Joint 3D Shape and Motion Estimation from Rolling Shutter Light-Field Images

Hermes McGriff [1,3] Renato Martins[1,2] Nicolas Andreff [3] Cédric Demonceaux[1,2]
[1]Université de Bourgogne, CNRS UMR 6303 ICB [2]Université de Lorraine, CNRS, Inria, LORIA
[3]Université de Franche-Comté, CNRS UMR 6174 FEMTO-ST
{hermes.mc-griff,renato.martins,cedric.demonceaux}@u-bourgogne.fr, nicolas.andreff@univ-fcomte.fr

Abstract

In this paper, we propose an approach to address the problem of 3D reconstruction of scenes from a single image captured by a light-field camera equipped with a rolling shutter sensor. Our method leverages the 3D information cues present in the light-field and the motion information provided by the rolling shutter effect. We present a generic model for the imaging process of this sensor and a two-stage algorithm that minimizes the re-projection error while considering the position and motion of the camera in a motion-shape bundle adjustment estimation strategy. Thereby, we provide an instantaneous 3D shape-and-pose-and-velocity sensing paradigm. To the best of our knowledge, this is the first study to leverage this type of sensor for this purpose. We also present a new benchmark dataset composed of different light-fields showing rolling shutter effects, which can be used as a common base to improve the evaluation and tracking the progress in the field. We demonstrate the effectiveness and advantages of our approach through several experiments conducted for different scenes and types of motions. The source code and dataset are publicly available at: `https://github.com/ICB-Vision-AI/RSLF`.

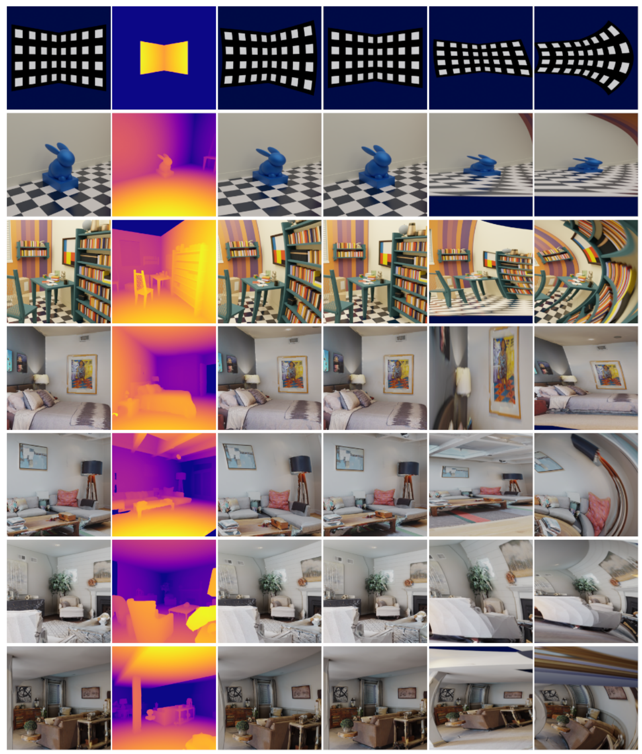

Figure 1. Some central views of the proposed RSLF dataset. From left to right: A global shutter view, the ground truth depth, and four different rolling shutter views with increasing camera motions.

1. Introduction

Light-field (LF) cameras (also known as plenoptic), introduced by Adelson and Wang [1] and prototyped by Ng [25], consist of a conventional camera with a microlens array in front of the photosensitive sensor. This type of imaging sensor has the particularity of being able to capture a light field of a scene in a single capture. LF cameras are now an established solution used in computer vision, photogrammetry and robotics [7, 15, 37]. The miniaturization of these cameras, *e.g.* in the context of applications such as intra-corporeal micro-robotics, requires the choice of a rolling shutter (RS) photosensitive sensor. Conversely to global shutter (GS) cameras, where all pixels of the image are acquired at the same time, the image acquisition by rolling shutter is sequential [24]. Notably, a RS sensor creates image deformations in the case of dynamic scenes or when the camera is moving, as depicted in Fig. 1. The rolling shutter then often degrades the performance and challenge existing reconstruction and pose estimation approaches. Ait-Aider *et al.* [3] have shown that, in the case of a conventional perspective monocular camera, these deformations can be leveraged in order to compute the motion of the scene with respect to the camera. However, their proposed model and subsequent improved strategies [4, 18] have the strong limitation of requiring the shape of the object/scenes to be known. Conversely, this paper proposes to jointly estimate the motion and the structure of a scene from a single view shot and in less constrained conditions. Although the possibilities given by RS, when properly modelled, has been shown for several computer vision and graphics problems, the combination with LF has not been yet exploited in a unified approach. One important motivation of this paper is to show the possibilities that this sensor modality presents, such as of being able to allow the estimation of the camera motion (or from the scene/object) from a single view without priors on the scene shape. We are particularly motivated by showing the interest of a unified approach (and its properties) that is capable of leveraging RS with existing light-field consumer devices. Indeed, such sensors are available, like the entry level cameras of Raytrix (R8, R42, R10μ, R20) or any camera array with RS sensors (like Pelican Imaging), but unfortunately no public dataset is available to be the best of the authors' knowledge. In this context, another core motivation of this paper is to present a suitable and challenging LF dataset collected with a RS camera with different motion levels and scene geometries. For that, we have generated new models and leveraged existing scene models (from Matterport3D) into an adapted rendering engine (based on Blender) to create LF data affected by RS distortions in different conditions (e.g., from mild to strong motions). The main contributions of this paper are as follows:

- We propose a generic projection model of a rolling shutter light-field (RSLF) camera. This model is capable to represent a light-field with both global shutter and rolling shutter settings.
- A non-linear bundle adjustment strategy is designed to estimate jointly the 3D shape and motion for this sensor modality. We also design a linear initialization strategy in order to recover a first coarse estimate of the 3D shape. This initialization is essential for the convergence of our approach as shown in the ablation studies.
- We also present a new dataset composed of Rolling Shutter Light Fields (RSLF) paired with ground truth depth maps, on several synthetic scenes and with different types and levels of motion. We aim this dataset to be used as a common base to improve the evaluation and help tracking the progress in the field.

