Depth Estimation From a Light Field Image Pair With a Generative Model

YAN, Tao; ZHANG, Fan; MAO, Yiming; YU, Hongbin; QIAN, Xiaohua; LAU, Rynson W. H.

Published in:
IEEE Access

Published: 01/01/2019

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1109/ACCESS.2019.2893354

Publication details:
https://doi.org/10.1109/ACCESS.2019.2893354

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.

Download date: 07/11/2022
Depth Estimation From a Light Field Image Pair With a Generative Model

TAO YAN, FAN ZHANG, YIMING MAO, HONGBIN YU, XIAOHUA QIAN, AND RYNSON W. H. LAU, (Senior Member, IEEE)

1Jiangsu Key Laboratory of Media Design and Software Technology, School of Digital Media, Jiangnan University, Jiangsu 214122, China
2Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China
3Department of Computer Science, City University of Hong Kong, Hong Kong

Corresponding authors: Tao Yan (yantao.ustc@gmail.com) and Xiaohua Qian (xiaohua.qian@sjtu.edu.cn)

This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20170197, and in part by the Fundamental Research Funds for the Central Universities under Grant JUSRP1141.

ABSTRACT In this paper, we propose a novel method to estimate the disparity maps from a light field image pair captured by a pair of light field cameras. Our method integrates two types of critical depth cues, which are separately inferred from the epipolar plane images and binocular stereo vision into a global solution. At the same time, in order to produce highly accurate disparity maps, we adopt a generative model, which can estimate a light field image only with the central subaperture view and corresponding hypothesized disparity map. The objective function of our method is formulated to minimize two energy terms/differences. One is the difference between the two types of previously extracted disparity maps and the target disparity maps, directly optimized in the gray-scale disparity space. The other indicates the difference between the estimated light field images and the input light field images, optimized in the RGB color space. Comprehensive experiments conducted on real and virtual scene light field image pairs demonstrate the effectiveness of our method.

INDEX TERMS Light field, depth estimation, disparity map, epipolar plane image, stereo matching, generative model.

I. INTRODUCTION

With the emergence of commercial light field cameras [1], [2], light field image processing has become a very popular research topic in computer vision and computer graphics communities [3]. A raw light field image can be decoded into a regular array of subaperture views/images, i.e., multi-perspective images of the same scene from slightly different viewpoints. A light field image can also be used to synthetically generate a dense focal stack. In post-processing, light field images can be used to overcome many challenges, such as scene structure inference, image refocusing, and novel view synthesis, which are difficult to solve in traditional 2D or stereoscopic images.

Depth estimation is a fundamental problem for light field image processing. Some existing methods explore the features of EPIs [4]–[7], occlusion and photo-consistency [8], symmetry in focal stack [9], [10], and multi-view stereo vision [11], and some even adopt the generative model [12] or an empirical Bayesian framework [13] to formulate their depth inference framework. Existing methods always formulate the depth estimation as a conventional optimization problem. There are also some studies [14]–[18] have designed special convolutional neural networks to infer depth from light field images. To the best of our knowledge, these works usually estimate depth from a single light field image without utilizing the binocular depth cues. Recently, two novel deep learning-based works [19], [20] have been proposed, which combine both the defocus cues and binocular depth cues for light field image depth inference.

Depth inference from either a light field image or a stereoscopic image pair has intrinsic limitations. Depth estimation from a single light field image usually with a narrow disparity range has the advantages of robustly occlusion inferring and depth estimation for occluded regions. However, since the small baseline and narrow disparity range, such disparity is always not accurate enough. In contrast, depth estimation from a stereoscopic image pair with a large baseline can return more accurate disparity map with a large disparity range, but cannot restore accurate disparity for complex thin structures (Fig. 1b) or very large textureless regions (Fig. 2b).
Therefore, a method to combine both kinds of disparity may overcome these limitations and produce more accurate disparity maps even for challenging scenes (Fig. 1h and 2h).

In this paper, we focus on estimating highly accurate disparity maps from a light field image pair captured by a pair of light field cameras. We propose a method to simultaneously utilize the two types of depth cues separately inferred from each single light field image and binocular stereo cues, i.e., correspondence between a light field image pair. The structure tensor technique [4] is adopted to extract initial coarse disparity maps from the EPIs of a single light field image. This approach usually returns rough disparity maps that contain accurate disparity values for the edges of the subaperture views, leaving significant noise for homogeneous regions of the subaperture views. In addition, the slanted plane fitting-based state-of-the-art stereo matching method [21] is adopted to estimate disparity maps from the two central subaperture views of the input light field image pair. Our method attempts to take advantages of these two types of disparity maps for accurate depth map estimation.

