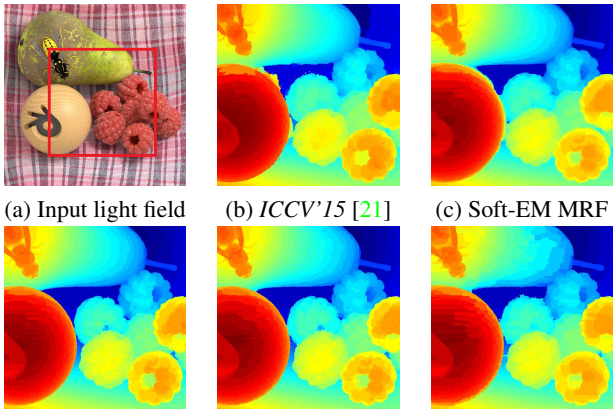
Robust Pseudo Random Fields for Light-Field Stereo Matching

Chao-Tsung Huang
National Tsing Hua University, Taiwan
chaotsung@ee.nthu.edu.tw

Abstract

Markov Random Fields are widely used to model light-field stereo matching problems. However, most previous approaches used fixed parameters and did not adapt to light-field statistics. Instead, they explored explicit vision cues to provide local adaptability and thus enhanced depth quality. But such additional assumptions could end up confining their applicability, e.g. *algorithms designed for dense light fields are not suitable for sparse ones.*

In this paper, we develop an empirical Bayesian framework—Robust Pseudo Random Field—to explore intrinsic statistical cues for broad applicability. Based on pseudo-likelihood, it applies soft expectation-maximization (EM) for good model fitting and hard EM for robust depth estimation. We introduce novel pixel difference models to enable such adaptability and robustness simultaneously. We also devise an algorithm to employ this framework on dense, sparse, and even denoised light fields. Experimental results show that it estimates scene-dependent parameters robustly and converges quickly. In terms of depth accuracy and computation speed, it also outperforms state-of-the-art algorithms constantly.



(a) Input light field (b) *ICCV'15* [21] (c) Soft-EM MRF

(d) RPRF (This work) (e) RPRF on 3×3 views (f) RPRF on 3 × 3 denoised views

Figure 1. **Robust Pseudo Random Fields (RPRFs).** (a) A challenging light field *StillLife* (9×9 views) in HCI dataset [23]. (b)-(f) The depth maps produced by (b) Wang *et al*. [21], (c) MRF using conventional soft-EM energy, (d) RPRF using robust hard-EM energy, (e) RPRF on a more sparse 3×3 light field, and (f) RPRF on a distorted 3×3 light field which is first corrupted by Gaussian noise ($\sigma = 10$) and then denoised by BM3D [4].

1. Introduction

Light-field stereo matching is an effective way to infer depth maps from color images. It is based on two properties: photo-consistency across views and depth continuity between pixels. They are often formulated by Markov Random Fields (MRFs) [2], a statistical graph model, for global optimization: the former as data energy and the latter as smoothness energy.

However, most previous approaches applied global optimization heuristically, not statistically. For example, the energy functions were not inferred from statistics but, instead, devised based on practical experience. They were often given in robust clipping forms (with constant parameters), such as truncated linear [25] and negative Gaussian [14], to preserve correct depth edges. Recent work has further explored advanced vision cues, such as depth consistency [22], line segments [25], and occlusion in angular patches [21], to achieve better depth quality. But these additional cues also narrow applicable scope correspondingly. For example, features for dense light fields ([22, 21]) may not work for sparse ones. Also, image denoising which is commonly used in low-light conditions could invalidate textural cues ([25]). In this paper, we aim to construct MRFs in a statistical way to infer robust energy functions for good depth accuracy and estimate scene-dependent parameters for broad applicability.

MRF parameter estimation via maximum likelihood is usually intractable because the normalization factor for unity is hard to calculate. Instead, pseudo-likelihood [1] is a classical approximation by exploring local dependence. One global MRF (likelihood) can be separated into lots of local neighborhoods (pseudo-likelihood) to collect statistics and perform distribution fitting. Nevertheless, this approach has a major issue for stereo matching: empirical distribu-

tions usually do not have robust clipping forms. Therefore, good distribution fitting will result in non-robust energy and thus over-smooth depth (Fig. 1(c)). On the other hand, keeping robust energy will lead to inaccurate fitting results.

Contributions. In this paper, we address this issue by developing a novel framework—Robust Pseudo Random Field (RPRF). Inspired by [7], we model pixel differences by scale mixtures with soft-edge hidden priors. For parameter estimation, we apply soft expectation-maximization (EM) by marginalizing out the hidden priors to achieve good pseudo-likelihood fitting. For MRF formulation, we employ *hard EM* by maximizing energy with respect to the priors to derive robust energy functions. Based on the proposed RPRF (Section 3), we devise an empirical Bayesian algorithm for light-field stereo matching in Section 4. It is, to our best knowledge, the first work of MRF parameter estimation for a single light field.

Extensive experimental results in Section 5 will show that this work has good statistical adaptability and produces great depth maps for not only dense light fields but also sparse and denoised ones. The scene-dependent parameters can also be estimated robustly with fast convergence. Finally, we demonstrate better depth accuracy and faster computation speed than previous work [22, 25, 21].

2. Related Work

Pseudo-likelihood. It assumes that local neighborhoods give independent observations; therefore, we can estimate parameters by maximizing the pseudo-likelihood that aggregates all the local observations. This approach has been widely used to learn MRF parameters for many different applications from training datasets [13, 16]. The reader is referred to [2] for further details. In this paper, we estimate parameters from a single light field.

Single-scene parameter estimation. Previous work for similar purposes focused on stereo image pairs and used the conventional framework: identical likelihood functions for distribution fitting and MRF inference. Zhang *et al*. [26] aimed to build robust energy functions and achieved that by performing soft EM on linear mixture models. However, the modeled distributions do not fit the histograms of pixel difference well. Also, it takes six iterations to converge between parameters and depth maps. In contrast, Liu *et al*. [15] and Su *et al*. [17] introduced advanced models to fit natural images, but the inferred depth maps do not have good accuracy. In this paper, we develop a new framework with quick convergence in which separate likelihood functions are used: soft-EM ones for good model fitting and hard-EM ones for robust energy functions.