2. Related Work

Depth estimation from light-fields. Light-field contains rich information cues about the geometry of the scene. The seminal work of Adelson and Wang [1] for the plenoptic camera exploit this ability for "single lens stereo". They used sub-aperture images (SAI) to perform a standard two frame displacement analysis with multiple pairs horizontally and vertically. In the same direction, multi-view stereo matching-based methods try to reproduce the results of classical stereo with plenoptic images [12, 16, 27, 38]. In this context, Georgiev and Lumsdaine [12] introduced the focused plenoptic camera and proposed a complete setup in order to recover depth with this new design. The method simultaneously render the image and estimate a per microlens depth map by computing the cross correlation between patches in micro images. Similarly, Perwass and Wietzke [27] introduced a multi-focused plenoptic camera model alongside a depth estimation algorithm based on point correspondences between micro images. Jeon and Park [16] explored the phase-shift theorem of the Fourier transform to estimate an accurate sub-pixel disparity map by computing a matching cost volume between SAI. Zeller *et al*. [38] proposed a filtering method for the estimation of semi-dense probabilistic depth maps for focused plenoptic cameras, with a Kalman filter like approach preserving discontinuities in the depth map. Ferreira and Goncalves [11] proposed a similar but faster depth map estimation method, with SIFT correspondences and through epipolar lines on the micro images. Bok *et al*. [5] proposed a calibration of the light-field camera based on a bundle adjustment method and Zhang *et al*. [39] proposed a generic multi-projection model (along with its calibration algorithm) for LF cameras. Most of these techniques rely on generating SAI and then applying classic stereo matching algorithms to estimate the depth of the scene. However, they assume GS cameras (or with slow moving objects and camera motions). Our approach, on the other hand, can handle scenes with a camera in movement and is far less affected by RS distortions due to camera motions.

Epipolar plane images and learning-based LF analysis. The scene structure can also be extracted from Epipolar Plane Images (EPI) [6, 8, 32, 36]. These approaches estimate depth information from the slopes of the lines observed in the Epipolar planes. Tao et al. [34] improved the accuracy of the depth estimation with a weighted sum between the defocus and correspondence cues present in EPIs. Zhang *et al*. [40] proposed a spinning parallelogram operator to determine the line slopes. Lin *et al*. [21] leveraged the refocus capability of light-fields and the possibility to use Shape-From-Focus. Closely related to our work, Srinivasan *et al*. [33] proposed a motion estimation from a single view with a light-field camera based on motion blur. Heber and Pock [13] first used a Convolutional Neural Network to compute depth from light-field images. Shin *et al*. [31] proposed a fast and accurate light field depth estimation method based on a fully-convolutional neural network and a light-field image-specific data augmentation. These techniques suffer by the lack of generalization to new/unseen scenes and often dependence on significant amount of data.

Rolling shutter structure-from-motion estimation. The potential of RS images received increased attention for scene analysis. Meingast *et al*. [24] developed a general projection equation for a rolling shutter camera and also proposed a calibration to estimate the rate of the rolling shutter. Ait-Aider *et al*. [3] first showed that the rolling shutter effect could be leveraged to estimate the motion of an object, but of known shape, when the majority of previous studies on the rolling shutter were about compensating it [17, 20]. This is notably done for blur compensation with both model and learning-based approaches [10, 23]. Saurer *et al*. [30] and Ait-Aider *et al*. [4] investigated RS effects for stereo vision. Recently Lao *et al*. [19] proposed a analogy with non-rigidity to solve shape estimation with a monocular rolling shutter image. Different than these previous methods, we address the ambiguity between shape and motion inherent to RS images exploiting the properties of the LF. We show that a micro-lens array in front of the RS sensor allows to model the RS effect in the case of 3D scenes and to estimate the movement of the scene with respect to the camera without prior knowledge of the scene geometry.

3. Method

Our joint 3D scene reconstruction and camera motion estimation approach has two main stages. Firstly, a coarse linear solution is computed to provide an initialization for a non-linear bundle adjustment method. This method is designed to handle the geometric and temporal constraints that are present in the Rolling Shutter Light-Field setting.

Light-field modeling and RS projection. To provide a comprehensive theoretical framework for our proposed approach, we begin by presenting an overview of the light field projection modeling. Subsequently, we use this framework to introduce a projection model formulation that is specifically designed for the RSLF setting. A more detailed description of the projection model formulation and theoretical analysis are given in the Supplementary material. An overview of the adopted light-field modeling and geometry is shown in Fig. 2. The pose of the camera with respect to the scene expressed in the world coordinates frame $(\mathbf{O}_w, X_w, Y_w, Z_w)$ is $[\mathbf{R} \mid \mathbf{T}] \in \mathbb{SE}(3)$. The camera position defines a new coordinate frame $(\mathbf{O}_c, X_c, Y_c, Z_c)$ with origin placed in the optical center of the main lens. The view plane (MLA plane) has coordinate frame $(\mathbf{O}, X, Y, Z)$ expressed in relation to the camera frame by a pure translation (O_x, O_y, d) expressed by the transformation matrix $\mathbf{D} \in SE(3)$, with $\mathbf{O} = (O_x, O_y, 0)^\top$ the intersection of the optical axis and the view plane, and d the distance between the optical center of the main lens and the view plane. The micro-image local frames (x, y) are attached to the image plane and are dependent of the considered viewpoint, as shown in Fig. 2. Given a point in the world homogeneous coordinates frame ${}^w\widetilde{\mathbf{p}} = (x_w, y_w, z_w, 1)^\top$ and the matrices ${}^c\mathbf{M}_w$ (the transformation between the camera to world coordinates) and $\mathbf{K}_c$ (thin lens projection matrix), we can obtain the virtual projection of the 3D point inside the camera as:

$$\lambda_c \widetilde{\mathbf{p}} = \mathbf{D}\mathbf{K}_c {}^c\mathbf{M}_w {}^w\widetilde{\mathbf{p}}, \tag{1}$$

with λ_c a scaling factor. For a given viewpoint $\mathbf{c} = (s, t, 0)^\top$, *i.e.* a projection center, the projection of the point $\widetilde{\mathbf{p}}$ onto the image plane is given by:

$$\lambda_s^{s,t} \widetilde{\mathbf{m}}^{s,t} = \mathbf{K}_s^{s,t} \widetilde{\mathbf{p}} = \begin{bmatrix} f & 0 & 0 & -fs \\ 0 & f & 0 & -ft \\ 0 & 0 & 1 & 0 \end{bmatrix} \widetilde{\mathbf{p}}, \tag{2}$$

with $\widetilde{\mathbf{m}}^{s,t} = (x^{s,t}, y^{s,t}, 1)^\top$ the final LF image points, f the focal length of the micro-lenses and λ_s a scaling factor.