To produce highly accurate disparity maps, we adopt the concept of generative model proposed by Sajjadi et al. [12] for our depth estimation method. This generative model formulates a estimated light field image with only the central subaperture view and corresponding hypothesized disparity map. Then, by minimizing the difference between the estimated light field image and the input light field image, the disparity map for the central subaperture view can be obtained. Since a light field image usually contains a large number of subaperture views, a method utilizing such a generative model can build strong correlations for the target disparity map inference.

The main contributions of our work can be summarized as follows:

1) An effective method is proposed to estimate depth from a light field image pair by adopting a generative model and a convex optimization technique.

2) Two types of depth cues inferred from EPIs and binocular stereo vision are integrated into a unified solution for highly accurate depth estimation.

The rest of this paper is organized as follows. In Sec. II, we review the existing works related to light field image depth estimation. In Sec. III, we introduce how our method estimates the disparity maps from the input light field image pair. Our experiments conducted on synthetic light field image pairs (Fig. 1, 2, 3, 4, 6, 7) and real scene light field image pairs with a Lytro Illum camera (Fig. 8, 9, 10, 11) are shown in Sec. IV. Finally, we conclude our work described in this paper and discuss some future works in Sec. V.
II. RELATED WORK

In the last few years, some methods have been proposed to estimate depth from light field images. These methods can be divided into two categories: conventional optimization-based methods, and deep learning-based methods. In this section, we will review these two categories of works. In addition, we will summarize the stereo matching methods related to our work.

A. CONVENTIONAL OPTIMIZATION-BASED METHODS

Wanner and Goldluecke [4] proposed a structure tensor technique to estimate the direction of slopes in EPIs. After obtaining the noisy disparity map, this method refines the rough disparity maps via a fast total variation optimization procedure. Though the disparity for the pixels on the edges is accurate, other pixels in homogeneous regions may be assigned incorrect disparity values, which is difficult to rectify in the subsequent total variation optimization procedure, such as consistent EPI depth labeling and total variation optimization.

Lin et al. [9] proposed a method to recover the depth map from a light field image by exploiting two features of the light field focal stack. One feature is the property that non-occluding pixels exhibit symmetry along the focal depth dimension centered at the in-focus slice. The other is the data consistency metric of Markov random field (MRF), which is used to measure the difference between the focal stack synthesized from the hypothesized disparity map and the all-in-focus central image and that computed directly from the light field image.

Wang et al. [8] first explicitly modeled occlusions by developing a modified version of the photo-consistency condition on pixels of angular patches for light field images. If there is an occlusion for an angular patch, the angular patch can be divided into two regions, where only one of them obeys photo-consistency. The line separating the two regions in the angular domain (correct depth vs. occluder) has the same orientation as the occlusion edge does in the spatial domain. This method modifies the photo-consistency condition and the means/variances in the two regions to estimate occlusion-aware depth.

Jeon et al. [11] used the phase shift theorem in the Fourier domain to design a phase-based sub-pixel shift (PSS) method for light field image disparity estimation. Then, the disparity map estimation problem was formulated as a sub-pixel-wise multi-view stereo matching-based MRF optimization problem.

Sajjadi et al. [12] proposed a generative model for depth estimation, which was fully parameterized by the central subaperture view and its corresponding disparity map.
FIGURE 5. Quantitative evaluation of our method for disparity estimation on the virtual scenes (Fig. 1-4). (a) Error statistics for Fig. 1. (b) Error statistics for Fig. 1. (c) Error statistics for Fig. 2. (d) Error statistics for Fig. 2. (e) Error statistics for Fig. 3. (f) Error statistics for Fig. 3. (g) Error statistics for Fig. 4. (h) Error statistics for Fig. 4.

FIGURE 6. Comparison of our method under different condition. From left to right: (a) results of our method without utilizing the generative model (GM) discussed in Sec. III-C; (b) and (c) are results produced by the GM separately utilizing the disparity estimated by [6] and [21]; (d) results of our method utilizing both types of disparity maps. From top to bottom: the first row shows disparity maps for the central subaperture of $L_1$, and the second row shows the absolute residual value of disparity map in first row subtracting the ground-truth disparity map. Similarly, the third row and fourth row show the disparity maps and residual maps for the central subaperture view of $L_2$.

The disparity maps produced by this method are more preferable than those obtained by other methods, especially for light field images captured by the Lytro camera in real scenes.

