Empirical Bayesian Light-Field Stereo Matching by Robust Pseudo Random Field Modeling

Chao-Tsung Huang, *Member, IEEE*

Abstract—Light-field stereo matching problems are commonly modeled by Markov Random Fields (MRFs) for statistical inference of depth maps. Nevertheless, most previous approaches did not adapt to image statistics but instead adopted fixed model parameters. They explored explicit vision cues, such as depth consistency and occlusion, to provide local adaptability and enhance depth quality. However, such additional assumptions could end up confining their applicability, e.g. algorithms designed for dense view sampling are not suitable for sparse one. In this paper, we get back to MRF fundamentals and develop an empirical Bayesian framework—Robust Pseudo Random Field—to explore intrinsic statistical cues for broad applicability. Based on pseudo-likelihoods with hidden soft-decision priors, we apply soft expectation-maximization (EM) for good model fitting and perform hard EM for robust depth estimation. We introduce novel pixel difference models to enable such adaptability and robustness simultaneously. Accordingly, we devise a stereo matching algorithm to employ this framework on dense, sparse, and even denoised light fields. It can be applied to both true-color and grey-scale pixels. 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.

Index Terms—Stereo matching, light field, Markov Random Field, empirical Bayesian method

1 Introduction

Light-field stereo matching is an effective way to infer depth maps from pictures captured in different viewpoints. It is based on two properties: photo-consistency across views and depth continuity between pixels. They are often formulated by Markov Random Fields (MRFs) [1], 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 [2] and negative Gaussian [3], to preserve correct depth edges. Recent work has further explored advanced vision cues, such as depth consistency [4], line segments [2], phase shift [5], occlusion in angular patches [6], spinning parallelogram operator [7], and constrained angular entropy [8] to achieve better depth quality. But these additional cues also narrow applicable scope correspondingly. For example, features for dense light fields ([4], [6], [8]) may not work for sparse ones. Also, image denoising which is commonly used in low-light conditions could invalidate textural cues ([2], [5], [7]). 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 modeling [9] 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 distributions usually do not have robust clipping forms. Therefore, good distribution fitting will result in non-robust energy and thus over-smooth depth (Fig. 1f), especially near object boundaries. On the other hand, keeping robust energy will lead to inaccurate fitting results.

In this paper, we address this issue by developing a novel statistical method for both MRF modeling and inference. There are two major contributions:

1) An empirical Bayesian framework—Robust Pseudo Random Field (RPRF)—which estimates scene-dependent parameters accurately and infers edge-preserved depth robustly;
2) A stereo matching algorithm which incorporates RPRF and generalizes to light fields of different configurations.

We model pixel differences by scale mixtures with soft-decision 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. After introducing such RPRF modeling in Section 3, we present the stereo matching algorithm in Section 4. 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 (Fig. 1g with accurate depth edges) but also sparse, denoised, and even grey-scale ones (Figs. 1h-1j). The scene-dependent parameters can also be estimated robustly with fast convergence. Finally, we demonstrate better depth accuracy and faster computation speed than the state-of-the-art algorithms.

- *C.-T. Huang is with the Department of Electrical Engineering, National Tsing Hua University, Taiwan.*
E-mail: chaotsung@ee.nthu.edu.tw

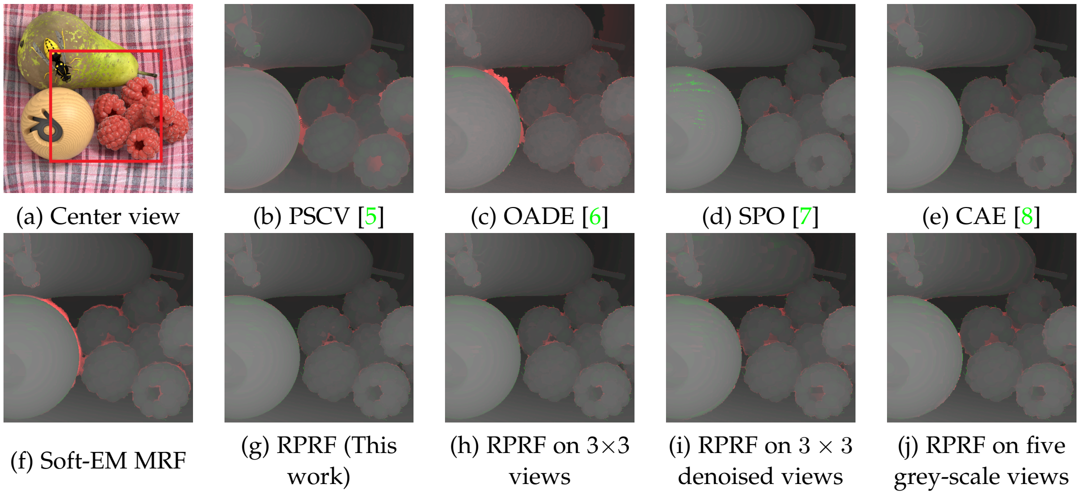


Fig. 1. Depth estimation results from different algorithms and view configurations. Disparity values are displayed in grey-scale intensity. To highlight positive and negative disparity errors, we add corresponding values to red and green channels respectively. (a) A challenging light field *StillLife* (9×9 views) in HCI dataset [10]. (b)-(g) The depth maps produced with the 9×9 views by (b) phase-shift cost volume (PSCV) [5], (c) occlusion-aware depth estimation (OADE) [6], (d) spinning parallelogram operator (SPO) [7], (e) constrained angular entropy (CAE) [8], (f) MRF using conventional soft-EM energy, and (g) the proposed Robust Pseudo Random Field (RPRF) using robust hard-EM energy. (h)-(j) The depth maps produced by RPRF with more difficult configurations: (h) a more sparse 3×3 light field, (i) a distorted 3×3 light field which is first corrupted by Gaussian noise ($\sigma = 10$) and then denoised by BM3D [11], and (j) a five-view crosshair light field with only one grey-scale channel.

This paper extends our previous work [12] with more theoretical and technical discussions on RPRF modeling and stereo matching implementation. In addition, we added experimental results for discussing model properties, including subjective quality, parameter variation, soft-EM energy, and energy function types, and also for generalizing to difficult settings, such as five-view light fields and grey-scale pixels. Also, to our best knowledge, this is the first work to estimate MRF parameters for a single light field.

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 explore such spatial dependency and learn corresponding MRF parameters from training datasets for many different applications [13], [14]. The reader is referred to [1] for further details. In this paper, we estimate parameters from a single light field adaptively.

Single-scene parameter estimation. Previous work for similar purposes focused on stereo image pairs and used a conventional framework: identical likelihood functions for distribution fitting and MRF inference. Zhang *et al.* [15] 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.* [16] 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 [18]. In [19], Huang proposed a neighborhood noise model (NNM) to estimate parameters statistically for bilateral filters and non-local means. 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.

Binocular stereo matching. The data energy collected from only two views are not reliable, so cost aggregation [20] or cost-volume filtering [21] is required to enhance robustness. Also, occlusion issues need to be detected and handled explicitly [22], such as left-right consistency checking. In this paper, we deal with light fields which have at least five crosshair views. Experimental results show that the data energy is sufficiently robust without cost aggregation in this case; in addition, our implicit occlusion handling via the hidden priors is able to provide accurate depth edges without explicit occlusion detection.

Light-field stereo matching. Light fields possess lots of depth information, and previous work has explored many vision cues to retrieve it. Many of them employed specific features for dense light fields. For example, the depth consistency in epipolar line image (EPI) was utilized in [4]. Also, the abundant information of densely angular sampling is useful to handle occlusion. Kim *et al.* [23] implicitly formulated it into reliable data terms with iterative mean shifts, and Chen *et al.* [24] used it to construct bilateral statistics of surface cameras. Wang *et al.* [6] further explicitly built occlusion-aware data terms using angular statistics, and Sheng *et al.* [25] employed the similarity between local and angular patches. In addition, Si *et al.* [26] proposed a pixel-wise plane model for detail refinement. However, these methods may not adapt well to sparse view sampling.

On the other hand, many approaches focused on physical or textural cues. In addition to the conventional pixel-wise photo-consistency, Kolmogorov *et al.* [27] explicitly formulated pixel-wise occlusion between views, and Yu *et al.* [2] further considered geometry structures of 3D line segments in light fields. Also, Tao *et al.* [28] incorporated the depth-dependent defocus blurs into data terms, and Johannsen *et al.* [29] used sparse light-field representation for correspondence matching. In particular, Zhang *et al.* [7] devised a spinning parallelogram operator which measures histogram distances between the two EPI regions separated by each depth label. Although great depth quality can be achieved, these approaches are mainly applicable to clean images and may fail if the texture is noisy or distorted.

Finally, some research works applied noise-resistant cues for noisy light fields. For example, Lin *et al.* [3] utilized color symmetry in simulated focal stacks. Williem *et al.* [8] devised robust data costs using constrained adaptive defocus and constrained angular entropy to provide invariant capability over noise and also occlusion, and this work presents the state-of-the-art quality. However, these methods, as well as most of the previous work, adopted many heuristic parameters, and this will degrade their generalization capability. In this paper, we get back to fundamentals in MRF inference by exploring intrinsic statistical cues for parameter estimation with wide application scope and achieve great depth quality using simple pixel-difference data costs. In addition, Heber *et al.* [30] trained a deep convolutional network on EPI volumes using a large database. In contrast, this work estimates MRF parameters for each single light field.

3 RPRF Modeling

For a given light field, we consider the problem of disparity (inverse depth) estimation 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.

We reason and infer the MRF in (1) by constructing pseudo-likelihoods of pixel differences for 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. Some important properties of these models will also be discussed.

