process [19]. However, their estimate of which edges are depth edges can be inaccurate, leading to errors in diffusion. Yucer et al. [35] present a diffusion-based method that uses image gradients to estimate a sparse label set. However, their method is designed to work for light fields with thousands of views. Chen et al. [5] estimate accurate occlusion boundaries from superpixels to regularize the depth estimation process. In general, densification methods [6, 31, 34] largely seek to recover accurate metric depth without considering occlusion boundaries.

Many methods have sought to use data-driven methods to learn priors to avoid the cost of dealing with a large number of images, and to overcome the loss of spatial information induced by the spatio-angular tradeoff in lenslet images. Huang et al.'s [13] work can handle an arbitrary number of uncalibrated views. Alperovich et al. [2] use an encoder-decoder architecture to perform an intrinsic decomposition of a light field, and also recover disparity for the central cross-hair of views. Jiang et al. [16, 17] fuse the disparity estimates at four corner views estimated using a deep learning optical-flow method, and Shi et al. [25] build on this by adding a refinement network to the fusion pipeline. Li et al. use oriented relation networks to learn depth from local EPI analysis [22]. In general, learned prior methods have been successful in estimating depth [6, 9, 26, 34]; we show that a method without any learned priors or training data requirements can be efficient and effective.

3 Our Approach

Our goal is to estimate disparity at a sparse set of points such that their labels can be efficiently diffused to generate occlusion-accurate depth maps. Based on this requirement we populate our sparse set for diffusion by selecting points around light field edges (Section 3.1). However, while

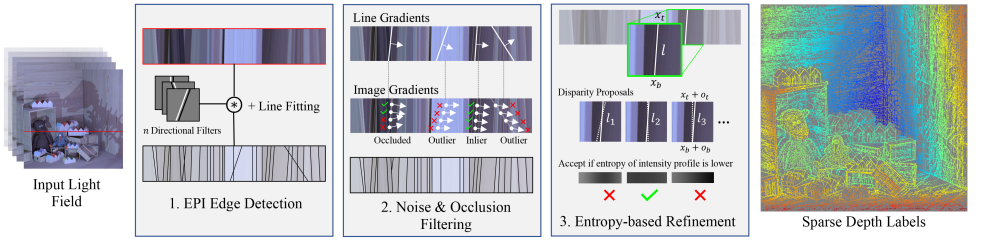


Figure 3: EPI edges provide both the location and disparity labels of a sparse point set $\mathcal{P}$. Thus, the first stage of our sparse labeling pipeline consists of EPI edge detection and line fitting. In the second stage, we compare the direction of each EPI line with underlying image gradients to remove noisy labels and points that are occluded in the central view. Finally, we improve the disparity estimates of the sparse set through an entropy-based random search.

past work on image reconstruction has shown that edges are sufficient for recovering a perceptually accurate representation of the original image [8], labels at edges are poorly localized at the intersection of surfaces (Fig. 2). Hence, we use a bi-directional diffusion process to determine the propagation direction that generates the most accurate occlusion boundaries (Section 3.2).

3.1 Sparse Depth Labels from EPI Edges

An EPI (Epipolar-Plane Image) provides an angular slice through a 4D light field, and has a linear structure resulting from epipolar geometry constraints: points in world space become lines in an EPI, with the slope of each line corresponding to the depth of the point. The regularity of an EPI makes it easy to identify salient edges and their depth at the same time.

Noise & Occlusion Filtering. For each EPI I in the central cross-hair of views, we use large Prewitt filters [18] to recover a set $\mathcal{L}$ of parametric lines representing edge points in 4D space. This process, while fast, tends to generate many false-positives. To filter these out we use a gradient-based alignment scheme: each line $l \in \mathcal{L}$ is sampled at n locations to generate the set of samples $S_l = \{(x_i, y_i)\}$. The line l is considered a false-positive if the local image gradient of I does not align with the line direction at a minimum k number of samples:

$$\sum_{\mathbf{s} \in S_l} \mathbb{1}\left(\frac{\nabla I(\mathbf{s})(\nabla l)^T}{\|\nabla I(\mathbf{s})\| \|\nabla l\|} > cos(\tau_f) \right) < k, \tag{1}$$

where $\mathbb{1}(\cdot)$ is the indicator function that counts the set of aligned samples, ∇I is the first-order image gradient approximated using a 3×3 Sobel filter, and ∇l is perpendicular to the line. The parameters τ_f and k are constants with $\tau_f = \pi/13$ and $k = (\text{EPI height})/c$, with $1 \leq c \leq \text{EPI height}$. To determine the constant value c, we consider two factors: 1) the accuracy of EPI line fitting, and 2) the expected minimum number of views a point is visible in. In the case of perfect alignment between the line and EPI gradients, $c = 1$. This means that a line with even a single misaligned sample is rejected. However, if a point is occluded in some views, the corresponding EPI line will be hidden and misalignment of samples in those views is inevitable. If we set $c = 1$ we risk discarding such lines. We determine empirically that $c = 4$ provides good results across the synthetic and real world scenes, and across the narrow and wider baseline light fields that we evaluate on.

The parametric definition of EPI lines does not carry any visibility information for a point across light field views. We determine visibility $v(l)$ of a point $l \in \mathcal{L}$ in the central view as:

$$v(l) = \mathbb{1}\left(\frac{\nabla I(\mathbf{s}_c)(\nabla l)^T}{\|\nabla I(\mathbf{s}_c)\|\|\nabla l\|} > cos(\tau_v)\right), \tag{2}$$

where $\mathbf{s}_c$ is the EPI sample corresponding to the central view and $\tau_v = \pi/10$.

Entropy-based Disparity Refinement. Notice that the number of discrete disparity values of points in $\mathcal{L}$ is bounded by the number of large Prewitt filters used for EPI line fitting. Computational efficiency considerations prevent this number from becoming too large. Moreover, numerical precision and sampling errors result in the granularity of depth estimates plateauing beyond a certain number of filters. Thus, to enable the calculation of sub-pixel disparity values we fine-tune the initial estimates through random search and filtering. Let $\mathcal{L}_c = \{l \in \mathcal{L} \mid v(l) = 1\}$. Then for each $l \in \mathcal{L}_c$ and image samples $S_l = \{(x_i, y_i)\}$ along the line we minimize the energy function defined by the entropy of normalized intensity values:

$$E(l) = \sum_{\mathbf{s} \in S_l} -P(I(\mathbf{s}))\log_2(P(I(\mathbf{s}))), \tag{3}$$

where $I(\mathbf{s})$ is the intensity value at $\mathbf{s}$ and $P(\mathbf{s})$ is estimated from a histogram.

We minimize $E(l)$ by performing a random search in the 2D parameter space defined by the x-intercepts of l on the top and bottom edge of the EPI, $l = (x_t, x_b)$: at the jth iteration of the search we generate uniform random numbers $(o_t, o_b) \sim U(-1,1)(\alpha t^j)$, to generate a proposal $l_j = (x_t + o_t, x_b + o_b)$ (Fig. 3). This is accepted with probability one if $E(l_j) < E(l_{j-1})$. We use $t = 0.88$, $\alpha = 0.15$ and run the search for a maximum of 10 iterations.