FIGURE 7. Comparison of our method. The description for this set of experimental results is similar to that of Fig. 6. (a) without GM. (b) GM&EPIs. (c) GM&stereo. (d) our results.

Zhang et al. [7] proposed a spinning parallelogram operator (SPO) integrated into a depth estimated framework to locate lines and calculate their slopes in an EPI for local depth estimation. The spinning parallelogram operator maximizes the distribution distances between the two parts of the parallelogram window to extract depth information. This method
combines the depth information inferred in the horizontal and vertical slices/EPIs. It is insensitive to occlusion, noise, spatial aliasing, and limited angular resolution.

Huang [13] proposed an empirical Bayesian framework, the robust pseudo random field (RPRF), to provide statistical adaptability and good depth quality for light field depth inference. In this work, the author focused on exploring intrinsic statistical cues for MRF inference. Based on the proposed model with hidden soft-decision priors, this method applies the soft expectation-maximization for good model fitting, and performs hard EM for robust depth estimation. This method can estimate scene-dependent parameters robustly and converges quickly for depth estimation.

All these methods estimate disparity maps from a single light field image based on multi-label optimization, as there...
are always significant noise/errors in the obtained disparity maps. Our method also adopts such an optimization strategy to obtain the optimal disparity maps and remove possible noise/errors.

B. DEEP LEARNING-BASED METHODS

Hazirbas et al. [17] proposed an auto-encoder-style convolutional neural network to estimate depth from a focal stack of a real scene light field image. To train the proposed convolutional neural network, the authors built a dataset including a large number of light-field images and the corresponding registered ground-truth depth maps recorded with an RGB-D sensor.

Heber et al. [16] presented a novel U shaped auto-encoder-style deep learning network to extract depth from light field images. This network takes the 3D subsets of the 4D light field, i.e., 3D EPI volumes, as input data. The network utilizes 3D convolutional layers to propagate information from two spatial dimensions and one direction of the light field image. This method can reduce depth artifacts and at the same time maintain clear depth discontinuities.

Alperovich et al. [18] proposed a fully convolutional auto-encoder network to jointly solve the disparity regression and reflectance separation in light fields. This network employs 3D convolutions to calculate features integrated over the whole range of both vertical and horizontal 3D EPIs volumes to deal with complex occlusions and reflections. This network implements the auto-encoder path to reconstruct the input, two decoders for the diffuse and specular components, and a separate decoder for the disparity map. The auto-encoder path of the network is jointly trained by unsupervised learning, and the other decoder paths are trained by supervised training. This method can recover reliable depth in the presence of strong specularity.

Guo et al. [19] proposed a unified learning-based technique that utilizes both binocular stereo cues and monocular focusness cues for depth inference. This network adopts a pair of focal stacks as input to emulate the human perception. The authors constructed three individual networks: a FocusNet to extract depth from a single focal stack, an EDoFNet to obtain the extended depth of a field image from the focal stack, and a StereoNet to conduct stereo matching. The EDoF image from EDoFNet serves to both guide the refinement of the depth from FocusNet and provide inputs for StereoNet. Later, these networks were integrated into a unified solution to obtain the final depth maps.

Deep learning-based disparity estimation methods are usually trained on light field images captured from a limited number of synthetic or real scenes. Since the ground-truth depth maps for light field images of real scenes are difficult to obtain, such methods are mainly trained on some virtual scenes that are always simpler than real scenes. While processing a light field image of a new scene, such methods cannot always produce disparity maps that are sufficiently accurate and cannot outperform the optimization-based methods.

C. STEREO MATCHING METHODS

Stereo matching is the most studied research topic in the computer vision community. We suggest that readers refer to the study [22] for an overview. Recently with the rapid development of deep learning, some methods have been proposed to estimate depth from stereoscopic image pair by adopting convolutional neural networks [23]–[27].

Taniai et al. [21] proposed an accurate and efficient method for stereo matching. This method utilizes 3D plane labels to establish a new stereo model based on a pairwise MRF. This method takes advantage of two recent breakthroughs, slanted patch matching, and tangent-based curvature regularization. The slanted patch matching means that the disparity for each pixel is over-parameterized by a local disparity plane defined on the image domain, and the triplet to parameterize a plane is estimated for each pixel instead of directly estimating its disparity. The curvature regularization is represented in pairwise terms similar to conventional linear and truncated linear models. Thus, the method can handle smooth surfaces beyond planes. A new move making scheme, local expansion moves, is introduced to replace global and expensive graph-cut optimization for the whole image by many small local α-expansions optimization [28] for different small regions of image content. The local expansion moves enable spatial propagation in graph-cut optimization. This method shows advantages both in efficiency and accuracy.

