Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.DOI

Multi-Flash Light Field Photography

YINGLIANG ZHANG[1,2,3,†], (Student Member, IEEE), CHENG SHEN[4,†], WEI YANG[5], AND JINGYI YU[1] (Member, IEEE)

[1]School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China
[2]Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]Columbia University, New York, NY 10027, USA
[5]DGene Inc, Shanghai 201210, China

Corresponding author: Jingyi Yu (e-mail: yujingy i@shanghaitech.edu.cn).

† These authors contributed to the work equally.

ABSTRACT We present a novel multi-flash light field photography (MFLF) technique that couples a light field camera with a ring of four flash lights to combine the benefits of both imaging systems. The multi-flash photography technique enables reliable detection of occlusion edges at individual subaperture images (light field views) in the light field image. We then separately treat occlusion vs. non-occlusion edges in light field stereo matching. For non-occlusion edge pixels, we conduct correspondence matching and depth-from-focus. For pixels on or near the occlusion boundary, we estimate the shape of the occlusion boundary using a novel Support Vector Machine (SVM) approach. We collect a MFLF dataset of synthetic and real scenes that exhibit strong occlusions. Extensive experiments show that our method significantly outperforms state-of-the-art light field stereo matching methods in accuracy and robustness.

INDEX TERMS Multi-flash photography, light field imaging, stereo matching, depth estimation.

I. INTRODUCTION

DEPTH estimation is a long-standing problem in computer vision. High fidelity depth maps are beneficial to 2D segmentation, localization, 3D reconstruction, autopiloting, etc. Existing depth estimation techniques can be classified into two categories, passive or active. Passive methods such as binocular or multi-view stereo matching compute dense pixel correspondences between different views and use triangulation to determine the depth of corresponding pixels [1]–[6]. A major drawback there is the requirement of textures for establishing reliable correspondences. For objects with no or even repeat-textures such as flowers or plants, passive methods can easily break down. Active methods such as structured light or time-of-flight (ToF) 3D sensing project specially designed lighting patterns (in either spatial or frequency domains) and directly analyze the observed pattern for depth inference. By far most active methods still suffer from low resolution and low fidelity. In particular, active methods cannot recover accurate object boundaries which is important for high quality shape reconstruction.

The goal of this paper is to develop a computational photography technique that combines passive and active 3D sensing. We employ two emerging imaging techniques, multi-flash photography and light field imaging. The former serves as an active sensing tool for occlusion boundary detection and the latter as a multi-view reconstruction tool for high fidelity stereo matching.

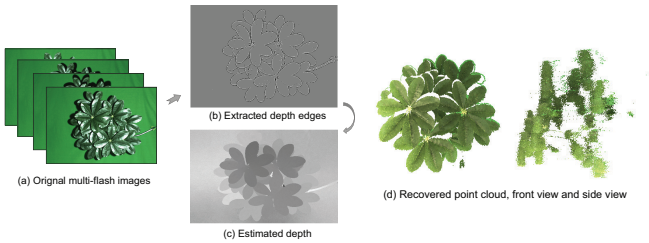


FIGURE 1. Our Multi-flash light field photography technique on a flower scene. (a) shows the multi-flash color images extracted from an LF, (b) shows our extracted depth edges, (c) and (d) show our recovered depth map and point cloud.

The multi-flash (MF) imaging solution [7] exploits shadow variations under different lighting directions: a flash light generates sharp cast shadows attached to the occlusion boundary when the flash-camera baseline is small. By strategically varying the location of the flash lights, one can then analyze shadow variations and reliably detect shadow boundaries. However, the depth of these boundaries cannot be reliably estimated despite efforts on shape-from-shadows [8]–[10]. In contrast, a light field (LF) camera captures views from different viewpoints within a single image, forming

a de facto multi-view acquisition system where multi-view stereo matching can be applied for depth estimation. With the availability of hand-held light field cameras for example Lytro [11] and Raytrix, one can easily capture such light field images of real scene. However, same as traditional multi-view reconstruction, the problem of lacking texture persists. Furthermore, when capturing objects composed of parts of similar color or textures, light field stereo matching fails to recover occlusion boundaries.