3.1 Pixel Difference Model for Data Energy

Let $\mathbf{X}_{\mathrm{d}}$ be a random vector for the view-wise color difference $\mathbf{x}_{\mathrm{d}} \triangleq \mathbf{z}_p - \mathbf{y}_{vp}(l_p)$. The empirical distribution of its L2-norm is usually heavy-tailed as shown in Fig. 4a, i.e. decays much slower than Gaussian distribution (a downward parabola in log-scale), and we employ a model similar to the NNM for good model fitting. We introduce a scale random variable W to model occlusions using soft decisions and formulate $\mathbf{X}_{\mathrm{d}}$ in a Gaussian scale mixture (GSM):

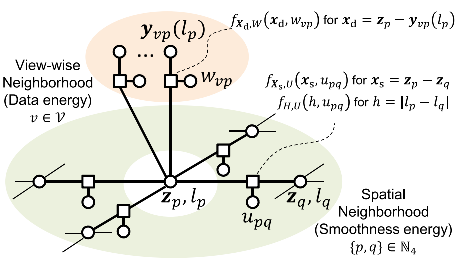

$\boldsymbol{y}_{vp}(l_p)$
…
$f_{X_{\mathrm{d}},W}(\boldsymbol{x}_{\mathrm{d}}, w_{vp})$ for $\boldsymbol{x}_{\mathrm{d}} = \boldsymbol{z}_p - \boldsymbol{y}_{vp}(l_p)$
View-wise Neighborhood (Data energy) $v \in \mathcal{V}$
w_{vp}
$f_{X_{\mathrm{s}},U}(\boldsymbol{x}_{\mathrm{s}}, u_{pq})$ for $\boldsymbol{x}_{\mathrm{s}} = \boldsymbol{z}_p - \boldsymbol{z}_q$
$f_{H,U}(h, u_{pq})$ for $h = |l_p - l_q|$
$\boldsymbol{z}_p, l_p$
$\boldsymbol{z}_q, l_q$
u_{pq}
Spatial Neighborhood (Smoothness energy) $\{p,q\} \in \mathbb{N}_4$

Fig. 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-occlusion hidden variable w_{vp}. For the spatial neighborhood, color difference $\mathbf{x}_{\mathrm{s}}$ and disparity difference h are formulated by two separate scale mixtures with an identical hidden soft-edge u_{pq}.

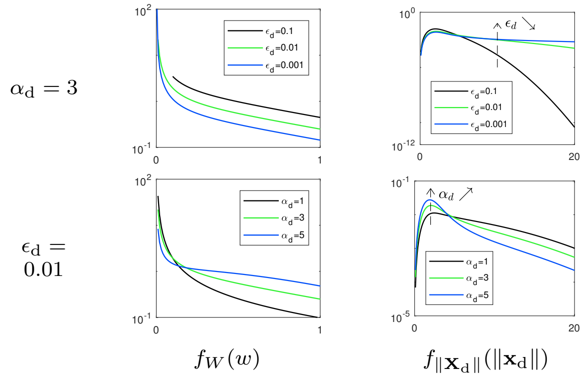


Fig. 3. Distributions of the soft-occlusion variable w and color difference $\mathbf{x}_{\mathrm{d}}$. $G(w)$ is of Reciprocal type and $\sigma_{\mathrm{d}} = 1$.

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

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

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. Note that the σ_{d} here represents only distribution scaling, not for the noise intensity in the NNM.

A smaller w will result in a more flat distribution for $\mathbf{x}_{\mathrm{d}}$ and thus implies occlusion more likely happens; on the other hand, a larger w will lean to non-occlusion. Therefore, its distribution directly affects the one of $\mathbf{x}_{\mathrm{d}}$. These distributions are controlled by the prior function $G(w)$ through parameters ϵ_{d} and α_{d} as shown in Fig. 3. The ϵ_{d} dictates the occlusion probability (on small w) and thus the distribution tail thickness of $\mathbf{x}_{\mathrm{d}}$. In contrast, the α_{d} sets the non-occlusion ratio (on large w) and determines the distribution peak height of $\mathbf{x}_{\mathrm{d}}$. By varying ϵ_{d} and α_{d}, we can fit heavy-tailed $f_{\|\mathbf{x}_{\mathrm{d}}\|}(\|\mathbf{x}_{\mathrm{d}}\|)$ in different shapes.

3.1.1 *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 $t_\mathrm{d} \triangleq \|\mathbf{x}_\mathrm{d}\|_2$ which follows a chi scale mixture (CSM) distribution. 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 $\hat{\boldsymbol{\theta}}_\mathrm{d}$ can be derived by model fitting to empirical observations $\tilde{t}_\mathrm{d} = \|\mathbf{z}_p - \mathbf{y}_{vp}(\tilde{l}_p)\|_2$ through global energy optimization:

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

However, the summation over all $\tilde{t}_\mathrm{d}$ is computation-intensive due to the massive numerical integrals. Instead, we apply a histogram-based expectation as in [31] to accelerate the fitting process:

$$\hat{\boldsymbol{\theta}}_\mathrm{d} = \arg\min_{\boldsymbol{\theta}_\mathrm{d}} \mathbf{E}_{\tilde{t}_\mathrm{d}} \left[E_{T_\mathrm{d}}^{\mathrm{soft}}(\tilde{t}_\mathrm{d}; \boldsymbol{\theta}_\mathrm{d}) \right], \tag{6}$$

which is also equivalent to minimizing the cross entropy $H(\tilde{t}_\mathrm{d}, t_\mathrm{d})$ between the empirical $\tilde{t}_\mathrm{d}$ and its modeled t_d in CSM. Then the parameter set $\boldsymbol{\theta}_\mathrm{d}$ can be estimated by iterative updates for solving (6) until converged.

3.1.2 Robust Data Energy by Hard EM

Define an energy function $E(x)$ as the convex conjugate of the prior $G(w)$:

$$E(x) \triangleq \min_{w} \left(wx - G(w)\right), x \geq 0. \tag{7}$$

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):

$$E_{\mathbf{X}_\mathrm{d}}^{\mathrm{hard}}(\mathbf{x}_\mathrm{d}) = -\log \max_{w} f_{\mathbf{X}_\mathrm{d},W}(\mathbf{x}_\mathrm{d}, w) \tag{8}$$

$$= \alpha_\mathrm{d} E\left(\frac{\|\mathbf{x}_\mathrm{d}\|_2^2}{2\alpha_\mathrm{d}\sigma_\mathrm{d}^2}\right) + C, \tag{9}$$

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

$$E_{pv}^{\mathrm{d}}(l_p) = E_{\mathbf{X}_\mathrm{d}}^{\mathrm{hard}}(\mathbf{z}_p - \mathbf{y}_{vp}(l_p)). \tag{10}$$

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 hidden random variable. To fit the heavy-tailed distribution of H, we propose a scale mixture for f_H using generalized Gaussian distributions for $f_{H|U}$. Along with a GSM for $\mathbf{X}_\mathrm{s}$, we construct the following joint model:

$$f_{H|U}(h|u) = \frac{2}{\gamma(\frac{1}{\beta}+1)\delta} u^{\frac{1}{\beta}} e^{-u(\frac{h}{\delta})^{\beta}}, \tag{11}$$

$$\mathbf{X}_\mathrm{s}|U = u \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma_\mathrm{s}^2}{u}\mathbf{I}_k\right), \tag{12}$$

$$f_U(u) = \frac{1}{N_\mathrm{s}} u^{-(\frac{k}{2}+\frac{1}{\beta})} e^{\alpha_\mathrm{s} G(u)},\ u \in [\epsilon_\mathrm{s}, 1], \tag{13}$$

where δ and β are the scale and shape parameters for H, and the β is fixed to 1.5 in this paper. The σ_s, α_s, ϵ_s and N_s serve the same purposes as the σ_d, α_d, ϵ_d and N_d separately.

The disparity difference H shares the same soft-edge prior U with the color difference $\mathbf{X}_\mathrm{s}$ 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 $f_{\mathbf{X}_\mathrm{s},H,U}$. As a result, the hard-EM smoothness energy can have a simple form similar to the case of the data energy (9).

3.2.1 Parameter Estimation by Soft EM

Consider the parameter set $\boldsymbol{\theta}_\mathrm{s} = (\delta, \sigma_\mathrm{s}, \alpha_\mathrm{s}, \epsilon_\mathrm{s})^T$. We update it using the empirical $\tilde{t}_\mathrm{s}$ for a sufficient statistic $t_\mathrm{s} \triangleq \|\mathbf{x}_\mathrm{s}\|_2$ 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 $f_{T_\mathrm{s},H,U}$, will induce a computation issue: it needs to evaluate the non-analytic energy for every pair of $\tilde{t}_\mathrm{s}$ and $\tilde{h}$. Instead, we derive the updated $\hat{\boldsymbol{\theta}}_\mathrm{s}$ by using their marginal energy to speed up the process:

$$\hat{\boldsymbol{\theta}}_\mathrm{s} = \arg\min_{\boldsymbol{\theta}_\mathrm{s}} \left(\mathbf{E}_{\tilde{t}_\mathrm{s}} \left[E_{T_\mathrm{s}}^{\mathrm{soft}}(\tilde{t}_\mathrm{s}; \boldsymbol{\theta}_\mathrm{s}) \right] + \eta \mathbf{E}_{\tilde{h}} \left[E_{H}^{\mathrm{soft}}(\tilde{h}; \boldsymbol{\theta}_\mathrm{s}) \right] \right), \tag{14}$$

where η controls the cross entropy ratio for model fitting.

3.2.2 Robust Smoothness Energy by Hard EM

Similarly to (8)-(9), we maximize the joint distribution $f_{\mathbf{X}_\mathrm{s},H,U}$ with respect to the hidden U and have the hard-EM energy for pixel differences $\mathbf{x}_\mathrm{s}$ and h as follows:

$$E_{\mathbf{X}_\mathrm{s},H}^{\mathrm{hard}}(\mathbf{x}_\mathrm{s}, h) = \alpha_\mathrm{s} E\left(\frac{\|\mathbf{x}_\mathrm{s}\|_2^2}{2\alpha_\mathrm{s}\sigma_\mathrm{s}^2} + \frac{h^{\beta}}{\alpha_\mathrm{s}\delta^{\beta}}\right). \tag{15}$$

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

$$E_{pq}^{\mathrm{s}}(l_p, l_q) = E_{\mathbf{X}_\mathrm{s},H}^{\mathrm{hard}}(\mathbf{z}_p - \mathbf{z}_q, |l_p - l_q|), \tag{16}$$

which constitutes a conditional random field.

3.3 Properties of Pixel Difference Models

3.3.1 On Core Functions

Given the definition of $E(x)$ in (7), the best guessed hidden $\hat{w} \triangleq \arg\min_w (wx - G(w))$ for an observed x should satisfy

$$x - G'(\hat{w}) = 0 \Rightarrow \hat{w} = K(x), \tag{17}$$

where the kernel function $K(x)$ is defined by

$$K(x) \triangleq \left(G'\right)^{-1}(x). \tag{18}$$

Then it can be shown using integration by parts (in Appendix A) that $K(x)$ is the derivative of $E(x)$. Therefore, the three core functions are directly connected by

$$E' = K = \left(G'\right)^{-1}, \tag{19}$$

which is also a property of convex conjugates. Table 1 shows two examples, Reciprocal and Gaussian, for a robust energy function $E(x)$. Note that in these models $E(x)$ and $G(w)$ can be offset by any constant without affecting their properties, i.e. only the derivatives $E'(x)$ and $G'(w)$ really matter.

Three interesting results of the above formulations can give us useful properties for robust MRF inference and

TABLE 1
Core functions for robust hard-EM energy.

