Learning Reliable Gradients from Undersampled Circular Light Field for 3D Reconstruction

Zhengxi Song, Xue Wang, Hao Zhu, Guoqing Zhou and Qing Wang, *Senior Member, IEEE*

Abstract—The paper presents a 3D reconstruction algorithm from an undersampled circular light field (LF). With an ultra-dense angular sampling rate, every scene point captured by a circular LF corresponds to a smooth trajectory in the circular epipolar plane volume (CEPV). Thus per-pixel disparities can be calculated by retrieving the local gradients of the CEPV-trajectories. However, the continuous curve will be broken up into discrete segments in an undersampled circular LF, which leads to a noticeable deterioration of the 3D reconstruction accuracy. We observe that the coherent structure is still embedded in the discrete segments. With less noise and ambiguity, the scene points can be reconstructed using gradients from reliable epipolar plane image (EPI) regions. By analyzing the geometric characteristics of the coherent structure in the CEPV, both the trajectory itself and its gradients could be modeled as 3D predictable series. Thus a mask-guided CNN+LSTM network is proposed to learn the mapping from the CEPV with a lower angular sampling rate to the gradients under a higher angular sampling rate. To segment the reliable regions, the reliable-mask-based loss that assesses the difference between learned gradients and ground truth gradients is added to the loss function. We construct a synthetic circular LF dataset with ground truth for depth and foreground/background segmentation to train the network. Moreover, a real-scene circular LF dataset is collected for performance evaluation. Experimental results on both public and self-constructed datasets demonstrate the superiority of the proposed method over existing state-of-the-art methods.

Index Terms—3D reconstruction, Circular light field, CNN+LSTM, Circular epipolar plane volume (CEPV).

1 Introduction

High-fidelity 3D reconstruction, which is expected to recover real-world objects efficiently, accurately, and more importantly omnidirectionally, has extensive applications in movie and game industries as well as in architecture, archaeology, arts, and many other areas [1]. As an emerging light field photography technique, the circular light field forms an image volume with regular grids in a circular arrangement. Compared with camera array based light fields, the circular LF can determine an object with a full 360°view (Fig.1 (a)). With a high angular sampling rate (i.e., usually 720 views at least to enable the redundancy), every captured scene point corresponds to a continuously sine-shaped curve in the circular epipolar plane image (CEPI). When a standard perspective camera is considered, the trajectory (Fig.1 (b)) is not just confined in a single CEPI but also moves in a certain 3D circular epipolar plane volume (CEPV) [2]. By exploiting the geometric characteristics of the coherent structure in the CEPV, current methods [3], [4], [5], [6] achieve remarkable 3D reconstruction performances from densely-sampled circular LFs.

To maintain the continuous CEPV-trajectories, the disparity of an image feature must be smaller than the frequency of texture around the feature [5], [6]. To accurately estimate the local gradients of the coherent structure, several methods [5], [6] require at least 3600 views located on a circle. Similarly, the Hough transform-based methods [3], [4], which rely on the local structure tensor and the binary edge map of the CEPIs, also require at least 720 views (Fig.1 (c)). However, establishing a circular LF imaging system with a high angular sampling rate is cumbersome and expensive, severely limiting practical applications. Moreover, when the angular sampling rate decreases, the original smooth trajectories will be replaced with discontinuous segments. The discontinuation will cause ambiguity for local orientation estimation, especially near the intersections or gaps in the trajectory (Fig.1 (d)), and further hinder accurate depth estimation.

To overcome these issues, we focus on estimating reliable gradients along CEPV-trajectories to reconstruct 3D objects from an undersampled circular LF. Despite the discontinuities and aliasing effect revealed in the CEPIs, the coherent structure still can be recovered due to the relation between motion parallax of different angular sampling LF. Inspired by the work for light field super resolution [7], we formulate the coherent structure as *3D predictable series*, such that both the trajectories and their gradients are differentiable. Therefore, it is possible to learn the mapping from the CEPV with a lower sampling rate to the gradients under a higher rate.

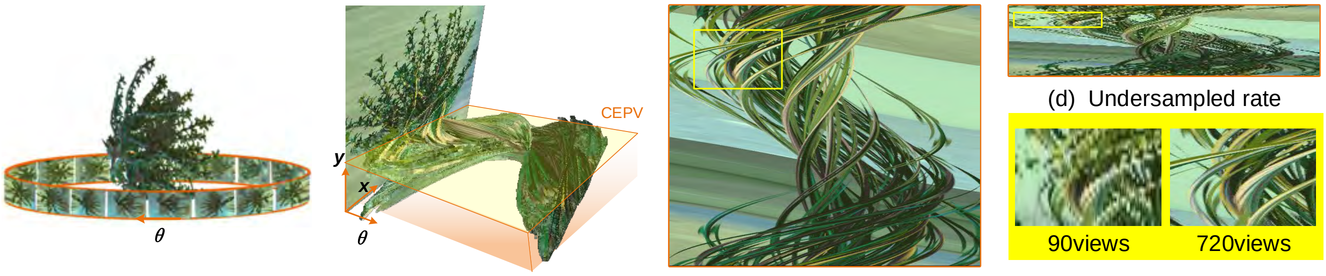
Specifically, we design a mask-guided CNN+LSTM network to compute the local gradients of the CEPV-trajectories by learning the above mapping. The core idea is that the use of reliable segments only, rather than the full CEPV-trajectories, helps increase the quality of the final reconstruction. To this end, a reliable mask is adopted to enable attention on reliable regions by ignoring intersections near the occlusions or gaps with undersampling in the trajectories.

This paper makes the following main contributions.

(1) We propose to formulate the trajectory and its local

- *Z. Song, X. Wang, G. Zhou, Q. Wang (corresponding author) are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China. E-mail: qwang@nwpu.edu.cn.*
- *H. Zhu is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China. E-mail: zhuhao_photo@nju.edu.cn.*

Manuscript received April 19, 2005; revised August 26, 2015.



(a) The Circular LF (b) CEPV-trajectories (c) High angular sampling rate (d) Undersampled rate (e) Different sampling insets

Fig. 1. The slices of CEPV with different angular sampling rates. (c) and (d) refer to the CEPI with fixed y (the yellow plane in (b)). With decreasing angular sampling rate, the smooth CEPV-trajectories (c) degenerate into discrete segments (d). Moreover, the visibility becomes indistinguishable around the intersection of CEPV-trajectory (e) caused by occlusion.

gradients in the CEPV as 3D predictable series, which are differentiable and could be utilized to reconstruct 3D objects from an undersampled circular LF.

(2) We design a novel prediction scheme based on CNN and LSTM to estimate the local gradients of the CEPV-trajectories. To improve the quality of 3D reconstruction, we integrate a reliable mask-based loss to alleviate the impact of unreliable CEPV-regions.

(3) We render a challenging synthetic circular LF dataset and capture a real-scene circular LF dataset for performance evaluation. The datasets will be released to inspire more research in this direction.

2 Related Work

The previous work on 3D reconstruction here we refer to includes different image capture techniques. For each type, we first discuss traditional methods and then learning-based methods.

2.1 Multi-view Stereo

The feature-based multi-view stereo methods utilize coherent features on different views for 3D reconstruction. Furukawa and Ponce [8] adapt an iterative framework by first reconstructing an initial seed of patches from feature matching and then expanding it with coherent constraints. However, the iterative framework leads to large irregular holes in the region where coherent feature points are hard to be detected and matched [9]. To achieve a complete 3D reconstruction, Geosele *et al.* [10] estimate high-quality depth maps and merge them into 3D scene reconstructions. Schonberger et al. propose COLMAP [11], [12] through pixel-wise depth estimation, which achieves good accuracy for diverse scenarios and public benchmarks.

Recent advances in deep learning have interested a series of learnable systems for solving MVS problems. Huang *et al.* [13] propose a DeepMVS system, which adopts the encoder-decoder architecture for feature extraction and formulates depth estimation as a multi-class classification problem. To infer the depth, Yao *et al.* [14] build a 3D cost volume upon the reference camera frustum via the differentiable homography warping. To combine local and global structures, Chen *et al.* [15] propose a feature pyramid structure to fuse multi-level information and then to generate a smooth depth map for 3D reconstruction. Wang *et al.* [16] introduce an iterative multi-scale Patchmatch in an end-to-end trainable architecture for 3D reconstruction with high computation speed and low memory requirement.

2.2 The Linear Light Field

Gortler *et al.* [17] and Levoy [18] reduce the 7-dimensional light field into the 4D Lumigraph and adopt the two-parallel-plane (TPP) model to describe a 4D light field. The structure of the frame-to-frame pixel motion in the EPI provides cues for geometry estimation [19]. Depth can be estimated from the slope of the linear feature in the densely sampled light field. Baker *et al.* [20] show that the geometric characteristic cannot be uniquely determined from the light field when the intensity of light radiated from the scene is constant over an extended region. They prove that the gradient is related to the scene depth via a one-to-one basis. Wanner and Goldluecke [21] employ a higher-order structure tensor of an EPI to obtain a fast and robust local disparity estimation. The EPI-based analysis shows significant advantages for 3D reconstruction [22], such as regular sampling pattern [23], [24], dense angular sampling [25], [26], sub-pixel disparity accuracy [21], [27], [28], and thus becomes increasingly common for LF-based 3D reconstruction.

The ConvNet-based method further improves the light field depth estimation performance. It is adopted by Heber and Pock [29] to learn the mapping between the 4D light field and the corresponding 4D depth field representation in terms of the 2D hyperplane orientation. Then they extend the U-shaped network structure to perform an additional spatial regularization [30]. The EPINET [31] is constructed by exploiting the geometric characteristics of EPIs to estimate depth in narrow-baseline scenarios. Wu *et al.* [32], [33] analyze the aliasing problem caused by large disparity and non-Lambertian effect in undersampled LFs. Li *et al.* [34] propose the LLF-Net for wide-baseline scenarios by incorporating a cost volume and an attention module. The above methods can only reconstruct one side of the object facing the linear LF. Although we can obtain depth from different sides, the 3D point matching and merging procedures are usually time-consuming.