The resulting disparity estimates are then refined by joint filtering in the spatial, disparity, and LAB color space. Let $\mathcal{P}$ represent the spatial projection of $\mathcal{L}_c$ into the central view, and let p_s, p_d, and p_c be the spatial position, disparity, and color of a point $p \in \mathcal{P}$. The filtered disparity estimate $f(p_d)$ is calculated via a spatial neighborhood $\mathcal{S}$ around p:

$$f(p_d) = \frac{1}{W}\sum_{q \in \mathcal{S}} \mathcal{N}_{\sigma_s}(\|p_s - q_s\|)\mathcal{N}_{\sigma_d}(p_d - q_d)\mathcal{N}_{\sigma_c}(\|p_c - q_c\|)p_d,$$

where the normalization factor W is given by

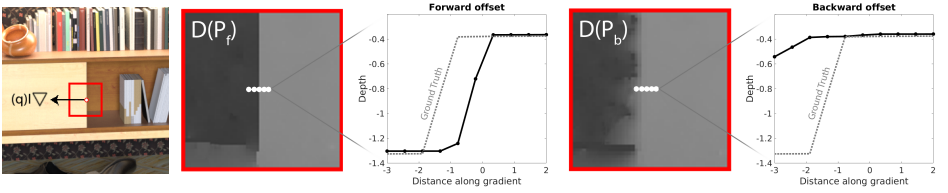
$$W = \sum_{q \in \mathcal{S}} \mathcal{N}_{\sigma_s}(\|p_s - q_s\|)\mathcal{N}_{\sigma_d}(p_d - q_d)\mathcal{N}_{\sigma_c}(\|p_c - q_c\|). \tag{4}$$

We found that the combination $\sigma_s = 10$, $\sigma_d = 0.1$ and $\sigma_c = 0.5$ works for all scenes.

3.2 Occlusion Edges via Diffusion Gradients

We want to diffuse the sparse set of disparity labels in $\mathcal{P}$ to a dense grid of pixels $\hat{D}$ such that $\nabla\hat{D}$ accurately represents all occlusion edges in the scene. However, this is a chicken-and-egg problem as we need the occlusion edges to determine the diffusion direction at each $p \in \mathcal{P}$. As Figure 2 shows, the disparity for a point lying on an EPI edge alone is not sufficient to determine the surface direction in which to perform diffusion. Directly propagating the sparse disparity estimates to generate a dense depth map results in significant errors around edges (supplemental Fig. 10).

As all potential occlusion edges are also depth edges, one way to determine diffusion direction is by distinguishing depth and texture edges. Yucer et al. [35] do this by comparing the variation



(a) (b) (c)

Figure 4: **(a)** Given an edge point p with image gradient $\nabla I(p)$ and depth label p_d we would like to determine which side of the edge to propagate p_d. We generate images $\hat{D}[\mathcal{P}_f]$ (**b**), and $\hat{D}[\mathcal{P}_b]$ (**c**) by solving a Poisson optimization problem with diffusion direction $p+\nabla I(p)$ and $p-\nabla I(p)$ respectively. The correct diffusion direction (**b**) generates an intensity profile resembling a step function. In the example shown, p_d corresponds to the surface on the right of the edge as $p+\nabla I(p)$ generates a profile more closely resembling a step function.

in texture on both sides of an edge as the view changes: the background seen around a depth edge will change more rapidly than the foreground, leading to a larger variation in texture along one side of the edge. The correct diffusion direction is to the side with lower variation. This method works for light fields with thousands of views (3000+ images), but proves ineffective on datasets that are captured using a lenslet array or camera rig (Fig. 7). This is because the assumption fails to hold in cases where 1. the background lacks texture, and 2. the light field has a small baseline with relatively few views, which is common for handheld cameras. Here, occlusion is minimal and image intensity variation is caused more by sensor noise than by background texture variation.

Our proposed solution to the depth edge identification problem works for light fields with few views (e.g., 7×7 from a Lytro). We use $S[\mathcal{A}]$ to represent the image created by splatting sparse points in a set $\mathcal{A}$ onto a $w\times h$ raster grid, and D to be a dense $w\times h$ disparity map. Diffusion is formulated as a constrained quadratic optimization problem:

$$\hat{D}[\mathcal{A}] = \underset{D}{\operatorname{argmin}} \sum_{p\in\mathcal{A}} E_d(p) + \sum_{(p,q)\in\mathcal{S}} E_s(p,q), \tag{5}$$

where $\hat{D}[\mathcal{A}]$ is the optimal disparity map given the sparsely labeled image $S[\mathcal{A}]$ and $\mathcal{S}$ is the set of all four-connected neighbors in D. The data term $E_d(p)$ and smoothness term $E_s(p,q)$ are defined as:

$$E_d(p) = \lambda_d(p)\big\|S[\mathcal{A}](p) - D(p)\big\|, \quad \text{and} \quad E_s(p,q) = \lambda_s(p)\big\|D(p) - D(q)\big\|, \tag{6}$$

with $\lambda_d(\cdot)$ and $\lambda_s(\cdot)$ being the spatially-varying data and smoothness weights.

Equation (5) represents a standard Poisson problem, and we solve it using an implementation of the LAHBPCG solver [27] by posing the constraints in the gradient domain as proposed by Bhat et al. [4]. We begin by defining two sets formed from opposite offset directions $\nabla I(p)$ and $-\nabla I(p)$:

$$\mathcal{P}_f = \{p+\nabla I(p) \;\forall\; p\in\mathcal{P}\}, \quad \text{and} \quad \mathcal{P}_b = \{p-\nabla I(p) \;\forall\; p\in\mathcal{P}\}, \tag{7}$$

where $\nabla I(p)$ is the gradient of the central light field view at point p. Then, we solve Equation (5) for both offset directions $\hat{D}[\mathcal{P}_f]$ and $\hat{D}[\mathcal{P}_b]$ using data and smoothness weights:

$$\lambda_d(p) = \begin{cases} 10^6 & \text{if } p\in\mathcal{A}, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \lambda_s(p) = \frac{1}{\|\nabla I(p)\| + \varepsilon}. \tag{8}$$

Given both solutions, we compare the normalized depth profile around each point $p\in P$ along $\nabla I(p)$ in $\hat{D}[\mathcal{P}_f]$ and $\hat{D}[\mathcal{P}_b]$. Figure 4 shows that the profile for the correct offset direction ($\nabla I(p)$