Soft and Hard EM. They are conventional approaches for maximum likelihood estimation with unobserved data and usually used for different purposes. For example, the EM algorithm (soft EM) for clustering minimizes likelihood and the K-means (hard EM) optimizes data distortion [10]. For neighborhood filters, Huang [7] proposed a neighborhood noise model (NNM) to estimate parameters. The NNM fits heavy-tailed empirical distributions by soft EM and reasons robust range-weighted filters by hard EM. In this paper, we apply this approach to infer RPRFs for robust light-field stereo matching. We employ a similar model for the data energy with a new kernel function and propose a novel model for the smoothness energy to include depth labels.

Light-field stereo matching. Light fields possess lots of information for depth estimation, and previous work explored different vision cues for specific applications. Some approaches employed features for dense light fields, such as depth consistency [22], reliable data terms [11], bilateral consistency [3], phase shift [8], and occlusion in angular patches [21]. Some others focused on clean physical or textural cues, such as explicit occlusion [12], 3D line segments [25], defocus [19], sparse presentation [9], and convolutional networks [6]. In contrast, some targeted noisy light fields and applied noise-resistant cues, such as focal stack [14] and adaptive refocus response [24]. In this paper, we get back to fundamentals in MRF inference and explore intrinsic statistical cues for wide application scope.

3. RPRF Modeling

For a given light field, we estimate the disparity (inverse depth) map for the center view in which each pixel p has a k-channel color signal $\mathbf{z}_p$ and an unknown disparity label l_p. The disparity map $\mathcal{D} = \{l_p\}$ is derived by optimizing the global MRF energy

$$\sum_{p} \left(\sum_{v \in \mathcal{V}} E^{\mathrm{d}}_{pv}(l_p) + \lambda \sum_{q \in \mathbb{N}_4(p)} E^{\mathrm{s}}_{pq}(l_p, l_q) \right). \qquad (1)$$

The view-wise data energy E^{d}_{pv} measures photo-consistency by color difference between $\mathbf{z}_p$ and the corresponding signal $\mathbf{y}_{vp}(l_p)$ in a surrounding view v for the disparity l_p. The edge-wise smoothness energy E^{s}_{pq} evaluates depth continuity by color-conditioned disparity difference between a pixel pair p and q in 4-connected neighborhood $\mathbb{N}_4$. Finally, the weight parameter λ determines their ratio of contribution to the global energy.

We reason and infer the MRF in (1) by constructing pseudo-likelihood for pixel differences of each pixel p from its view-wise and spatial neighborhoods as shown in Fig. 2. In the following, we will introduce the view-wise pixel difference model for the data energy first and then the novel spatial model for the smoothness energy.

3.1. Pixel difference model for data energy

Let $\mathbf{X}_{\mathrm{d}}$ be a random vector for the view-wise color difference $\mathbf{x}_{\mathrm{d}} \triangleq \mathbf{z}_p - \mathbf{y}_{vp}(l_p)$. Its empirical distribution is

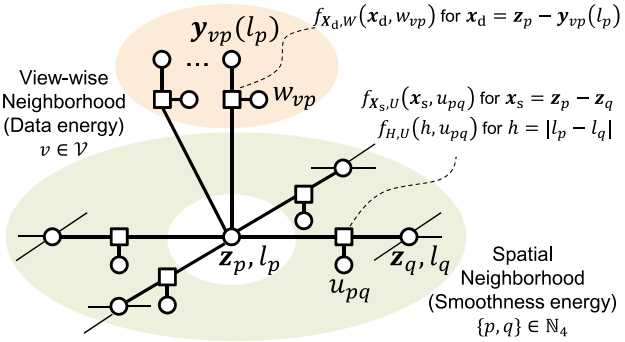


Figure 2. **Pseudo-likelihood modeling for RPRFs.** For the view-wise neighborhood, color difference $\mathbf{x}_\mathrm{d}$ is modeled by a scale mixture with a soft-edge hidden variable w_{vp}. For the spatial neighborhood, color difference $\mathbf{x}_\mathrm{s}$ and disparity difference h are formulated by different scale mixtures with an identical hidden u_{pq}.

usually heavy-tailed, and we employ a model similar to the NNM for good model fitting. We introduce a scale random variable W to model color edges using soft decisions and formulate $\mathbf{X}_\mathrm{d}$ in a Gaussian scale mixture (GSM):

$$\mathbf{X}_\mathrm{d}|W = w \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma_\mathrm{d}^2}{w}\mathbf{I}_k\right), \tag{2}$$

$$f_W(w) = \frac{1}{N_\mathrm{d}} w^{-\frac{k}{2}} e^{\alpha_\mathrm{d} G(w)},\ w \in [\epsilon_\mathrm{d}, 1], \tag{3}$$

where σ_d is a scale parameter for $\mathbf{X}_\mathrm{d}$, α_d and ϵ_d are shape parameters for W, and N_d is the normalization factor for unity. The hidden prior function $G(w)$ controls the distribution of W and will be shown its direct link to the energy function. Note that the σ_d here represents only distribution scaling, not for the noise intensity in the NNM.

Parameter estimation by soft EM Given an estimated disparity map $\tilde{\mathcal{D}} = \{\tilde{l}_p\}$, we can use the corresponding signals from surrounding views $\{\mathbf{y}_{vp}(\tilde{l}_p)\}$ to update the parameter set for data energy $\boldsymbol{\theta}_\mathrm{d} = (\sigma_\mathrm{d}, \alpha_\mathrm{d}, \epsilon_\mathrm{d})^T$. Consider a sufficient statistic $t_\mathrm{d} \triangleq \| \mathbf{x}_\mathrm{d} \|_2$. By marginalizing out the hidden W from the joint distribution $f_{T_\mathrm{d},W}$, we can have its soft-EM energy:

$$E_{T_\mathrm{d}}^{\mathrm{soft}}(t_\mathrm{d}; \boldsymbol{\theta}_\mathrm{d}) = -\log \int_{\epsilon_\mathrm{d}}^{1} f_{T_\mathrm{d},W}(t_\mathrm{d}, w; \boldsymbol{\theta}_\mathrm{d}) dw. \tag{4}$$

Note that it is a non-analytic function and requires numerical integrals for evaluation. The updated parameter $\hat{\boldsymbol{\theta}}_\mathrm{d}$ can be derived by model fitting or, equivalently, global energy optimization using empirical observations:

$$\hat{\boldsymbol{\theta}}_\mathrm{d} = \arg\min_{\boldsymbol{\theta}_\mathrm{d}} \sum_{\tilde{t}_\mathrm{d}} E_{T_\mathrm{d}}^{\mathrm{soft}}(\tilde{t}_\mathrm{d}; \boldsymbol{\theta}_\mathrm{d}), \tag{5}$$

where $\tilde{t}_\mathrm{d} = \| \mathbf{z}_p - \mathbf{y}_{vp}(\tilde{l}_p) \|_2$. Then the parameter set $\boldsymbol{\theta}_\mathrm{d}$ can be estimated by iterative updates until converged.