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}	$w(1-\log w)$

model fitting. First, the range of $K(x)$ is equal to that of w, i.e. $K(x) \in (0,1]$. Thus $E(x)$ is an increasing function, which meets our expectation for a proper energy function. Second, $K(x)$ is invertible and thus strictly monotonic. In this paper, we make a further but reasonable assumption: $K(x)$ is strictly decreasing with $K(0) = 1$ and $K(\infty) = 0$. Then we have a robust form for $E(x)$ because its derivative will approach to zero when $x \to \infty$. Finally, as the inverse function of $K(x)$, $G'(w)$ ranges between $[0, \infty)$ and is also strictly decreasing. Therefore, $G(w)$ has the same good properties as that in the NNM [19] for fitting heavy-tailed distributions.

3.3.2 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 Figs. 4a-4b. The Reciprocal type in Table 1 can provide better fitting accuracy than the conventional Gaussian one, especially for spatial difference $\|\mathbf{x}_\mathrm{s}\|_2$. Therefore, we adopt it for the stereo matching algorithm in this paper. Detailed comparisons including depth quality will be given in Section 5.1.1.

In addition, Figs. 4c-4d 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. 1f. 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. 1g shows. More comparisons will be presented in Section 5.1.2.

4 Implementation for Stereo Matching

We designed an empirical Bayesian stereo matching algorithm using the proposed RPRF. In the following, we will introduce implementation details for RPRF parameter estimation and MRF inference, hyper-parameters, and the algorithm itself, respectively.

4.1 Efficient RPRF Implementation

4.1.1 Parameter Estimation for RPRF

Given a disparity map, we estimate the parameters $\boldsymbol{\theta}_\mathrm{d}$ and $\boldsymbol{\theta}_\mathrm{s}$ for data and smoothness energy from the corresponding view-wise and spatial pixel differences. For the $\boldsymbol{\theta}_\mathrm{d}$, we applied the EM+ fitting method designed for NNM in [19] to our soft-EM update formulation (6). Small changes were made for the model difference. For the $\boldsymbol{\theta}_\mathrm{s}$ updated by (14), we modified the fitting method to include the disparity difference with its additional parameter δ and statistics $\tilde{h}$. Further details are given in Appendix B. In addition, for accelerating the fitting processing, we applied the equal-frequency merging technique in [31] with 20 bins for the pixel differences.

4.1.2 Belief Propagation for MRF Inference

After deciding the parameters, we use belief propagation (BP) [32] to implement MRF inference on (1) iteratively for solving the disparity map. We adopted the BP-M approach in [33] 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.

4.1.3 Energy Approximation for MRF Inference

To further accelerate message propagation, we use the linear-time algorithm in [32]. To achieve this, we approximated (15) for the Reciprocal smoothness energy using a truncated linear function that has the least squared error:

$$E^{\mathrm{hard}}_{\mathbf{X}_\mathrm{s},H} \simeq \alpha_\mathrm{s} \min\left(\frac{0.3726}{\delta\alpha_\mathrm{s}^{\frac{1}{\beta}} b^{\frac{1}{\beta}+1}} h, \frac{1}{b}\right) + \mathrm{const}, \qquad (20)$$

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

4.2 Hyper-Parameter Setting

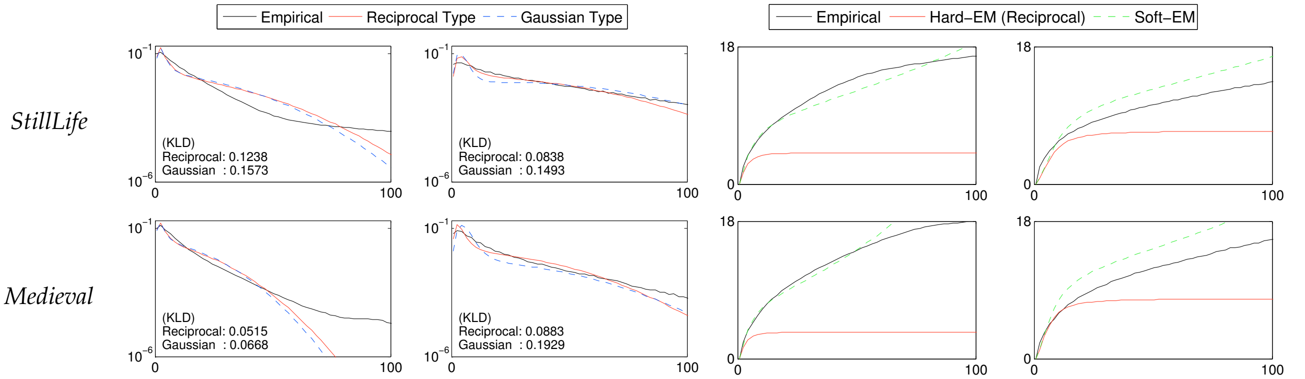
There are two hyper-parameters which cannot be explained and estimated by RPRF. One is the energy ratio λ for global MRF inference in (1). The other one is the cross entropy ratio η for parameter estimation of smoothness energy in (14). In the following, we will introduce our heuristics for their settings.

4.2.1 On Adaptive Selection of Energy Ratio λ

Configuration adaptability. Although the parameters α_d and α_s can statistically determine the dynamic ranges of the data and smoothness energy, we further use λ to weight importance between them for different configurations. For example, denoised light fields rely on smoothness energy more than clean 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.

Scene adaptability. We also found that λ should adapt to different scenes. For example, there are two typical performance trends for λ as shown in Fig. 5: one needs a small λ, as *StillLife*, 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 entropy reduction ratio to adaptively assign λ^{weak} and $\lambda^{\mathrm{strong}}$. Table 2 summarizes our adaptive selection schemes of λ for true-color cases. In greyscale cases, the value is further halved to accommodate their stronger smoothness energy.

Entropy threshold. Finally, we need to decide the entropy threshold. As shown in Fig. 5, the weak scenario, e.g. *StilLife*, will have large penalties on depth error if assigned a strong λ. On the contrary, the strong scenario is more tolerant to a weak λ. Therefore, we prefer a higher entropy threshold to avoid misjudgement for weak scenarios. In this



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

Fig. 4. 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 $\|\mathbf{x}_\text{d}\|_2$ and (b) spatial difference $\|\mathbf{x}_\text{s}\|_2$ with their Kullback-Leibler divergence (KLD) values shown at the corners. (c)-(d) Corresponding energy functions of the Reciprocal type for (c) data energy and (d) smoothness energy. Note that the energy values are adjusted with constant offsets such that they are all aligned at the origin for comparison.

TABLE 2
Adaptive selection of λ (true-color, k=3).

Configuration	λ^weak	λ^strong
Clean	$\max(\lvert\mathcal{V}\rvert/8, 2)$	$\max(3\lvert\mathcal{V}\rvert/2, 12)$
Denoised	$\max(\lvert\mathcal{V}\rvert/4, 3)$	$\max(3\lvert\mathcal{V}\rvert/2, 24)$

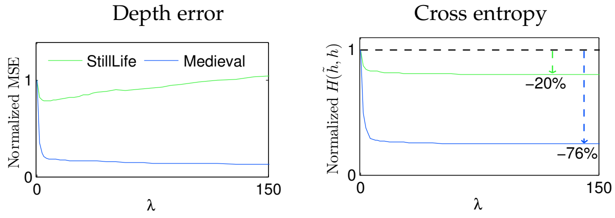
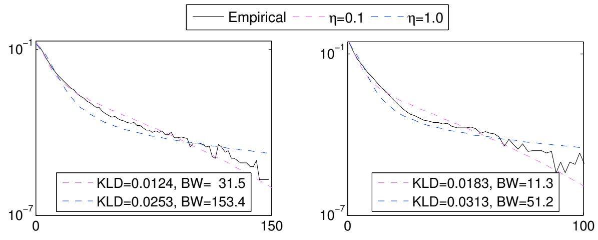


Fig. 5. Depth estimation quality vs. λ. Depth error is represented by mean squared error (MSE) in disparity. Cross entropy measures the distance between the distributions of the empirical disparity difference $\tilde{h}$ and the modeled h. Their values are both normalized with respect to the case of $\lambda = 0$ for comparison.

paper, this threshold r is set to 50% for true-color cases (k=3) and 75% for grey-scale ones (k=1).

4.2.2 On Selection of Cross Entropy Ratio η

When estimating parameters for smoothness energy, the η weighs the importance of the cross entropy $H(\tilde{h}, h)$ over that of $H(\tilde{t}_\text{s}, t_\text{s})$. In true-color cases, we simply set $\eta = 1.0$ for balanced weighting, and the fitting results are reliable. Note that the three-degree CSM t_s has a distribution peak, as shown in Fig. 4b, which helps to discriminate between distributions. In contrast, the one-degree CSM t_s in grey-scale cases has no such peak, and the cross entropy of t_s becomes less discriminative. Fig. 6 shows such examples to demonstrate that we need to weigh more on $H(\tilde{t}_\text{s}, t_\text{s})$, i.e. use a smaller η, for accurate model fitting. For example, even if the KLD is as small as 0.0253 for *Buddha2* ($\eta = 1.0$), its estimated parameters can be significantly different to those for KLD $= 0.0124$ ($\eta = 0.1$). As a result, we set $\eta = 0.1$ for fitting robustness in grey-scale cases.



Buddha2 *Mona*

Fig. 6. Model fitting in grey-scale cases ($k = 1$) over η. Fitting results of spatial difference $t_\text{s} = \|\mathbf{x}_\text{s}\|_2$ are shown for two light fields, *Buddha2* and *Mona*, along with their KLD values and the corresponding bandwidth (BW) parameters $\alpha_s\sigma_s^2$. KLD $\triangleq H(\tilde{t}_\text{s}, t_\text{s}) - H(\tilde{t}_\text{s})$.

4.3 Stereo Matching Algorithm

Based on the above mentioned details, we devise an empirical Bayesian algorithm as summarized in Algorithm 1. At first, we initialize a disparity map $\mathcal{D}^\text{ini}$ using MRF inference with default parameters: $\lambda^\text{ini} = 300$, $\boldsymbol{\theta}_\text{d}^\text{ini} = (\sqrt{2/3}, 6, 0.1)$, and $\boldsymbol{\theta}_\text{s}^\text{ini} = (0.05, \sqrt{8/3}, 9, 0.1)$. Then we update parameters and estimate depth iteratively. In each iteration, we use λ^strong to infer the disparity map first. If the cross entropy $H(\tilde{h}, h)$ is reduced by less than the threshold ratio r compared to the case of $\lambda = 0$, we will switch to the weak scenario and use λ^weak for inference instead. Also, we found depth convergence is fast and, therefore, set the default iteration number M to one.