Rolling shutter modeling. We follow a similar formalism to Ait-Aider *et al*. [3] to represent an RS imaging process. The main insight is to define a projection model dependent of the camera pose and as a function of the micro-image line t being observed. We adopt the hypothesis that the speeds $(\mathbf{v}, \Omega)$ are constant during the LF acquisition. Adapting the initial projection defined in Eq. (1) for the RS we have:

$$\lambda_c \widetilde{\mathbf{p}} = \mathbf{D}\mathbf{K}_c \begin{bmatrix} \delta\mathbf{R}^t {}^c\mathbf{R}_w & {}^c\mathbf{T}_w + \delta\mathbf{T}^t \\ \mathbf{0}^\top & 1 \end{bmatrix} {}^w\widetilde{\mathbf{p}}, \tag{3}$$

with $\delta\mathbf{R}^t = \mathbf{a}\mathbf{a}^\top(1 - \cos(\Omega\tau t)) + \mathbf{I}\cos(\Omega\tau t) + [\mathbf{a}]_\wedge \sin(\Omega\tau t)$, and $\delta\mathbf{T}^t = \mathbf{v}\tau t$, where $\mathbf{a}$ (axis of rotation), Ω (angular velocity) and $\mathbf{v}$ (linear velocity) describe the uniform movement of the camera coordinate frame with respect to the world coordinates frame and τ the time between the acquisition of two lines of the micro-images. The full Rolling Shutter LF projection from Eq. (3) that projects the 3D point ${}^w\widetilde{\mathbf{p}}_i$ to an image point $\mathbf{m}_i^{s,t} \in \mathbb{P}^2$, given a center of projection $\mathbf{c} = (s, t, 0)^\top$ is then

$$\lambda \mathbf{m}_i^{s,t} = \mathbf{K}_s^{s,t}\mathbf{D}\mathbf{K}_c[\delta\mathbf{R}^t {}^c\mathbf{R}_w \mid {}^c\mathbf{T}_w + \delta\mathbf{T}^t]^w\widetilde{\mathbf{p}}_i, \tag{4}$$

where $\mathbf{K}_s^{s,t}\mathbf{D}\mathbf{K}_c$ can be represented as a single compact intrinsic Rolling Shutter LF tensor:

$$\mathbf{K}_s^{s,t}\mathbf{D}\mathbf{K}_c = \begin{bmatrix} f & 0 & -\frac{f}{F}(O_x - s) & f(O_x - s) \\ 0 & f & -\frac{f}{F}(O_y - t) & f(O_y - t) \\ 0 & 0 & 1 - \frac{d}{F} & d \end{bmatrix}, \tag{5}$$

with F the focal length of the main lens. This formulation has the strong advantage of being generic and represent both GS and RS configurations. Another advantage is that all parameters of this unified model can be calibrated with existing techniques such as Bok *et al*. [5] for the intrinsic parameters and Meingast *et al*. [24] for the rolling shutter rate.

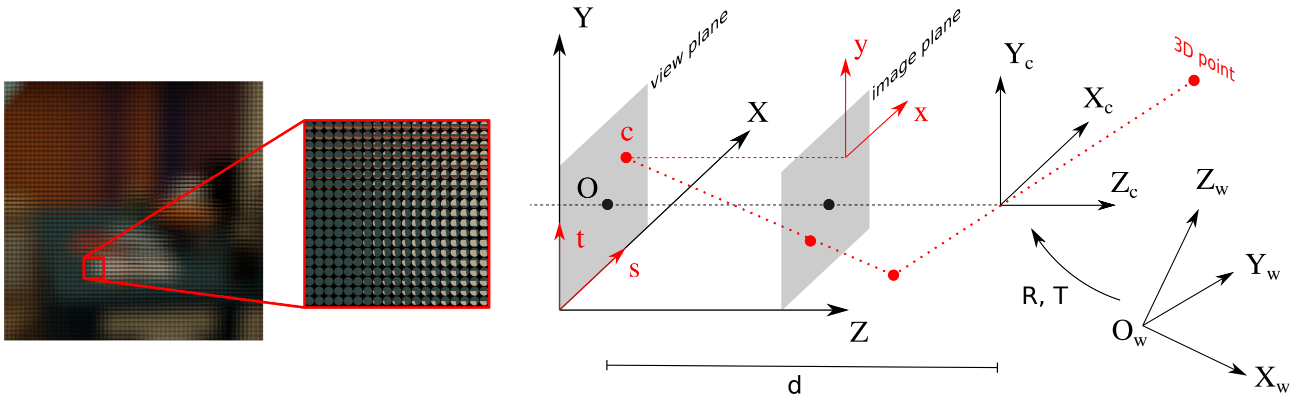

Figure 2. *left* - A raw plenoptic image from a near viewpoint in the scene shown in Fig. 1 and a detail of the micro-images. *right* - The adopted LF coordinate frames: The 3D point is projected in a 3D virtual scene by thin lens projection, then on the 2D image plane by pinhole projection which coordinate frame depends on the considered viewpoint.

Generalization and particular cases. When $\tau = 0$ (*i.e.*, no temporal delay between two consecutive lines), this model can be simplified to a GS light-field camera as the position of the sensor with respect to the scene will be identical for any t. The situation where the camera has no velocity with respect to the scene can also be seen as GS for similar reasons. The proposed projection model in Eq. (4) generalizes to a conventional pinhole camera projection in the case where the MLA is composed of a unique lens. More details are given in the Supplementary material.

3.1. Scene Structure and Motion Estimation

For a given set of matching points inside a calibrated LF and assuming that all points belongs to the same rigid scene in a uniform movement with respect to the camera, we can recover the position of the points in the 3D world as well as their motion at a given time. We will use a re-projection error minimization in order to find jointly these 3D coordinates ${}^{w}\widetilde{\mathbf{p}}_i$ and the dynamic parameters describing the movement of the camera.

Linear initialization. A classical multi-view stereo strategy is applied to provide a first estimate of the 3D points in the scene. In order to reduce the influence of the RS effect, we apply the multi-view stereo only horizontally, thereby ensuring that each measured point ${}^{w}\widetilde{\mathbf{p}}_i \in \mathbb{R}^3$ is captured at the same instant. From the experiments, this first estimate is essential to allow convergence of the following non-linear optimization.