2.3 The Circular Light Field

The circular LF sampling is suitable for omni-directional object reconstruction. Bolles *et al.* [35] start to build an image volume to analyze the 3D position of an object and the spatio-temporal event such as occlusion. Further, they extend their analysis to a wider class of camera motions such as circular or hand-held camera motions. Then Feldmann *et al.* [36] focus on the circular camera motion and carry out the depth-corrected EPI analysis called the image cube trajectory (ICT) analysis. By detecting variation-based color consistency, they try to extract the trajectory of a CEPV curve. Yücer *et al.* [5] analyze local gradient information in the 2D EPI slice with high spatio-angular sampling. They propose a confidence measure to segment the foreground and background based on local gradients. They further assess the reliability of depth estimates using a novel two-sided photo-consistency measure [6]. However, the above methods are based on local feature propagation relying on smooth trajectories and thus require ultra-high angular sampling, i.e., usually thousands of input frames/views.

Vianello *et al.* [4] concentrate on the curve in each CEPI slice and solve the curve function in the Hough transform space. The geometry of the CEPI-curve is analyzed for both orthographic and perspective camera projection models. To properly fit the trajectories using camera parameters, their method requires at least 720 views and is sensitive to noise or degeneration of continuous curves.

To alleviate the distortion problem from the pin-hole projection, Cserkaszky *et al.* [37] simulate a spherical lens to fit the distorted CEPI-curve into a standard sinusoid curve for the synthetic scene. Instead of focusing on the 2D curve in a single CEPI slice, Song *et al.* [2], [38] analyze the 3D feature in the CEPV to improve the robustness. Still, these methods require a high angular sampling rate, i.e., at least 180 views are needed in [2], to preserve continuous and smooth trajectories in the CEPV.

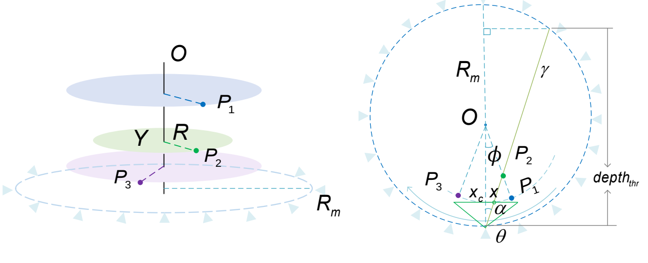
Compared with the initial work at ICME [2], this paper studies the predictable properties of the coherent structure in the undersampled CEPV, whose sampled rate decreases from 180 views to 90 views. A prediction scheme is further proposed to take the gradients from a densely sampled LF as supervision. By adding the attention masks of reliable regions, the ambiguity in the local direction of CEPV-trajectory caused by occlusion and undersampling is mitigated. In addition, more thorough experiments and discussions are provided.

3 Problem Analysis and Formulation

Due to the circular layout, a scene point is projected onto all images with different relative depths. By stacking the images on top of each other, the projected pixels of the same scene point form a 3D trajectory, which lives on a 2D manifold in the CEPV [5].

3.1 3D Predictable Series with Large Disparity

As shown in Fig.2, the scene point P can be expressed using polar coordinates (R, ϕ, Y), where R is the distance between P and the rotational shaft O, ϕ is the phase offset and Y is the vertical coordinate. If the generic scene point P is considered, its trajectory in the CEPV $L(\theta, x, y)$ (see Fig.1(b)) can be derived as [4]:



(a) Side view (b) Top view

Fig. 2. The side view and top view of the circular LF model. P_1, P_2, P_3 are three scene points. P_2 can be expressed using polar coordinates (R, ϕ, Y). λ represents the ray coming from P_2, of which the projected pixel on the view is $p(\theta, x, y)$ in the CEPV L.

$$x(\theta) = f \cdot \frac{R\sin(\theta+\phi)}{R_m - R\cos(\theta+\phi)} + x_c \tag{1a}$$

$$y(\theta) = f \cdot \frac{Y}{R_m - R\cos(\theta+\phi)} + y_c, \tag{1b}$$

where f is the focal length and R_m is the distance between the camera center C and the rotational shaft O. (x_c, y_c) denotes the camera's principal point.

If the scene point P is projected to two adjacent cameras C_{θ_i} and C_{θ_j} at the views θ_i and θ_j (see Fig.3), the disparity in the CEPV between the projected positions $p(\theta_i, x_i, y_i)$ and $p(\theta_j, x_j, y_j)$ can be expressed as follows:

$$\vec{d} = p(\theta_j, x_j, y_j) - p(\theta_i, x_i, y_i). \tag{2}$$

Similar to the EPI of a linear LF, when $|\vec{d}| < 1$, the CEPV-trajectory is continuous. Otherwise, there are gaps in the trajectory and the angular resolution of the circular LF is considered undersampled.

Since the CEPV-trajectory is differential, the missing pixels in the 3D series can be predicted by the partial deviates of Eq.1. Given the pixel $p_0 = (\theta_0, x_0, y_0)$ in the CEPV, the position (x_1^*, y_1^*) of the corresponding pixel p_1 in the predicted view θ_1 can be expressed as follows:

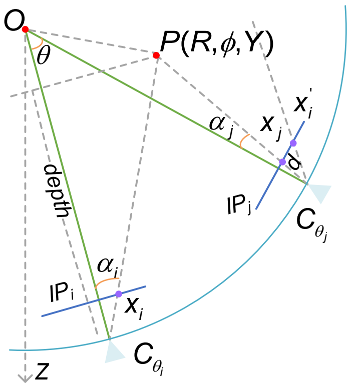
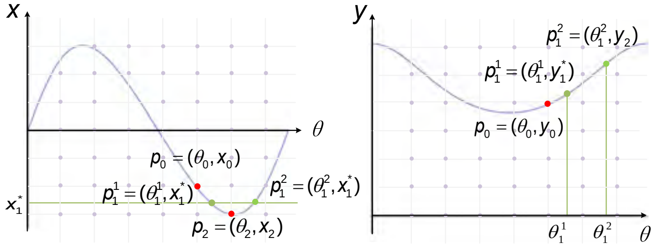


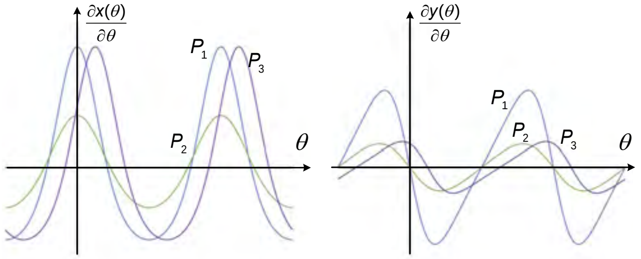
Fig. 3. Optical path of adjacent views in the circular LF. The point P is projected to camera C_{θ_i} with x_i and C_{θ_j} with x_j, respectively. When we draw x_i in C_{θ_j} with $x_i^{'}$, the d between two purple pixels is the disparity.



(a) The *x*-*θ* plane

(b) The y-*θ* plane

Fig. 4. Each CEPV-trajectory can be projected to the $x-\theta$ plane or $y-\theta$ plane by fixing the y-axis or x-axis, respectively. p_0 and p_2 are the pixels in the CEPV; p_1 is the predicted pixel at discontinuity. With a specific x_1^*, both θ_1^1 and θ_1^2 can be solved on the $x-\theta$ plane. Such ambiguity can be removed on the $y-\theta$ plane, where θ_1^1 and θ_1^2 relate to y_1^* and y_2 respectively.



(a)

(b)

Fig. 5. The partial derivatives of $x(\theta)$ and $y(\theta)$ of CEPV-trajectories of the 3 scene points in Fig.2. P_1 and P_2 share the same phase offset but different radii, while P_1 and P_3 share the same radius but different offsets.

$$x_1^* = \int_{\theta_0}^{\theta_1} \frac{\partial x}{\partial \theta} d\theta + x_0 \tag{3a}$$

$$y_1^* = \int_{\theta_0}^{\theta_1} \frac{\partial y}{\partial \theta} d\theta + y_0, \tag{3b}$$

where $\frac{\partial x}{\partial \theta}$, $\frac{\partial y}{\partial \theta}$ are the partial derivatives of x, y with respect to θ, respectively.

To better understand the *3D predictable series*, the CEPV-trajectory is projected to the $x-\theta$ plane and $y-\theta$ plane in the CEPV respectively, as shown in Fig.4. When the pixel position x_1^* is fixed, two corresponding pixels p_1^1, p_1^2 drop into the views θ_1^1, θ_1^2 respectively [38] (see Fig.4(a)). However, such ambiguity can be distinguished by considering the y component. Therefore, we propose to model the CEPV-trajectories as 3D predictable series.

3.2 Depth from 3D Gradients Series

The 3D gradients along the CEPV-trajectory depict the disparity $\vec{d}$ between adjacent views, which is tangential to the CEPV-trajectory on the 2D manifold in the CEPV L. According to Eq.1, the 3D gradients in x and y directions can be modeled by the partial derivatives,

$$\frac{\partial x}{\partial \theta} = f \cdot \frac{R_m R \cos(\theta+\phi) - R^2}{(R_m - R\cos(\theta+\phi))^2} \tag{4a}$$

$$\frac{\partial y}{\partial \theta} = f \cdot \frac{YR\sin(\theta+\phi)}{(R_m - R\cos(\theta+\phi))^2}. \tag{4b}$$

The trajectory function is continuous and differentiable [4], and its second derivative is also differentiable. Following Eq.3a, given the pixel $p=(\theta_0, x_0)$, its partial derivatives in view θ_1 can be predicted as,

$$\left.\frac{\partial x(\theta)}{\partial \theta}\right|_{\theta=\theta_1} = \int_{\theta_0}^{\theta_1} \frac{\partial^2 x(\theta)}{\partial \theta^2} d\theta + \left.\frac{\partial x(\theta)}{\partial \theta}\right|_{\theta=\theta_0}. \tag{5}$$

Fig.5 illustrates the distribution of gradients under different θ, *i.e.*, $(\frac{\partial x}{\partial \theta}, \frac{\partial y}{\partial \theta})$. Both $\frac{\partial x}{\partial \theta}$ and $\frac{\partial y}{\partial \theta}$ move along the CEPV-trajectory rather than stay on a single slice. Therefore, the 3D gradients could be considered as 3D predictable series.