5 Experiments

We perform extensive experiments on four datasets for objective evaluation: *HCI Blender* [10], *Berkeley* [6], and *HCI-UK Training* [34] are synthetic; *HCI Gantry* [10] contains real pictures. They are all dense light fields of 9×9 views, and we subsample their viewpoints to produce sparse 5×5, 3×3, and five-view crosshair test cases. We also generate denoised light fields by adding Gaussian noise and then performing BM3D [11]. For grey-scale experiments, we average RGB channels to obtain single-channel pixel intensity

Algorithm 1 Empirical Bayesian Stereo Matching

```
Input: Center view $\{\mathbf{z}_p\}$, surrounding views $\{\mathbf{y}_{vp}(\cdot)\}$
Output: Disparity map $\tilde{\mathcal{D}}$
  Initialization: $\boldsymbol{\theta}_{\rm d} = \boldsymbol{\theta}_{\rm d}^{\rm ini}$, $\boldsymbol{\theta}_{\rm s} = \boldsymbol{\theta}_{\rm s}^{\rm ini}$
  Data energy: get $\{E_{pv}^{\rm d}(l_p; \boldsymbol{\theta}_{\rm d})\}$
  MRF inference: $\tilde{\mathcal{D}} = \mathcal{D}^{\rm ini} \leftarrow$ BP-M$(\lambda^{\rm ini}, \boldsymbol{\theta}_{\rm s})$
  for $m = 1$ to $M$ do    $\triangleright$ Iterative parameter-depth update
    Parameter estimation: update $(\boldsymbol{\theta}_{\rm d}, \boldsymbol{\theta}_{\rm s})$ from $\tilde{\mathcal{D}}$
    Data energy: get $\{E_{pv}^{\rm d}(l_p; \boldsymbol{\theta}_{\rm d})\}$
    MRF inference: $\tilde{\mathcal{D}} \leftarrow$ BP-M$(\lambda^{\rm strong}, \boldsymbol{\theta}_{\rm s})$
    if $\triangle H(\tilde{h}, h) < r$ then
      MRF inference: $\tilde{\mathcal{D}} \leftarrow$ BP-M$(\lambda^{\rm weak}, \boldsymbol{\theta}_{\rm s})$
    end if
  end for
```

TABLE 3
Fitting accuracy and depth error of different energy types. Average numbers for *HCI Blender* dataset are reported.

Condition	View Type	KLD ($t_{\rm d}$)		KLD ($t_{\rm s}$)		DMSE	
		Reciprocal	Gaussian	Reciprocal	Gaussian	Reciprocal	Gaussian
Clean	9x9	**0.077**	0.106	**0.129**	0.206	**0.65**	0.69
	5x5	**0.080**	0.109	**0.129**	0.206	**0.65**	0.79
	3x3	**0.083**	0.120	**0.129**	0.206	**0.62**	1.65
Denoised (σ=10)	9x9	**0.049**	0.072	**0.157**	0.218	**0.99**	1.20
	5x5	**0.050**	0.074	**0.156**	0.218	**1.02**	2.17
	3x3	**0.054**	0.079	**0.157**	0.217	**1.17**	5.20
Denoised (σ=20)	9x9	**0.040**	0.055	**0.136**	0.201	**1.65**	2.03
	5x5	**0.043**	0.057	**0.138**	0.207	**1.72**	3.35
	3x3	**0.043**	0.061	**0.141**	0.205	**2.07**	7.25

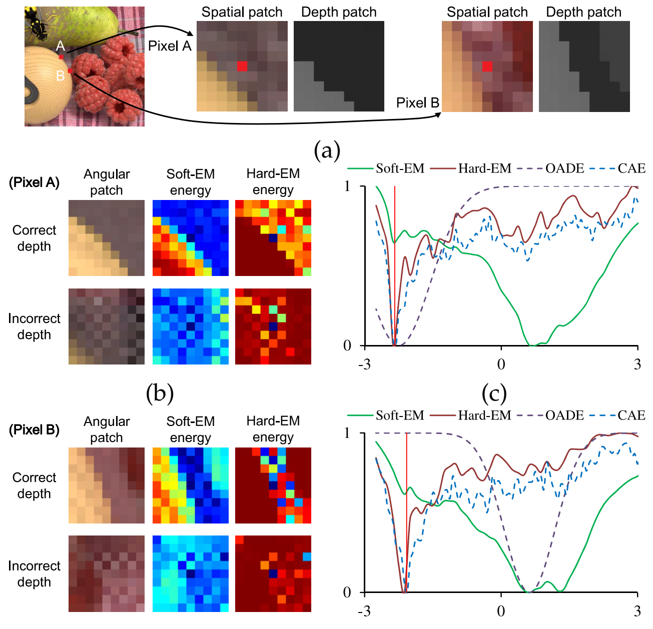
($k = 1$). For comparing objective quality across view types, we calculate mean squared errors (MSE) in disparity all with respect to the baselines of 9×9 light fields. The values are then multiplied by 100 to keep significant figures and denoted by DMSE.

We will also show results for real-scene light fields which are captured by Lytro ILLUM and provided by [23] for demonstrating generalization capability. The Lytro light-field RAW data from *EPFL* dataset [35] and our own pictures are processed using Lytro Power Tools Beta[1]. Our software is available online[2].

5.1 Energy Function

5.1.1 Reciprocal vs. Gaussian Energy

For comparing the Reciprocal and Gaussian types of $E(x)$, we use ground-truth disparity maps (ideal case) to estimate parameters for each of them and then perform MRF inference accordingly. Table 3 demonstrates the relationship between fitting accuracy in Kullback-Leibler divergence (KLD) and depth error in DMSE, i.e. smaller KLD results in smaller DMSE. Since the Reciprocal $E(x)$ shows clear advantages, we adopt it in the following experiments. In addition, the smoothness energy, which depends on the statistic $t_{\rm s}$, is more sensitive to different types of $E(x)$ than the data energy. Therefore, the DMSE gap becomes larger for sparser view sampling because the smoothness term becomes more important in the global MRF energy.



(a)

(b) (c)

(d) (e)

Fig. 7. Hard-EM vs. Soft-EM data energy. (a) Two selected pixels A and B from the tablecloth of *StillLife*, and their spatial and depth patches. They are both occluded by the wooden ball in left views, and pixel B is further occluded by the berry in right views. (b) Angular patches (9×9) and view-wise data energy for pixel A. The energy is normalized and displayed using MATLAB jet map, i.e. warmer colors represent higher energy. For the hard-EM case, the soft-occlusion variable $\hat{w}$ is exactly the square of negative energy according to (17) and Table 1, i.e. warmer colors represent smaller $\hat{w}$ (more likely occlusion). The incorrect depth is where soft-EM data energy has its minimum value. (c) Normalized data energy vs. disparity for pixel A. The correct depth is indicated by the vertical red line. (d) and (e), similarly to (b) and (c), for pixel B.

5.1.2 Hard-EM vs. Soft-EM Energy

Fig. 7 demonstrates the robustness of the hard-EM data energy for occlusion handling. The soft-EM data energy is prone to be biased by occluders, e.g. wooden ball, and thus not robust to handle occlusion. In contrast, the hard-EM one will be saturated for them and estimate depth based on small pixel differences. Note that the saturated energy results from small soft-occlusion $\hat{w}$, which therefore verifies the effectiveness of the proposed soft prior. We also compare with the advanced data costs of occlusion-aware depth estimation (OADE) [6] and constrained angular entropy (CAE) [8]. The data energy curves in Figs. 7c and 7e show that the hard-EM data energy is comparable to CAE and outperforms OADE in these challenging cases.

Similarly, the robustness of the hard-EM smoothness energy is exemplified by Fig. 8. High hard-EM energy, or equivalently small soft-edge $\hat{u}$, is mainly assigned near intensity edges (object boundaries in this example), and the rest regions have low energy inside to have smooth depth. In contrast, the soft-EM energy often misassigns low energy to intensity edges. This causes foreground depth to spread into background pixels and thus moves the depth edges (high energy) wrongly inside background regions.

In Table 4, we compare depth errors for using hard-EM and soft-EM energy, and Algorithm 1 is applied in each case. The DMSE gap is more obvious for sparser light fields, so we list the results of 3×3 test cases. Hard-

[1] https://www.lytro.com/imaging/power-tools
[2] http://www.ee.nthu.edu.tw/chaotsung/rprf

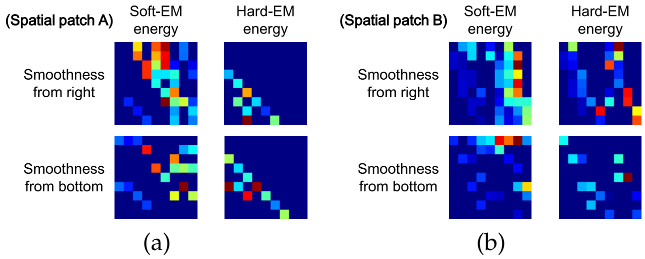


Fig. 8. Hard-EM vs. Soft-EM smoothness energy. The edge-wise energy from right and bottom pixels in the final-round BP is shown for the spatial patches of (a) pixel A and (b) pixel B in Fig. 7.

TABLE 4
Depth errors in DMSE for hard-EM and soft-EM energy.

Dataset	Light Field	Clean		Denoised (σ=10)		Denoised (σ=20)	
		Hard-EM	Soft-EM	Hard-EM	Soft-EM	Hard-EM	Soft-EM
HCI Blender (3x3)	*Buddha*	**0.32**	0.48	**0.53**	2.10	**0.86**	2.40
	Horses	0.57	**0.56**	0.93	**0.78**	1.38	**1.16**
	Papillon	**0.61**	3.34	**2.81**	5.27	**5.20**	7.87
	StillLife	**1.22**	3.82	**1.81**	4.46	**3.42**	8.01
	Buddha2	**0.54**	0.58	**0.31**	0.95	**0.54**	3.85
	Medieval	**0.76**	1.74	**0.96**	1.93	**2.21**	2.99
	Mona	**0.42**	0.48	**0.62**	2.29	**0.85**	3.03
	Average	**0.63**	1.57	**1.14**	2.54	**2.07**	4.19

EM energy outperforms soft-EM one in both clean and denoised conditions. Fig. 9 shows subjective comparisons. Through implicit and statistical occlusion handling, hard-EM energy can preserve better object boundaries and resist more denoising artifacts. This also confirms the robustness brought by hard-EM energy for MRF inference.

5.2 Parameter Estimation

5.2.1 *Robustness to Initial Conditions*

Different initial parameters lead to different initialized disparity maps $\mathcal{D}^{\text{ini}}$. For example, a large value of λ^{ini} prefers the smoothness energy and gives an over-smooth $\mathcal{D}^{\text{ini}}$. Note that an ideal $\mathcal{D}^{\text{ini}}$ is the ground-truth depth itself. However, in this work different $\mathcal{D}^{\text{ini}}$ can generate similar MRF parameters as shown in Fig. 10. Such robustness can be explained by the corresponding empirical distributions in Fig. 11. The empirical statistics 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 in tails mainly causes the variation of ϵ_{d} and ϵ_{s}, but these two parameters do not affect hard-EM energy. As a result, robust parameter estimation is achieved by including these two shape parameters for distribution adaptation, and robust depth inference is accomplished by removing them for energy saturation.