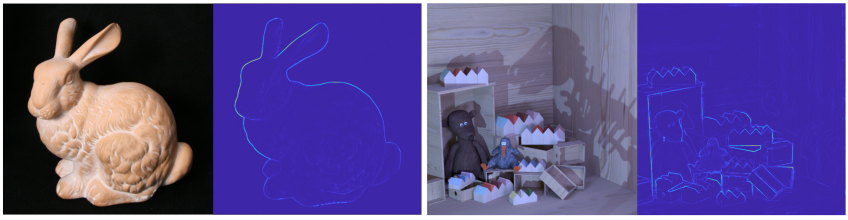


Figure 5: *Estimated depth edge confidence* λ_s. The bi-directional diffusion process allows us to identify depth edges by considering the mean gradient at each pixel across the backward-forward pass. Texture edge gradient remains low in both passes. For depth edges, the gradient is higher in one pass. For depth edges that are not meant to be sharp, the change in depth around that region from the bi-directional solve is small, and picking either offset leads to low error.

or $-\nabla I(p)$) more closely resembles a step function around p due to a strong depth gradient. This is because neighboring points in the correct offset direction will have a disparity value similar to p. The high data term together with the global smoothness constraint results in a small gradient around p when the incorrect offset pushes it to the wrong side of the edge. We estimate the profile around p in $\hat{D}[\mathcal{P}_f]$ and $\hat{D}[\mathcal{P}_b]$ by convolving the normalized value of a set N_p of pixels around p with the step filter $\mathbf{F}=[-1\ -1\ 1\ 1]$. We define:

$$\lambda_e(p) = \max_{\{\hat{D}[\mathcal{P}_f],\hat{D}[\mathcal{P}_b]\}} \|N_p \circledast \mathbf{F}\|. \tag{9}$$

The final map $\hat{D}[\mathcal{Q}]$ with the desired depth edges is generated using Equation (5) where $\mathcal{Q} = \{p \pm \nabla I(p)\ \forall\ p \in \mathcal{P}\}$ is a sparse set of points offset in the diffusion direction determined above. The final data and smoothness weights are:

$$\lambda_d(p) = \omega \exp(a\lambda_e(p)), \quad \text{and} \quad \lambda_s(p) = \frac{1}{\|\nabla I(p)\|\|\nabla \hat{D}[\mathcal{P}_f] + \nabla \hat{D}[\mathcal{P}_b]\|}, \tag{10}$$

where $\lambda_s(p)$ defines the depth edge confidence at every pixel (Fig. 5). The parameters in Equation (10) are set as $\omega = 1.5\times10^2$ and $a=3$. These values work for all scenes.

4 Experiments

Occlusion Edge Accuracy. Qualitatively, our method produces sharper and more accurate occlusion edges than state-of-the-art light field depth estimation methods. We compare our results to three non-learning-based methods: the defocus and correspondence cues methods by Jeon et al. [15] and Wang et al. [30], and the spinning parallelogram operator of Zhang et al. [36]. We also compare with the learning-based methods of Jiang et al. [16], Shi et al. [25], and Li et al. [22]. We do not compare to Holynski and Kopf [11]: this uses COLMAP, which fails on typical skew-projected light field data. In Figure 6, we show results on light fields from the EPFL MMSPG Light-Field Dataset [24] (7×7) and the Stanford Light Field Archive [1] (17×17). The latter dataset is captured with a camera rig and has a wider baseline than the EPFL light fields, which come from a Lytro Illum camera. Our method was implemented in MATLAB, as were the three traditional algorithms, and parts of Jiang et al. The network code of Jiang et al. Li et al., and Shi et al. was implemented in Tensorflow. All CPU code was run on an AMD Ryzen Threadripper 2950X 16-Core Processor, and GPU code on an NVIDIA GeForce RTX 2080Ti.

In Figure 7, we visualize occlusion boundaries as depth gradients. While the learning-based methods of Shi et al. and Li et al. generating spurious boundaries in textureless regions, the approach of Yucer et al. [35] fails entirely in the absence of thousands of views. We also evaluate our edges

quantitatively on four scenes from the synthetic HCI Light Field Dataset [12] via ground truth disparity maps for the central view (Fig. 8 and Tab. 1). Although our Q25 error is higher, our method has high boundary-recall precision, and a lower average mean-squared error than all baselines.

Our method works on 2D slices of a 4D light field. While jointly considering the 4D structure may improve accuracy, edge detection and diffusion become computationally expensive. In principle, the accuracy of our current edge detection can be improved by entropy-based refinement of labels (Sec. 3) in both vertical and horizontal EPIs. In practice, we found no advantage of doing so.

Diffusion Gradients as Self-supervised Loss. One way to think about bidirectional diffusion gradients is as a self-supervised loss function for depth edge localization. With this view, we compare its performance to *multi-view reprojection error*—a commonly used self-supervised loss in disparity optimization. We use the dense disparity maps $\hat{D}[\mathcal{P}_f]$ and $\hat{D}[\mathcal{P}_b]$ to warp all light field views onto the central view through an occlusion-aware inverse projection. A reprojection error map is calculated as the mean per-pixel L1 intensity error between the warped views and the central view. The offset direction at each point $p \in \mathcal{P}$ is then determined based on the disparity map that minimizes the reprojection error at the pixel location of p. Table 2 evaluates the result of calculating $\mathcal{Q} = \{p \pm \nabla I(p) \;\forall\; p \in \mathcal{P}\}$ based on the reprojection error maps instead of our bidirectional diffusion gradients. Our method has consistently lower MSE, indicating better edge performance. This intuition is qualitatively confirmed by supplemental Figure 17.

5 Discussion

Light Field Editing. As our method generates accurate depth edges that allow visibility to be handled correctly, our depth allows simple object insertion with few artifacts (Figs. 1 and 9).

Errors. Our method has consistently lower mean squared error (MSE), but suffers a higher number of erroneous pixels (Q25). As Q25 measures the first quantile of absolute error, this indicates that baseline methods must have more outliers: the errors that they do have must be considerably large. This intuition is confirmed by visualizing the absolute error (supplemental Fig. 18) which shows regions of large error around occlusion boundaries for the baseline methods.

Our method outperforms deep learning methods when they suffer from under-specification and over-fitting. Our three learning-based baselines are trained by supervision on the synthetic HCI dataset. When tested on real-world light fields, they produce artifacts along depth edges (Fig. 6). This failure to generalize is especially evident with Li et al. [22] which suffers severe artifacts on real world data (compare supplemental Figs. 16 and 19). Our method is not susceptible to these problems, producing comparable output on both synthetic and real-world light fields. Moreover, our method is robust to noise in the label set and low-gradient edges (supplemental Figs. 12 and 13).

In addition, our method explicitly optimizes depth-edge localization whereas the learning-based methods are trained to minimize mean depth error over all pixels—edge and non-edge. Incorporating hard edge information in CNNs is generally not straightforward due to its sparsity and non-differentiability. Shi et al. [25] include a Canny edge-based loss term in their training routine, and consequently their performance on edges is relatively higher. Nonetheless, they are unable to effectively differentiate depth and texture edges (Figure 7).

Hyperparameters. Supplemental Fig. 14 shows our method's stability to hyperparameter variation.

Limitations. As we estimate depth explicitly only around potential occlusion boundaries our method has lower accuracy in non-edge regions, reflected by the Q25 error (Table 1). A fundamental trade-off exists between dense processing of all pixels and the disambiguation of depth and color edges, as well as run-time. We focus on the latter attributes as they are more useful for scene editing.