Our multi-flash light field photography (MFLF) technique couples a light field camera with a ring of four flash lights to combine the benefits of both imaging system. For each capture, we subsequently capture 5 light field images, four images with only a specific flash on and one all flashes off. We adopt the multi-flash photography processing pipeline to detect occlusion edges at individual subaperture image (light field view) in the light field image. We then separately treat occlusion vs. non-occlusion edges for light field stereo matching. Specifically, for non-occlusion edge pixels, we conduct correspondence matching and depth-from-focus. For pixels on or near the occlusion boundary, we estimate the shape of the occlusion boundary using a novel Support Vector Machine (SVM) approach. Specifically, we use SVM as a classifier to determine the depth of the pixel by analyzing its angular patch. We further present a technique for fusing the results of the occlusion edge vs. non-occlusion edge pixels. We collect a MFLF dataset of synthetic and real scenes that exhibit strong occlusions. Extensive experiments show that our method significantly outperforms state-of-the-art light field stereo matching methods in effectiveness and accuracy.

II. RELATED WORK

Our work is closely related to multi-flash imaging and light field stereo matching. For the slope of our work, we only discuss the most relevant work in this section.

a: Multi-Flash Imaging

Flash emitted from different position generates direction-related shadows on the depth edges. One can then directly extract occlusion boundaries if multiple flashes from directly locations are used. Feris et al. [7] build a capture apparatus composed of one Canon DSLR camera and four flashes surrounding the camera. They utilize the multi-flash illumination to generate several feature maps about depth edges, the sign of the depth edge and the relative distances between objects. They then compute an occlusion map and estimate the depth map according to shadow widths. The depth map preserves high quality occlusion edge boundaries. However, the accuracy of depth maps relies heavily on the quality and reliability of shadow width detection, which is very challenging in heavily occluded scenes. More recent efforts include a mobile implementation of the multi-flash camera on smart phones [12].

b: Light Field Stereo

In our MFLF setup, we use a light field camera in place of a regular camera. Light field records abundant information of each ray in space. To represent a LF, it is common to use the two-plane-parameterization (2PP) [13], [14] for its efficiency. In 2PP, each ray is parameterized by the interaction with two parallel planes, e.g., the $s-t$ plane or Π_{st} and the $u-v$ plane or Π_{uv}. In this paper, we treat Π_{st} as our camera plane, and Π_{uv} as the image plane and use $[s,t,u,v]$ tuple to represent each ray and align the origin of light field coordinate system to the CoP of the central subaperture camera. Light field can be recorded either using a light field camera [11] or a light field camera array [15], corresponding to small or large baseline setups.

In this paper, we combine the light field camera with multi-flash photography. The light field camera captures the multi-view data in a single shot. The dense sampling of angular and spatial information enables reliable depth estimations. Wanner et al. [16] generate the epipolar plane images (EPI) using horizontal and vertical cuts on LF and solve a constraint labeling problem on EPI for depth estimation. Chen et al. [17] propose the surface camera (SCam) model that bundles rays passing through a 3D point observed by the light field camera. They further employ a new bilateral metric to analyze the occlusion probability. However, these methods are highly sensitive to the color/texture of the scene. Wang et al. [18] propose a technique to explore the occlusion boundary in angular patch or SCam. However, the boundary analysis can fail on object with repetitive textures or color discontinuity.

Zhang et al. [19] adopt the color-disparity correlation model where they represent disparity as a linear combination of RGB colors. Their method produces satisfactory results on textured scenes and manages to preserve sharp occlusion boundaries when the color contrast between the foreground and the background objects are strong. However, same as any stereo matching algorithm, if the foreground and the background layers exhibit similar appearances, LF stereo matching can easily fail. There is also a trend on employing deep learning in light field stereo matching [20]. The quality however relies on the training data and the problem of similar foreground/background is still not resolved. Our MFLF technique seeks to combine the advantages of both multi-flash and light field imaging: we use the former for detecting occlusion boundaries and the latter for stereo matching.

III. MULTI-FLASH ON LIGHT FIELD

Depth edges are related to depth boundaries and a main indicator of the relative position of objects in the camera view. Raskar et al. [21] first use the term "depth edges" to describe the $C0$ discontinuities in a depth map. They are fundamentally different from texture edges caused by color change. Taking two planes at different depth for example, the boundary of the front plane correspond to depth edges due to depth discontinuity. The depth edges are also signed, e.g., positive and negative. Along a traversal direction on the image (in our case left to right, top to bottom), positive side

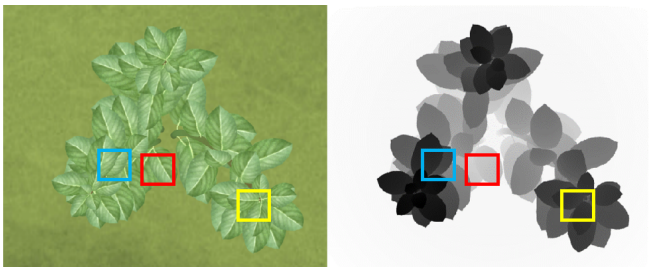


FIGURE 2. A plant scene exhibiting complex occlusions patterns formed by leaves and our recovered depth map using MFLF.

refers to the part of edge transitioning from the foreground to the background, and negative from the background to the foreground.