Non-linear bundle adjustment. Using this 3D initialization of the observed points in the light field and our projection model, we design a re-projection error in order to recover simultaneously a refined structure of the scene and the camera motion. From our projection in Eq. (4) we compute the point:

$$\left(u_i^{s,t}, v_i^{s,t}, w_i^{s,t}\right)^T = \mathbf{K}_s^{s,t}\mathbf{D}\mathbf{K}_c[\delta\mathbf{R}^t \mid \delta\mathbf{T}^t]\,{}^{w}\widetilde{\mathbf{p}}_i \tag{6}$$

and deduce the Euclidean pixel coordinates, the scalars $x_i^{s,t}$ and $y_i^{s,t}$, computed as:

$$\begin{aligned} x_i^{s,t} &= \frac{u_i^{s,t}}{w_i^{s,t}} := \xi_{(x)}^{s,t}({}^{w}\widetilde{\mathbf{p}}_i, \Omega, \mathbf{a}, \mathbf{v}), \qquad \text{and} \\ y_i^{s,t} &= \frac{v_i^{s,t}}{w_i^{s,t}} := \xi_{(y)}^{s,t}({}^{w}\widetilde{\mathbf{p}}_i, \Omega, \mathbf{a}, \mathbf{v}), \end{aligned} \tag{7}$$

with $\xi^{s,t}$ the projection function that, given a center of projection $\mathbf{c} = (s, t, 0)^\top$, return the coordinates of the image point with respect to its static position and its movement. The re-projection error function is obtained by computing the distance between the measured points $\tilde{\mathbf{m}}_i^{s,t_0}(\tilde{x}_i{}^{s,t}, \tilde{y}_i{}^{s,t})$ and the coordinates estimated with $\xi_{(x)}^{s,t}$ and $\xi_{(y)}^{s,t}$ from Eq. (7) as follows:

$$\begin{aligned} \epsilon = \sum_s \sum_t \sum_i &\left(\tilde{x}_i{}^{s,t} - \xi_{(x)}^{s,t}({}^{w}\widetilde{\mathbf{p}}_i, \Omega, \mathbf{a}, \mathbf{v})\right)^2 \\ &+ \left(\tilde{y}_i{}^{s,t} - \xi_{(y)}^{s,t}({}^{w}\widetilde{\mathbf{p}}_i, \Omega, \mathbf{a}, \mathbf{v})\right)^2. \end{aligned} \tag{8}$$

This problem has three unknowns for $\Omega\mathbf{a}$, three unknowns for $\mathbf{v}$, and three unknowns for every ${}^{w}\widetilde{\mathbf{p}}_i$. It can be solved if at least four non-coplanar 3D points can be observed, meaning that they need to be located at least an LF image in two different lines and at two different columns of micro-images.

Regularization. For the moment, the rotation axis $\mathbf{a}$ in Eq. (8) is defined to pass through the world origin, which corresponds to the optical center of the main lens. However, this is generally not the instantaneous center of rotation of the movement between the camera and the scene. To ease the description of the movement, we regularize the optimization by providing a "center of rotation" $\mathbf{g}$ to the point cloud. This allows to express all points ${}^{w}\mathbf{p}_i$ in a new coordinate frame centered on this center of rotation. It also allows to compute normalized points ${}^{n}\mathbf{p}_i$ from which the coordinates are lying in the range $[-1, 1]$. The final non-linear adapted re-projection error from Eq. (8) using the normalized points and the center of rotation regularization is then:

$$\epsilon = \sum_s \sum_t \sum_i \left(\tilde{x_i}^{s,t} - {}^{n}\xi_{(x)}^{s,t}({}^{n}\mathbf{p}_i, \mathbf{g}, \Omega, \mathbf{a}, \mathbf{v}) \right)^2 + \left(\tilde{y_i}^{s,t} - {}^{n}\xi_{(y)}^{s,t}({}^{n}\mathbf{p}_i, \mathbf{g}, \Omega, \mathbf{a}, \mathbf{v}) \right)^2, \tag{9}$$

where ${}^{n}\xi_{(x)}^{s,t}$ and ${}^{n}\xi_{(y)}^{s,t}$ are designed to handle the normalization, and $\mathbf{g}$ is also optimized in the loop so that the model is able to find the optimal center of rotation on-the-fly. Further details on the optimization are provided in the supplementary materials.

4. Rolling Shutter Light-Field Dataset

Despite the potential of rolling shutter plenoptic cameras, to the best of the authors' knowledge, all existing LF datasets are done assuming a global shutter hypothesis [2, 9, 26, 29]. Unfortunately, there is no public data available showing the rolling shutter effect on light-field images. Therefore, we have carefully designed and collected a new dataset with seven different synthetic scenes build on Blender, containing notably pseudo-real scenes created from Habitat-Matterport benchmark [28]. This new dataset (inspired by the HCI 4D LF benchmark [14]) is composed of four photo-realistic scenes from Matterport and three synthetic ones (as the examples depicted in Fig. 1 and Fig. 4). We provide, per scene, the following data:

(i) Config files with camera settings and disparity ranges.

(ii) Different motion scenarios:

- *GS*: This is the static configuration. It allows to have a good measure of the performance difference with or without RS distortion by having the same scene in both scenarios. It is equivalent to a GS light field.
- *slow*: The motions affect the image enough to affect largely the perception of the scene geometry.
- *fast*: The linear and angular camera velocities are about three times more important than for the *slow* motions.

We collect 11 light field sequences per scene (1 *GS*, 5 *slow*, 5 *fast*). Please see the table in the supplementary with the velocity intervals for each motion scenario.

(iii) Each light field is of dimension $9 \times 9 \times 512 \times 512 \times 3$, which is equivalent to a light field captured from a plenoptic camera with a 512×512 micro-lense array and 9×9 micro-images.

(iv) A depth map corresponding to the geometry of the scene at middle time of exposition (*i.e.*, the pose of the camera during the acquisition of the center line).

We believe this dataset has the potential to help the evaluation and to promote futher investigation of RS applications for scene analysis with light fields. Visualizations and additional details of the dataset are given in the Supplementary material.