Given a circular LF imaging system, usually R_m is far larger than Y and R. Thus the partial derivative $\frac{\partial x}{\partial \theta}$ dominates the variation (refer to Eq.4). Based on the above 3D predictable series analysis, the depth of a scene point in the reference view θ_i can be deduced from the gradient $\frac{\partial x(\theta)}{\partial \theta}|_{\theta=\theta_i}$ with triangulation,

$$\tan(\alpha_j) = \frac{(x-x_c)}{f} \tag{6a}$$

$$\tan(\alpha_i) = \frac{(x - x_c + \frac{\partial x(\theta)}{\partial \theta}|_{\theta=\theta_i})}{f} \tag{6b}$$

$$depth = \frac{2R_m \cos(\alpha_i)\sin(\theta/2)\cos(\alpha_j - \theta/2)}{\sin(\theta + \alpha_i - \alpha_j)}. \tag{6c}$$

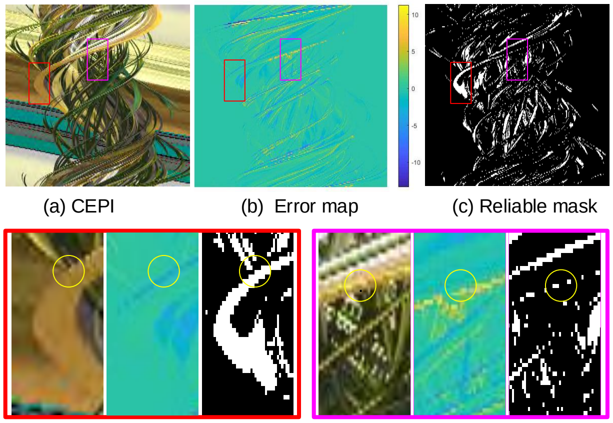
where α_i and α_j are illustrated in Fig.3.

3.3 Problem Formulation

Through triangulation, the depth of the 3D object can be derived from the partial derivatives of the CEPV-trajectory. However, such gradients based on local features suffer from the ambiguity of two folds.

(1) **The undersampling issue**. The smooth trajectories decay into discrete segments and other artifacts when the angular density is undersampled (see Fig.1(c)(d)). Then local gradients will be affected by the additional structure around the discrete segments. The disparities between adjacent views can be approximated by the local gradients from a smooth and continuous CEPV-trajectory. According to the structural property of the trajectory (see Eq.3a), it is possible to learn the mapping from a deteriorated CEPV-trajectory to its gradients from a denser circular LF.

(2) **The occlusion event**. The occlusion-free scene point results in a continuous coherent structure in the CEPV. The unaffected local regions on the CEPV are all reliably available for depth estimation. However, when the trajectories intersect with each other due to occlusion, the ground truth gradient in the reference view is calculated without considering the point's visibility in adjacent views (see Eq.6). Hence, the occlusion regions on the CEPV will bring ambiguity for the network to estimate local gradients (see Fig.6(d)). It is nature to consider excluding occlusion regions out for estimating local gradients.



(a) CEPI (b) Error map (c) Reliable mask

(d) Occlusion event (e) Undersampled situation

Fig. 6. An example of the reliable mask. (a) shows the CEPV-trajectories with 90 views. (b) shows the error map between predicted gradients and the ground truth after 50 training epochs. (c) is the reliable mask with $\epsilon = 0.1$. (d) and (e) correspond to the ambiguity discussed in Sec.3.3.

To reconstruct a 3D object from an undersampled circular LF, we define the ground truth segmentation s as the imaging region of the rays from the objects inside the camera rig. For each ray γ in the observation space to the view (see Fig.2(b)), the foreground threshold $depth_{thr}$ is,

$$depth_{thr} = 2R_m(\cos\alpha)^2. \tag{7}$$

The ground truth segmentation s is computed as follows:

$$s = \begin{cases} 1 & depth_{gt} < depth_{thr} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Consider a given CEPV $L \in \mathbb{R}^{\Theta \times X \times Y}$, where X and Y denote the embedded space defined by the spatial coordinates, and Θ denotes the angular embedded space. The learning-based local gradients can be estimated as follows:

$$[\hat{g}(\theta, x, y), \hat{s}(\theta, x, y)] = f(L(\theta, x, y); \Psi), \tag{9}$$

where $\Psi = \left\{\psi^{(0)}, \psi^{(1)}, ..., \psi^{(K-1)}\right\}$ represents the parameters of the networks, and $f(\cdot)$ describes the learned mapping from the CEPV to trajectory gradients at a higher angular rate. $\hat{g}$ and $\hat{s}$ are the output gradients and segmentation respectively.

All parameters of the model are optimized to reduce the loss $\mathcal{L}(\cdot)$, which is defined in Sec.4.3. Thus, the problem can be formulated as follows:

$$\Psi^* = \arg\min_{\Psi} \ \mathcal{L}([g, s], f(L; \Psi)), \tag{10}$$

where g is the ground truth gradients from a higher angular sampling rate with continuous trajectories.

The network directly learns the mapping $f(\cdot)$ and estimates the reliable gradients in a single feed-forward propagation.

4 Method

To address the challenges posed by undersampling and occlusion, we propose a mask-guided CNN-LSTM network to learn the mapping from the CEPV of an undersampled circular LF to the CEPV-trajectory gradients of the LF with a higher angular rate.

4.1 Training Pairs

We use the publicly available software for creating 3D graphics to render circular LFs with ground truth depths for all the views. With the known depths, the per-pixel ground truth gradients g of CEPV-trajectory corresponding to any specific angular resolution can be computed by Eq.6. Then the ground truth foreground mask s is computed by Eq.8. The gradients from a denser LF convey the information helpful for reconstructing continuous trajectories from broken segments. Since the CEPV-trajectories and their gradients are both predictable series, a neural network can be designed to learn the mapping between them.

Considering the trajectory of a point moving in a limited range in the y-direction during the rotation, we use the sliding window strategy along the y-direction. Given a circular LF with 90 views, we set Y to 11 and calculate the gradient and segmentation of the central $x\theta$-plane in the y-stacked CEPV. The supervision is the ground truth gradients and foreground segmentation computed from the corresponding CEPV with 180 views.

4.2 Network Architecture

The architecture of the proposed network is illustrated in Fig.7. Based on the 3D predictable property, a CEPV $(90 \times 400 \times 33)$ stacks in the y-axis as input, then it is upsampled to volume of $180 \times 400 \times 33$. Firstly, the up-sampled volume is fed into a U-shaped CNN-LSTM network. After that, the learned gradients and the up-sampled volume are concatenated. Then, they are exploited by two dense-block layers to distinguish a reliable foreground segmentation.

The CNN-LSTM network has four levels, each used to analyze the CEPV at different resolutions. Four convolutional layers are applied to encode local information on the top three levels. Then two convolutional LSTM layers are cascaded to extract the series features in top-down and left-right directions sequentially. Noting that, given a $h \times w \times c$ volume (h, w and c refer to the height, width and channels, respectively), it is separated as $h \times 1 \times c$ with length w and $1 \times w \times c$ with length h series in left-right and top-down LSTM layers, respectively. The kernel sizes of the convolutional LSTM layers are 3×1 and 1×3 in left-right and top-down directions, respectively. The channel size is 100 (Fig.7). Later, the outputs from top-down LSTM layers are concatenated with the up-convolutional results from a higher level, and then are decoded with another four convolutional layers. There are six convolutional layers in the highest 4-th level. Neighboring levels of the CNN-LSTM are connected with down and up-convolutional layers. The kernel size of all convolutional layers is 3×3. The channel size for the i-th layer is $\min(60 \times i, 180), i = 1, 2, 3, 4$. After the CNN-LSTM network, the $\frac{\partial x}{\partial \theta}$ is extracted. Finally, two standard dense-blocks are applied to segment reliable foreground regions. Each dense-block contains 4 layers.

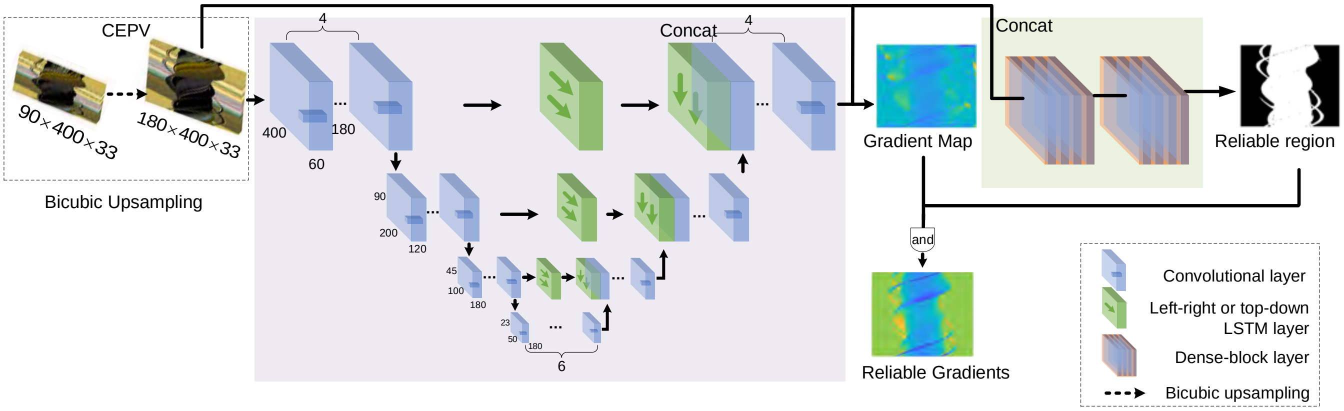


Fig. 7. The architecture of the proposed network. The pipeline contains three parts. The prediction scheme starts with an upsampling processing. The second part is to learn gradients from the CEPV. The third part concatenates original CEPV and gradients for predicting the reliable region.

4.3 Loss Function

The occlusion is still challenging from two folds. First, the ground truth foreground mask can only prevent disturbance from the background. The self-occlusion is not considered. Moreover, as discussed in Sec.3.3, the ground truth gradients are computed from $depth_{gt}^{\theta_i}$ without accounting for the visibility in adjacent views. If a scene point is visible in view θ_i but occluded in view θ_j, the ground truth gradients still correspond to the occluded pixel.