Type	**Energy** $E(x)$	Kernel $K(x)$	Hidden Prior $G(w)$
Reciprocal	$-\frac{1}{x+1}$	$\frac{1}{(x+1)^2}$	$2\sqrt{w} - w$
Gaussian	$-e^{-x}$	e^{-x}	$w(1 - \log w)$

Table 1. **Examples of robust hard-EM energy functions.** We adopt the Reciprocal energy as detailed in Section 3.3.

Robust data energy by hard EM Define an energy function $E(x)$ related to the prior $G(w)$ by

$$E(x) \triangleq \min_w \left(wx - G(w)\right), x \geq 0. \tag{6}$$

Then we can construct the hard-EM energy for color difference $\mathbf{X}_\mathrm{d}$, given a parameter set $\boldsymbol{\theta}_\mathrm{d}$, by maximizing the joint distribution $f_{\mathbf{X}_\mathrm{d},W}$ with respect to the hidden W (select the best guess of w):

$$E_{\mathbf{X}_\mathrm{d}}^{\mathrm{hard}}(\mathbf{x}_\mathrm{d}) = -\log \max_w f_{\mathbf{X}_\mathrm{d},W}(\mathbf{x}_\mathrm{d}, w) \tag{7}$$

$$= \alpha_\mathrm{d} E\left(\frac{\| \mathbf{x}_\mathrm{d} \|_2^2}{2\alpha_\mathrm{d}\sigma_\mathrm{d}^2}\right) + C, \tag{8}$$

where C is a constant offset and will be discarded because only the energy difference matters for MRF inference. Therefore, we have the view-wise data energy in (1) as

$$E_{pv}^\mathrm{d}(l_p) = E_{\mathbf{X}_\mathrm{d}}^{\mathrm{hard}}(\mathbf{z}_p - \mathbf{y}_{vp}(l_p)). \tag{9}$$

For exploring the relationship between $E(x)$ and $G(w)$, we define a kernel function $K(x)$ by

$$K(x) \triangleq (G')^{-1}(x). \tag{10}$$

Then it can be shown using integration by parts that $K(x)$ is exactly the derivative of $E(x)$. As a result, the three core functions are directly connected by

$$E' = K = (G')^{-1}. \tag{11}$$

Table 1 shows two examples, Reciprocal and Gaussian, for a robust energy function $E(x)$.

3.2. Pixel difference model for smoothness energy

Let H be a random variable for the spatial disparity difference $h \triangleq |l_p - l_q|$, $\mathbf{X}_\mathrm{s}$ be a random vector for the corresponding color difference $\mathbf{x}_\mathrm{s} \triangleq \mathbf{z}_p - \mathbf{z}_q$, and U be a soft-edge scale random variable. We propose a scale mixture of generalized Gaussian distributions for $f_{H|U}$ to fit the thick-tailed distribution of H. Along with a GSM for $\mathbf{X}_\mathrm{s}$, we construct the following joint model:

$$f_{H|U}(h|u) = \frac{2}{\gamma(\frac{1}{\beta} + 1)\delta} u^{\frac{1}{\beta}} e^{-u(\frac{h}{\delta})^\beta}, \tag{12}$$

$$\mathbf{X}_\mathrm{s}|U = u \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma_\mathrm{s}^2}{u}\mathbf{I}_k\right), \tag{13}$$

$$f_U(u) = \frac{1}{N_\mathrm{s}} u^{-(\frac{k}{2} + \frac{1}{\beta})} e^{\alpha_\mathrm{s} G(u)},\ u \in [\epsilon_\mathrm{s}, 1], \tag{14}$$

where δ and β are the scale and shape parameters for H, and the β is fixed to 1.5 in this paper. The $\sigma_{\rm s}$, $\alpha_{\rm s}$, $\epsilon_{\rm s}$ and $N_{\rm s}$ serve the same purposes as the $\sigma_{\rm d}$, $\alpha_{\rm d}$, $\epsilon_{\rm d}$ and $N_{\rm d}$ separately.

The disparity difference H shares the same edge prior U with the color difference $\mathbf{X}_{\rm s}$ for inferring color-conditioned MRF. Also, the additional factor $u^{-\frac{1}{\beta}}$ in $f_U(u)$, compared to $f_W(w)$, is devised to cancel out the $u^{\frac{1}{\beta}}$ in $f_{H|U}(h|u)$ for the joint distribution $f_{\mathbf{X}_{\rm s},H,U}$. Thus the hard-EM smoothness energy can have a simple form similar to the case of the data energy (8).

Parameter estimation by soft EM Consider the parameter set $\boldsymbol{\theta}_{\rm s} = (\delta, \sigma_{\rm s}, \alpha_{\rm s}, \epsilon_{\rm s})^T$. We update it using the empirical $\tilde{t}_{\rm s}$ for a sufficient statistic $t_{\rm s} \triangleq \| \mathbf{x}_{\rm s} \|_2$ and also the empirical disparity difference $\tilde{h}$ from a given disparity map $\tilde{\mathcal{D}}$. However, using their soft-EM joint energy, which marginalizes out U for $f_{T_{\rm s},H,U}$, will induce a computation issue: it needs to evaluate the non-analytic energy for every pair of $\tilde{t}_{\rm s}$ and $\tilde{h}$. Instead, we derive the updated $\hat{\boldsymbol{\theta}}_{\rm s}$ by using their marginal energy to speed up the process:

$$\hat{\boldsymbol{\theta}}_{\rm s} = \arg\min_{\boldsymbol{\theta}_{\rm s}} \left(\sum_{\tilde{t}_{\rm s}} E^{\rm soft}_{T_{\rm s}}(\tilde{t}_{\rm s}; \boldsymbol{\theta}_{\rm s}) + \sum_{\tilde{h}} E^{\rm soft}_{H}(\tilde{h}; \boldsymbol{\theta}_{\rm s}) \right). \tag{15}$$

Robust smoothness energy by hard EM Similarly to (7)-(8), we maximize the joint distribution $f_{\mathbf{X}_{\rm s},H,U}$ with respect to the hidden U and have the hard-EM energy for pixel differences $\mathbf{x}_{\rm s}$ and h as follows:

$$E^{\rm hard}_{\mathbf{X}_{\rm s},H}(\mathbf{x}_{\rm s}, h) = \alpha_{\rm s} E \left(\frac{\| \mathbf{x}_{\rm s} \|_2^2}{2\alpha_{\rm s}\sigma_{\rm s}^2} + \frac{h^\beta}{\alpha_{\rm s}\delta^\beta} \right). \tag{16}$$

At last, we have the edge-wise smoothness energy in (1):

$$E^{\rm s}_{pq}(l_p, l_q) = E^{\rm hard}_{\mathbf{X}_{\rm s},H}(\mathbf{z}_p - \mathbf{z}_q, |l_p - l_q|), \tag{17}$$

which constitutes a conditional random field.