Traditional occlusion edge extraction methods rely heavily on color discontinuity. In reality, many scenes such as flowers or plants exhibit few textures but rich occlusions. Fig.2 shows the depth map of a plant scene with many mutually occluding leaves. Since each leaf can be approximated as a planar surface, depth edges can be efficiently used to model depth discontinuity caused by occlusions between leaves. The figure also shows the differences between depth discontinuity and color discontinuity. For leaves in the yellow rectangle of the figure, the color difference between different leaves at same depth generates sharp texture edges. In contrast, leaves in the blue rectangle are at different depth and have different color. In this case, the occlusion edge and texture edge coincide with each other. However, color and depth edges are independently: leaves at different depths but with same color can still exhibit edge edges as shown in the red rectangle of Fig.2.

Although multi-flash can be used to detect the occlusion edges, it cannot directly be used to determine the exact depth. In [22], the authors attempt to use shadow widths to infer the depth gradient and then use Poisson gradient integration to obtain the scene depth. In practice, shadow widths are particularly difficult to estimate under occlusions and complex self-shadowing. In our case, we resort to the LF camera and apply reliable multi-view depth estimation on these edges. In our setting, we substitute the DSLR camera in the original multi-flash camera [21] with a light field plenoptic camera (Lytro Illum).

Recall that shadow location also follows the standard epipolar geometry. Given a light source L at position P_l and a camera at O, we form an epipolar plane with P_l, O and a 3D point SP at S_l. The epipolar plane intersects with the image plane of the camera at an epopolar line e_l. It is obvious that if L casts a shadow caused by point SP, the shadow must line on the epipolar line. We therefore can directly traverse along the epipolar line to detect the shadow, as shown in Fig.3.

Now that consider we use a light field camera. A LF image captured by Lytro is composed of dense subaperture images (13×13). We first detect shadows in the center subaperture image. We use $I_n^+(x)$ to represent a image lit by the n-th light, $n = 1, 2, 3, 4$ in our case. And $I_n(x) = I_n^+(x) - I_a$

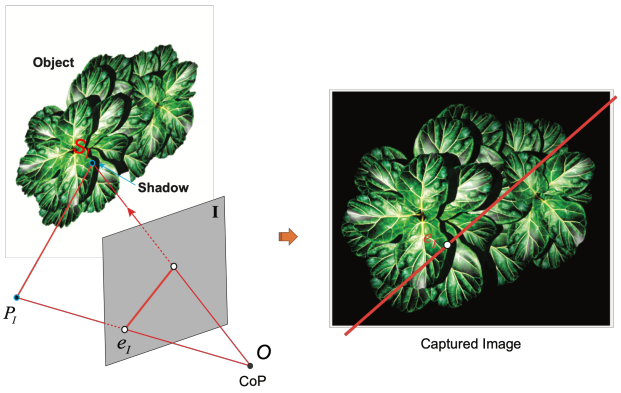


FIGURE 3. Geometric relationship between a 3D scene point and it shadow.

represents the image has the ambient component removal, where the image I_a is captured under ambient light. Given a 3D point X on the surface and its corresponding pixel x on the image, we can express the intensity of X under lambertian assumption if X is lit by the n-th light:

$$I_n(x) = \rho(x) I_n \mathbf{n_x} \cdot \mathbf{D_n} \tag{1}$$

where $\mathbf{D_n}$ reflects the light direction from X to the light, $\mathbf{n_x}$ is the surface normal and I_n is the light intensity. We use $\rho(x)$ to represent the reflectance at X. Therefore, we derive the ratio at x as follows:

$$R_n(x) = \frac{I_n(x)}{I_{max}(x)} = \frac{\rho(x) I_n \mathbf{n_x} \cdot \mathbf{D_n}}{max(\rho(x) I \mathbf{n_x} \cdot \mathbf{D})} \tag{2}$$

where we capture a maximum composite image I_{max}, which is approximated by distributing the 4 flashes around camera with the same magnitude and the extremely small baseline than depth of the scene. And we have:

$$R_n(x) \approx \begin{cases} 1 & \text{area lit by the light} \\ 0 & \text{shadow area} \end{cases} \tag{3}$$