Based on the observation that the learned gradients deviate significantly at the intersections of CEPV-trajectories, the network output provides a clue for the occlusion event. The basic idea is that, if a region tends to bring ambiguity for computing gradient due to occlusion, we just discard it and refer to other unaffected gradients for depth estimation. We define a reliable mask $\tilde{s}$ to discount a dynamical region during training,

$$\tilde{s} = \begin{cases} 1 & |g - \hat{g}| < \epsilon \\ 0 & \text{otherwise}, \end{cases} \tag{11}$$

where ϵ is the threshold for identifying reliable regions of which the gradients are close to the ground truth.

Discounting the occluded region does not affect depth estimation. Each CEPV-trajectory corresponds to a scene point. Therefore its per-pixel gradients are redundant for depth estimation. Fig.6 illustrates an example of reliable mask. The reliable mask can eliminate the occlusion event region (the yellow circle of (d)) where the CEPV-trajectory intersects. The undersampling region can also be removed (the yellow circle of (e)) since the difference is large when CEPV-trajectories fade into discrete segments.

The loss function contains four terms, 2 of which are for gradients optimization and 2 for segmentation.

$$\mathcal{L} = \mathcal{L}_{\hat{s}}(g) + \lambda_1 \mathcal{L}_s(g) + \lambda_2 \mathcal{L}_s(\hat{s}) + \lambda_3 \mathcal{L}_{\tilde{s}}(\hat{s}), \tag{12}$$

where $\lambda_i, (i = 1, 2, 3)$ are the weights of the loss terms. $\mathcal{L}_{\hat{s}}(g)$ and $\mathcal{L}_s(g)$ are the $\mathcal{L}_1$ loss of gradients predictions with the output foreground mask and the ground truth mask respectively,

$$\mathcal{L}_{\hat{s}}(g) = \frac{1}{|\hat{s}|} \sum_{p \in \{\hat{s}=1\}} |g(p) - \hat{g}(p)|,$$

$$\mathcal{L}_s(g) = \frac{1}{|s|} \sum_{p \in \{s=1\}} |g(p) - \hat{g}(p)|.$$

$\mathcal{L}_s(\hat{s})$ and $\mathcal{L}_{\tilde{s}}(\hat{s})$ are the cross entropy losses of the output segmentation with the ground truth and the reliable mask respectively,

$$\mathcal{L}_s(\hat{s}) = \sum_p s(p) * log(\hat{s}(p)),$$

$$\mathcal{L}_{\tilde{s}}(\hat{s}) = \sum_p \tilde{s}(p) * log(\hat{s}(p)).$$

4.4 Algorithm

The whole procedure of our algorithm is shown in Algorithm 1. Specifically, given a 3D CEPV as input, we first calculate the gradients and their corresponding segments of the central $x\theta$-plane in the y-stacked CEPV. Based on Eq.6, the depth can be computed by the learned gradients. We

Algorithm 1 3D reconstruction algorithm from the CEPV

```
Input: The 3D CEPV $L(\theta, x, y)$ with parameters of the CLF
Output: The reconstructed point cloud $PC$
for $y$ = 1:$Y$ do
    Stacked CEPV: $L(\theta, x, (y - \Delta y) : (y + \Delta y))$
    Obtain $\hat{g}(\theta, x, y)$ and $\hat{s}(\theta, x, y)$ via Eq.9
    $depth_\theta \leftarrow D(\hat{g}(\theta, x, y), \hat{s}(\theta, x, y))$ via Eq.6
for $\theta$ = 1:$\Theta$ do
    $PC_\theta \leftarrow depth_\theta$ via Eq.13
    for $\beta$ = $(\theta - \Delta\theta) : (\theta + \Delta\theta)$ do
        $PC_\theta \leftarrow depth_\beta \cdot \hat{s}(\beta, x, y)$
    $PC \leftarrow PC_\theta \cup PC$
return $PC$
```

then map the $depth$ of each view θ into a 3D point cloud (PC) as,

$$PC_\theta : \begin{cases} R &= \sqrt{(\tan(\alpha)depth)^2 + (R_m - depth)^2} \\ \phi &= \arccos\left(\frac{R_m - depth}{R}\right) - \theta \\ Y &= depth\frac{y - y_c}{f}. \end{cases} \tag{13}$$

Then we project PC_θ to depth on neighbour views $\beta = (\theta - \Delta\theta) : (\theta + \Delta\theta)$ (we set $\Delta\theta = 30^\circ$ in the experiment) by,

$$depth_\beta = (R_m - Rcos(\beta + \phi))\hat{s}, \tag{14}$$

where $\hat{s}$ is used to segment the reliable region, and further back project the depth in the PC_θ. For each view, we repeat this step and merge all PC_θ in to the final PC.

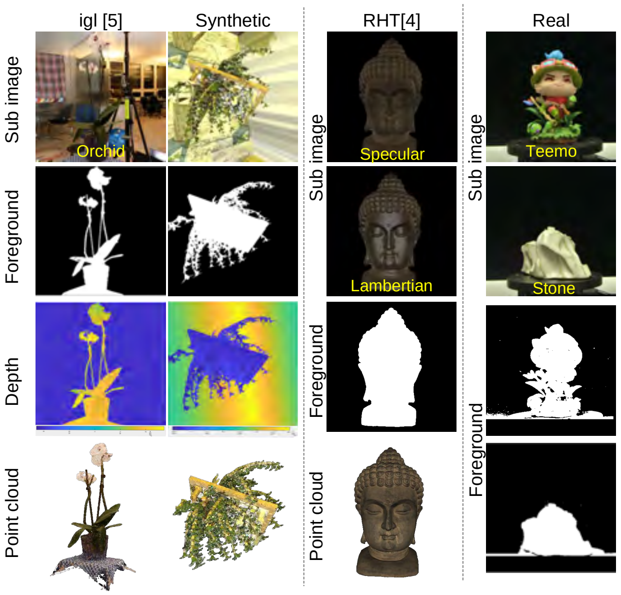


Fig. 8. Examples of 4 circular LF datasets. igl [5] is captured with the camera mounted on a boom rotating around the object. Our Synthetic data is rendered by POV-Ray [39]. Here shows the chessboard-plant scene. RHT [4] is rendered by Blender [40] with both specular and Lambertian surfaces. For our real dataset, the reconstructed object is staged on a high precise turntable with a controllable rotation interval.

5 Experiment

5.1 Dataset

Four datasets are used to evaluate the performance compared to different baseline methods, whose details are summarized in Table 1. The igl [5], captured from the natural scene, provides the sub-images, the calibration parameters on each, and the point cloud computed from a 3600-frame Circular LF. We take this point cloud as ground truth and generate the ground truth depth and foreground mask with camera parameters. RHT [4] renders the synthetic budda dataset with specular and Lambertian surfaces. The ground truth point cloud is also provided. We further compute the ground truth foreground mask for this dataset. We also self-construct a novel synthetic dataset and a real dataset. The synthetic dataset provides the ground truth foreground mask, depth, and point cloud. The real dataset provides only the sub-image and foreground mask.

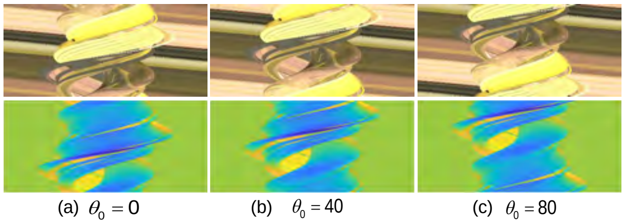

(a) $\theta_0 = 0$ (b) $\theta_0 = 40$ (c) $\theta_0 = 80$

Fig. 9. Data augmentation for a slice of CEPV by shifting the start view. The first row is a slice of CEPV and the second row is the corresponding ground truth gradients g.

We design the synthetic datasets for the training and ablation study and capture the real-scene datasets for the SOTAs comparison. For synthetic datasets, we render 150 circular LFs using the POV-Ray, 100 for training, and 50 for testing, which contain various challenging environments such as occlusion, shadowing, reflection, and structures with fine details. Fig.8 shows examples of the sub-view image, the foreground mask, depth map, and GT point clouds for four datasets. Since the extended circular LF L_{θ_0} is periodic, any view can be chosen as the start view θ_0 and the CEPV-trajectory is still continuous but with a diversified pattern. We augment the training data by shifting the start view θ_0. Fig.9 demonstrates the augmentation process by changing start views, e.g., $\theta_0 = 0$, $\theta_0 = 40$, and $\theta_0 = 80$. Both the CEPV and their corresponding gradients are augmented simultaneously. Additionally, the training data is augmented by exchanging RGB channels.

TABLE 1
Details of 4 circular LF datasets used in experiments. The term 'Num.' refers to the number of circular LFs. The term 'Res.' refers to the spatial resolution. The term σ refers to the pitch size of the sensor. The term f refers to the focal length.

Setting	igl [5]	Synthetic	RHT [4]	Real
Num.	5	150	2	5
Views	> 3600	180	720	360
Res.	1920×1080	400×400	1001×1001	4256×2832
$f(mm)$	24	20	18	48
$\sigma(\mu m)$	28	57.8	6	8.4
GT Mask	✓	✓	✓	✓
GT depth	✓	✓	×	×
GT PC	✓	✓	✓	×

We put objects on a high-precision rotation stage for the real-world scene data. The calibration is performed to remove distortion and determine the correct rotation center. The real-world and synthetic circular LFs datasets will be available to the research community to facilitate reproducible research.

5.2 Training Configuration

The network is trained with the input CEPV with 90 views and the gradients from 180 views. The network training is conducted by using the TensorFlow framework [41], and the parameter λ_1 is set as 1. λ_2 and λ_3 change with the epoch, that $\lambda_2 = 3 \times 0.8^{\lfloor i/10 \rfloor}$, $\lambda_3 = 0$ in the previous 50 epochs and becomes $1.15^{\lfloor i/10 \rfloor}$ after the 50-th epoch; here i is the epoch index. Note that λ_3 is initially zero because the network in the first few epochs is unstable. We use the Adam [42] optimizer. The learning rate is $1e-4$ initially and decreases 0.99. All convolutional kernels and bias are initialized using the Xavier method [43].