5.2.2 *Quick Convergence*

The robustness to initialized disparity maps also results in the quick convergence of parameter-depth update iterations. Fig. 12 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 improves the DMSE by only 0.2% on average. Therefore, we set the default iteration number M to one.

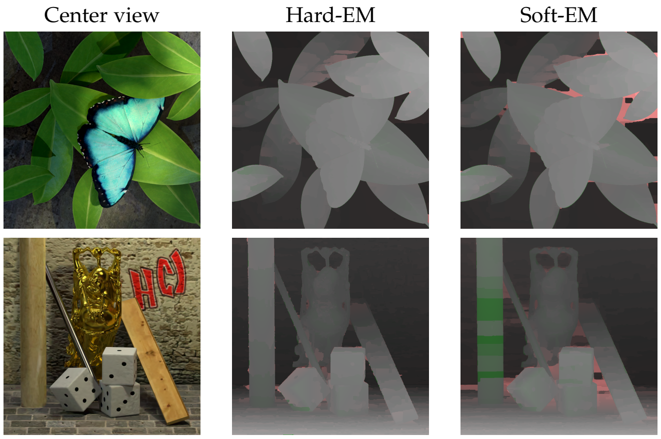


Fig. 9. Depth estimation using hard-EM and soft-EM energy. Top: Clean *Papillon* 3×3. Bottom: Denoised *Buddha2* 3×3 ($\sigma = 20$). The reader is encouraged to zoom in the paper for comparing details.

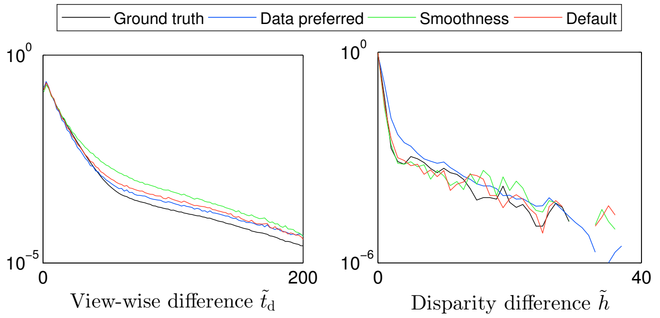


Fig. 11. Empirical distributions. They are collected using the initialized disparity maps in Fig. 10. Note that the distributions of $\tilde{t}_{\text{s}}$ are not related to disparity maps and are all the same.

5.2.3 *Parameter Variation*

Consider the two essential bandwidth parameters for MRF inference: $\alpha_d\sigma_d^2$ (data) and $\alpha_s\sigma_s^2$ (smoothness). Their values vary a lot across different light fields as shown in Fig. 13. For example, their value ranges are [2.2, 8.2] and [10.0, 63.5], respectively, for the clean *HCI Blender* dataset. Note that for sparser light fields each scene has similar parameters. For denoised light fields (σ=10), the data bandwidth becomes 4.7x larger on average and the smoothness one is 0.4x smaller. This reflects the fact that pixel differences across views become larger and those between denoised pixel edges turn smaller. All these variations are well captured by this work.

5.3 Depth Estimation

We compare our results (RPRF) with the globally consistent depth labeling (GCDL) [4], line-assisted graph cut (LAGC) [2], phase-shift cost volume (PSCV) [5], OADE, spinning parallelogram operator (SPO) [7], and CAE. We use the codes provided by the authors. We set the disparity step between labels to dyadic rationals, i.e. 2^{-n}, and around 64 disparity labels are searched for each light field. We apply the same settings to GCDL, LAGC, and PSCV, but double the numbers of labels for OADE and CAE to achieve comparable accuracy.

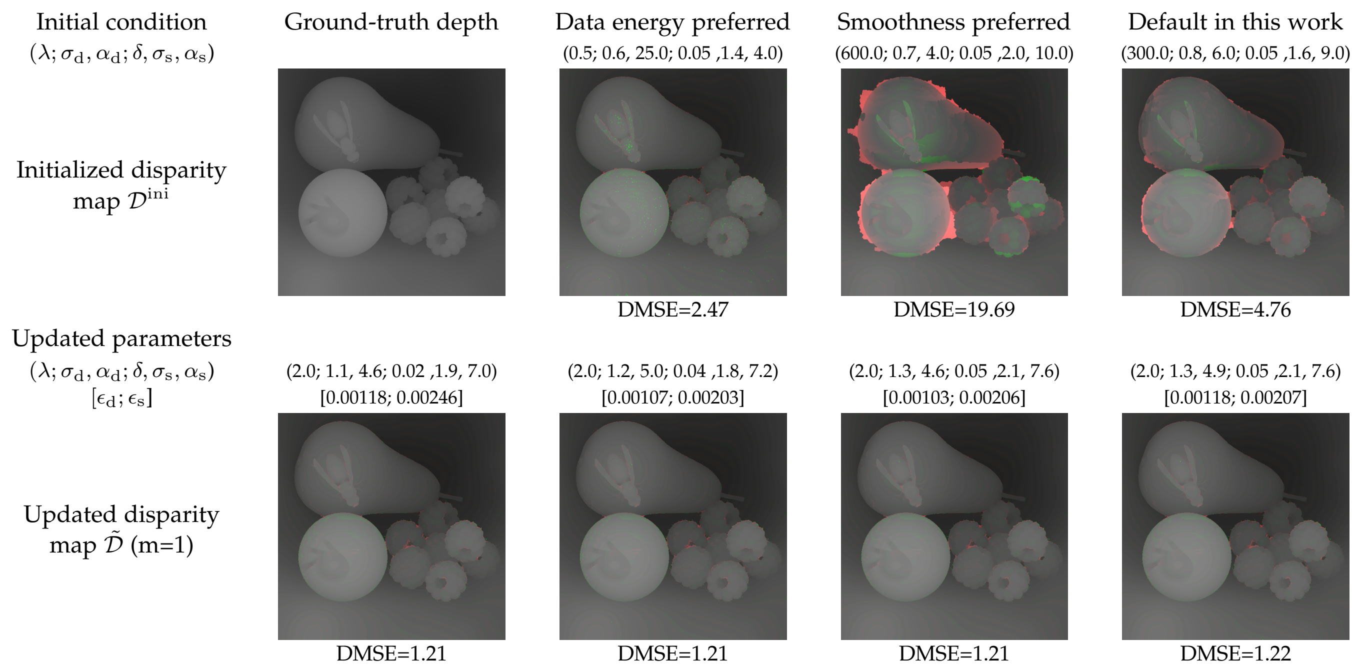


Fig. 10. Robust update for parameters and depth. Four initial disparity maps $\mathcal{D}^\mathrm{ini}$ for the 3×3 case of *StillLife* are used for comparison. One uses ground-truth depth (ideal). The others are initialized by different parameter sets: one is noisy by weighting data energy more (λ=0.5), 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.

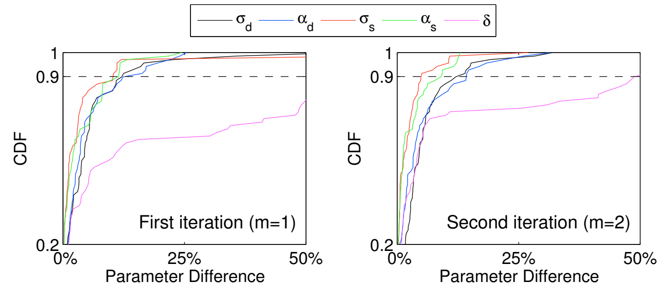


Fig. 12. Parameter estimation accuracy. Accuracy is measured by relative absolute difference, and cumulative distribution functions (CDFs) are derived from all test cases for *HCI Blender* dataset.

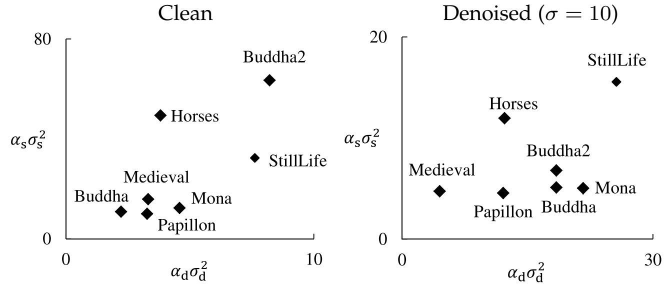


Fig. 13. Variation of bandwidth parameters for *HCI Blender* 9×9 dataset. Left: clean light fields. Right: denoised ones.

TABLE 5
Depth errors in DMSE for dense 9×9 light fields.

Dataset	Light Field	GCDL	LAGC	PSCV	OADE	SPO	CAE	RPRF
HCI Blender	*Buddha*	0.68	3.04	1.13	0.91	0.54	0.64	**0.28**
	Horses	4.98	2.73	1.70	1.36	1.37	0.79	**0.50**
	Papillon	2.68	19.51	5.98	1.00	0.66	**0.63**	0.66
	StillLife	4.01	2.28	2.10	4.29	1.51	1.24	**1.09**
	Buddha2	1.02	2.71	0.45	1.18	1.02	**0.35**	0.75
	Medieval	1.24	4.49	1.40	1.15	0.91	0.97	**0.79**
	Mona	0.93	1.70	0.66	0.73	0.55	0.50	**0.47**
	Average	2.22	5.21	1.92	1.52	0.94	0.73	**0.65**
HCI Gantry	*Couple*	**0.39**	0.62	0.64	1.00	0.56	0.60	0.50
	Cube	0.73	1.27	1.88	1.02	0.91	1.26	**0.67**
	Maria	0.15	0.47	0.23	0.16	**0.13**	0.19	0.18
	Pyramide	0.45	1.38	0.69	0.61	0.51	0.58	**0.44**
	Statue	1.75	2.47	0.35	0.55	**0.23**	0.45	0.78
	Average	0.69	1.24	0.76	0.67	**0.47**	0.61	0.51
Berkeley	*Bedroom*	0.52	1.70	0.48	0.59	0.41	0.39	**0.38**
	LivingRoom	2.12	10.46	2.62	1.91	**1.71**	1.80	1.77
	Outdoor	0.60	2.62	0.93	**0.34**	0.52	0.45	0.38
	Plant	5.00	3.43	5.25	4.10	4.21	2.57	**2.50**
	Average	2.06	4.55	2.32	1.73	1.71	1.30	**1.26**
HCI-UK Training	*Boxes*	11.02	20.35	14.26	9.55	9.37	10.09	**8.68**
	Cotton	3.74	75.31	9.99	1.34	1.35	1.29	**0.81**
	Dino	1.36	3.63	1.36	1.09	**0.34**	0.52	0.50
	Sideboard	2.71	3.11	2.65	2.36	**1.01**	1.31	1.35
	Average	4.71	25.60	7.06	3.58	3.02	3.30	**2.83**

5.3.1 *Dense and Sparse Light Fields*

Table 5 details the depth estimation errors for dense 9×9 test cases, and our work has better performance. Fig. 14 further compares the results from dense to sparse light fields. The GCDL and OADE fail in the sparse cases since they are designed for dense ones; on the contrary, the LAGC is better for sparse light fields because the line textures outweigh additional views. The SPO and CAE have quality drops for some sparse cases. In contrast, our work, as well as PSCV, has constant quality over these view types. This also points out that sparse views can generate great depth maps as dense ones do if robust energy from clean images is used. Figs. 15 and 16 show examples for subjective evaluation.