3.3. On selection of energy function $E(x)$

The proposed models can fit light fields of different characteristics and track their heavy tails well as shown in Fig. 3(a)-(b). The Reciprocal type in Table 1 can provide better fitting accuracy than the conventional Gaussian one, especially for spatial difference $\| \mathbf{x}_{\rm s} \|_2$. Therefore, we adopt it for the stereo matching algorithm in this paper. In addition, Fig. 3(c)-(d) show the corresponding energy functions. The soft-EM ones are close to the empirical ones due to the model fitting; however, they induce bad depth edges as shown in Fig. 1(c). In contrast, the hard-EM ones have similar values for small color difference but saturate quickly as robust metrics for large difference. As a result, the depth edges can be well preserved as Fig. 1(d) shows.

4. Implementation for Stereo Matching

Parameter estimation Given an disparity map, we estimate the parameters $\boldsymbol{\theta}_{\rm d}$ and $\boldsymbol{\theta}_{\rm s}$ for data and smoothness energy from the corresponding view-wise and spatial pixel differences. For the $\boldsymbol{\theta}_{\rm d}$, we applied the EM+ fitting method designed for NNM in [7] to our soft-EM update formulation (5). Small changes were made for the model difference. For the $\boldsymbol{\theta}_{\rm s}$ updated by (15), we modified the method to include the disparity difference with its additional parameter δ and statistics $\tilde{h}$.

Belief propagation Given the global MRF energy (1), we use belief propagation (BP) [5] to solve the disparity map iteratively. We adopted the BP-M approach in [18] for its efficient message propagation. We stop the BP-M if the global energy is not decreased by more than 1%, and around four iterations on average are performed in our experiments.

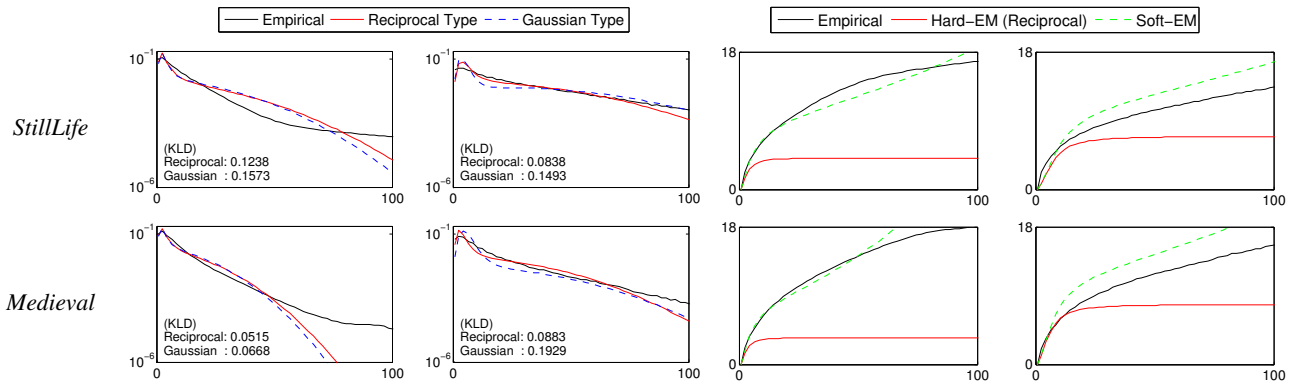
Energy approximation We use the linear-time algorithm in [5] for fast message updating. To achieve this, we approximated the Reciprocal smoothness energy (16) by the truncated linear function with the least squared error:

$$E^{\rm hard}_{\mathbf{X}_{\rm s},H} \simeq \alpha_{\rm s} \min \left(\frac{0.3726}{\delta \alpha_{\rm s}^{\frac{1}{\beta}} b^{\frac{1}{\beta}+1}} h, \frac{1}{b} \right) + \text{const}, \tag{18}$$

where $b = 1 + \frac{\|\mathbf{x}_{\rm s}\|_2^2}{2\alpha_{\rm s}\sigma_{\rm s}^2}$ which can be calculated in advance for each pixel pair p and q before running BP-M.

Adaptive selection for λ The parameters $\alpha_{\rm d}$ and $\alpha_{\rm s}$ statistically determine the dynamic ranges of the data and smoothness energy. We further use the heuristic parameter λ to weight importance between them for different conditions. For example, denoised light fields relies on smoothness energy more than normal ones, so we assign larger values to λ for them. Also, we set the values proportional to the numbers of surrounding views, $|\mathcal{V}|$. But they are lower truncated because the data energy using few views becomes unreliable. Lastly, we use two categories, $\lambda^{\rm weak}$ and $\lambda^{\rm strong}$, for scene-adaptive assignment as listed in Table 2.

Entropy-based scene adaptability There are two typical scenarios as shown in Fig. 4: one needs a small λ, as *Still-Life*, to minimize depth errors, and the other prefers a large λ, as *Medieval*. We found that the cross entropy $H(\tilde{h}, h)$ between the empirical and modeled disparity differences is a good indicator for them. A small reduction of $H(\tilde{h}, h)$ for λ from zero to a large value means the data energy outweighs the smoothness one, so a weak λ is preferred. On the contrary, a significant reduction encourages a strong λ. Based on this observation, we use the reduction ratio to adaptively assign $\lambda^{\rm weak}$ and $\lambda^{\rm strong}$.