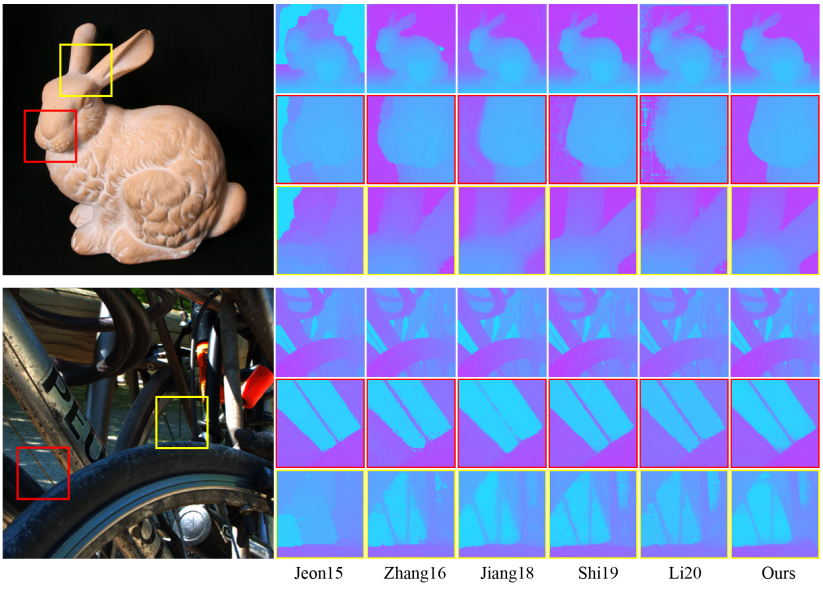


Figure 6: Occlusion edges in disparity maps. *Top:* Stanford dataset light field captured with a camera rig. *Bottom:* EPFL light field from a Lytro Illum. *Left to right:* Jeon et al. [15], Zhang et al [36], Jiang et al. [16], Shi et al. [25], and ours.

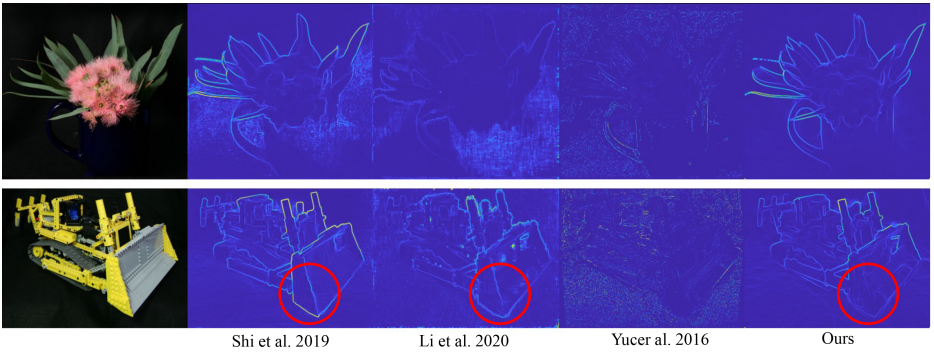


Figure 7: Visualizing occlusion edges as gradients of disparity maps. *Left to right:* Shi et al. [25], Li et al [22], Yucer et al. [35], and ours. *Bottom row, red circle:* the learning-based methods hallucinate a strong depth edge on the plow even though it is in contact with the black ground cloth at the same depth (supplemental Sec. A.1). Yucer et al.'s method fails in the absence of many views.

Table 1: Quantitative comparison of our method and the baselines on the synthetic HCI light fields. The top three results are highlighted in gold, silver and bronze. MSE is the mean squared error; Q25 is the 25th percentile of the absolute error.

Light Field	MSE × 100							Q25							Run time (s)						
	[15]	[36]	[16]	[25]	[35]	[22]	Ours	[15]	[36]	[16]	[25]	[35]	[22]	Ours	[15]	[36]	[16]	[25]	[35]	[22]	Ours
Sideboard	3.21	1.02	1.96	1.12	1.89	13.3	1.03	0.61	1.15	0.37	0.48	0.66	2.46	1.22	754	537	507	72.3	77.1	635	35.5
Dino	1.73	0.36	0.47	0.43	3.28	4.19	0.45	1.07	1.40	0.25	0.31	0.50	2.02	0.85	805	531	500	59.3	76.8	609	37.7
Cotton	12.5	1.81	0.97	0.88	1.95	9.56	0.70	0.50	1.01	0.21	0.36	0.59	2.30	0.74	748	530	500	79.8	76.9	612	34.0
Boxes	16.0	7.90	11.6	8.48	4.67	12.5	7.52	0.75	1.64	0.42	0.69	0.78	2.21	1.41	736	541	491	56.2	78.0	667	34.3
Average	8.37	2.77	3.75	2.72	2.94	9.91	2.43	0.73	1.3	0.31	0.46	0.63	2.25	1.05	761	535	500	66.9	77.2	631	35.4

Light Field	MSE × 100		Q25	
	Reproj	Ours	Reproj	Ours
Sideboard	1.39	**1.03**	**1.20**	1.22
Dino	0.64	**0.45**	**0.81**	0.85
Cotton	1.04	**0.70**	**0.68**	0.74
Boxes	9.32	**7.52**	1.65	**1.41**
Average	3.10	**2.43**	1.08	**1.05**

Table 2: Evaluating disparity maps with depth edges identified via reprojection error and via our approach of diffusion gradients on the synthetic HCI dataset. MSE is the mean squared error; Q25 is the 25th percentile of absolute error.

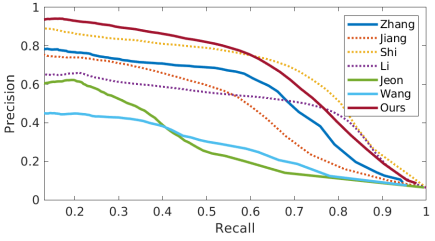


Figure 8: Average precision-recall curves of depth boundaries for all baseline algorithms (HCI dataset). Learning-based methods are shown as dotted lines. Our approach consistently outperforms traditional algorithms [15, 30, 36] and the learning-based method of Jiang et al. [16], while outperforming Shi et al. [25] and Li et al. [22] at medium-to-low recall rates.

Light Field	Mean F1†	Peak F1†	AUC†
Zhang [36]	3.41	6.07	5.07
Jiang [16] *	2.64	5.23	4.52
Shi [25] *	3.29	6.85	6.30
Li [22] *	3.72	5.78	4.60
Jeon [15]	2.14	3.67	3.06
Wang [30]	2.02	3.56	2.70
Ours	3.42	6.52	6.30

† $\times 10^{-1}$ * Learning-based

Table 3: Mean and Peak F1 across all thresholds, and the area under the precision-recall Curve (AUC) for the HCI dataset. Our method has the second-highest F1 score and, along with the learning-based method of Shi et al., the highest AUC.