Therefore, we compute a ratio image from the image captured with one flash on. The ratio image emphasizes the depth edges instead of texture edges. Traversing along the epipolar line, the entry to the shadow corresponds to the depth edge. If the side of the depth edge locates at the part same to the epipole, the side indicates the positive, and the other side indicates the negative. Specifically, when we put the flash on horizontal and vertical side of center of camera, the direction of epipolar ray for the light is also horizontal or vertical. Fig.4(a) shows one ratio map, and alone the horizontal red line, we draw its corresponding ratio values in (b). In one flash image, we can detect depth edge on the opposite direction of epipolar ray. With four epipolar rays with different directions, we can get the whole depth edges via combining the depth edges from four directions. Our algorithm is showed as follows:

S is the edge detector kernel, for example, to detect positive edge from right to left, the kernel is $[-1, 1]$ or positive

Algorithm 1 Depth edge detection

for n= 1 to 4 **do**
  $R_n = I_n/I_{max}$
  T_n^+ = conv(R_n,S) //Detect positive edge
  T_n^- = conv(R_n,S/)
  $P_n = T_n^+ \geq \epsilon$
  $N_n = T_n^- \leq -\epsilon$
  $C_n^+ = P_n \cdot T_n^+$
  $C_n^- = N_n \cdot T_n^-$
end for
$P = P_1|P_2|P_3|P_4$
$N = N_1|N_2|N_3|N_4$
$C^+ = \sum C_n^+$
$C^- = \sum C_n^-$
if P and N has conflict **then**
  if $C_{i,j}^+ \geq C_{i,j}^-$ **then**
    assign positive
  else
    assign negative
  end if
end if

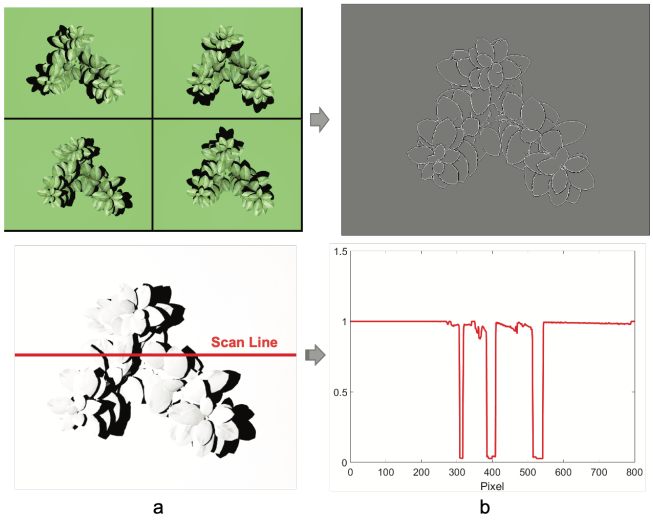


FIGURE 4. First row shows our extracted depth edges. Second shows sample ratio images. (a) using multi-flash photography; (b) shows how ratio changes along the traversal direction (red line).

detector kernel, and S' is $[1, -1]$ or the negative detector kernel, and can detect negative edge. We set a threshold ϵ to get the binary positive and negative edges, e.g., P_n^+ and P_n^-. We record the correspondence confidence value C_n for each edge in different sign. Since we use 4 flash lights in our setup, each edge may be assigned positive and negative signs at different flash images. To resolve conflicting signs, we accumulate confidence value C_n of the same sign, and use the sign of the highest confidence value as the final sign for the edge.

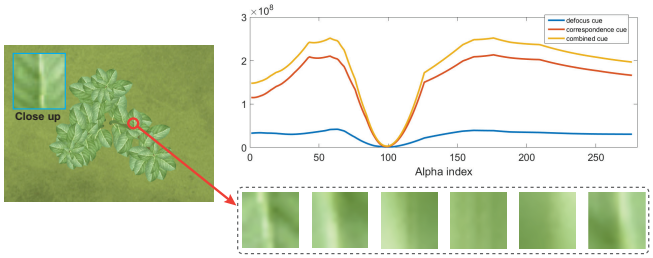


FIGURE 5. This figure shows the different cue values corresponding to the change of α.