(a) Model fitting of $t_{\mathrm{d}} = \| \mathbf{x}_{\mathrm{d}} \|_2$ (b) Model fitting of $t_{\mathrm{s}} = \| \mathbf{x}_{\mathrm{s}} \|_2$ (c) Data energy vs. $\| \mathbf{x}_{\mathrm{d}} \|_2$ (d) Smoothness energy vs. $\| \mathbf{x}_{\mathrm{s}} \|_2 \ \ (h = 0)$

Figure 3. **Modeling fitting and robust energy.** The top row is for the light field *StillLife* and the bottom for *Medieval*. Ground-truth disparity maps are used to generate empirical distributions. (a)-(b) Distribution fitting results using the Reciprocal and Gaussian types for (a) view-wise color difference $\| \mathbf{x}_{\mathrm{d}} \|_2$ and (b) spatial difference $\| \mathbf{x}_{\mathrm{s}} \|_2$ with their Kullback-Leibler divergence (KLD) shown at the corners. (c)-(d) Corresponding energy functions of the Reciprocal type for (c) data energy and (d) smoothness energy.

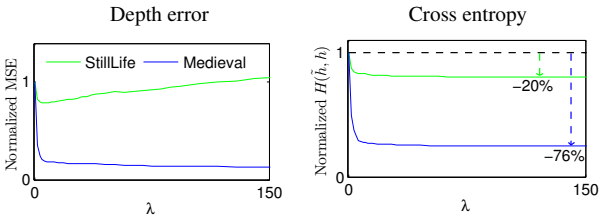


Figure 4. **Depth estimation quality vs. λ.** Depth error is represented by mean squared error (MSE) in disparity. Cross entropy measures the distance between the distributions of the empirical disparity difference $\tilde{h}$ and the modeled h. Their values are both normalized with respect to the case of $\lambda = 0$ for comparison.

Stereo matching algorithm At first, we initialize a disparity map $\mathcal{D}^{\mathrm{ini}}$ using MRF inference with default parameters: $\lambda^{\mathrm{ini}} = 300$, $\boldsymbol{\theta}_{\mathrm{d}}^{\mathrm{ini}} = (\sqrt{2/3}, 6, 0.1)$, and $\boldsymbol{\theta}_{\mathrm{s}}^{\mathrm{ini}} = (0.05, \sqrt{8/3}, 9, 0.1)$. Then we update parameters and estimate depth iteratively. In each iteration, we use the $\lambda^{\mathrm{strong}}$ in Table 2 to infer the disparity map first. If the cross entropy $H(\tilde{h}, h)$ is reduced by less than a ratio r, 50% in this paper, compared to the case of $\lambda = 0$, we will switch to the weak scenario and use λ^{weak} for inference instead.

4.1. Parameter estimation

5. Experiments

We perform extensive experiments on three datasets for objective evaluation: *HCI Blender* [23] and *Berkeley* [21] are synthetic, and *HCI Gantry* [23] contains real pictures. They are all dense light fields of 9×9 views, and we subsample them to produce sparse 5×5 and 3×3 test cases. We also generate denoised light fields by adding Gaussian noise and then performing BM3D [4]. For comparing objective

Condition	λ^{weak}	$\lambda^{\mathrm{strong}}$
Normal	$\max(\lvert\mathcal{V}\rvert/4, 4)$	$\max(3\lvert\mathcal{V}\rvert, 24)$
Denoised	$\max(\lvert\mathcal{V}\rvert/2, 6)$	$\max(3\lvert\mathcal{V}\rvert, 48)$

Table 2. **Adaptive selection of λ.** Two operating conditions are considered: normal and denoised light fields. Two scene-adaptive categories, λ^{weak} and $\lambda^{\mathrm{strong}}$, are further defined.

quality across view types, we calculate mean squared errors (MSE) in disparity all with respect to the baselines of 9×9 light fields. The values are then multiplied by 100 to keep significant figures and denoted by DMSE.

We will also show results for light fields captured by Lytro ILLUM for demonstrating generalization capability. Light-field RAW data from EPFL dataset [20] and our own pictures are processed using Lytro Power Tools Beta. Our software and more results are available online[1].

Robustness to initial conditions Different initial parameters lead to different initialized disparity maps $\mathcal{D}^{\mathrm{ini}}$. For example, a large value of λ^{ini} prefers the smoothness energy and gives an over-smooth $\mathcal{D}^{\mathrm{ini}}$. Note that an ideal $\mathcal{D}^{\mathrm{ini}}$ is the ground-truth depth itself. In this work, different $\mathcal{D}^{\mathrm{ini}}$ can generate similar MRF parameters as shown in Fig. 5. Such robustness can be explained by the corresponding empirical distributions in Fig. 6. They differ mainly in distribution tails but behave similarly for small pixel difference; therefore, the inferred hard-EM energy functions are also similar. Note that the difference of the tails results in the variation of ϵ_{d} and ϵ_{s}, but these two parameters do not affect the hard-EM energy and thus the MRF inference.

[1] http://www.ee.nthu.edu.tw/chaotsung/rprf

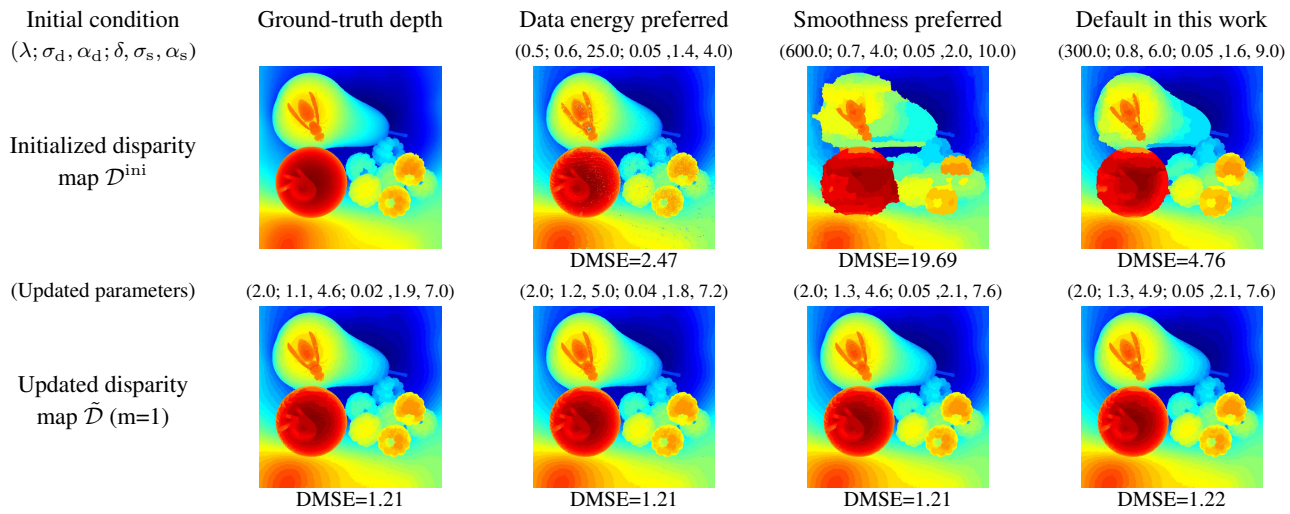


Figure 5. **Robust update for parameters and depth.** Four initial disparity maps $\mathcal{D}^{\mathrm{ini}}$ for the 3×3 case of *StillLife* are used for comparison. One uses ground-truth depth (ideal). The others are initialized by different parameter sets: one is noisy by weighting data energy more (λ=0.5), one is over-smooth by weighting smoothness term more (λ=600), and the last one is moderate using the default setting (λ=300). The updated parameters for MRF energy are all similar in one iteration, and the accordingly estimated disparity maps show little difference.