6 Conclusion

Estimating occlusion-accurate depth maps from light fields is useful for scene editing and AR applications. Our approach is based around a bidirectional diffusion process that can disambiguate depth from color edges and estimate a correct depth edge offset to provide accurate gradient information for diffusion. We also contribute a faster method to find sub-pixel disparity labels at a sparse set of points via an entropy-based depth refinement process. The effectiveness of this strategy is shown with results on synthetic and real world light fields, producing competitive or better mean squared error accuracy while being significantly faster than other non-learning-based methods.

Acknowledgments. Numair Khan acknowledges an Andy van Dam PhD Fellowship, Min H. Kim acknowledges the partial support of Korea NRF grants (2019R1A2C3007229) and Samsung Electronics, and James Tompkin acknowledges support from Cognex and NSF CNS-203889.

Figure 9: Adding a BMVC'21 tarot card to the scene. *Left:* input scene. *Center:* Our editing results. *Right, clockwise from top-left*: Detail of the unmodified light field image, Zhang et al. [36]'s editing result, Shi et al. [25]'s editing result, and our result with fewer artifacts.

References

[1] Andrew Adams, Vaibhav Vaish, Bennett Wilburn, Neel Joshi, and Marc Levoy. The New Stanford Light Field Archive, 2008. URL http://lightfield.stanford.edu/.

[2] A. Alperovich, O. Johannsen, and B. Goldluecke. Intrinsic light field decomposition and disparity estimation with a deep encoder-decoder network. In *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, 2018.

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), August 2009.

[4] Pravin Bhat, Larry Zitnick, Michael Cohen, and Brian Curless. Gradientshop: A gradient-domain optimization framework for image and video filtering. In *ACM Transactions on Graphics (TOG)*, 2009.

[5] Jie Chen, Junhui Hou, Yun Ni, and Lap-Pui Chau. Accurate light field depth estimation with superpixel regularization over partially occluded regions. *IEEE Transactions on Image Processing*, 27(10):4889–4900, 2018.

[6] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–119, 2018.

[7] A. Chuchvara, A. Barsi, and A. Gotchev. Fast and accurate depth estimation from sparse light fields. *IEEE Transactions on Image Processing*, 29:2492–2506, 2020.

[8] James H Elder. Are edges incomplete? *International Journal of Computer Vision*, 34(2-3): 97–122, 1999.

[9] Abdelrahman Eldesokey, Michael Felsberg, and Fahad Shahbaz Khan. Confidence propagation through cnns for guided sparse depth regression. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2423–2436, 2019.

[10] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The Lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54. ACM, 1996.

[11] Aleksander Holynski and Johannes Kopf. Fast depth densification for occlusion-aware augmented reality. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 37(6), 2018.

[12] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*, pages 19–34. Springer, 2016.

[13] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

[14] Aaron Isaksen, Leonard McMillan, and Steven J Gortler. Dynamically reparameterized light fields. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 297–306, 2000.

[15] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon. Accurate depth map estimation from a lenslet light field camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1547–1555, 2015.

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

[17] Xiaoran Jiang, Jinglei Shi, and Christine Guillemot. A learning based depth estimation framework for 4d densely and sparsely sampled light fields. In *Proceedings of the 44th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[18] Numair Khan, Qian Zhang, Lucas Kasser, Henry Stone, Min Hyuk Kim, and James Tompkin. View-consistent 4d light field superpixel segmentation. In *International Conference on Computer Vision (ICCV) 2019*. IEEE, 2019.

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

[20] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *European conference on computer vision*, pages 82–96. Springer, 2002.

[21] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996.

[22] 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.

[23] In Kyu Park, Kyoung Mu Lee, et al. Robust light field depth estimation using occlusion-noise aware data costs. *IEEE transactions on pattern analysis and machine intelligence*, 40(10): 2484–2497, 2017.

[24] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[25] 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*, 28(12):5867–5880, 2019.

[26] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4748–4757, 2018.

[27] Richard Szeliski. Locally adapted hierarchical basis preconditioning. In *ACM SIGGRAPH 2006 Papers*, pages 1135–1143. ACM, 2006.

[28] Michael W Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 673–680, 2013.

[29] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3487–3495, 2015.

[30] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. Depth estimation with occlusion modeling using light-field cameras. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2170–2181, 2016.

[31] Tsun-Hsuan Wang, Fu-En Wang, Juan-Ting Lin, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. Plug-and-play: Improve depth estimation via sparse data propagation. *arXiv preprint arXiv:1812.08350*, 2018.

[32] Sven Wanner and Bastian Goldluecke. Globally consistent depth labeling of 4d light fields. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48. IEEE, 2012.

[33] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. In *ACM SIGGRAPH 2005 Papers*, pages 765–776. ACM, 2005.

[34] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. Depth completion from sparse lidar data with depth-normal constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2811–2820, 2019.

[35] Kaan Yucer, Changil Kim, Alexander Sorkine-Hornung, and Olga Sorkine-Hornung. Depth from gradients in dense light fields for object reconstruction. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 249–257. IEEE, 2016.

[36] 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.

A Supplemental Material

We include additional discussion covering the benefits over naive diffusion, consistency over views within the 4D light field, tolerance to depth label errors and edge blur, robustness to hyperparameter variation, details of dataset preprocessing, and an example of textures within dark backgrounds in the Stanford dataset (Section A.1). Next, we present error maps comparing reprojection loss versus our bidirectional diffusion approach (Section A.3), and error maps versus ground truth for the HCI dataset (Section A.4). Finally, we show additional qualitative results on the Stanford dataset (Section A.2) and an additional editing example (Figure 20).

A.1 Additional Discussion

Naive Diffusion. In Figure 10, we demonstrate visually that naively diffusing disparity labels can be problematic because edge localization is ambiguous.

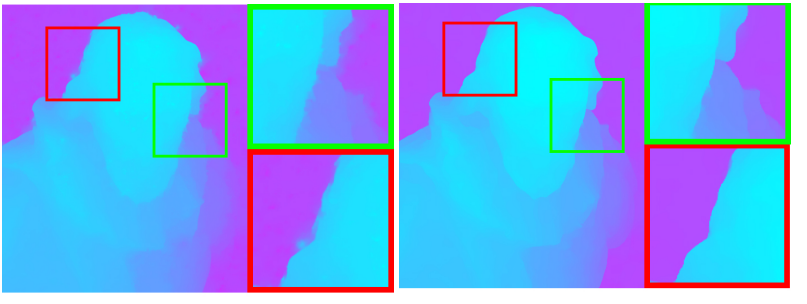


Figure 10: *Left*: Naïvely diffusing disparity labels causes artifacts around edges due to ambiguity in the localization of labels around edges. *Right*: Estimating the diffusion gradient removes this ambiguity and yields sharp depth edges.

Multi-view Depth and Error. As ground truth disparity is only provided for the central view of the HCI data set, and as the Stanford data set has no ground truth depth, we did not include quantitative error evaluation across '4D' views. Qualitatively, our method tends to produce results that are consistent across views (Fig. 11).