IV. OCCLUSION AWARE STEREO MATCHING

Once we detect the depth edges, we apply a novel light field stereo matching scheme that combines defocus, correspondence and sharp depth edge cues. Recall that a light field is composed of densely sampled subaperture images, Ng et al. [11] conduct refocusing effect by shearing the 4D light field image as follows:

$$L_\alpha(u, v, s, t) = L(u + s(1 - \frac{1}{\alpha})), v + t(1 - \frac{1}{\alpha}), s, t) \quad (4)$$

where L is the input LF image. u and v are coordinates of spatial patch. s and t are coordinates of angular patch. And, we can do refocusing to different depth positions via adjusting α, where α is the ratio of current focus depth to the original focus depth.

We assume that the center subaperture image is located at $s = 0, t = 0$. Given a pixel coordinate $x = (u, v)$, we can generate an angular patch by querying the same x in different subaperture images. So the size N_x of an angular patch is determined by the number of subaperture images in the LF, e.g., 13×13 in our case. If there is no occlusion existing at the pixel, an angular patch keeps photo-consistency at correct focus depth. Given a refocused image $L_\alpha(u, v, s, t)$, our goal is to find the optimal α with the highest contrast at the pixel. Tao et al. [23] estimate the defocus and correspondence cues in the sheared EPI to recover the depth. For the defocus cue, we first calculate the average value by following:

$$\overline{L}_\alpha(x) = \frac{1}{N_x} \sum_{s,t} L_\alpha(x, s, t) \quad (5)$$

where s, t are coordinates of angular patch. The defocus cue is as following:

$$D_\alpha(x) = (\overline{L}_\alpha(x) - L(x, 0, 0))^2 \quad (6)$$

We also calculate the variance for angular patch of x:

$$V_\alpha(x) = \frac{1}{N_x - 1} \sum_{s',t'} (L_\alpha(x, s', t') - \overline{L}_\alpha(x))^2 \quad (7)$$

We use the variance as the correspondence term:

$$C_\alpha(x) = V_\alpha(x) \quad (8)$$

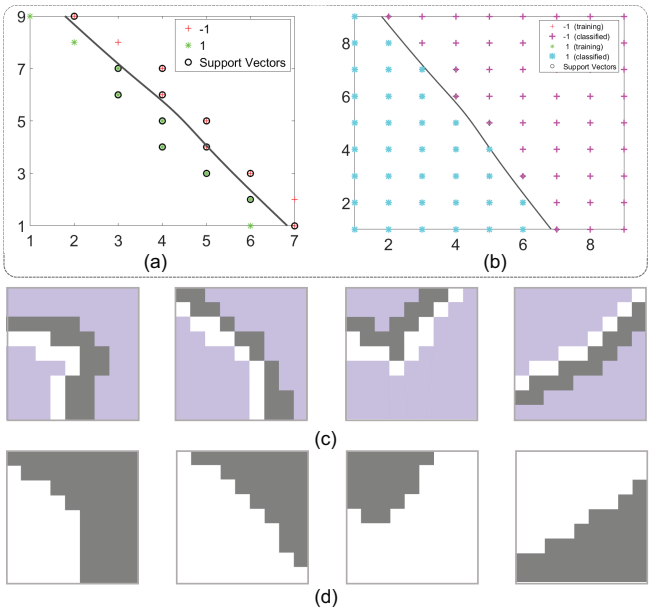


FIGURE 6. (a) and (b) show the training and classification process of SVM. (c) and (d) show the labeling results before and after the SVM process.

Fig.5 show the different cue values corresponding to the change of the disparity value. Since the correct depth for pixel x has the smallest variance in the angular path, the defocus and correspondence should also be the smallest. We have following target function:

$$\alpha^*(x) = \underset{\alpha}{\operatorname{argmin}}(C_\alpha(x) + D_\alpha(x)) \tag{9}$$

However, the photo-consistency of angular patches is only valid if the point is visible to all views. If the pixel x in central subaperture image is on or near the occlusion boundary, the angular patch will contain two parts, e.g., one part of pixels comes from the occluder and the other comes from different scene components being occluded. We name the region including x as the target region, and the rest as the non-target region. In other words, the target region maintains photo-consistency while the non-target region doesn't.