We adopt this method to estimate the disparity maps for the two central subaperture views of a light field image pair. The obtained disparity maps will be utilized as the binocular depth cues for our method.

III. OUR METHOD

Our method takes a light field image pair as input, which can be captured by two horizontally arranged light field cameras or a light field camera shifting on a shelf. The overview of our method can be described as follows.

First, we estimate the initial rough disparity maps separately from the EPIs of each single light field image and the two central subaperture views of the input light field image pair.

Second, we register the two types of disparity maps to merge them together as prior depth cues for producing accurate disparity maps.
Finally, we optimize the target disparity maps by minimizing the differences between the optimized disparity and the initially inferred rough disparity, and between the input light field images and the estimated light field images.

A. PREPROCESSING FOR INITIAL DEPTH CUES

We observed that depth estimation methods, such as [4], can obtain more accurate disparity for edge pixels than homogeneous region pixels. In contrast, the state-of-the-art stereo matching method [21] based on slanted plane fitting can obtain more accurate disparity for homogeneous regions. Therefore, we adopt the structure tensor technique [4] to gain high-confidence disparity for the edge pixels. Additionally, we use the stereo matching method [21] to obtain reliable disparity for the homogeneous region pixels from the two central subaperture views of a light field image pair. Then, we take advantage of the two types of disparity maps to produce more accurate disparity maps for the input light field image pair.

Let \( L_1 \) and \( L_2 \) represent the input light field image pair. \( \hat{D}_1 \) and \( \hat{D}_2 \) denote the cleaned disparity maps for the two central subaperture view of \( L_1 \) and \( L_2 \) obtained by method [4]. \( C_1 \) and \( C_2 \) denote the corresponding confidence maps for \( \hat{D}_1 \) and \( \hat{D}_2 \). \( \hat{D}_1 \) and \( \hat{D}_2 \) denote the disparity map for the central subaperture view of \( L_1 \) and \( L_2 \) estimated by the stereo matching method [21]. To remove the inconsistent/incorrect disparity values, we clean the initial coarse disparity maps. \( M_1 \) and \( M_2 \) denote the mask maps after cross-checking with \( \hat{D}_1 \) and \( \hat{D}_2 \), which indicate the valid and consistent disparity values in \( \hat{D}_1 \) and \( \hat{D}_2 \). In additional, we compute the gradient map for the two central subaperture views of \( L_1 \) and \( L_2 \), denoted as \( G_1 \) and \( G_2 \).

B. REGISTRATION FOR A LIGHT FIELD IMAGE PAIR

We register and merge the two types of disparity maps for our disparity estimation. Since all the subaperture views of a light field image take a converged (toed-in) stereo camera model [29] and the proposed light field camera pair ensures that their main optical axes are in parallel, it is easy to build the relationship between the two types of disparity maps, e.g., \( \hat{D}_1 \) and \( \hat{D}_1 \), as shown in the following equation.

\[
\tilde{D}_1 = \frac{b}{B} \hat{D}_1 - d_s \tag{1}
\]

where \( b \) is the small baseline for the nearest neighboring subaperture views in the horizontal or vertical direction in the same light field image; \( B \) is the baseline for the light field image pair, which is a constant parameter calibrated in the experiment; and \( d_s \) represents the shift of the projected subaperture view on the photosensor of the light field camera relative to the optical axis of the subaperture view, which is a constant parameter for a specified light field image.

In this work, we adopt the decoding method [30] for real scene light field images recorded by the Lytro Illum camera [1]. This method utilizes a parallel stereo camera model, which means that the parameter \( d_s \) is equal to zero (Eqn. 1). The other parameter \( \frac{b}{B} \) can be obtained by adopting the least squares technique. For generating the virtual scene light field image pairs, we adopt the toed-in stereo camera model as described in [29].

Once the relationship between the two types of disparity maps have been determined, \( \tilde{D}_1^1 \) and \( \tilde{D}_2^2 \) are used to indicate the scaled disparity maps of \( D_1 \) and \( D_2 \), which means that they are mapped to the disparity range same to the input light field image pair.