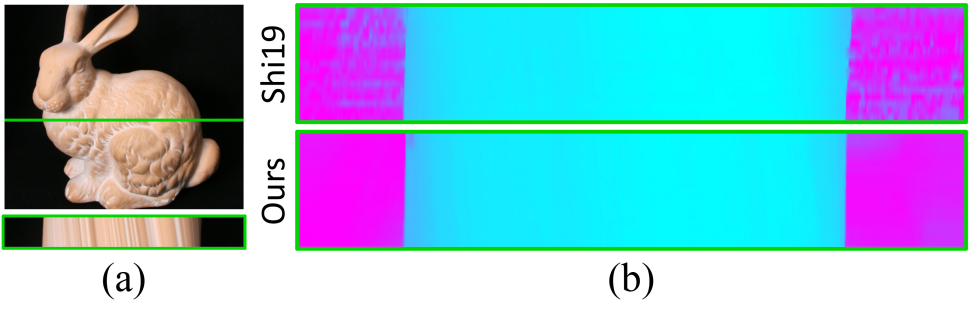


Figure 11: **(a)** We visualize depth consistency for the highlighted epipolar line. **(b)** Our results are more consistent than Shi et al. [25] across views (EPIs are scaled vertically for clarity).

Disparity Noise and Blur Tolerance. To show our robustness, we evaluate our method on noisy disparity labels (Fig. 12) and low-gradient edges (Fig. 13). Our method provides greater robustness to disparity errors than naive diffusion, and provides greater robustness via MSE to low-gradient (or blurry) edges than two learning-based baselines.

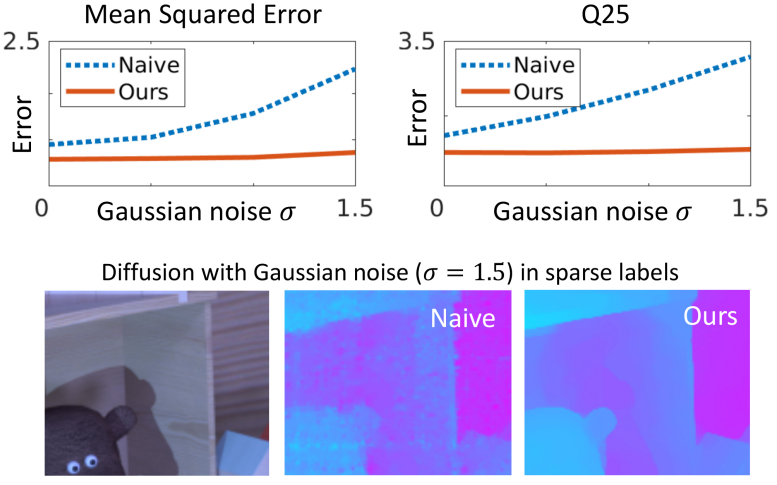


Figure 12: Robustness of our method to noise in disparity labels (*Dino* light field; we compare with naive diffusion.).

Hyperparameter variation. Figure 14 demonstrates the variation in error as hyperparameter values change. Across all parameters, our approach is stable around our declared values.

Lenslet Distortion and EPFL Lytro Dataset. The Lytro light fields in the EPFL dataset are provided decoded as MATLAB files. In general, while our method can handle small amounts of distortion, the EPI-based edge detection stage expectedly fails when EPI features are no longer linear. This is true for the edge views of Lytro light fields. As such, we only use the central 7×7 views of the EPFL scenes for all experiments.

Black Backgrounds and Stanford Dataset. Our EPI edge detector aggregates information from all three channels in CIE LAB color space, which allows it to detect even faint edges. Thus, it captures the subtle background texture on the black cloth in the Stanford dataset examples of single objects; typically, this detail is not visible to the naked eye. This feature of our work also explains why we do not incorrectly detect false edges in the Lego Technic Plow scene, as shown in Figure 7 of the main paper.

A.2 Expanded Results

We present qualitative results on the HCI dataset in Figure 16, and expanded results on the real-world light fields of the Stanford dataset in Figure 19. Our method produce stronger depth edges compared to the baselines, and our smoothness regularization (Equation 10, main paper) leads to fewer artifacts in textureless regions.

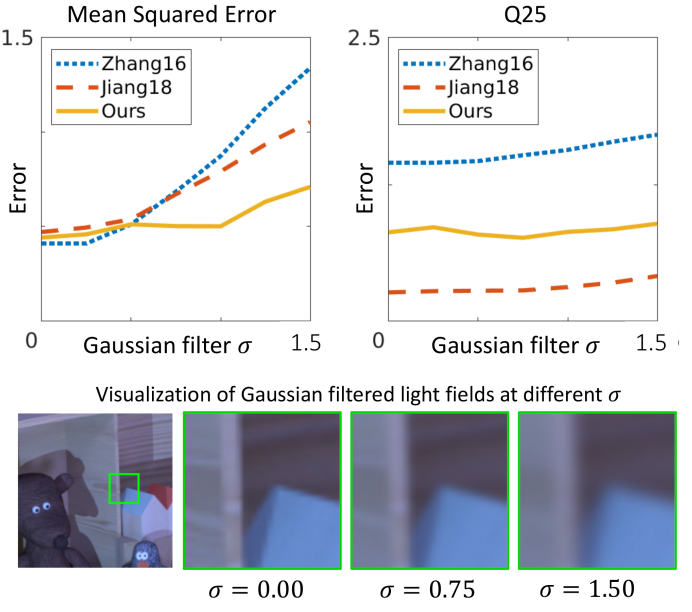


Figure 13: Robustness of our method to low-gradient edges (*Dino* light field; we compare to the methods of Zhang et al. [36] and Jiang et al. [16] which have the best MSE and Q25 performance on this light field, respectively).

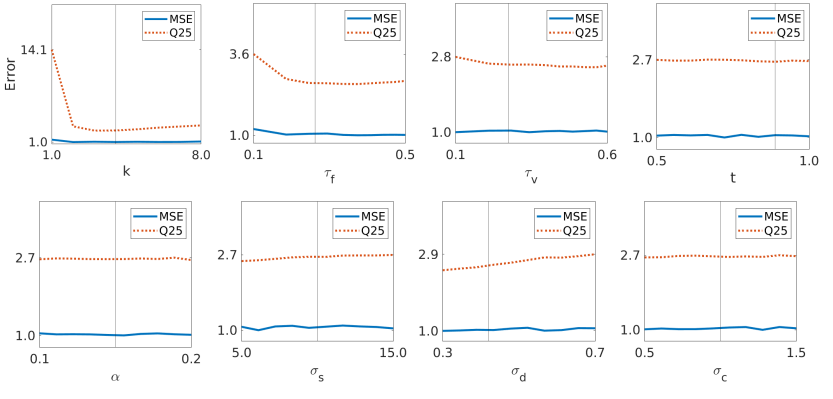


Figure 14: Effect of hyperparameter values on the MSE and Q25, averaged across the HCI dataset: k and τ_f (Eqn. 1), τ_v (Eqn. 2), t and α (entropy-based refinement), and σ_s, σ_d and σ_c (Eqn. 4). The vertical lines indicate our chosen values. The stochasticity of our algorithm means the chosen values may not be optimal in all cases. However, the method is stable to variation around these values.