Quick convergence The robustness to initialized disparity maps also results in the quick convergence of parameter-depth update iterations. Fig. 7 shows the accuracy of the parameters estimated by the default initial condition compared to the ideal one. All MRF parameters can be well estimated and converged in one iteration except δ. But such inaccuracy of δ only affects depth quality slightly, e.g. the second iteration increases its accuracy but only improves the DMSE by 0.2% on average. Therefore, we set the default iteration number to one.

Parameter variation Consider the two bandwidth parameters of MRF inference: $\alpha_d\sigma_d^2$ (data) and $\alpha_s\sigma_s^2$ (smoothness). Over 9x9 light fields, their value ranges are [2.2, 8.2] and [10.0, 63.5], respectively. They distribute more uniformly than a normal distribution, and their excess kurtosis are -0.9 and -0.7. For denoised light fields (σ=10), the data bandwidth becomes 4.7x larger and the smoothness one 0.4x smaller. All these variations can be estimated well by this work.

5.1. Depth estimation

We compare our results (RPRF) with the globally consistent depth labeling (GCDL) [22], line-assisted graph cut (LAGC) [25], phase-shift cost volume (PSCV) [8], and occlusion-aware depth estimation (OADE) [21]. We use the codes provided by the authors.

Dense and sparse light fields Table 3 details the depth estimation errors for dense 9×9 test cases, and our work has better performance. Fig. 8 further compares the results

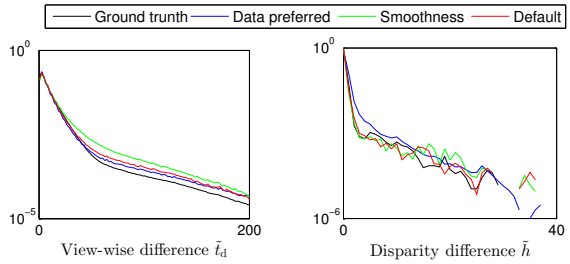


Figure 6. **Empirical distributions.** They are collected using the initialized disparity maps in Fig. 5. Note that the distributions of $\tilde{t}_{\mathrm{s}}$ are not related to disparity maps and are all the same.

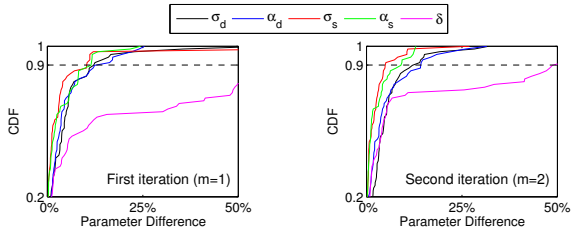


Figure 7. **Parameter estimation accuracy.** Accuracy is measured by relative absolute difference, and cumulative distribution functions (CDFs) derived from all test cases for *HCI Blender* dataset.

from dense to sparse light fields. The GCDL and OADE fail in sparse cases, while the LAGC becomes worse in denser ones. In contrast, PSCV and our work have constant quality over these view types. We demonstrate that sparse views can generate depth maps as great as dense ones do if robust energy from clean images is used. Fig. 9 shows an example for subjective evaluation.

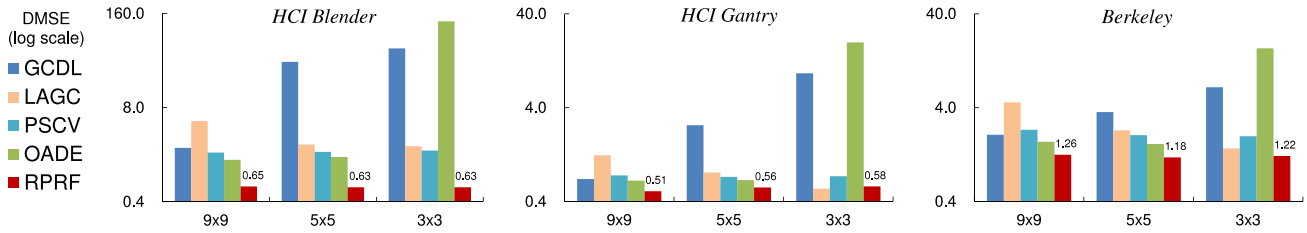


Figure 8. **Depth estimation error vs. Light-field view type.** The average errors are represented in DMSE (log scale).

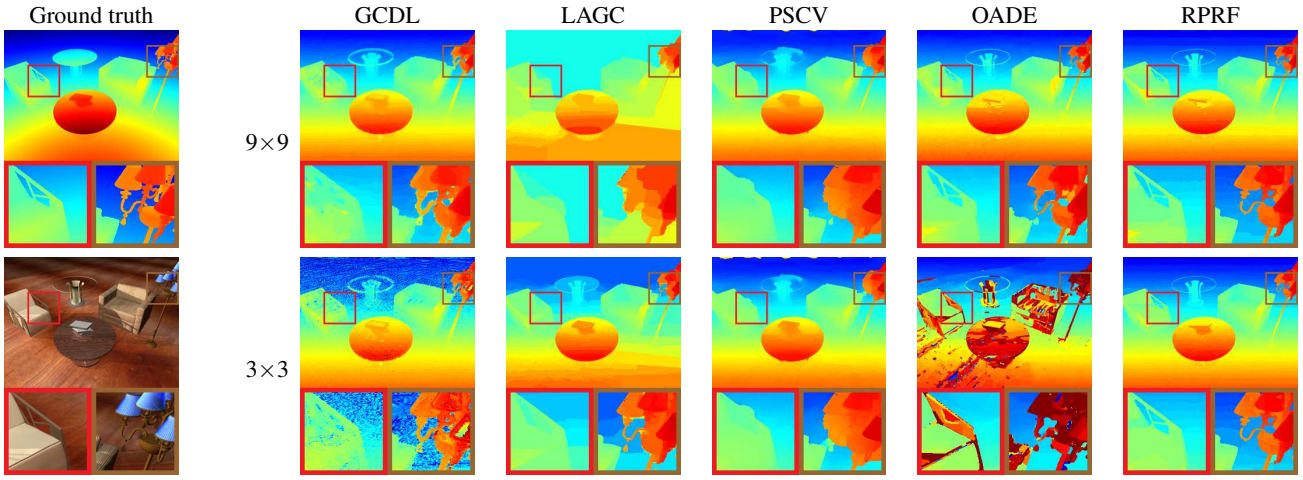


Figure 9. **Estimated depth maps for *LivingRoom*.** This work infers good depth edges in both 9×9 and 3×3 test cases, especially around the chair arm and the lamps. GCDL and OADE fail in the 3×3 case, and LAGC and PSCV produce over-smooth depth.