5.3.2 *Denoised Light Fields*

For noisy light fields, RPRF can produce comparable depth quality compared to CAE which is explicitly devised as a

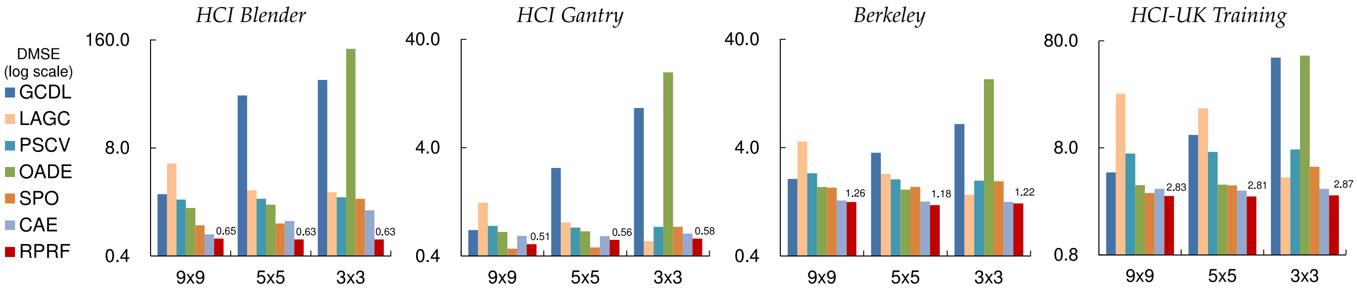


Fig. 14. Depth estimation error vs. Light-field view type. The average errors are represented in DMSE (log scale).

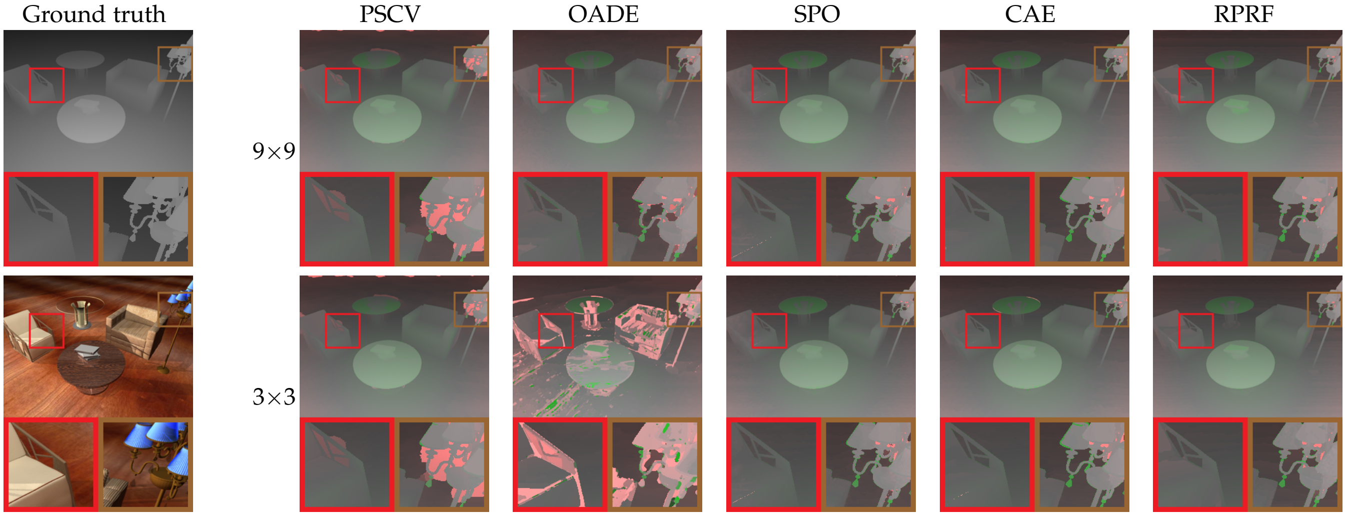


Fig. 15. Estimated depth maps for *LivingRoom*. This work, as well as the state-of-the-art SPO and CAE, infers good depth edges in both 9×9 and 3×3 test cases, especially around the chair arm and the lamps. OADE fails in the 3×3 case, and PSCV produces over-smooth depth.

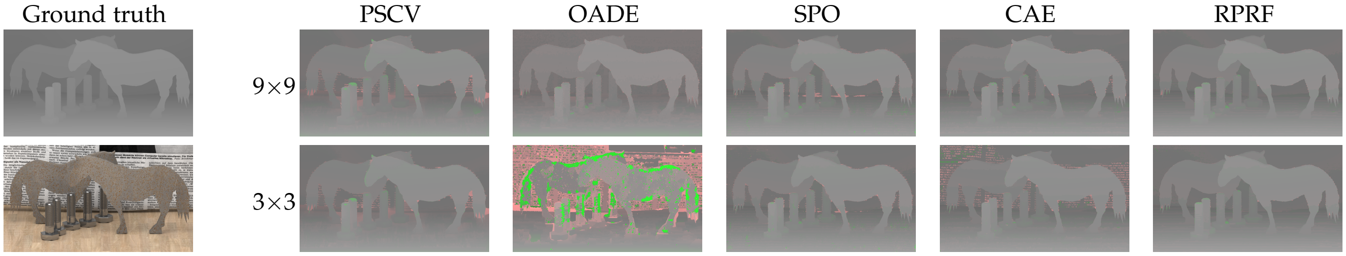


Fig. 16. Estimated depth maps for *Horses*. In particular, CAE fails for the high-frequency texts in the 3×3 case due to insufficient view sampling.

noise-aware data cost. In these cases, RPRF can still capture the statistics successfully; in particular, the σ_d and σ_s resemble the noise intensity well. Furthermore, applying RPRF to denoised light fields can produce much better depth maps as shown in Table 6. The results regarding denoising conditions are summarized in Fig. 17. In these cases, fewer views will decrease depth quality owing to the less reliable data energy. However, as shown in Fig. 18, the depth edges and gradients can still be preserved by our work.

5.3.3 Real Scenes

Fig. 19 shows results for light fields captured by Lytro ILLUM. Using only 3×3 views, our work constantly produces similarly good depth maps compared to Lytro software which uses raw light fields. Other algorithms also perform well but occasionally cause obvious artifacts. We also show the results for the dense 3-D light fields of [23] in Fig. 20. These results also demonstrate that our stereo matching algorithm has good generalization capability to real scenes.

5.3.4 Execution Time

We implement parameter estimation in MATLAB and the other parts in C++. For *HCI Blender* dataset, the parameter estimation and BP-M take about 7 and 10 seconds respectively for one light field on average. The remaining computation is mostly contributed by computing data energy, and its complexity is proportional to the number of surrounding views $|\mathcal{V}|$. The run times for *HCI Blender* dataset are summarized in Table 7. Our work runs much faster for its simple but efficient MRF formulation.

5.3.5 Crosshair View Sampling and Grey-Scale Pixels

Regarding efficient implementation for stereo matching, we may consider a lightweight camera-array system with only

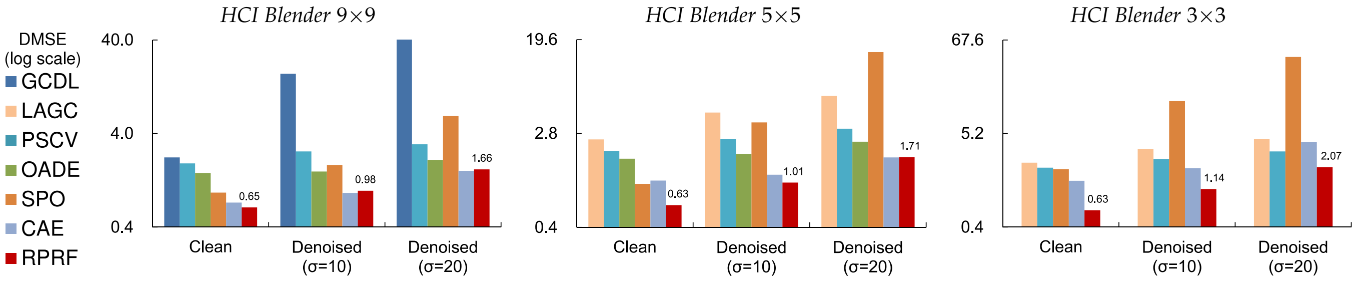


Fig. 17. Depth estimation error vs. Denoising conditions. The cases with too large DMSE are omitted for clarity.

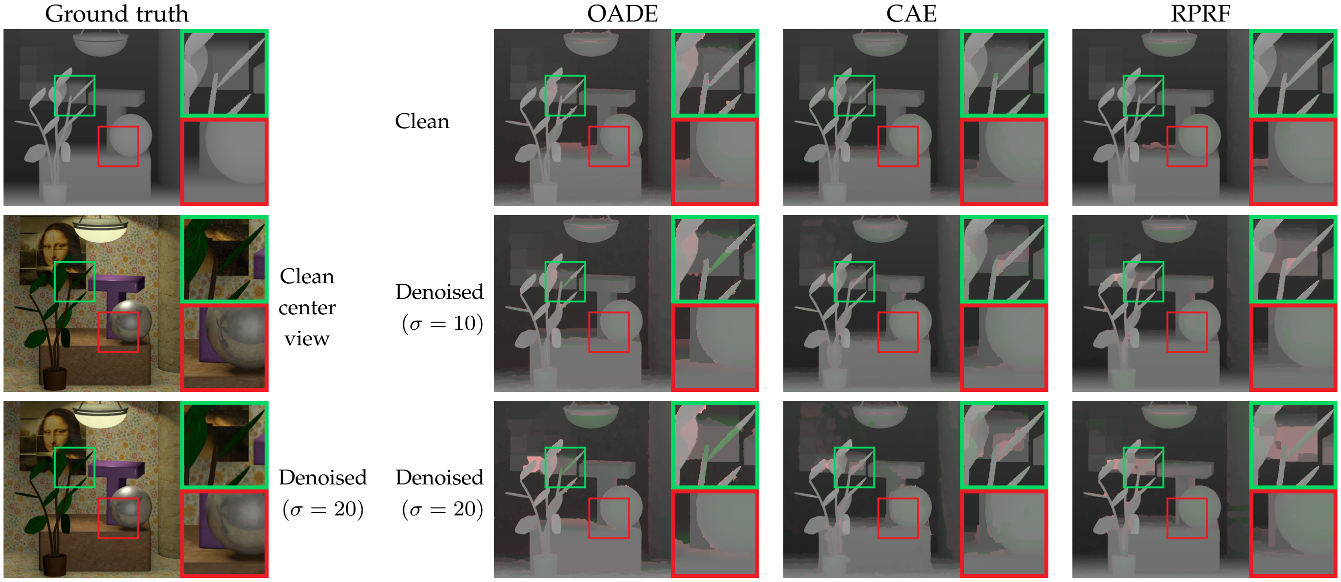


Fig. 18. Estimated depth maps for *Mona* with 5×5 views. This work preserves good depth edges for the case $\sigma = 10$ as well as CAE. When the noise intensity is increased to 20, this work still keeps depth gradients, e.g. near the ball and table, while others produce over-smooth depth.