A.3 Diffusion Gradients as Self-supervised Loss

As in main paper Section 4, we compare our method to a reprojection error loss. In Figure 17, to complement the quantitative MSE numbers in the main paper, we demonstrate the qualitative improvement from our bidirectional diffusion gradient approach in comparison.

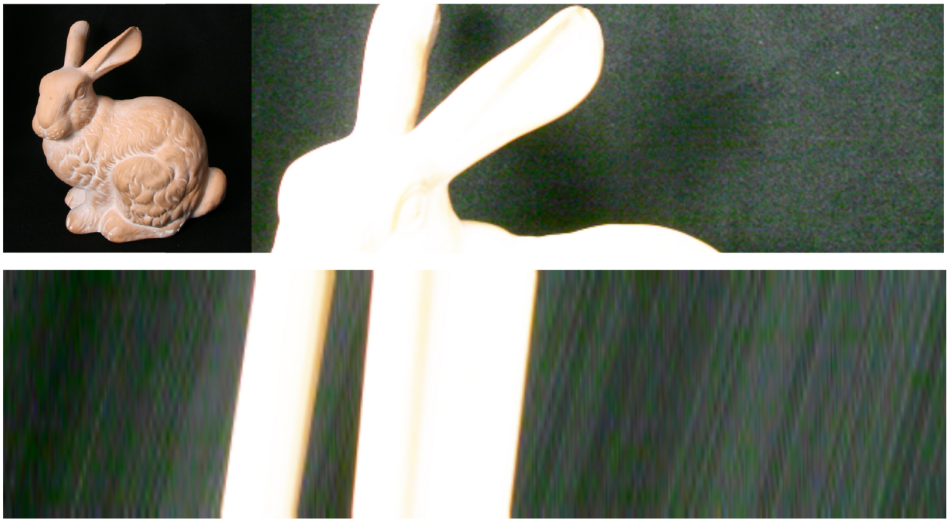


Figure 15: *Top:* On the Stanford Bunny scene, enhanced image contrast shows the texture of the cloth in the seemingly black background. *Bottom:* In EPI space (scaled vertically for clarity) the texture appears as sloped lines, providing background disparity to methods that can exploit this subtle information.

A.4 Error Maps

We visualize the absolute disparity error of all baselines and our method in Figure 18. The baseline methods produce larger errors around depth edges compared to our approach. This can be seen in the fewer regions of red for our method compared to the baselines. The corresponding dense disparity maps are shown in Figure 16. Qualitatively, our results are comparable to the learning-based baselines [16, 22, 25] with fewer extreme errors around edges.

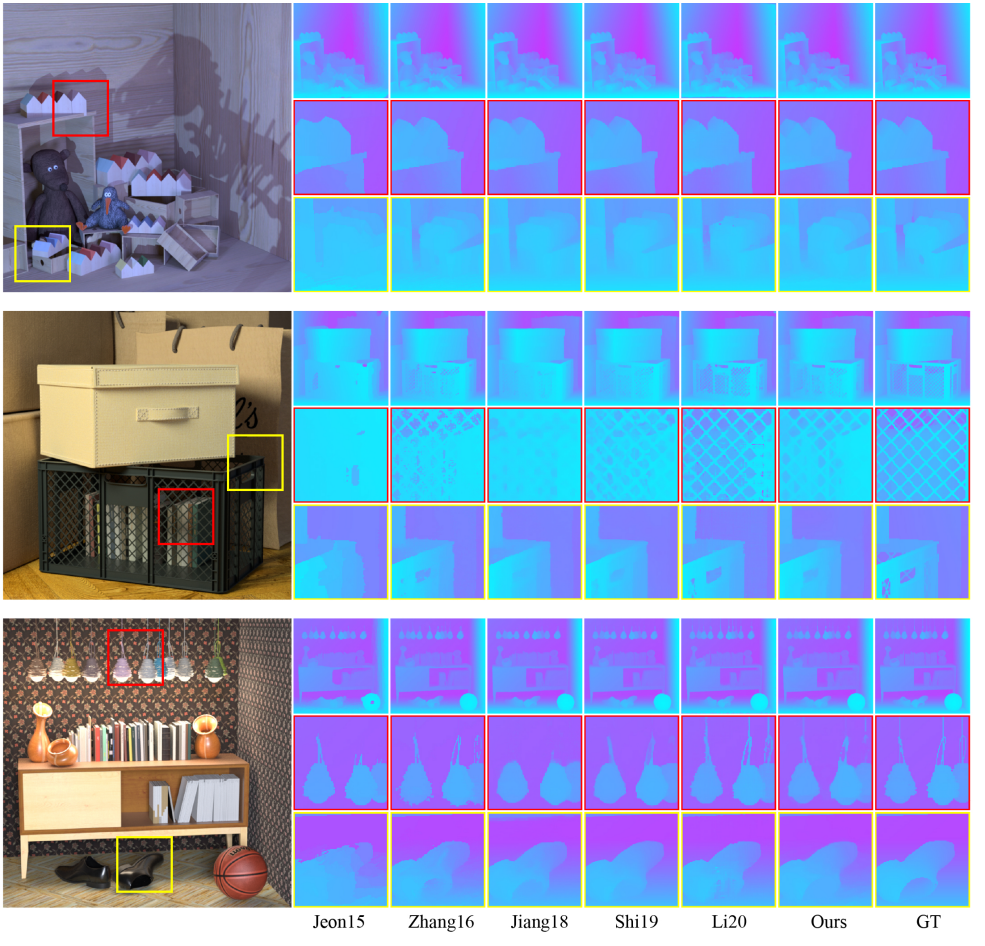


Figure 16: Results on the synthetic light fields of the HCI dataset. *Left to right:* Jeon et al. [15], Zhang et al [36], Jiang et al. [16], Shi et al. [25], Li et al. [22], our method, and finally, the ground truth. Qualitatively, our results are comparable to the learning-based baselines [16, 22, 25] with fewer extreme errors around edges.