C. DISPARITY MAP OPTIMIZATION BY ADOPTING A GENERATIVE MODEL

We take the two types of previously inferred disparity maps as prior depth cues, and define the following energy function to merge them and propagate high confidence disparity values to nearby regions.

\[
E_d(D_1, D_2) = \lambda_1(E_{de}(D_1) + E_{de}(D_2)) + \lambda_2(E_{ds}(D_1) + E_{ds}(D_2)) + \lambda_3(E_{dr}(D_1) + E_{dr}(D_2)) + \lambda_4(E_{de}(D_1, D_2) + E_{de}(D_2, D_1)) \tag{2}
\]

where \( D_1 \) and \( D_2 \) are the desired disparity maps to be optimized by our method for the two central subaperture views, \( L_1(s_c, t_c) \) and \( L_2(s_c, t_c) \). \( (s_c, t_c) \) denotes the central subaperture viewpoint of a light field image. \( E_{de} \), \( E_{ds} \), \( E_{dr} \) and \( E_{de} \) are energy terms designed for our disparity map optimization, which will be discussed later in this section. \( \lambda_1, \lambda_2, \lambda_3 \) and \( \lambda_4 \) are constant weighting parameters.

The energy terms \( E_{ds} \) and \( E_{de} \) are used to enforce the disparity maps \( D_1 \) and \( D_2 \) to be similar to the two types of disparity maps previously predicted. These two energy terms are defined as:

\[
E_{de}(D_1) = \sum_p \sum_{q \in N(p)} \omega_{1,pq} C_1(p) \|D_1(p) - \hat{D}_1(p)\|^2 \tag{3}
\]

\[
E_{de}(D_2) = \sum_p \sum_{q \in N(p)} \omega_{2,pq} C_2(p) \|D_2(p) - \hat{D}_2(p)\|^2 \tag{4}
\]

\[
E_{ds}(D_1) = \sum_p (1 - G_1(p)) \|D_1(p) - \hat{D}_1(p)\|^2 \tag{5}
\]

\[
E_{ds}(D_2) = \sum_p (1 - G_2(p)) \|D_2(p) - \hat{D}_2(p)\|^2 \tag{6}
\]

where \( N(p) \) refers to the neighboring pixels around \( p \), e.g., in an \( 11 \times 11 \) window around \( p \). \( \omega_{1,pq} \) (\( \omega_{2,pq} \)) is the weighting term to measure the similarity of the color of the \( 3 \times 3 \) windows around \( p \) and \( q \) in \( L_1 (L_2) \), defined as follows:

\[
\omega_{1,pq} = \exp(-\sum_{\overset{p' \in N(p)}{q' \in N(q)}} \frac{(I_1(p') - I_1(q'))^2}{\sigma_c^2} + \frac{(G_1(p') - G_1(q'))^2}{\sigma_g^2}) \tag{7}
\]

where \( I_1 = L_1(s_c, t_c) \). \( \sigma_c^2 \) and \( \sigma_g^2 \) are the color and gradient variance for the whole image \( L_1(s_c, t_c) \), respectively, which are set very differently from the previous model [12].
N′(p) and N′(q) denote the 3 × 3 windows centered at p and q, p′ and q′ refer to the corresponding pixels with the same local coordinates within N′(p) and N′(q).

The regularization term, \(E_{dr}\), which is used to enforce the smooth disparity change for neighboring pixels, is defined as:

\[
E_{dr}(D_1) = \sum_{p} \sum_{q \in N(p)} \omega_{1,pq} \| D_1(q) - D_1(p) \|_2^2
\]

\[
E_{dr}(D_2) = \sum_{p} \sum_{q \in N(p)} \omega_{2,pq} \| D_2(q) - D_2(p) \|_2^2
\]

We use the energy term \(E_{dc}\) to enforce the coherence between \(D_1\) and \(D_2\), which can be defined as:

\[
E_{dc}(D_1, D_2) = \sum_{p} M_1(p) \| D_1(p) - D_2(p - \hat{D}_1(p)) \|_2^2
\]

\[
E_{dc}(D_2, D_1) = \sum_{p} M_2(p) \| D_2(p) - D_1(p + \hat{D}_2(p)) \|_2^2
\]

where \(p - \hat{D}_1(p)\) are coordinates of the corresponding pixel pairs separately within the two central subaperture views \(L_1(s_c, t_c)\) and \(L_2(s_c, t_c)\). Similarly, \(p + \hat{D}_2(p)\) are coordinates of corresponding pixel pairs separately within \(L_2(s_c, t_c)\) and \(L_1(s_c, t_c)\). \(M_1\) and \(M_2\) are the mask maps defined in Sec. III-A for the cross checked disparity maps.