5.3 Ablation Study

This section evaluates the performance of each component proposed in the mask-guided CNN+LSTM on our 50 randomly generated synthetic scenes. Table 2 gives the average quantitative statistic of all 50 datasets.

We use three evaluation metrics for comparison. The first two are between ground truth and reconstructed PC. For evaluation, the obtained PC is aligned with the ground truth utilizing the iterative closest point (ICP) algorithm [44]. We compute (1) the RMSE[1] (Root Mean Square Error) [45]; (2) the percentage of bad matching points (BP) through the Hausdorff distance[2] [3], between two aligned PCs. The Root Mean Square Error of Depth (RMSE(D)) is the third evaluation metric, measuring the difference between the ground truth and the learned depth map on all views.

The ablation study is firstly carried out progressively to show the improvements gained by each proposed component. The baseline network (CNN+LSTM) is trained from the CEPV to ground truth gradients with the same angular resolution (90 views). With this undersampled circular LF (90 views), the CEPV-trajectory and its local gradients turn into discrete segments. Then it is hard to estimate the proper depth of reconstructed object between these two deteriorated trajectories, as the baseline results are shown in Table 2.

To justify the effectiveness of our proposed network architecture, we also conduct the ablation study by substituting each element in our design, as eliminating the LSTM module, replacing 3D convolution kernel with a 2D kernel, and reducing the four-layer network to a two-layer one. Our proposed method (Baseline+Prd+RM) achieves the best performance on both point cloud and depth map.

TABLE 2
The ablation study on our synthetic dataset. The term ‘RM’ refers to the reliable-mask-based loss.The term ‘Prd’ refers to the prediction scheme.

	RMSE	BP	RMSE(D)
Baseline (3D+4Layers+LSTM)	1.2100	0.6194	1.2238
Prediction (Baseline+Prd)	0.4731	0.1934	0.4042
Ours (Baseline+Prd+RM)	**0.4334**	**0.1893**	**0.3318**
Ours (3D+4Layers+Prd+RM)	0.4523	0.1941	0.3334
Ours (2D+4Layers+LSTM+Prd+RM)	0.4500	0.1924	0.3388
Ours (3D+2Layers+LSTM+Prd+RM)	0.5377	0.2292	0.3576

5.3.1 Prediction scheme

The prediction scheme will up-sample the CEPV from 90 to 180 views by Bicubic sampling. Furthermore, the network is trained with the CEPV at 90 views and ground truth gradients from 180 views. Even though the CEPV-trajectory falls into discrete segments, the series of local gradients at 180 views reveals increasing continuity. The prediction scheme motivates the network to predict the local gradients in a higher angularly sampled circular LF, which is more reliable to estimate depth. As shown in Table 2, the prediction scheme improves the performance significantly on all three metrics.

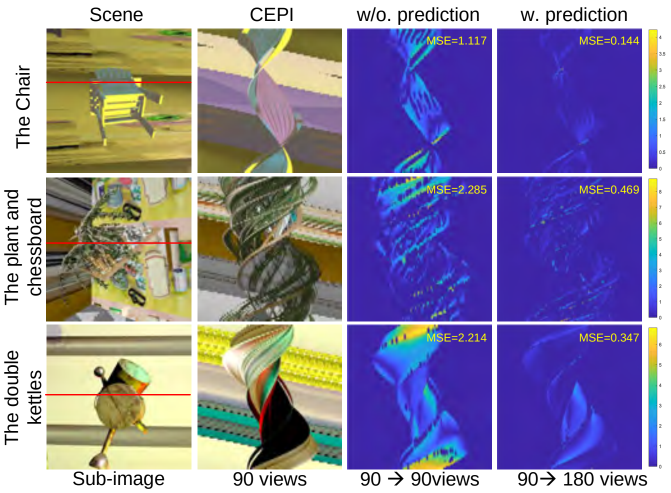


Fig. 10. Comparisons between the ground truth and the estimated gradients of our networks with or without the prediction scheme. Here are examples of 3 scenes with one sub-image and one $x\theta$-slice of the CEPV. The error map between the predicted and ground truth gradients are illustrated in columns III and IV.

Fig.10 visualizes the error maps of 3 different scenes in rows from networks with or without the prediction scheme. The MSE is marked on the top-right of each error map. The first scene is related to a chair with a texture-less surface. The second scene is more challenging, composed of a plant and a chessboard with a reflective surface. Moreover, the third scene has two kettles occluding with each other.

With undersampled circular LF at 90 views, the smooth CEPV-trajectories disappear due to aliasing and are replaced by discrete segments and additional structures arising from scene texture. The MSE is large for the result without the prediction scheme, especially in regions with discrete segments (see the 3rd column of Fig.10). Based on the predictable series discussed in Section 3.1, both the CEPV-trajectory and its partial derivative are differential and further predictable. The gradients from continuous CEPV-trajectories are prone to the proper depth of reconstructed objects. Thus, the prediction scheme improves the performance by learning the gradients from a 180-view circular LF (see the 4th column of Fig.10).

5.3.2 Reliable-mask-based loss

By combining the reliable-mask-based loss, the quantitative results of both PC and depth map are further improved, as shown in Table 2. The ground truth gradients are computed with known depth from Eq.6, in which the occlusion is not considered. The difference between the learned and ground truth gradients varies where CEPV-trajectories intersect. Thus, the reliable mask is designed to neglect these regions.

We illustrate the CEPIs and error maps between the learned and ground truth gradients of two scenes in Fig.11. The MSE of learned gradients with reliable-mask-based

[1] We use the function *packepcregistericp* in Matlab2021

[2] This operation is performed with Meshlab [46] to compute the percentage of PCs larger than $5mm$

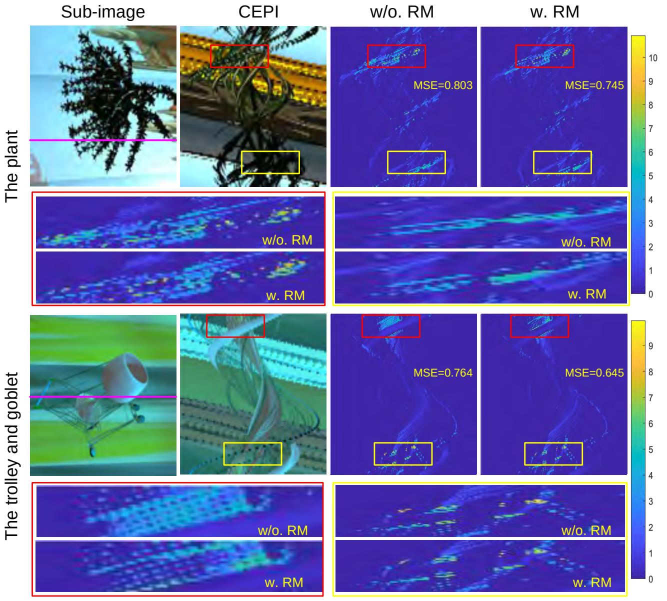


Fig. 11. Comparisons between the ground truth and the estimated gradients of our networks with or without the reliable-mask-based loss. Here are examples of 2 scenes with one sub-image and one $x\theta$-slice of the CEPV. The error map between the predicted and ground truth gradients are illustrated with insect focusing on the intersection regions.

loss is smaller than that of the one without reliable-mask-based loss. Both scenes suffer from the thin structure, and the corresponding CEPV-trajectories are twined. The insets (the yellow and red box) in Fig.11 illustrate that the error is smaller from the result with the reliable-mask-based loss, especially when the broken CEPV-trajectories intersect. Therefore, the the proposed method that adds reliable-mask-based loss is more robust in the occlusion events.

5.4 Comparison with SOTAs

5.4.1 *Synthetic scene*

We perform experiments on the data obtained from RHT, as shown in Table 1. Since the original circular LF has 720 views, we downsample it to 90 views, which degenerates the CEPV-trajectory into broken segments with aliasing.

The comparison is performed with seven SOTA methods, including traditional MVS-based methods, EPI-based methods and learning-based 3D reconstruction. All the results are obtained using the codes released by the authors, except the Robust Hough Transform method (RHT) [4]. We self-implement the algorithm since the code has not been released yet. We also compare with our previous network [2], which requires the circular LF with 180 views. For CLSTM, we directly up-sample the original CEPV from 90 to 180 views and feed them into the trained network. For our method without the reliable-mask based loss, we re-train the network using the circular LF with 90 views.

Table 3 demonstrates the RMSE, chamfer distance [47], and BP with two thresholds ($0.2mm$ and $0.5mm$) of the whole PCs in both Lambertian and Specular cases. Their corresponding distributions of the Hausdorff distance are illustrated in Fig.13.

All the MVS-based methods provide sparse reconstructed PCs. CMVS fails on both datasets and generates lots of wrong PCs. Due to occlusion and reflection, MVE shows large numbers of mismatched features. Colmap provides the most accurate result with the Hausdorff distance under $0.2mm$. However, comparing to the learning-based approaches, it is still a sparse result which only reconstructs 3.37% of the GT points (4675307 in total).

RHT is based on estimating the depth of each CEPI and merging them into PC with standard circular parameters. To detect and model the CEPV-trajectories, RHT [4] requires circular LF with 720 views. It relies on the smooth and continuous structure of the CEPI for the Hough voting. With undersampled circular LF, the broken segments and the additional structure of the CEPI affect both the feature detection and geometry estimation.

MVSNet is robust in the Lambertian case, with the minimum BP and chamfer distance. However, it generates the sparsest result. PmNet performs best under BP-$0.5mm$ in the Lambertian case. However, its performance degenerates significantly for specular surfaces.

CLSTM also suffers from the degeneration of CEPV-trajectories. It directly feeds the up-sampled CEPV to the trained network. The two versions of our methods, i.e. with and without the reliable-mask based loss, both learn the mapping from a 90-view CEPV to the ground truth gradients from a 180-view circular LF by adding the prediction scheme. The result is further improved for the occlusion situation (near the noise region) by adding the reliable-mask-based loss.