Dataset	Light Field	GCDL	LAGC	PSCV	OADE	RPRF
HCI Blender	*Buddha*	0.68	3.04	1.13	0.91	**0.28**
	Horses	4.98	2.73	1.70	1.36	**0.50**
	Papillon	2.68	19.51	5.98	1.00	**0.66**
	StillLife	4.01	2.28	2.10	4.29	**1.09**
	Buddha2	1.02	2.71	**0.45**	1.18	0.75
	Medieval	1.24	4.49	1.40	1.15	**0.79**
	Mona	0.93	1.70	0.66	0.73	**0.47**
	Average	2.22	5.21	1.92	1.52	**0.65**
HCI Gantry	*Couple*	**0.39**	0.62	0.64	1.00	0.50
	Cube	0.73	1.27	1.88	1.02	**0.67**
	Maria	**0.15**	0.47	0.23	0.16	0.18
	Pyramide	0.45	1.38	0.69	0.61	**0.44**
	Statue	1.75	2.47	**0.35**	0.55	0.78
	Average	0.69	1.24	0.76	0.67	**0.51**
Berkeley	*Bedroom*	0.52	1.70	0.48	0.59	**0.38**
	LivingRoom	2.12	10.46	2.62	1.91	**1.77**
	Outdoor	0.60	2.62	0.93	**0.34**	0.38
	Plant	5.00	3.43	5.25	4.10	**2.50**
	Average	2.06	4.55	2.32	1.73	**1.26**

Table 3. **Depth errors in DMSE for dense 9×9 light fields.**

Denoised light fields Applying the RPRF to the light fields denoised by BM3D produces better depth than directly applying to the noisy ones. It can even outperform the algorithms designed for noisy light fields in [14, 24] as shown in Table 4. The results regarding denoising conditions are summarized in Fig. 10. In these cases, fewer views will decrease depth quality owing to the less reliable energy.

Real scenes Fig. 11 shows results for light fields captured by Lytro ILLUM. Our work constantly produces similarly good depth maps using only 3×3 views compared to Lytro software that uses raw light fields. Other algorithms also perform well but occasionally cause obvious artifacts.

Execution time We implement parameter estimation in MATLAB and the other parts in C++. For one light field, the parameter estimation and BP-M take about 7 and 10 seconds respectively on average. The remaining computation is mostly contributed by computing data energy. The run times are summarized in Table 5. Our work runs much faster for its simple but efficient MRF formulation.

6. Discussion and Limitation

Occlusion handling. Instead of explicit handling as [21], we show great depth quality can be achieved by implicit modeling with the soft-edge prior w. A value toward zero represents more likely occlusion. In this case, hard-EM energy will saturate data cost, which equivalently separates the occlusion pixel in a soft way. Therefore, the fact that hard-EM energy is better than soft-EM one also confirms the necessity of occlusion formulation.

Scene statistics. Consider a scene has two regions of different statistics, e.g. tablecloth and fruits in *StillLife*. In

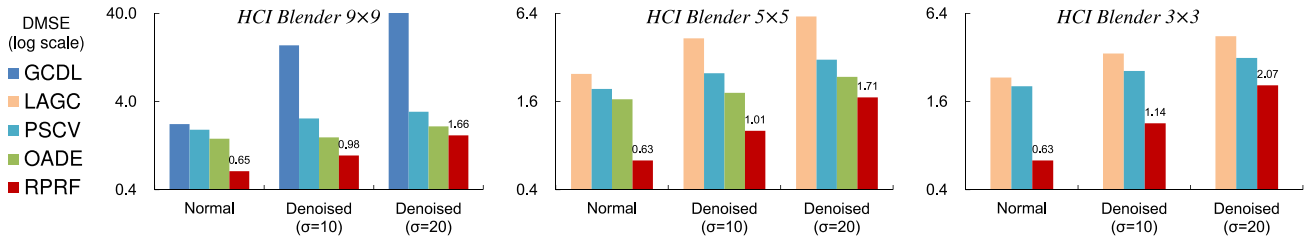


Figure 10. **Depth estimation error vs. Denoising conditions.** The cases with too large DMSE are omitted for clarity.

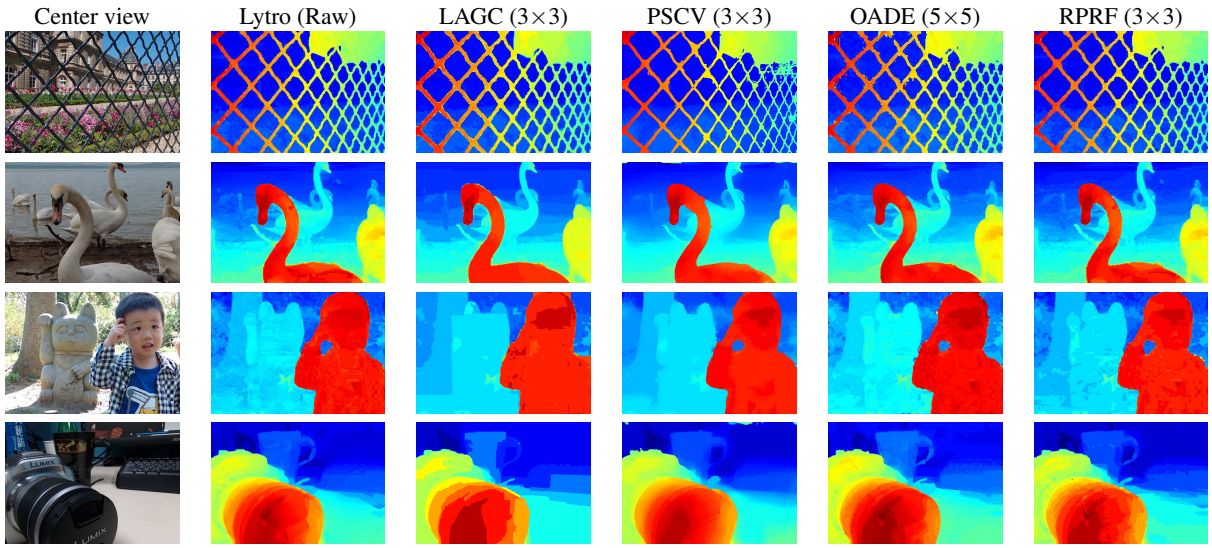


Figure 11. **Estimated depth maps for real scenes.** The first two light fields are from EPFL dataset, and the last two captured by us.