**Generative model for depth estimation:** to obtain highly accurate disparity maps, we adopt the concept of the generative model described in [12], which formulate the estimated light field image based only on the central subaperture view and its disparity map. Then, by minimizing the difference between the estimated light field images and the original light field images, we can make our method produce more accurate disparity maps than just adopting the objective function Eqn. 2.

The estimated light field image corresponding to \(L_1\) can be defined as:

\[
\tilde{L}_1(s, t, x, y) = \sum_{x_c} \sum_{y_c} L_1(s_c, t_c, x_c, y_c) W(x - (x_c + d_t) \times D_1(x_c, y_c)) W(y - (y_c + d_t \times D_1(x_c, y_c)))
\]

where \((x_c, y_c)\) denotes a pixel on the central subaperture view of \(L_1\), \(d_t = s - s_c\) and \(d_t = t - t_c\). \(W\) is a weighting function that can distribute the color of shifted sub-pixel \(L_1(s_c, t_c, x_c, y_c)\) on the estimated subaperture view \(\tilde{L}_1(s, t)\) to its four nearest neighboring pixels.

\[
W(x) = \begin{cases} 
1 - x, & \text{if } 0 \leq x < 1 \\
1 + x, & \text{if } -1 < x < 0 \\
0, & \text{if } |x| \geq 1 
\end{cases}
\]

The estimated light field image \(\tilde{L}_2\) corresponding to \(L_2\) can be defined similarly. For more details, we suggest that readers refer to the previous study [12].

Then, we can minimize the RGB difference between the estimated light field images, \(\tilde{L}_1\) and \(\tilde{L}_2\), and the input light field images, \(L_1\) and \(L_2\), by optimizing the following energy function:

\[
E_l(D_1, D_2) = \sum_{(s, t) \not\in \tilde{L}_1} \| \tilde{L}_1(s, t) - L_1(s, t) \|_2^2 + \| \tilde{L}_2(s, t) - L_2(s, t) \|_2^2
\]

**Final objective function:** therefore, the global objective function for our disparity estimation method can be defined as:

\[
(D_1, D_2) = \arg \min E_l(D_1, D_2) + E_l(D_1, D_2)
\]

This objective function can be efficiently solved by adopting the convex optimization L-BFGS-B algorithm [31], [32].

**IV. EXPERIMENTAL RESULTS AND DISCUSSION**

Our method is implemented in MATLAB. We evaluated our method on both real and virtual scene light field images. Since there is no existing dataset for light field image pair processing and evaluation, we have built a dataset\(^1\) by capturing some challenging real scene light field image pairs and synthesizing some challenging virtual scene light field images. For virtual scene light field image pairs generated with Blender [33] accompanying with ground-truth disparity maps, we quantitatively compared our method with the existing state-of-the-art methods by running their code on our dataset. For real scene light field images captured by a Lytro Illum camera [1], we qualitatively compared our method with the existing methods. The spatial resolution is set as 512 × 512 for virtual scene light field images, and 625 × 434 for real scene light field images. The angular resolution for both kinds of light field image is set as 9 × 9.

The constant weighting parameters in Eqn. 2 are set as \(\lambda_1 = 10^5\), \(\lambda_2 = 10^5\), \(\lambda_3 = 2 \times 10^5\), \(\lambda_4 = 5 \times 10^4\) for virtual scene light field image pairs, and \(\lambda_1 = 10^5\), \(\lambda_2 = 10^5\), \(\lambda_3 = 2 \times 10^4\), \(\lambda_4 = 10^4\) for real scene light field image pairs. The difference in the parameter setting for the two kinds of light field images can be explained as follows. Since real scene light field images always contains some noise introduced in the capturing and decoding steps, the initial rough disparity maps estimated from real scene light field image pairs are less reliable than those inferred from the synthetic light field images. The time cost for our method processing a virtual scene light field image pair is about 15 minutes. And the time cost for processing a real scene light field image pair is about 20 minutes.

**A. QUANTITATIVE EVALUATION**

The camera parameters for light field images shown in Fig. 1-4 are listed in Tab. 1. For each light field image pair \((L_1, L_2)\), we show the central subaperture view of \(L_1\) and corresponding ground-truth disparity map. To demonstrate the effectiveness of our method, we compare the disparity

\(^1\)https://github.com/yantaocv/Depth-Estimation-from-a-Light-Field-Image-Pair-with-a-Generative-Model
map obtained by our method with those produced by other methods \cite{4,7,8,11,13,21}. 