5.4.2 *Real scenes from igl [5]*

We use the public circular LF datasets from igl [5] to evaluate the performance of our proposed method on real scenes. The datasets provide the PCs computed from dense LFs with 3600+ views. Then the estimated results are used as ground truth for quantitative comparison. The datasets also provide calibration parameters. By projection the GT point clouds onto each view, we obtain the GT depth. According to Yucer *et al.* [5], a camera is mounted on a boom rotating around the object for capturing the views. The original 3600 views are down-sampled to 90 views to obtain an undersampled circular LF. We provide the BP-$5mm$, RMSE and Chamfer distance of the reconstructed point clouds in Table 4. Also, the evaluation on depth in term of L1, MSE, and BP-$2mm$ is shown in Table 5. In these two tables, the term 'Scare.' refers to the Scarecrow scene.

The CMVS method [8] achieves sparse PCs for only one side of the reconstructed object. The reconstructed result by MVE [10] tends to be affected by cluttered background. The background noise sticks to the reconstructed surface, causing significant errors for both reconstructed PCs and depth. Colmap is most competitive among the traditional MVS-based methods. However, all these MVS-based methods could only provide sparse reconstruction, and even Colmap can reconstruct 1.76% of the GT PCs on average.

The EPI-based methods, which require dense angular sampling, work well when CEPV-trajectories are continuous without distortion. RHT [4] fails on this dataset due to the undersampling and distortion from the capture device. The camera is mounted on a boom rotating in a circle with a slight vibration in the Y-direction. Such degeneration causes

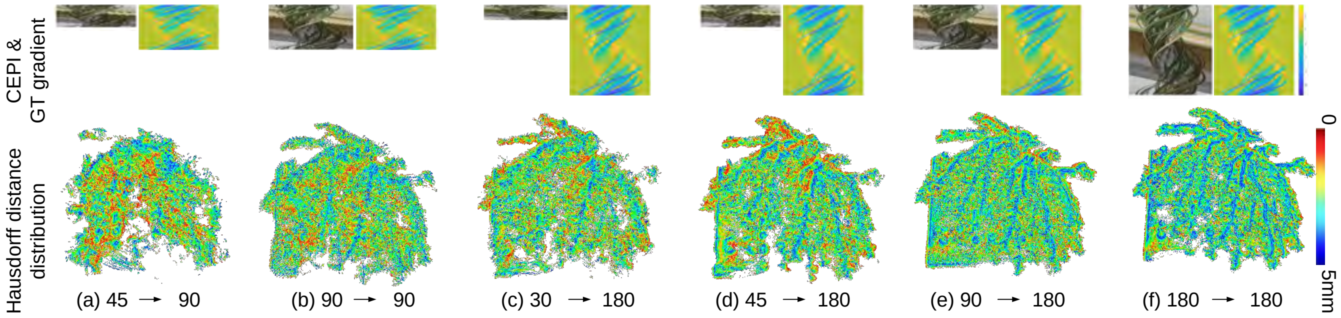


Fig. 12. Distributions of the Hausdorff distances of reconstructed point clouds at different angular resolutions. The first row illustrates training pairs of the CEPI and corresponding gradients under different angular resolutions.

TABLE 3
Data from RHT: the number of 3D points, RMSE, BP, and CD of whole point clouds of baseline methods and ours. The terms 'LBT' and 'SPC' refer to the Lambertian and the Specular respectively. The term 'CD' refers to the Chamfer distance.

	Metric	CMVS [8]	MVE [10]	Colmap [11]	RHT [4]	MVSNet [14]	PmNet [16]	CLSTM [2]	Ours w/o RM	Ours
LBT	Num.(×e+6)	0.0050	0.6767	0.1570	3.4867	0.3097	4.8966	4.1314	3.8863	3.6228
	RMSE	1.8879	1.6859	0.3723	4.0005	0.3837	0.4671	0.3810	0.3753	**0.3691**
	BP-0.5	0.7752	0.6386	0.0152	0.7054	0.0155	**0.0132**	0.0663	0.0372	0.0149
	BP-0.2	0.9896	0.8133	**0.2487**	0.9099	0.2896	0.2683	0.4662	0.3713	0.3142
	CD	3.3591	2.6486	0.6474	6.1308	0.4724	0.6501	0.5837	0.5779	**0.4160**
SPC	Num.(×e+6)	0.0052	0.3255	0.1190	3.2232	0.3076	4.7970	3.9681	3.7318	3.5049
	RMSE	3.0626	1.8850	0.3931	4.5517	0.3915	0.4713	0.5577	0.3953	**0.3871**
	BP-0.5	0.7769	0.7091	0.0455	0.7333	**0.0257**	0.1337	0.2034	0.0507	0.0373
	BP-0.2	0.9961	0.8391	0.2747	0.9412	**0.2065**	0.3770	0.5424	0.5774	0.3141
	CD	3.4511	2.9155	0.6656	6.6878	**0.4642**	0.6613	0.7390	0.5838	0.5779

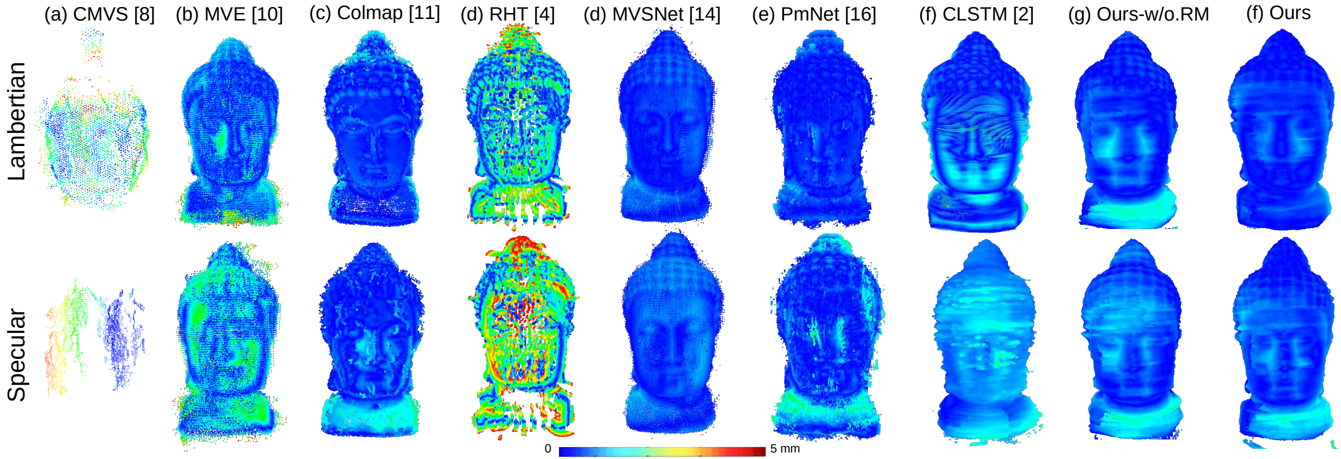


Fig. 13. Reconstruction results of 3D object in dataset with Lambertian and specular surfaces [4] and Hausdorff distances under $5mm$ obtained by different methods. From left to right: the results by CMVS, MVE, Colmap, RHT, MVSNet, PmNet, CLSTM, our method without the reliable-mask based loss and the our complete method.

irregular features in trajectories of each 2D CEPI, leading the Hough to vote for improper parameters.

The learning-based methods are good at finding proper feature correspondence. Based on patch matching, the results from the PmNet achieve the densest reconstruction among all other methods. The MVSNet fails on the scene of Orchid, which contains a tiny thin reconstructed object against a complex background.

Our method performs better than CLSTM [2] and Ours-w/o. RM since both the prediction scheme and reliable-mask-based loss are designed for undersampled circular LF. By predicting the gradients on CEPV, we can achieve per-pixel depth estimation. We show competitive performance on the PCs and also depth maps.

5.4.3 Real scenes from our dataset

We also conduct experiments on our collected real-scene datasets and carry out visual comparisons with SOTAs, as shown in Fig.14. For the circular LF with 90 views, the MVS-based methods suffer from low accuracy and severe noise. The results by CMVS degenerate into sparse PCs, and the results by MVE are noisy on the reconstructed surface. Colmap performs best among traditional MVS-based methods. But it only reconstructs sparse PCs, which would lose details for tiny structure.

RHT also fails due to undersampling and brings sparse results near the rotational shaft. The broken trajectories are

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3206207

TABLE 4
Data from igl: the number of 3D points, BP, RMSE and CD of point clouds by baseline methods and ours. The term 'PmNet' is short for PatchmatchNet.

	Metric	Orchid	Plant	Scare.	Ship	Statue
GT	Num.(×e+6)	2.1795	7.2881	4.9407	6.5992	6.5571
CMVS [8]	Num.(×e+3)	1.5880	3.2760	2.3430	6.2360	3.1600
	BP	0.0309	0.8834	0.2808	0.0438	0.1307
	RMSE	0.3070	2.9196	0.5397	0.2342	0.3416
	CD	0.4022	2.8863	0.8601	0.4633	1.0000
MVE [10]	Num.(×e+5)	0.0349	1.6477	1.8897	2.8313	3.2871
	BP	0.7149	0.3512	0.4726	0.1404	0.6023
	RMSE	1.4969	0.5177	0.7187	0.3880	1.1551
	CD	1.6126	0.7263	0.8603	0.4646	1.1879
Colmap [11]	Num.(×e+5)	0.5045	1.3657	0.6671	1.1829	0.9631
	BP	0.0123	0.0732	0.0952	0.0085	0.1235
	RMSE	0.1114	0.4002	0.2805	0.5661	0.4178
	CD	0.1579	0.2428	0.1503	0.2551	0.3829
RHT [4]	Num.(×e+6)	1.3710	1.9915	2.6639	2.9897	2.2522
	BP	0.9123	0.8745	0.8931	0.8595	0.9019
	RMSE	7.1229	6.7809	6.6534	6.5539	6.7722
	CD	7.3212	7.1438	7.0419	7.6485	7.9185
MVS-Net [14]	Num.(×e+5)	NaN	3.7338	0.8275	1.1112	0.9695
	BP	NaN	0.3959	0.4154	0.2229	0.4228
	RMSE	NaN	0.2678	0.2265	0.1918	0.3245
	CD	NaN	**0.1993**	0.1629	0.2332	0.1364
PmNet [16]	Num.(×e+6)	0.9077	4.8004	2.0564	3.8230	3.9043
	BP	0.0028	0.0029	0.0434	0.0046	0.2438
	RMSE	0.0901	0.3820	0.2096	0.1713	0.4315
	CD	0.1482	0.2533	**0.0506**	0.1923	**0.1193**
CLSTM [2]	Num.(×e+6)	0.6949	2.4621	1.1070	1.2646	1.8047
	BP	0.0474	0.0580	0.0056	0.0056	0.0048
	RMSE	0.1161	0.2806	0.1921	0.1984	0.1818
	CD	0.0883	0.2777	0.5014	0.1781	0.3095
Ours w/o RM	Num.(×e+6)	0.6054	2.3601	0.9236	1.0894	1.6017
	BP	0.1036	0.0771	0.0481	0.0077	0.0602
	RMSE	0.1533	0.4364	0.2039	0.1881	0.2709
	CD	0.0928	0.3129	0.4455	0.1851	0.2626
Ours	Num.(×e+6)	0.5658	2.2378	0.8975	0.7060	1.5218
	BP	**0.0017**	**0.0372**	**0.0002**	**0.0021**	**0.0024**
	RMSE	**0.0731**	**0.2591**	**0.1691**	**0.1714**	**0.1627**
	CD	**0.0822**	0.2757	0.0710	**0.1731**	0.1640

TABLE 5
Data from igl: the BP, MSE, and L1 on depth maps by baseline methods and ours. The term 'DpMVS' is short for DeepMVS, which only provides the result of depth map. CMVS only provides the result of point clouds.