We present a novel technique to separate the angular patch according to occlusion boundary. Compared to the method in [18], our method can handle more complex boundaries in different shapes. We observe that occlusion boundary in angular patch should maintain a similar shape as in the spatial domain. One pixel in spatial patch may correspond to several pixels in angular patch, so there exits a scaling relationship of the shape. For example, a straight line keeps the same slop in spatial and angular patch, as described in [18]. For a curve, the correlation also includes a scaling factor. This implies that we can recover the boundary in angular patch by rescaling the boundary in spatial domain. The rescaling ratio is determined by the baseline of pixels in spatial domain to the baseline of pixels in angular domain.

To elaborate, we first extract a spatial window centered at the current pixel in spatial domain from the depth edge

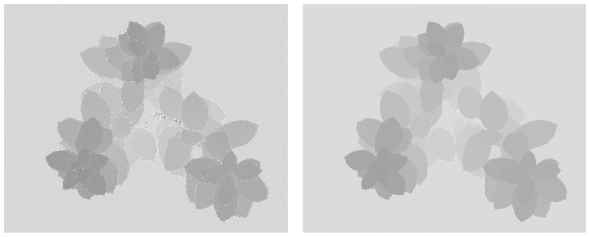


FIGURE 7. Left shows the original depth map with noise. Right shows our propagation that achieves much less noisy result using additional geometry constraints.

image. If the center pixel is on or near the positive edge, we simply deem the pixel as on the occluder (the foreground). If the center pixel on or near the negative edge, we deem the pixel in on the occludee (the background). In the spatial window, the pixels can be uniquely classified into thee categories: the ones on the positive edge, on the negative edge or not on any edge.

We assign 1 to pixels on the positive edge and -1 to pixels on the negative edge. Other pixels have no labels. Based on this labeling, we adopt a polynomial kernel SVM to learn the shape of occlusion boundary. The SVM is a supervised learning method used for data classification. We have two labeling results in our case and the SVM training algorithm can build a model to assign labels to the rest pixels without labels as a binary linear classifier. The polynomial kernel measures the similarity of vectors and can enhance SVM to learn non-linear models. After training, we use the SVM to separate the angular patch with rescaled coordinates. We show our technique can handle various complex occlusion patterns where the occlusion boundaries can be straight lines, curves or other shapes, e.g., the boundaries in the plant scene. Fig.6 shows our SVM process. After SVM, we assign each pixel a label of foreground or background.

Furthermore, in order to avoid reversed patch in Wang et al. [18], e.g., a case where the region on a angular patch shows low variance when focusing at incorrect depth. We add a penalty to wrong refocusing. Given the separated region, we calculate the individual average of regions in spatial patch, p_1 and p_2 and the average of regions in angular patch $\overline{L}_{\alpha 1}, \overline{L}_{\alpha 2}$.

$$\frac{|\overline{L}_{\alpha,1} - p_1| + |\overline{L}_{\alpha,2} - p_2|}{|\overline{L}_{\alpha,1} - p_2| + |\overline{L}_{\alpha,2} - p_1|} > \epsilon \tag{10}$$

When reversed patches occur, the ratio will be larger than the threshold ϵ. We add a penalty to avoid this incorrect case. In our case, $\epsilon = 1$. Finally, we solve Eq.9 to calculate the depth map.

A. DEPTH PROPAGATION

Recall that pixels separated by an depth edge exhibit depth discontinuity, and pixels on the same side of the depth edge should have similar depth values. Similar to XXX, we apply an iterative dampened-spring diffusion technique [24] for interpolating the depth values of non-edge pixels. Specifically, we employ a depth field D with a velocity field V initialized