5. Experiments

Metrics and competitors. We have selected two representative algorithms for comparison: the model-based approach of Jeon *et al*. [16], and a recent learning-based 3D estimation from LF of Wang *et al*. [35]. The comparison is done in both GS and RS scenarios for all methods with the aim of fair conditions for the competitors. Six commonly used metrics are selected for the evaluation *abs rel*, *abs diff*, *RMS*, $\delta < 1.25$, $\delta < 1.25^2$ and $\delta < 1.25^3$. *abs rel* is the absolute difference between the estimation and the ground truth (gt), normalized by the gt. *abs diff* is the absolute difference between the estimation and the gt. *RMS* is the Root Mean Square Error between the estimation and the gt. $\delta < 1.25$, $\delta < 1.25^2$ and $\delta < 1.25^3$ are respectively the proportion of the points in a range of 1.25 times the gt, 1.25^2 times the gt and 1.25^3 times the gt.

5.1. Results

The evaluation and averaged metrics for all scenes (and different motion conditions) are shown in Tab. 1. We can observe the proposed method achieves the best scores overall in several of the considered metrics (e.g., "abs rel" and "abs diff"), and notably for all metrics of the "fast" sequences' split. We can also notice that it has even a competitive performance to the recent competitors in the GS scenario. This aspect will be further investigated in the ablation and sensitivity analysis. As we can observe, the two competitors perform far worse when motion is present. The detailed metrics for three representative scenes considering the eleven light fields sequences per scene (1 *GS*, 5 *slow*, 5 *fast*) are shown in Tab. 2, where we can see that our method performs better in most cases. Please check some qualitative examples of the obtained shape reconstructions for these three scenes shown in Fig. 4. We alternate, for these three scenes, the GS case and a RS case with high velocity (motion scenario number 9). We can clearly see the capacity of our algorithm to model the RS deformations. In the scene "bedroom", motion scenario 9, (the last line of Fig. 4), one can clearly notice from visual inspection the compensation

	abs rel ↓			abs diff ↓			RMS ↓		
Method	GS	slow	fast	GS	slow	fast	GS	slow	fast
Jeon-CVPR [16]	**0.040**	0.053	0.110	**0.027**	0.036	0.071	**0.035**	**0.048**	0.092
OACC-Net [35]	0.143	0.171	0.196	0.091	0.109	0.125	0.109	0.128	0.144
Ours	**0.040**	**0.041**	**0.059**	0.031	**0.032**	**0.044**	0.046	0.051	**0.064**

	$\delta < 1.25$ ↑			$\delta < 1.25^2$ ↑			$\delta < 1.25^3$ ↑		
Method	GS	slow	fast	GS	slow	fast	GS	slow	fast
Jeon-CVPR [16]	**0.993**	**0.976**	0.894	**1.000**	**0.999**	0.973	**1.000**	**1.000**	0.998
OACC-Net [35]	0.767	0.720	0.676	0.959	0.945	0.933	**1.000**	0.997	0.997
Ours	0.958	0.961	**0.949**	0.989	0.988	**0.982**	0.999	0.999	**0.999**

Table 1. Average reconstruction error metrics in different scenarios for all dataset sequences: *GS* (global shutter, equivalent to a static camera scenario), *slow* (RS with small camera linear and angular velocities), and *fast* (RS with camera motion three times higher velocities than in the *slow* case). The upward arrow means that a higher score is better. Our approach is significantly better than the considered methods, and with competitive results even for the GS case. Please see the text for details.

	abs rel ↓										
rabbit	0	1	2	3	4	5	6	7	8	9	10
Jeon-CVPR [16]	0.06	0.08	0.07	0.07	0.07	0.1	0.19	0.12	0.13	0.34	0.39
OACC-Net [35]	0.4	0.48	0.5	0.44	0.38	0.49	0.47	0.48	0.44	0.5	0.5
Ours	**0.03**	**0.03**	**0.03**	**0.02**	**0.03**	**0.03**	**0.03**	**0.03**	**0.02**	**0.03**	**0.03**
table	0	1	2	3	4	5	6	7	8	9	10
Jeon-CVPR [16]	0.03	0.03	0.04	0.03	0.05	0.03	0.05	0.09	0.05	0.17	0.07
OACC-Net [35]	0.17	0.21	0.2	0.19	0.19	0.2	0.19	0.24	0.15	0.25	0.2
Ours	**0.02**	**0.02**	**0.02**	**0.03**	**0.02**	**0.03**	**0.04**	**0.02**	**0.04**	**0.03**	**0.04**
bedroom	0	1	2	3	4	5	6	7	8	9	10
Jeon-CVPR [16]	**0.02**	0.03	**0.02**	0.04	0.06	0.03	0.07	**0.02**	0.11	**0.03**	0.07
OACC-Net [35]	0.03	0.05	0.03	0.06	0.1	0.05	0.13	0.03	0.13	0.05	0.13
Ours	0.03	**0.03**	0.03	**0.03**	**0.03**	**0.03**	**0.04**	0.03	**0.04**	0.04	**0.05**

	$\delta < 1.25$ ↑										
rabbit	0	1	2	3	4	5	6	7	8	9	10
Jeon-CVPR [16]	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	0.82	**1.0**	0.91	0.59	0.35
OACC-Net [35]	0.26	0.08	0.06	0.14	0.29	0.09	0.1	0.1	0.1	0.13	0.1
Ours	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
table	0	1	2	3	4	5	6	7	8	9	10
Jeon-CVPR [16]	**1.0**	**1.0**	0.99	**1.0**	**1.0**	1.0	0.97	0.96	0.99	0.76	0.94
OACC-Net [35]	0.69	0.6	0.64	0.64	0.63	0.59	0.67	0.55	0.79	0.5	0.65
Ours	0.995	0.995	**0.99**	**1.0**	0.99	**1.0**	**1.0**	**0.99**	**0.99**	**0.98**	**1.0**
bedroom	0	1	2	3	4	5	6	7	8	9	10
Jeon-CVPR [16]	**1.0**	1.0	**1.0**	**1.0**	0.97	1.0	0.99	**1.0**	0.89	**1.0**	0.94
OACC-Net [35]	1.0	0.98	1.0	0.99	0.93	0.98	0.8	1.0	0.77	**1.0**	0.79
Ours	**1.0**	**1.0**	**1.0**	0.999	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**0.99**

Table 2. Detailed reconstruction error metrics for three representative scenes “rabbit”, “table” and “bedroom” considering the eleven different motion scenarios (from 0 to 10) of the dataset. The upward arrow means that a higher score is better.

done on the painting (the rectangle is less stretched). Unfortunately, this qualitative observation is not highlighted in the detailed quantitative metrics Tab. 2. Indeed, the painting is stretched in the estimation given by the competitors, but is still close to the wall plane, resulting in similar scores. However, the proposed formulation is at least twice as accurate than the competitors for the other two scenes detailed in Tab. 2 for all motion profiles, accordingly to the average scores for all sequences shown in Tab. 1. The detailed results for all sequences and scenes are included for completeness in the Supplementary materials due to page space limitations.