	Metric	Orchid	Plant	Scare.	Ship	Statue
MVE [10]	BP	0.9938	0.9647	0.9559	0.9820	0.9120
	MSE	27.8006	17.9377	6.6326	14.2523	5.3812
	L1	5.2285	4.3574	2.5478	3.7087	2.2920
Colmap [11]	BP	0.1741	0.0036	0.2726	0.0444	0.1190
	MSE	5.4871	1.9603	2.8525	0.7794	1.7488
	L1	1.9056	1.3697	1.4259	0.5390	1.1575
RHT [4]	BP	0.9373	0.8906	0.8868	0.9443	0.9006
	MSE	24.1109	10.8422	5.2899	12.6579	5.1396
	L1	4.8499	3.3048	2.2130	3.4741	2.2266
MVS-Net [14]	BP	Nan	0.0037	0.2524	0.0465	0.1648
	MSE	Nan	1.9912	2.5519	0.8389	1.9276
	L1	Nan	1.3936	1.3446	0.6009	1.2431
PmNet [16]	BP	0.1624	0.0037	0.3120	0.0791	0.1850
	MSE	3.1125	1.9917	3.6278	1.2595	2.1224
	L1	1.7401	1.3939	1.6745	0.6758	1.3026
DpMVS [13]	BP	0.1453	0.2702	0.0402	0.0152	0.0210
	MSE	3.0486	3.3176	1.4081	0.4811	1.5605
	L1	1.6106	1.7843	1.0976	0.4542	1.1933
CLSTM [2]	BP	0.0279	0.0053	0.0444	0.0024	0.0988
	MSE	1.0540	1.1756	1.3760	0.3170	2.0097
	L1	0.9902	1.0536	1.1063	0.4627	1.3409
Ours w/o RM	BP	0.0036	0.0039	0.0141	0.0018	0.0198
	MSE	0.9953	1.0128	0.9557	0.2676	1.2882
	L1	0.9657	0.9430	0.9236	0.4336	1.0867
Ours	BP	**0.0024**	**0.0032**	**0.0078**	**0.0015**	**0.0141**
	MSE	**0.9614**	**0.9116**	**0.8105**	**0.2570**	**1.1931**
	L1	**0.9441**	**0.8637**	**0.8393**	**0.4289**	**1.0448**

TABLE 6
Quantitative comparison on the network trained with the CEPV and ground truth gradients from LFs with various angular rates. A→B refers to a network trained with the CEPV of A views and supervised by the ground truth gradients of B views.

CEPV → gradients	RMSE	BP	RMSE (D)
45→90	1.2150	0.7121	1.2502
90→90	1.0100	0.6078	1.1478
30→180	0.9085	0.6364	0.5716
45→180	0.4776	0.4107	0.4650
90→180	0.4334	0.1893	0.3318
180→180	0.3727	0.1586	0.2227

twined together near the rotational shaft, where is mainly occluded. It increases the challenge to estimate the proper local direction for structure tensor. Then the Hough voting will accumulate to improper parameters for depth estimation.

MVSNet suffers from holes in the reconstructed surface, especially where the visibility is complex. PmNet performs best among learning-based MVS methods. However, it fails to reconstruct boundary details, such as the ear of the Ziggs and the foot of the Optimus Prime.

The results by CLSTM also suffer from distortion and noise to a certain compared to ours. Using the prediction scheme improves the reconstructed outcomes with lower noise. After training with the reliable-mask based loss, our method achieves accurate and dense results, even on the Stone scene, which is challenging due to homogenous textures and reflective surfaces.

5.5 Discussions

5.5.1 Different angular resolutions

To analyze the influence of the angular resolution of the circular LF on the reconstruction results, we conduct experiments with various angular resolutions to evaluate the generalization ability of our proposed method. Table 6 shows the average RMSE, BP, and RMSE(D) on the randomly generated 50 synthetic circular LFs. Correspondingly, Fig.12 visualizes the error distribution at different angular resolutions. With the same angular rate for gradients, all metrics indicate that the noise increases with the decreasing angular rate of CEPV. However, the gradients from a dense circular LF could guide the network to estimate reliable gradients for 3D reconstruction. The performance of the networks trained from 30 to 180, 45 to 180 outperform the networks trained from 45 to 90 and even 90 to 90.

5.5.2 The threshold of reliable mask

Different thresholds of the reliable mask are also analyzed. We experiment with different values of ϵ, including 0.05, 0.1, and 0.2. Table 7 demonstrates the quantitative result. When $\epsilon = 0.05$, the result has the minimum RMSE and the percentage of bad pixels. Nevertheless, fewer point clouds are reconstructed. Therefore, the noise of the depth map is larger since limited pixels of depth are estimated. When $\epsilon = 0.2$ and $\epsilon = 0.1$, the results show smaller errors with similar numbers of reconstructed points. Therefore, $\epsilon = 0.1$ is chosen as the threshold in all other experiments.

TABLE 7
Quantitative results by training the proposed network using different thresholds of reliable mask. The term 'Num.' refers to the average number of reconstructed points on 50 test scenes.

ϵ	Num.	RMSE	BP	RMSE (D)
0.05	1248610.9	0.3899	0.1749	0.3333
0.1	1763588.9	0.4334	0.1893	0.3318
0.2	1768850.7	0.4782	0.2088	0.3425

6 Conclusion

We propose a 3D reconstruction framework by learning the reliable gradients from the undersampled CEPV-trajectory. The 3D reconstruction accuracy deteriorates severely when the CEPV-trajectory turns into discrete segments. Our key idea is that coherence is still embedded, and depth can be estimated from reliable segments in a denser LF with less noise and ambiguity. By formulating both the CEPV-trajectory and its local gradients as *3D predictable series*, we design a mask-guided CNN+LSTM network, which is capable of learning the mappings from the undersampled CEPV-trajectory to gradients of denser LF. A reliable mask loss is further integrated to alleviate the impact of both occlusion and undersampling. The ablation study shows that both the prediction scheme and the reliable mask loss promote the performance of the 3D reconstruction. Our results outperform existing state-of-the-art 3D reconstruction methods in undersampled circular LF on four existing available datasets.

However, our method still requires a circular capturing with little distortion, either by a turntable stage or a boom rotating around the object. The rotation shaft and the optical axis should be orthogonal and intersect. Also, the boom should move in a relatively standard circle. Any deterioration will distort our reconstruction to a certain.

In the future, we will develop our research to more flexible scenes, such as more sparse LFs or a near circular case. With fewer views needed, it is possible to build a circular LF for dynamic scene capturing. We will study the deformation of the dynamic coherent structure for robust 3D estimation. The properties analyzed with the findings in this paper may provide new insights for Multi-view Stereo, novel view synthesis, and other related areas.

Acknowledgments

The work was supported by NSFC under Grants 62031023 and 61801396. The authors would like to thank Prof. Hongdong Li for helpful suggestions and Zhongyun Hu, Libing Yang, Qi Wu, Chengcheng Qin, Guangcan Xu and Qing Cui for their help. We also thank anonymous reviewers for their valuable feedback.

References

[1] Y. Zhang, Z. Li, W. Yang, P. Yu, H. Lin, and J. Yu, "The light field 3d scanner," in *IEEE International Conference on Computational Photography*, 2017, pp. 67–75.

[2] Z. Song, H. Zhu, Q. Wu, X. Wang, H. Li, and Q. Wang, "Accurate 3d reconstruction from circular light field using cnn-lstm," in *IEEE International Conference on Multimedia and Expo*, 2020, pp. 1–6.

[3] A. Vianello, "Robust 3d surface reconstruction from light fields," Ph.D. dissertation, Heidelberg University, 2017.

[4] A. Vianello, J. Ackermann, M. Diebold, and B. Jähne, "Robust hough transform based 3d reconstruction from circular light fields," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7327–7335.

[5] K. Yucer, A. Sorkine-Hornung, O. Wang, and O. Sorkine-Hornung, "Efficient 3d object segmentation from densely sampled light fields with applications to 3d reconstruction," *Acm Transactions on Graphics*, vol. 35, no. 3, pp. 1–15, 2016.

[6] K. Yucer, C. Kim, A. Sorkine-Hornung, and O. Sorkine-Hornung, "Depth from gradients in dense light fields for object reconstruction," in *International Conference on 3D Vision*, 2016, pp. 249–257.

[7] H. Zhu, M. Guo, H. Li, Q. Wang, and A. Robles-Kelly, "Revisiting spatio-angular trade-off in light field cameras and extended applications in super-resolution," *IEEE transactions on visualization and computer graphics*, vol. 27, no. 6, pp. 3019–3033, 2019.

[8] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.

[9] Z. Li, W. Zuo, Z. Wang, and L. Zhang, "Confidence-based large-scale dense multi-view stereo," *IEEE Transactions on Image Processing*, vol. 29, pp. 7176–7191, 2020.