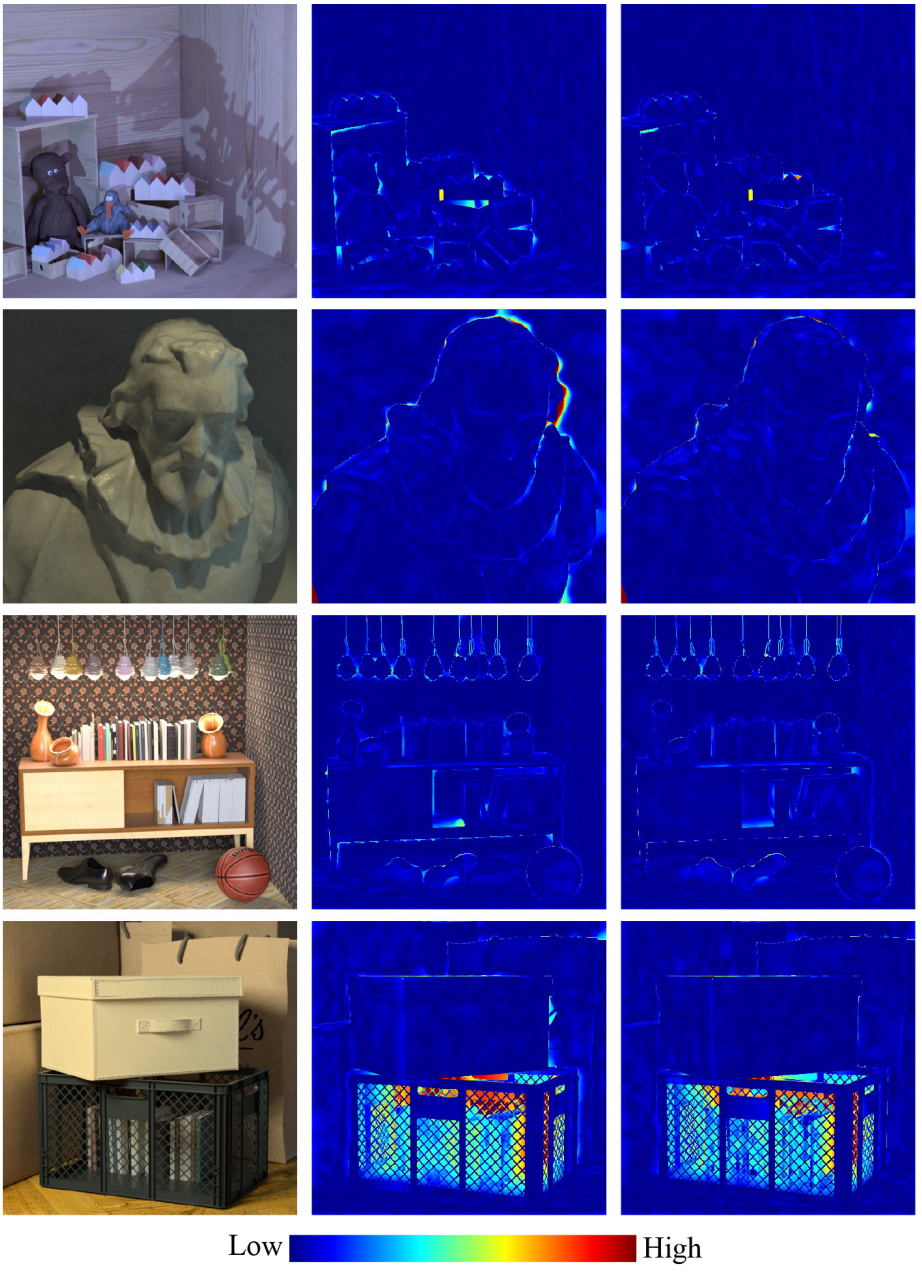


Figure 17: Multiview reprojection error (*center*) as self-supervised loss for depth edge localization, compared to our bidirectional diffusion gradients (*right*). We show absolute disparity error. Our method has lower error around edges.

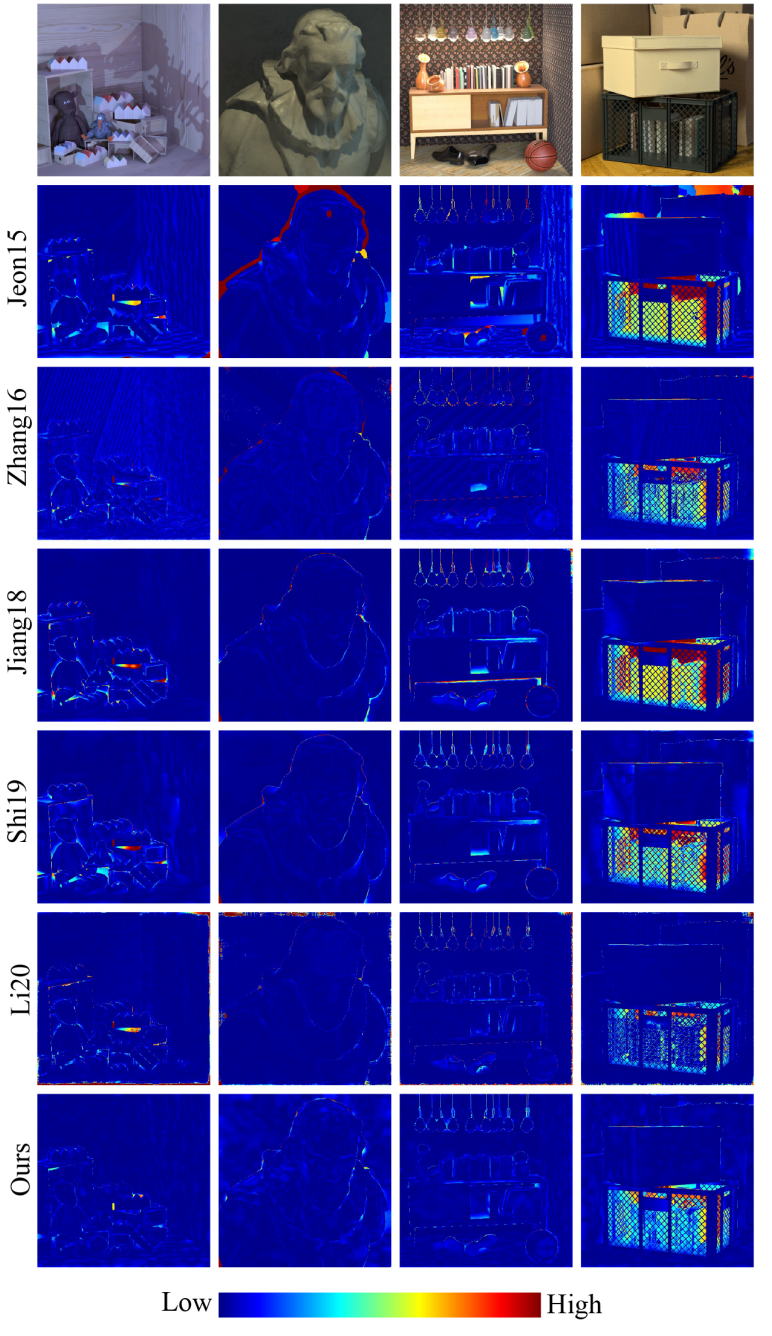


Figure 18: A visualization of the absolute disparity error for all baselines. *Top to bottom:* Jeon et al. [15], Zhang et al [36], Jiang et al. [16], Shi et al. [25], Li et al. [22], and our method.

Figure 19: Results on light fields from the Stanford dataset. *Top to bottom:* Jeon et al. [15], Zhang et al [36], Jiang et al. [16], Shi et al. [25], Li et al. [22] and our method.

Figure 20: Additional light field editing result. *Left:* input scene. *Center:* Our editing results. *Right, clockwise from top-left*: Detail of the unmodified light field image, Zhang et al. [36]'s editing result, Shi et al. [25]'s editing result, and our result with fewer artifacts.