Finally, we analyse the performance of the approaches in the easy to understand ‘‘chart” scene as shown in the quantitative results from Tab. 3 and visualizations in Fig. 3. Similarly to all other scenes, it is composed of eleven light fields (1 *GS*, 5 *slow*, 5 *fast*), where a double checkerboard pattern is joint in a 90° angle configuration. Our method achieves the best scores for every metric in both the *slow* and *fast* scenarios. However, we can also obtain competitive results to both strong competitors in the case of GS. We can also notice that sometimes our obtained estimation is more accurate when the camera is moving slowly than when the camera is static. This will be discussed in the ablation study Sec. 5.2. Tab. 3 also indicates that our method slightly degrades with the augmentation of the camera speed, but it still considerably outperforms all the competitors in the *fast* scenarios for the four first metrics. Fig. 3 shows some qualitative examples of the three methods in the different scenarios and the associated point clouds. We can observe how our method is still capable of fitting the object shape even with the presence of RS and fast camera motions. Looking at the object 3D reconstruction results obtained by the other methods, we can clearly observe deformation effects caused by the misinterpretation of the RS checkerboard images. These degradation of performance can be explained if we observe that the computed disparity maps of the competitors map the distortions of the scene due to RS from the center view. They also interpret the movement of the camera between vertically distant views only as spatial disparity. Thus, if a point moves vertically downwards during acquisition, it will have a bigger disparity than it should (between two viewpoints, where the point is moving because of changes in point of view but also because of its own movement). Inversely, if a point moves vertically upwards during acquisition, it will have a smaller disparity than it should. These two effects contribute to degrade the performance of GS-designed algorithms in the estimation of the 3D geometry of the scene.

Method	abs rel ↓			abs diff ↓			RMS ↓		
	GS	slow	fast	GS	slow	fast	GS	slow	fast
Jeon-CVPR [16]	**0.003**	0.013	0.049	**8.464**	30.293	76.824	**17.489**	47.647	129.720
OACC-Net [35]	**0.003**	0.013	0.051	12.214	30.882	79.938	26.197	54.799	140.215
Ours	0.004	**0.003**	**0.003**	13.692	**15.395**	**23.754**	22.146	**25.327**	**44.791**

Method	$\delta < 1.25$ ↑			$\delta < 1.25^2$ ↑			$\delta < 1.25^3$ ↑		
	GS	slow	fast	GS	slow	fast	GS	slow	fast
Jeon-CVPR [16]	**1.000**	0.923	0.745	**1.000**	0.992	0.898	**1.000**	0.995	0.991
OACC-Net [35]	**1.000**	0.922	0.730	**1.000**	0.993	0.895	**1.000**	0.996	0.939
Ours	0.988	**0.982**	**0.973**	0.996	**0.999**	**0.995**	**1.000**	**1.000**	**0.998**

Table 3. Detailed reconstruction error metrics in different scenarios for the "chart" sequence: *GS* (global shutter, equivalent to a static camera scenario), *slow* (RS with small camera linear and angular velocities), and *fast* (RS with camera motion three times higher velocities than in the *slow* case). The upward arrow means that a higher score is better. Our approach performed significantly better than the two recent considered methods.

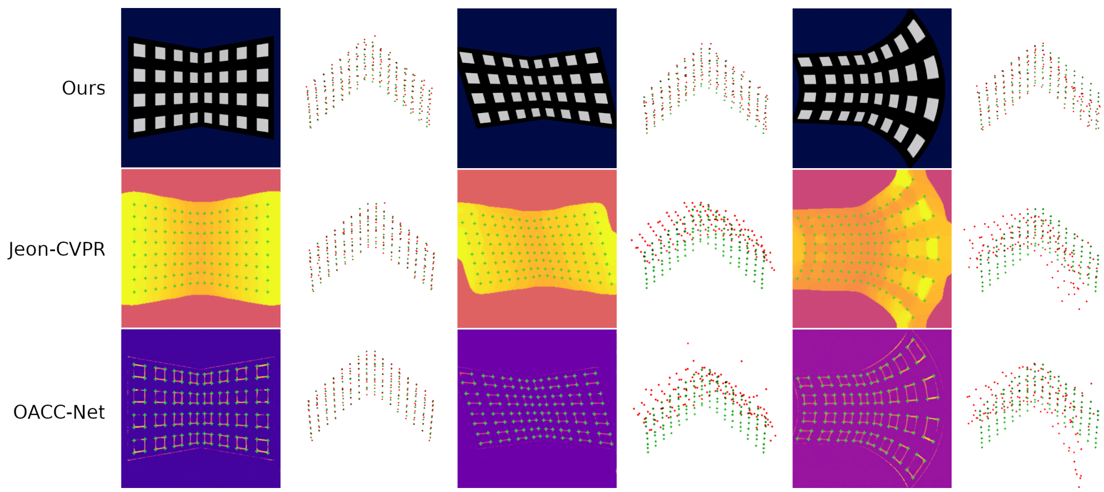

Figure 3. Qualitative examples of reconstruction for different motion scenarios for the "chart" sequence. The "GS" scenario on the left. A "slow" scenario in the middle. A "fast" scenario on the right. - *first column*: The central view of the LF, the disparity map of Jeon-CVPR [16], the disparity map of OACC-Net [35]. - *Second column*: The 3D point clouds (red dots) obtained for our method, Jeon-CVPR [16], OACC-Net [35]. Despite the fact that the images look different, due to the rolling shutter effect, the reconstruction is supposed to give the same result (in green crosses in the point clouds).