[10] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz, "Multi-view stereo for community photo collections," in *IEEE International Conference on Computer Vision*, 2007, pp. 825–834.

[11] J. L. Schönberger and J. M. Frahm, "Structure-from-motion revisited," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[12] J. L. Schönberger, E. Zheng, J. M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision*. Springer, 2016, pp. 501–518.

[13] P. H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "Deepmvs: Learning multi-view stereopsis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2821–2830.

[14] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *15th European Conference on Computer Vision (ECCV)*, 2018, pp. 785–801.

[15] P. H. Chen, H. C. Yang, K. W. Chen, and Y. S. Chen, "Mvsnet++: Learning depth-based attention pyramid features for multi-view stereo," *IEEE Transactions on Image Processing*, vol. 29, pp. 7261–7273, 2020.

[16] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "Patchmatchnet: Learned multi-view patchmatch stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 194–14 203.

[17] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH 1996. ACM, 1996, p. 43–54.

[18] M. Levoy, "Light fields and computational imaging," *Computer*, vol. 39, no. 8, pp. 46–55, 2006.

[19] A. Criminisi, S. B. Kang, R. Swaminathan, R. Szeliski, and P. Anandan, "Extracting layers and analyzing their specular properties using epipolar-plane-image analysis," *Computer vision and image understanding*, vol. 97, no. 1, pp. 51–85, 2005.

[20] S. Baker, T. Sim, and T. Kanade, "When is the shape of a scene unique given its light-field: A fundamental theorem of 3d vision?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 1, pp. 100–109, 2003.

[21] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606–619, 2014.

[22] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 926–954, 2017.

[23] 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.

[24] M. W. Tao, P. P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi, "Depth from shading, defocus, and correspondence using light-field angular coherence," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1940–1948.

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

[26] 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. 73–86, 2013.

[27] 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 Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1547–1555.

[28] H. Zhu, Q. Zhang, Q. Wang, and H. Li, "4d light field superpixel and segmentation," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 85–99, 2020.

[29] S. Heber and T. Pock, "Convolutional networks for shape from light field," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3746–3754.

[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.

[31] C. Shin, H. G. Jeon, Y. Yoon, I. So Kweon, and S. Joo Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4748–4757.

[32] G. Wu, Y. Liu, L. Fang, and T. Chai, "Revisiting light field rendering with deep anti-aliasing neural network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[33] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on epi," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 6319–6327.

[34] Y. Li, Q. Wang, L. Zhang, and G. Lafruit, "A lightweight depth estimation network for wide-baseline light fields," *IEEE Transactions on Image Processing*, vol. 30, pp. 2288–2300, 2021.

[35] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International Journal of Computer Vision*, vol. 1, no. 1, pp. 7–55, 1987.

[36] I. Feldmann, P. Kauff, and P. Eisert, "Optimized space sampling for circular image cube trajectory analysis," in *International Conference on Image Processing*, 2007, pp. 1947–1950.

[37] A. Cserkaszky, P. A. Kara, A. Barsi, M. G. Martini, and T. Balogh, "Light-fields of circular camera arrays," in *IEEE European Signal Processing Conference*, 2018, pp. 241–245.

[38] Z. Song, L. Yang, Q. Wu, H. Zhu, and Q. Wang, "Effective 3d object reconstruction from densely sampled circular light fields," in *Optoelectronic Imaging and Multimedia Technology VI*, vol. 11187. SPIE, 2019, pp. 133 – 141.

[39] P. of Vision Raytracer Pty. Ltd, "Pov-ray," http://www.povray.org/.

[40] B. Foundation., "Blender." http://www.blender.org/, 2017.

[41] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[42] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[43] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[44] P. J. Besl and H. D. Mckay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[45] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image and Vision Computing*, vol. 10, no. 3, pp. 145–155, 1992.

[46] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, and G. Ranzuglia, "Meshlab: an open-source mesh processing tool," in *Eurographics Italian Chapter Conference*, 2008.

[47] G. Borgefors, "Distance transformations in arbitrary dimensions," *Computer Vision, Graphics, and Image Processing*, vol. 27, no. 3, pp. 321–345, 1984.

Zhengxi Song is now a PhD candidate at the School of Computer Science and Technology, Northwestern Polytechnical University. She received the BS and MS degrees from Northwestern Polytechnical University in 2012 and 2015, respectively. Her research interests include 3D reconstruction and computational photography.

Dr. Xue Wang is currently an Associate Research Fellow with the School of Computer Science, Northwestern Polytechnical University. She received the BS and PhD degrees from Northwestern Polytechnical University, in 2007 and 2017, respectively. She studied in University of Pennsylvania as a visiting PhD student financed by China Scholarship Council. Her research interests include computer vision, computational photography and machine learning.

Dr. Hao Zhu is currently an Associate Researcher with the School of Electronic Science and Engineering, Nanjing University. He received the BS and PhD degrees from Northwestern Polytechnical University in 2014 and 2020, respectively. He was a visiting scholar at the Australian National University. His research interests include computational photography and optimization for inverse problems.

Dr. Guoqing Zhou is now an Associate Professor with the School of Computer Science, Northwestern Polytechnical University. In 2009 and 2013 he obtained MS and PhD degrees in the School of Computer Science, Northwestern Polytechnical University. He worked as a visiting scholar in Center for Imaging Science (CIS) of The Johns Hopkins University. His research interests include computer vision and computational photography.

Qing Wang graduated from the Department of Mathematics, Peking University, in 1991. He received PhD degree from the Department of Computer Science, Northwestern Polytechnical University in 2000. He is currently a Professor with the School of Computer Science, Northwestern Polytechnical University. He worked as Research Scientist at the Department of Electronic and Information Engineering, the Hong Kong Polytechnic University, from 1999 to 2002. He also worked as a Visiting Scholar at the School of Information Engineering, The University of Sydney, Australia, in 2003 and 2004. In 2009 and 2012, he visited the Human Computer Interaction Institute, Carnegie Mellon University, for six months and the Department of Computer Science, University of Delaware, for one month. He has published more than 100 papers in the international journals and conferences. His research interests include computer vision and computational photography, such as 3D vision, light field imaging and processing, novel view synthesis. He is a member of ACM. In 2006, he was awarded as Outstanding Talent Program of New Century by Ministry of Education, China.

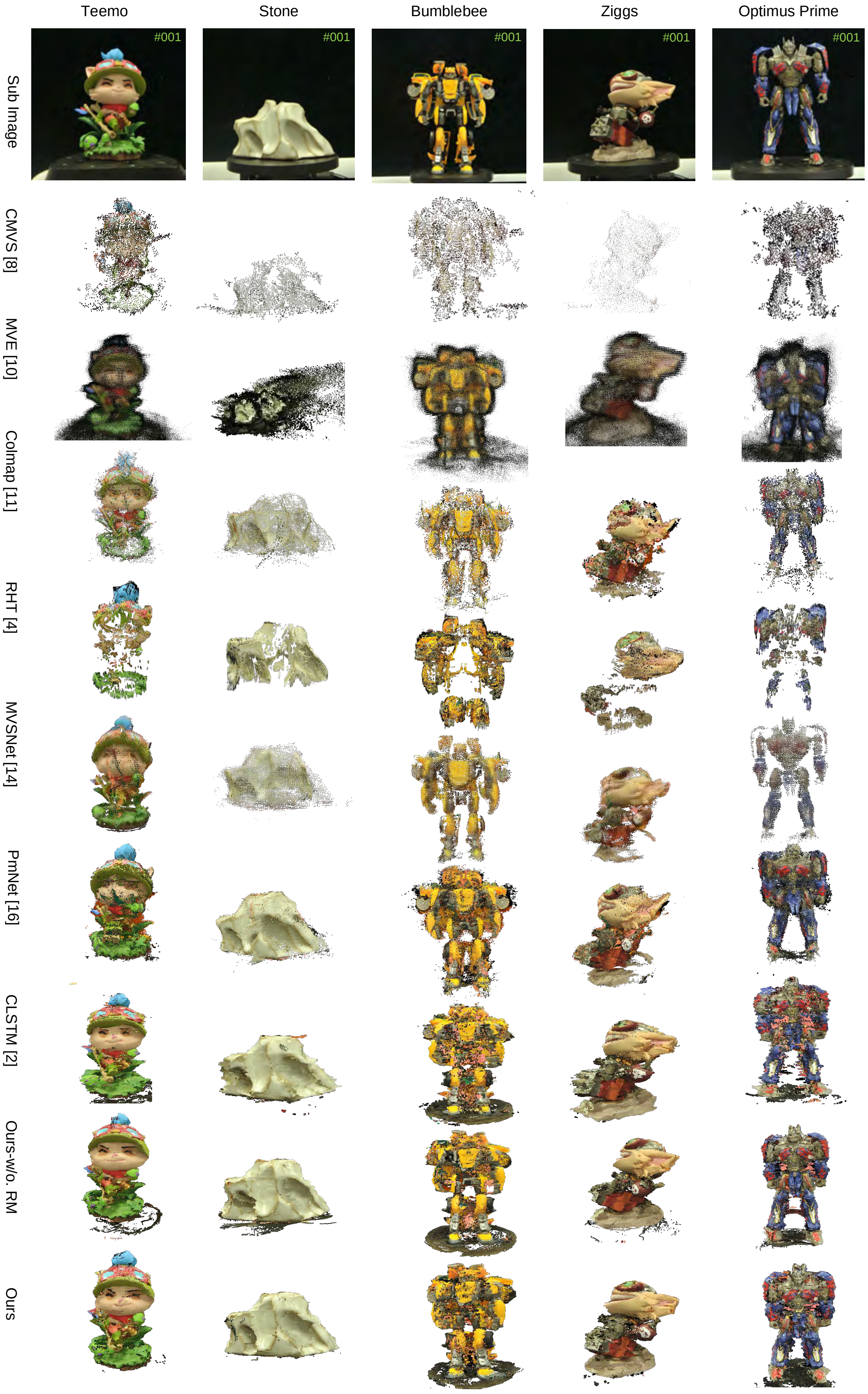


Fig. 14: 3D reconstruction results for real-scene circular LFs.