TABLE 6
Depth errors in DMSE for noisy and denoised light fields.

Light Field	Noise std.	CAE (9x9)	RPRF (9x9)	BM3D+ RPRF (9x9)	BM3D+ RPRF (3x3)
Buddha	σ=10	**1.59**	1.87	**0.44**	0.53
Buddha	σ=20	4.74	**3.63**	**0.71**	0.86
Mona	σ=10	**2.03**	2.10	**0.53**	0.62
Mona	σ=20	4.07	**3.48**	0.86	**0.85**
StillLife	σ=10	**2.43**	3.17	**1.55**	1.81
StillLife	σ=20	5.80	**4.32**	**2.21**	3.42

TABLE 7
Average run time in seconds per light field. GCDL ran on GeForce GT 630 hosted by a 3.5 GHz CPU. Others ran on a 3.4 GHz CPU.

LF type	GCDL	LAGC	PSCV	OADE	SPO	CAE	RPRF
9x9	3,516	334,816	1,951	672	1,773	990	**87**
5x5	3,518	38,233	1,040	357	1,200	559	**33**
3x3	3,562	7,379	778	331	912	437	**24**

TABLE 8
Depth errors in DMSE for special settings. The left part is for crosshair views and grey-scale pixels, and five-view crosshair view sampling is denoted by ”5+”; the right part is for only using data energy.

Setting	Clean				Denoised, σ=20				Data term only		
Algorithm	RPRF				RPRF				OADE	CAE	RPRF
Light Field	True-color		Grey-scale		True-color		Grey-scale		True-color		
	3x3	5+	3x3	5+	3x3	5+	3x3	5+	9x9	9x9	9x9
Buddha	**0.32**	0.46	0.38	0.44	**0.86**	1.08	1.23	1.81	0.97	**0.45**	0.54
Horses	**0.57**	0.64	0.65	0.77	1.38	1.60	**1.29**	1.72	9.39	**3.20**	5.98
Papillon	**0.61**	0.67	0.90	1.00	**5.20**	5.92	5.54	6.08	2.61	**1.60**	2.67
StillLife	**1.22**	1.26	1.51	1.56	3.42	4.59	**3.22**	10.88	6.24	**1.75**	1.98
Buddha2	**0.54**	0.68	0.59	0.77	**0.54**	0.69	0.66	1.22	1.88	**1.06**	2.75
Medieval	**0.76**	0.84	1.07	1.30	2.21	**1.82**	2.26	2.29	8.41	**4.83**	14.17
Mona	**0.42**	0.57	0.43	0.76	**0.85**	0.95	1.28	1.45	0.83	**0.72**	0.92
Average	**0.63**	0.73	0.79	0.94	**2.07**	2.38	2.21	3.64	4.33	**1.95**	4.14

five crosshair views. Also, we may use grey-scale intensity, instead of three-channel RGB, for faster computation and smaller memory footprint. Therefore, we apply our algorithm to these particular cases to evaluate quality trade-offs and also generalization capability. The results are summarized on the left of Table 8. The case of true-color 3×3 light fields is used as the baseline for comparison. It is interesting to find that the quality drops of true-color five crosshair views and grey-scale 3×3 light fields are similarly small. These two cases have comparable depth quality to the baseline as shown in Fig. 21. This demonstrates that this work can provide lightweight implementation for light-field stereo matching while maintaining high-accuracy depth. However, if five crosshair views and grey-scale pixels are applied together, the quality sometimes drops significantly; therefore, it has a quality-complexity trade-off in this case.

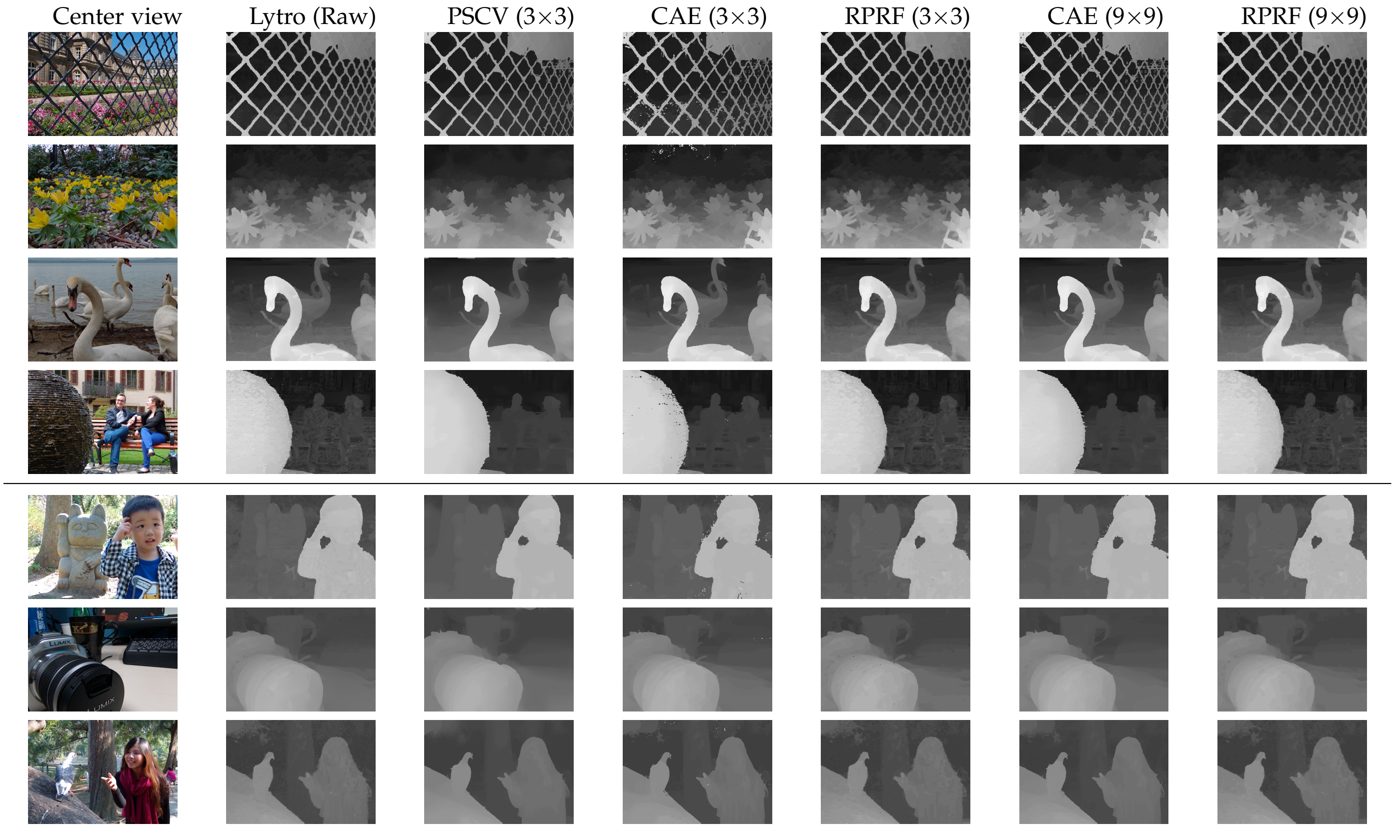


Fig. 19. Estimated depth maps for Lytro pictures. The four light fields on the top are from *EPFL* dataset [35], and the last three captured by us.

5.3.6 Data Energy Only

To compare the simple but statistically-optimized data energy in this work with advanced ones, we also perform depth estimation using only data terms. The results are summarized on the right of Table 8, and CAE gives the best quality. Note that our data energy provides similarly good depth edges near object boundaries. The quality difference mainly comes from textureless regions as shown in Fig. 22. CAE is more reliable in such areas while our data energy and OADE are noise-prone there. In conclusion, MRF optimization is required in our framework.

5.3.7 HCI-UK Test and Stratified Datasets

In [34], the authors also provide a dense synthetic *HCI-UK Test* dataset without giving ground-truth depth maps. The depth errors are only available via their 4D Light Field Benchmark[3] website. For comparison, we list the depth errors from the website in Table 9. For this dataset, CAE outperforms others for its superior data terms. In contrast, RPRF generates slightly over-smooth depth maps; in particular, for the light field *Herbs* its MRF energy cannot discriminate the leaves well due to their similar colors.

In addition, [34] also devises a non-photorealistic *Stratified* dataset with artificial noises for stress testing, and the results are also listed in Table 9. RPRF fails for the light field *Dots* because its spatially-varying and high-intensity noises are far beyond our assumption for the spatial neighborhood. In summary, these two datasets prefer delicate data terms, and RPRF sometimes cannot serve well because it highly relies on the accuracy of MRF modeling.

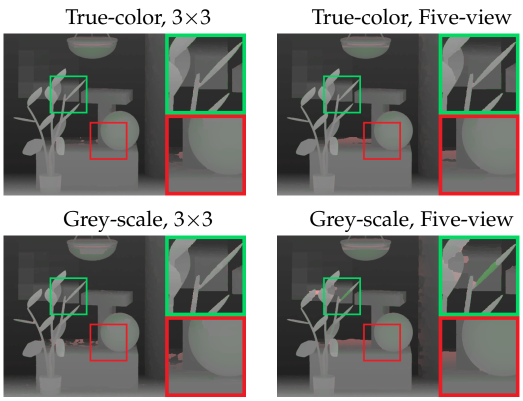
TABLE 9
Depth errors[3] in DMSE for *HCI-UK Test* and *Stratified* datasets (9×9).