Take Fig. 1 as an example. This example is challenging for disparity estimation, since there are many thin structures in the scene. Fig. 1a, from top to bottom, shows the ground-truth disparity map and the central subaperture view of \( L_1 \). In Fig 1b, the top figure shows the grey-scale disparity map estimate from the two central subaperture views of the input light field image pair by adopting the state-of-the-art stereo matching method \cite{21}, and the bottom figure shows the colored residual map, which represents the absolute value of the disparity map subtracting the ground-truth disparity map. Similarly, in Fig 1c, Fig. 1d, Fig. 1e, Fig. 1f, Fig. 1g and Fig. 1h, from top to bottom, the figures in the first row show the disparity maps estimated from the single light field image \( L_1 \) by separately adopting the methods \cite{4,7,8,11,13,21} and our method, and the figures in the second row show the color residual maps generated in the same way as in the bottom row of Fig. 1b. In the residual maps, the deep blue color represents that the estimated disparity is consistent with the ground-truth disparity. In contrast, the deep red color indicates that the estimated disparity is inconsistent with the ground-truth disparity, which means that there is a large gap between the estimated disparity and ground-truth disparity.

It can be seen that there are fewer obvious errors in the disparity maps produced by our method than those obtained by other methods. Our method can well deal with the complex thin structures of the bicycle and light pole to produce more accurate disparity maps.

The situations in Fig. 2 and Fig. 3 are similar to that in Fig. 1. Our method outperforms the other methods in these two examples. In Fig. 2, compared with the disparity maps obtained by other methods, there are few errors in the disparity map produced by our method. In this example, our method can restore disparity for the large textureless background and complex thin structure of the boat better than any other methods. In Fig. 3, although some other methods \cite{4,8} may also produce good disparity maps with few errors, our method obtains more excellent and favorable disparity map. There are no noticeable errors in the disparity maps obtained by our method. In contrast, there are some obvious errors in the disparity maps produced by other methods. In Fig. 4, the disparity maps obtained by our method and the methods \cite{7,13,21} are much better than those obtained by other methods \cite{4,8,11}. Our method is comparable with the most recent state-of-the-art methods \cite{7,13,21} in this example.

We also compare our method with other state-of-the-art methods in the way of error statistics. Fig. 5 shows the error statistics for the disparity maps produced in Fig. 1- 4. Fig. 5a plots the statistics of absolute error for disparity maps shown in Fig. 1. Fig. 5b is a subfigure of Fig. 5a by zooming in its bottom right part, with the absolute error ranging from 0.1 to 0.3. Obviously, our method outperforms other methods with fewer large errors, whose ratio stably converges to zero. The results in Fig. 5c and Fig. 5d for Fig. 2, Fig. 5e and Fig. 5f for Fig. 3, Fig. 5g and Fig. 5h for Fig. 4 are similar to the error statistics in Fig. 5a and Fig. 5b, which demonstrates the effectiveness and stability of our proposed method.

Additionally, we conduct an evaluation for our method under different conditions of objective function and initial depth cues. The light field image pairs utilized in Fig. 6 and Fig. 7 are same to the light field image pair used in Fig. 1 and Fig. 4. The central subaperture views of these light field image pair are shown in Fig. 1d and Fig. 1e.

Fig. 6a shows the disparity maps obtained by our method without adopting the generative model (defined in Eqn. 14) and their colored residual maps comparing with the ground-truth disparity maps. Fig. 6b shows the disparity maps produced by the generative model utilizing only the initial depth cues estimated by the method \cite{4} from the EPIs. Similarly, Fig. 6c shows the disparity maps optimized by the generative model utilizing only the initial disparity maps estimated from the two central subaperture views of the input light field image pair by adopting the stereo matching method \cite{21}. Fig. 6d shows our method’s normal disparity maps while simultaneously taking advantage of the generative model and two types of initial depth cues. The disparity maps shown in Fig. 6d are much better than those shown in other columns. In addition, from the experimental results, it can be seen that the coherence between the disparity maps for the two central subaperture views of the input light field image pair can be guaranteed.

We show another example in Fig. 7. The indoor scene can be seen as mainly consisting of some slanted planes. The generative model utilizing only the binocular depth cues can produce good disparity maps as shown in Fig. 7c. Our method’s norm disparity maps shown in Fig. 7d are comparable with the disparity maps shown in Fig. 7c, and much better than those shown in Fig. 7a and Fig. 7b.