5.2. Ablation study

We performed different ablation studies in order to evaluate the relevance of the different parts of the method. In the first ablation, we retained two major components for evaluation, the contribution of i) linear initialization strategy (No Init.), and ii) the regularization (No Reg.) as shown in Tab. 4. For the ablation of the initialization, we initialized the optimization Eq. (9) with all the points clustered in a position near the center of mass of the point cloud we should have found with the linear initialization. We show in Tab. 4 that, even after convergence, the solution is still far from correct. For the ablation of the regularization, we see that the method without the regularization gives worst results in the RS scenarios. These evaluations confirm the importance of these components in the designed method.

A second ablation study was designed to evaluate the performance of our method without the RS modelling (Ours No RS) depicted in Tab. 5. By modeling the RS effect we also have additional degrees of freedom that lead to a slight degradation of the results when compared to a GS scheme for the GS scenes. We performed an evaluation to verify the effect of constraining the dynamic degrees of freedom ($\Omega = 0$ and $\mathbf{v} = \mathbf{0}$) in case of GS would result in the estimation. The results in Tab. 5 show an improvement on all the metrics of up to about 6%. This concurs with the aforementioned hypothesis. The obtained performance is on par with the competitors which are specifically designed for GS settings.

5.3. Discussion

From the experiments, we can observe that our method is capable of handling different camera motions and provides

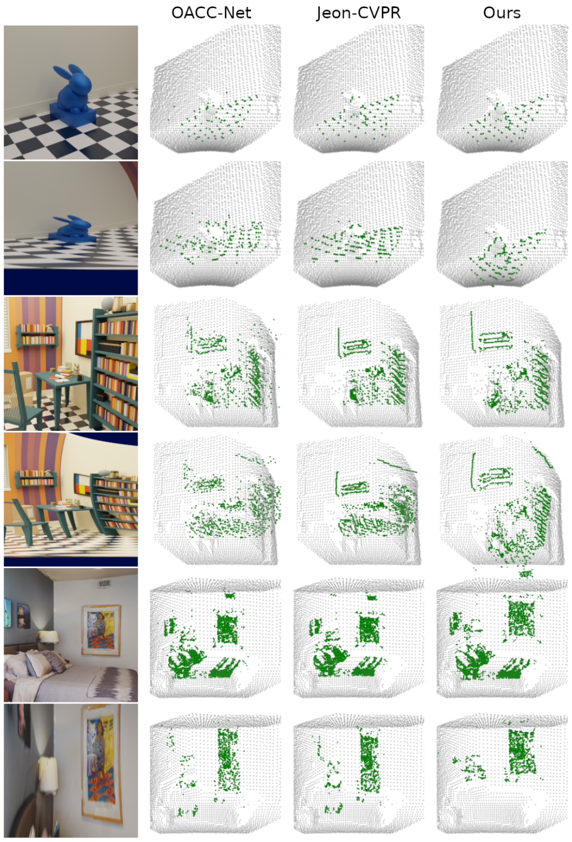

Figure 4. Some central views and associated point cloud reconstructions for the scenes and results shown in Tab. 2. From right to left, OACC-Net [35], Jeon-CVPR [16] and Ours. Ground truth points in gray and estimated in green.

improved scene structure estimates. The proposed model is designed to handle rigid scenes, yet it can estimate the structure and motion parameters for 3D scene points independently if at least four image points are available, *i.e.*, to compute a "3D scene flow" from a single LF image. We assumed rigidity in order to compute a common set of dynamic parameters to each point (corresponding to a camera motion in a rigid scene). We believe our strategy could be also extended to handle scenes with dynamic objects independently (or non-rigid) with multiple camera motion hypotheses. The adopted RS projection also assumes that both linear and angular velocities to be uniform during the LF image acquisition (i.e., zero acceleration). However, RS devices, while having a sequential acquisition, usually have a small time of total exposure per frame (about 0.1 s for a 4K image). Therefore the assumption of constant camera speeds during the frame acquisition holds in typical motion-scene scale scenarios. Nevertheless, the proposed approach could still be applied for accelerated motions with a piece-wise decomposition of the plenoptic image in horizontal bands. Such a strategy of piece-wise decomposition in horizontal bands for classic images has been investigated in [22] for a classic monocular RS sensor. The motion and shape estimation could then be done at different time instants and allow to recover more complex scenes (*e.g.*, non-rigid) and motion scenarios.

	RMS ↓		
Abl.	GS	slow	fast
No Init.	0.243	0.242	0.240
No Reg.	**0.045**	0.060	0.086
Full	0.046	**0.051**	**0.064**
	$\delta_1 < 1.25$ ↑		
Abl.	GS	slow	fast
No Init.	0.650	0.646	0.630
No Reg.	**0.969**	0.950	0.895
Full	0.958	**0.961**	**0.949**

Table 4. Reconstruction errors for the ablation study of our method for the initialization and regularization steps.

Abl.	abs rel ↓	abs diff ↓	RMS ↓	$\delta < 1.25$ ↑
Jeon-CVPR [16]	0.040	0.027	0.035	0.993
Ours Full	0.040	0.031	0.046	0.958
Ours No RS	0.040	0.029	0.041	0.976

Table 5. Ablation study of the dynamic motion parameters with a static GS scene.

6. Conclusion

In this paper, we proposed a projection model for a light-field camera equipped with a rolling shutter sensor. This model allows us to jointly estimate the shape and motion on unknown scenes from a single light field image. The approach has been evaluated on different motions and 3D scenes. Furthermore, it does not suffer from shape/motion ambiguity thanks to the relatively reasonable assumption of a row-wise GS. To fill the lack of publicly available rolling-shutter LF data, we created a dataset that includes simulated photo-realistic light fields with different motion scenarios, and we will make it publicly available. We plan to build upon this model to generate denser depth maps and extend the motion estimations to non-rigid scenes. Our proposed model shows improved 3D scene geometry estimates, and we believe that it will inspire further research in this area, notably for applications in the context of robot vision, manipulation and micro-robotics.

Acknowledgements. The authors would like to thank the funding from the French "Investissements d'Avenir" program, project ISITE-BFC, contract ANR-15-IDEX-03, by the Conseil Régional BFC from the project ANER-MOVIS and by "Grand Prix Scientifique 2018, Fond. Ch. Defforey-Institut de France".