HCI-UK Test					*HCI-UK Stratified*				
Light Field	OADE	SPO	CAE	RPRF	Light Field	OADE	SPO	CAE	RPRF
Bedroom	0.53	**0.21**	0.23	0.27	*Backgammon*	22.78	**4.59**	6.07	5.58
Bicycle	7.67	5.57	**5.14**	5.92	*Dots*	**3.19**	5.24	5.08	21.21
Herbs	22.96	**11.23**	11.67	14.12	*Pyramids*	0.08	**0.04**	0.05	0.06
Origami	2.22	2.03	**1.78**	1.94	*Stripes*	7.94	6.96	**3.56**	7.90
Average	8.35	4.76	**4.70**	5.56	Average	8.50	4.21	**3.69**	8.69

6 Discussion and Limitation

Robust Model Fitting and Depth Inference. In this paper, we achieve robust parameter and depth estimation using limited information in a single light field. In retrospect, the original problem is divided into two smaller ones: 1) how to formulate good MRF energy given ground-truth depth and 2) how to estimate those MRF parameters as if we have the ground-truth depth. Th key of the solution is to separate model parameters into two parts: $\boldsymbol{\theta}_{\mathrm{MRF}}$ for MRF inference and $\boldsymbol{\theta}_{\mathrm{tail}}$ only for fitting distribution tails. The latter sub-problem is addressed by $\boldsymbol{\theta}_{\mathrm{tail}}$ which explains the difference in distribution tails between the depth map we have and the ground truth. Then the former one is solved by modeling hard-EM energy without $\boldsymbol{\theta}_{\mathrm{tail}}$. This processing flow is summarized in Fig. 23. In this work, $\boldsymbol{\theta}_{\mathrm{tail}}$ is $\{\epsilon_{\mathrm{d}}, \epsilon_{\mathrm{s}}\}$

[3] http://hci-lightfield.iwr.uni-heidelberg.de



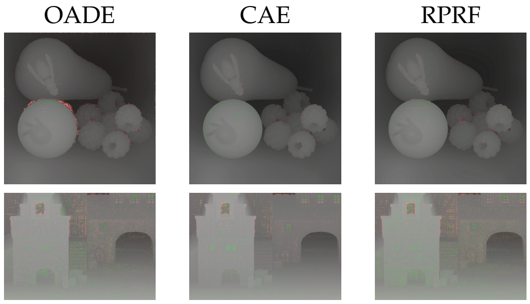
True-color, 3×3

True-color, Five-view

Grey-scale, 3×3

Grey-scale, Five-view

Fig. 21. Estimated depth maps for *Mona* using five-view crosshair sampling and/or grey-scale pixels.



OADE

CAE

RPRF

Fig. 22. Estimated depth maps using only data energy.

and $\boldsymbol{\theta}_{\mathrm{MRF}}$ covers the rest parameters. The above discussion is not limited to the simple pixel difference adopted in this work; therefore, we believe this framework can be extended to more sophisticated vision cues to further improve depth quality if they can be formulated by GSM.

Occlusion Handling. Instead of explicit handling as [6], we show that great depth quality can be achieved by implicit modeling with the soft-decision priors W and U. A value toward zero represents more likely occlusion or an edge. In this case, hard EM will saturate energy functions, which equivalently separates the occlusion or edge in a soft way. In contrast, soft-EM energy will consider all possible values of W and U (by integrating over them) and lack of such discrimination. Therefore, the fact that hard-EM energy is better than soft-EM one also confirms the necessity of occlusion formulation. In addition, a possible extension of this work is further enforcing inter-view depth consistency by adding other pixel difference models.

Scene Statistics. One limitation of the proposed framework is that we use one single model to explain a whole light field. Consider a scene that has two regions of different statistics, e.g. tablecloth and fruits in *StillLife*. In 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.

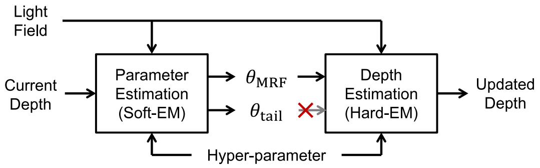


Fig. 23. Processing flow for robust model fitting and depth inference.

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 is to devise a more delicate heuristic to improve accuracy; however, overfitting should be avoided.

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-decision 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 MRF energy via hard EM. Accordingly, we build a stereo matching algorithm with efficient implementation. The effectiveness is demonstrated by experimental results on dense, sparse, denoised, and grey-scale light fields. Our work can estimate parameters robustly in one iteration. 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.

Acknowledgments

The author acknowledges support from the Ministry of Science and Technology, Taiwan, under Grant MOST 103-2218-E-007-008-MY3.

References

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

[2] 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*, 2013, pp. 2792–2799. 1, 3, 8

[3] 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*, 2015, pp. 3451–3459. 1, 3

[4] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 41–48. 1, 2, 8

[5] 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*, 2015, pp. 1547–1555. 1, 2, 8

[6] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *IEEE International Conference on Computer Vision*, 2015, pp. 3487–3495. 1, 2, 6, 7, 13

[7] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Computer Vision and Image Understanding*, vol. 145, pp. 148–159, Apr. 2016. 1, 2, 3, 8

[8] Williem, I. K. Park, and K. M. Lee, "Robust light field depth estimation using occlusion-noise aware data costs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, to appear. 1, 2, 3, 7

[9] J. Besag, "Statistical analysis of non-lattice data," *Journal of the Royal Statistical Society Series D (The Statistician)*, vol. 24, no. 3, pp. 179–195, Sep. 1975. 1

[10] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fiels," in *Vision, Modeling, and Visualization*, 2013, pp. 145–152. 2, 6

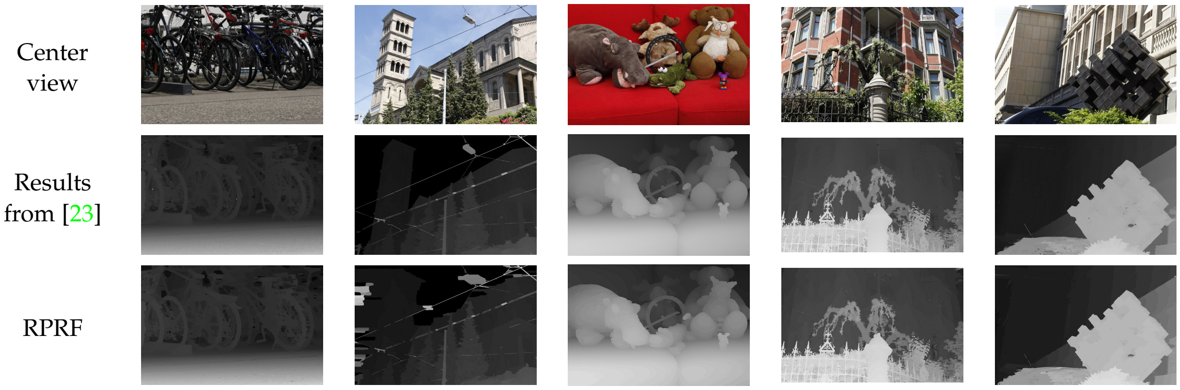


Fig. 20. Estimated depth maps for dense 3-D light fields from [23]. RPRF generates comparable depth maps, except for the sky in *Church*.

[11] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007. 2, 6

[12] C.-T. Huang, "Robust pseudo random fields for light-field stereo matching," in *IEEE International Conference on Computer Vision*, 2017. 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*, vol. 2, 2003, pp. 1150–1157. 2

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

[15] 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*, vol. 29, no. 2, pp. 331–342, Feb. 2007. 2

[16] 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*, vol. 20, no. 9, pp. 2515–2530, Sep. 2011. 2

[17] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Color and depth priors in natural images," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2259–2274, Jun. 2013. 2

[18] 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*, 1997, pp. 282–293. 2

[19] C.-T. Huang, "Bayesian inference for neighborhood filters with application in denoising," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4299–4311, Nov. 2015. 2, 5

[20] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang, "On building an accurate stereo matching system on graphics hardware," in *IEEE International Conference on Computer Vision Workshops*, 2011, pp. 467–474. 2

[21] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 504–511, Feb. 2013. 2

[22] G. Egnal and R. P. Wildes, "Detecting binocular half-occlusions: empirical comparisons of five approaches," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1127–1133, Aug. 2002. 2

[23] 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*, vol. 32, no. 4, pp. 73:1–73:12, Jul. 2013. 2, 7, 10, 14

[24] 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*, 2014, pp. 1518–1525. 2

[25] H. Sheng, S. Zhang, G. Zhu, and Z. Xiong, "Guided integral filter for light field stereo matching," in *IEEE International Conference on Image Processing*, 2015, pp. 852–856. 2

[26] L. Si and Q. Wang, "Dense depth-map estimation and geometry inference from light fields via global optimization," in *Asian Conference on Computer Vision*, 2016, pp. 83–98. 2

[27] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proceedings of the 7th European Conference on Computer Vision*, 2002, pp. 82–96. 3

[28] 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*, 2013, pp. 673–680. 3

[29] 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*, 2016, pp. 3262–3270. 3

[30] S. Heber, W. Yu, and T. Pock, "Neural EPI-volume networks for shape from light field," in *IEEE International Conference on Computer Vision*, 2017, pp. 2252–2260. 3

[31] C.-T. Huang, "Fast distribution fitting for parameter estimation of range-weighted neighborhood filters," *IEEE Signal Processing Letters*, vol. 23, no. 3, pp. 331–335, Mar. 2016. 4, 5

[32] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 41–54, Oct. 2006. 5

[33] 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*, vol. 30, no. 6, pp. 1068–1080, Jun. 2008. 5

[34] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Asian Conference on Computer Vision*, 2016. 6, 12

[35] M. Řeřábek and T. Ebrahimi, "New light field image dataset," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016. 7, 12

Chao-Tsung Huang (M'11) received the B.S. degree in electrical engineering and Ph.D. degree in electronics engineering from National Taiwan University in 2001 and 2005 respectively. He has been with National Tsing Hua University as an Assistant Professor since 2013. From 2005 to 2011, he was with Novatek Microelectronics Corp. for developing multi-standard image and video codecs. He then performed postdoctoral research at Massachusetts Institute of Technology and National Taiwan University for designing an HEVC decoder chip and light-field cameras respectively. His research interests include light-field signal processing and deep convolutional networks, especially from algorithm exploration to VLSI architecture design, chip implementation, and demo system.

Dr. Huang serves as Associate Editor for *IEEE Transactions on Circuits and Systems for Video Technology* and *Springer Circuits, Systems and Signal Processing*. He is also a member of DISPS Technical Committee of IEEE Signal Processing Society. He received 2017 Junior Faculty Research Award from College of EECS, National Tsing Hua University.