Light Field	Noise std.	[14] (9x9)	[24] (9x9)	BM3D+ RPRF (3x3)	BM3D+ RPRF (9x9)
Buddha	σ=10	1.73	1.70	0.53	**0.44**
Buddha	σ=20	3.03	N/A	0.86	**0.71**
Mona	σ=10	2.06	1.25	0.62	**0.53**
Mona	σ=20	3.83	N/A	**0.85**	0.86
StillLife	σ=10	N/A	2.51	1.81	**1.55**

Table 4. **Depth errors in DMSE for noisy light fields.** The numbers reported by Lin *et al*. [14] and Williem *et al*. [24] are used.

LF type	GCDL	LAGC	PSCV	OADE	RPRF
9x9	3,516	334,816	1,951	672	**87**
5x5	3,518	38,233	1,040	357	**33**
3x3	3,562	7,379	778	331	**24**

Table 5. **Average run time in seconds per light field.** GCDL ran on a GPU, GeForce GT 630, and others on a 3.4 GHz CPU.

this case, our model will capture mixed statistics, and the result could be sub-optimal. In this viewpoint, a possible extension of this work is to segment a scene into different regions and then learn parameters separately.

Hyper-parameter λ. It cannot be explained by RPRF and thus requires heuristic estimation. We found that depth quality is not sensitive to its small variation, so we applied an entropy-based heuristic for coarse-level adaptability. A possible extension of this work is to devise a more delicate heuristic to improve accuracy.

7. Conclusion

In this paper, we propose an empirical Bayesian framework—RPRF—to provide statistical adaptability and good depth quality for light-field stereo matching. Two scale mixtures with soft-edge priors are introduced to model the data and smoothness energy. We estimate scene-dependent parameters by pseudo-likelihood fitting via soft EM and infer depth maps using robust energy functions via hard EM. Accordingly, we build a stereo matching algorithm with efficient implementation. The effectiveness is demonstrated by experimental results on dense, sparse, and denoised light fields. It outperforms state-of-the-art algorithms in terms of depth accuracy and computation speed. We also believe that this framework can be extended in many possible ways to achieve better depth quality.

Acknowledgment

This work was supported by the Ministry of Science and Technology, Taiwan, R.O.C. under Grant No. MOST 103-2218-E-007-008-MY3.

References

[1] J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society Series D (The Statistician)*, 24(3):179–195, Sept. 1975. 1

[2] A. Blake, P. Kohli, and C. Rother. *Markov Random Fields for Vision and Image Processing*. The MIT Press, 2011. 1, 2

[3] C. Chen, H. Lin, Z. Yu, S. B. Kang, and J. Yu. Light field stereo matching using bilateral statistics of surface cameras. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1518–1525, 2014. 2

[4] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, Aug. 2007. 1, 5

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 70(1):41–54, Oct. 2006. 4

[6] S. Heber and T. Pock. Convolutional networks for shape from light field. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3746–3754, 2016. 2

[7] C.-T. Huang. Bayesian inference for neighborhood filters with application in denoising. *IEEE Transactions on Image Processing*, 24(11):4299–4311, Nov. 2015. 2, 4

[8] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon. Accurate depth map estimation from a lenslet light field camera. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1547–1555, 2015. 2, 6

[9] O. Johannsen, A. Sulc, and B. Goldluecke. What sparse light field coding reveals about scene structure. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3262–3270, 2016. 2

[10] M. Kearns, Y. Mansour, and A. Y. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 282–293, 1997. 2

[11] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Transactions on Graphics*, 32(4):73:1–73:12, July 2013. 2

[12] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proceedings of the 7th European Conference on Computer Vision*, pages 82–96, 2002. 2

[13] S. Kumar and M. Hebert. Discriminative random fields: a discriminative framework for contextual interaction in classification. In *IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, 2003. 2

[14] H. Lin, C. Chen, S. B. Kang, and J. Yu. Depth recovery from light field using focal stack symmetry. In *IEEE International Conference on Computer Vision*, pages 3451–3459, 2015. 1, 2, 7, 8

[15] Y. Liu, L. K. Cormack, and A. C. Bovik. Statistical modeling of 3-D natural scenes with application to Bayesian stereopsis. *IEEE Transactions on Image Processing*, 20(9):2515–2530, Sept. 2011. 2

[16] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007. 2

[17] C.-C. Su, L. K. Cormack, and A. C. Bovik. Color and depth priors in natural images. *IEEE Transactions on Image Processing*, 22(6):2259–2274, June 2013. 2

[18] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):1068–1080, June 2008. 4

[19] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *IEEE International Conference on Computer Vision*, pages 673–680, 2013. 2

[20] M. Řeřábek and T. Ebrahimi. New light field image dataset. In *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016. 5

[21] T.-C. Wang, A. A. Efros, and R. Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *IEEE International Conference on Computer Vision*, pages 3487–3495, 2015. 1, 2, 5, 6, 7

[22] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4D light fields. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 41–48, 2012. 1, 2, 6

[23] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4D light fiels. In *Vision, Modeling, and Visualization*, pages 145–152, 2013. 1, 5

[24] W. Williem and I. K. Park. Robust light field depth estimation for noisy scene with occlusion. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 4396–4404, 2016. 2, 7, 8

[25] Z. Yu, X. Guo, H. Ling, A. Lumsdaine, and J. Yu. Line assisted light field triangulation and stereo matching. In *IEEE International Conference on Computer Vision*, pages 2792–2799, 2013. 1, 2, 6

[26] L. Zhang and S. M. Seitz. Estimating optimal parameters for MRF stereo from a single image pair. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):331–342, Feb. 2007. 2