References

[1] Edward H Adelson and John YA Wang. Single lens stereo with a plenoptic camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2):99–106, 1992. 1, 2

[2] Vamsi Kiran Adhikarla, Marek Vinkler, Denis Sumin, Rafał Mantiuk, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. Towards a quality metric for dense light fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 5

[3] Omar Ait-Aider, Nicolas Andreff, Jean Marc Lavest, and Philippe Martinet. Simultaneous object pose and velocity computation using a single view from a rolling shutter camera. In *Eur. Conf. Comput. Vis.*, 2006. 1, 3

[4] Omar Ait-Aider and François Berry. Structure and kinematics triangulation with a rolling shutter stereo rig. In *Int. Conf. Comput. Vis.*, 2009. 2, 3

[5] Yunsu Bok, Hae-Gon Jeon, and In So Kweon. Geometric calibration of micro-lens-based light field cameras using line features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(2):287–300, 2016. 2, 3

[6] Robert C Bolles, H Harlyn Baker, and David H Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *Int. J. Comput. Vis.*, 1(1):7–55, 1987. 2

[7] Caroline Conti, Luís Ducla Soares, and Paulo Nunes. Dense light field coding: A survey. *Access*, 8:49244–49284, 2020. 1

[8] Don Dansereau and Len Bruton. Gradient-based depth estimation from 4d light fields. In *Int. Symposium on Circuits and Systems*. IEEE, 2004. 2

[9] Donald G. Dansereau, Bernd Girod, and Gordon Wetzstein. LiFF: Light field features in scale and depth. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 5

[10] Bin Fan, Yuchao Dai, and Mingyi He. Sunet: symmetric undistortion network for rolling shutter correction. In *Int. Conf. Comput. Vis.*, 2021. 3

[11] Rodrigo Ferreira and Nuno Goncalves. Fast and accurate micro lenses depth maps for multi-focus light field cameras. In *German Conf. on Pattern Recog.* Springer, 2016. 2

[12] Todor G Georgiev and Andrew Lumsdaine. Focused plenoptic camera and rendering. *J. of Electronic Imaging*, 19(2):021106, 2010. 2

[13] Stefan Heber and Thomas Pock. Convolutional networks for shape from light field. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2

[14] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conf. on Comput. Vis.*, pages 19–34. Springer, 2017. 5

[15] Ivo Ihrke, John Restrepo, and Lois Mignard-Debise. Principles of light field imaging: Briefly revisiting 25 years of research. *Sign. Proc. Magazine*, 33(5):59–69, 2016. 1

[16] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon. Accurate depth map estimation from a lenslet light field camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2, 5, 6, 7, 8

[17] Alexandre Karpenko, David Jacobs, Jongmin Baek, and Marc Levoy. Digital video stabilization and rolling shutter correction using gyroscopes. *CSTR*, 1(2):13, 2011. 3

[18] Yizhen Lao, Omar Ait-Aider, and Helder Araujo. Robustified structure from motion with rolling-shutter camera using straightness constraint. *Pattern Recognition Letters*, 111:1–8, 2018. 2

[19] Yizhen Lao, Omar Ait-Aider, and Adrien Bartoli. Solving rolling shutter 3d vision problems using analogies with non-rigidity. *Int. J. Comput. Vis.*, 129(1):100–122, 2021. 3

[20] Chia-Kai Liang, Li-Wen Chang, and Homer H Chen. Analysis and compensation of rolling shutter effect. *IEEE Trans. Image Process.*, 17(8):1323–1330, 2008. 3

[21] Haiting Lin, Can Chen, Sing Bing Kang, and Jingyi Yu. Depth recovery from light field using focal stack symmetry. In *Int. Conf. Comput. Vis.*, 2015. 2

[22] Ludovic Magerand and Adrien Bartoli. A generic rolling shutter camera model and its application to dynamic pose estimation. In *Int. symposium on 3D Data Proc., Visualiz. and Transmis.*, 2010. 8

[23] Maxime Meilland, Tom Drummond, and Andrew I Comport. A unified rolling shutter and motion blur model for 3d visual registration. In *Int. Conf. Comput. Vis.*, 2013. 3

[24] Marci Meingast, Christopher Geyer, and Shankar Sastry. Geometric models of rolling-shutter cameras. *arXiv preprint cs/0503076*, 2005. 1, 3

[25] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. *Light field photography with a hand-held plenoptic camera*. PhD thesis, Stanford University, 2005. 1

[26] Said Pertuz, Edith Pulido-Herrera, and Joni-Kristian Kamarainen. Focus model for metric depth estimation in standard plenoptic cameras. *J. of Photogrammetry and Remote Sensing*, 144:38–47, 2018. 5

[27] Christian Perwass and Lennart Wietzke. Single lens 3d-camera with extended depth-of-field. In *Human Vis. and Elect. imaging*. SPIE, 2012. 2

[28] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied ai. In *Adv. Neural Inform. Process. Syst.*, 2021. 5

[29] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *Int. Confe. on Qual. of Multimed. Exp.*, 2016. 5

[30] Olivier Saurer, Kevin Koser, Jean-Yves Bouguet, and Marc Pollefeys. Rolling shutter stereo. In *Int. Conf. Comput. Vis.*, 2013. 3

[31] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[32] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Adv. Neural Inform. Process. Syst.*, 2021. 2

[33] Pratul P Srinivasan, Ren Ng, and Ravi Ramamoorthi. Light field blind motion deblurring. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[34] Michael W Tao, Pratul P Srinivasan, Jitendra Malik, Szymon Rusinkiewicz, and Ravi Ramamoorthi. Depth from shading, defocus, and correspondence using light-field angular coherence. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2

[35] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Wei An, and Yulan Guo. Occlusion-aware cost constructor for light field depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 5, 6, 7, 8

[36] Sven Wanner and Bastian Goldluecke. Globally consistent depth labeling of 4d light fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2012. 2

[37] Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field image processing: An overview. *J. of Selected Topics in Sign. Proc.*, 11(7):926–954, 2017. 1

[38] Niclas Zeller, Franz Quint, and Uwe Stilla. Depth estimation and camera calibration of a focused plenoptic camera for visual odometry. *J. of Photogrammetry and Remote Sensing*, 118:83–100, 2016. 2

[39] Qi Zhang, Chunping Zhang, Jinbo Ling, Qing Wang, and Jingyi Yu. A generic multi-projection-center model and calibration method for light field cameras. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2539–2552, 2018. 2

[40] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Comput. Vis. and Image Underst.*, 145:148–159, 2016. 2