### B. QUALITATIVE EVALUATION

The content of the light field images captured from real scenes is usually more complex than those generated from virtual scenes. Additionally, complex lighting conditions in capturing step, noise and distortion introduced in the decoding step, make depth estimation from real scene light field images more challenging than that from virtual scene light field images. The calibrated camera parameters for these light field image pairs (as shown in Fig. 8) are listed in Tab. 2.

Fig. 8 shows a set of light field image pairs captured from various challenging real scenes. For each scene exhibited in Fig. 8, from left to right, we show the central subaperture

---

**TABLE 1.** The camera parameters and disparity range for each light field image (LFI) pair from the virtual scene.

<table>
<thead>
<tr>
<th>No.</th>
<th>Parameters for each LFI pair from the virtual scene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( f ) (pixel)</td>
</tr>
<tr>
<td>Fig. 1</td>
<td>385.14</td>
</tr>
<tr>
<td>Fig. 2</td>
<td>385.14</td>
</tr>
<tr>
<td>Fig. 3</td>
<td>385.14</td>
</tr>
<tr>
<td>Fig. 4</td>
<td>385.14</td>
</tr>
</tbody>
</table>
view of the light field image $L_1$ and corresponding disparity map inferred from the two central subaperture views by adopting the method [4], as well as the disparity map estimate from the single light field image $L_1$ by separately adopting [7], [8], [11], [13], [21] and our method. Since there are no ground-truth disparity maps for this set of light field images, we can only qualitatively compare the disparity maps obtained by our method with those estimated by the other methods. Although some existing methods may produce favorable disparity maps for some of the scenes, such as the first row of Fig. 8b obtained by [21], the fourth row and sixth row of Fig. 8d produced by [8] and Fig. 8f estimated by [7], these methods cannot produce favorable disparity maps for most of the scenes. In contrast, our method produces much better and more reasonable disparity maps for all the scenes, as shown in Fig. 8h.

We also conduct an evaluation on real scene light field image pairs for our method under different conditions of objective function and initial depth cues. The light field image pairs utilized for this evaluation are same to the light field image pairs used in the first row, fourth row and last row of Fig. 8. The central subaperture views of these light field image pair are shown in Fig. 12a, 12b and 12c. In Fig. 9, the normal disparity maps obtained by our method, as shown in Fig. 9d, are much better than those obtained by our method without adopting the generative model or utilizing only one type of the initial depth cues. In Fig 10, the normal disparity maps optimized by our method (as shown in Fig. 10d) are much better than those obtained under other conditions. While zooming the images shown in this example, we can find that there are many small noise/errors in the disparity maps (Fig. 10a) that are produced by our method without adopting the generative model. Fig.10b and Fig. 10c are disparity maps optimized by the generative model using only one type of the initial depth cues. These disparity maps are very smooth and contain less noise. In contrast, our method’s normal disparity maps (Fig. 10d) are more accurate and smooth than those shown in the Fig. 10a, 10b and 10c. We observed that while utilizing the depth cues inferred from EPIs, the disparity for the branches on the further background can be clearly restored by our method as shown in Fig. 10a, 10b and 10c. In contrast, while our method utilizing only the depth cues estimated by the stereo matching method [21], it cannot restore the disparity for such small details, as shown in Fig. 10c.

Similarly, in Fig. 11, the disparity maps (Fig. 11a) produced by our method without a generative model contain some obvious errors for the branches on the left of the background. The disparity maps shown in Fig. 11b contains much noise and not smooth enough. The disparity maps shown in Fig. 11c contain too many errors. The normal disparity maps of our method as shown in Fig. 11d are more accurate and smooth than those obtained by our method under other conditions.

V. CONCLUSION AND FUTURE WORK

In this work, we have proposed a novel method to estimate depth from a light field image pair. Our method takes advantage of two types of important depth cues that are separately inferred from EPIs and binocular stereo vision, i.e., the two central subaperture views of a light field image pair. To optimize the disparity maps for the two central subaperture views, we build an objective function that has two parts. One part of the energy function attempts to directly merge the two types of disparity maps and optimize the disparity maps for the two central subaperture views in the disparity grey-scale space. The other part of the energy function aims at optimizing the desired disparity map in the RGB color space by adopting a modified generative model to minimize the difference between the input light field images and the estimated light field images. The objective function can be efficiently solved by adopting a convex optimization technique. Extensive experiments conducted on real and virtual scenes demonstrate the effectiveness of our method.

REFERENCES


et al.


[32] L-BFGS-B-C. https://github.com/stephenbeckr/L-BFGS-B-C