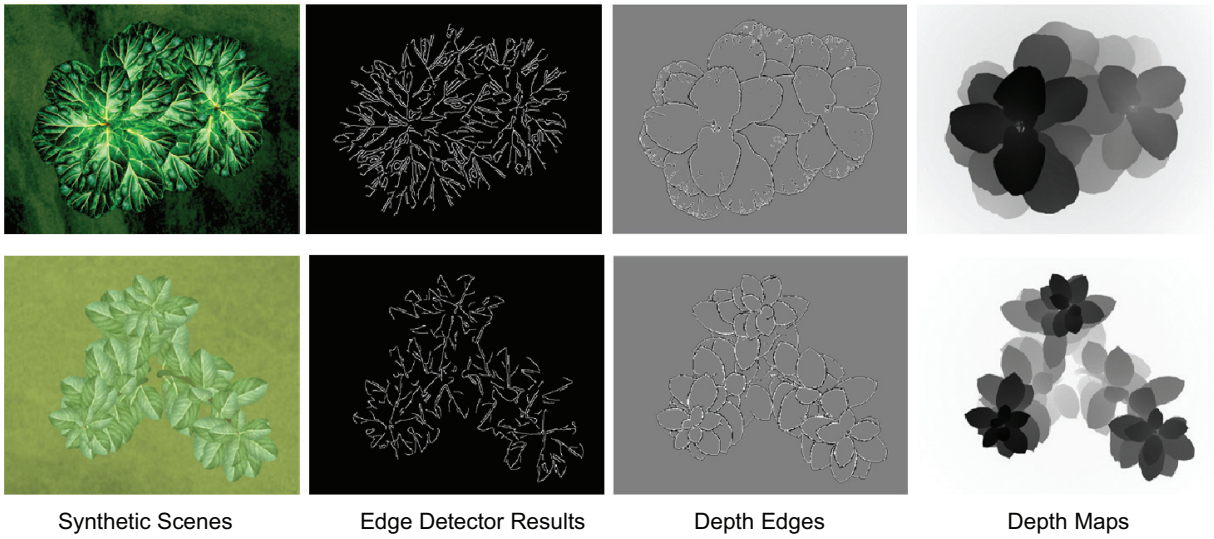


FIGURE 8. Results on two synthetic scenes using our depth edge detection vs. classical Canny edge detector.

as zero. V aims to measure the depth difference between neighboring pixels. At each iteration the depth and velocity fields are updated as follows:

$$V'_{u,v} = d{\cdot}V_{u,v}+k{\cdot}(D_{u-1,v}+D_{u+1,v}+D_{u,v-1}+D_{u,v+1}-4{\cdot}D_{u,v}) \tag{11}$$

$$D'_{u,v} = D_{u,v} + V'_{u,v} \tag{12}$$

where d and k are dampening and spring parameters. We iterate the process until the difference between consecutive rounds below a pre-determined threshold. We also forbid propagating depth values from pixels on negative edge to the positive edge, i.e., background object should have minimal influence to the foreground ones. To do so, we introduce $E_{u,v}$ that assigns pixels on the depth edge to have value 1 and the rest 0. We revised iteration works as follows:

$$\begin{aligned} V'_{u,v} = d \cdot V_{u,v} + k \cdot ((D_{u-1,v} - D_{u,v}) \cdot (1 - E_{u-1,v} \cdot E_{u,v}) \\ + (D_{u+1,v} - D_{u,v}) \cdot (1 - E_{u+1,v} \cdot E_{u,v}) \\ + (D_{u,v-1} - D_{u,v}) \cdot (1 - E_{u,v-1} \cdot E_{u,v}) \\ + (D_{u,v+1} - D_{u,v}) \cdot (1 - E_{u,v+1} \cdot E_{u,v})) \end{aligned} \tag{13}$$

Our occlusion aware depth interpolation scheme is simple but robust and efficient. Fig.7 shows the results before and after propagation. The propagation process decreases the noise level while maintaining smoothness of the final depth map.

V. EXPERIMENTAL RESULTS

We conduct experiments on both synthetic and real data. All our experiments run on a desktop computer with a Intel Xeon CPU, a 32GB memory and a NVidia 1080ti GPU card. We compare our methods with the state-of-the-art methods, including [19], [25], and [18][1].

A. SYNTHETIC DATA

Several open-access datasets are available for evaluating light field depth estimation algorithms, such as HCI dataset [26] and 4D light field benchmark [27]. However, these datasets do not have occlusion edge maps provided by our multi-flash cameras. We therefore use the 3ds Max software to render synthetic data for our evaluation. Specifically, we render each light field at a resolution of $13 \times 13 \times 539 \times 359$. Next, we emulate the multi-flash capture process by positioning the light field camera in the center and flash lights surrounding it. The baseline between the LF camera and the flash is set relatively small compared to the distance to the scene. We render five LFs, one under ambient light and four with individual flash on.

Fig.8 compares the depth edge extraction results between Canny detector and our algorithm. Directly applying Canny edge detector on the color image produces strong artifacts and mistakens many texture edges and occlusion edges. Instead, our multi-flash solution not only manages to separate texture vs. occlusion edges but produces very high quality occlusion edges, especially on complex plants. Our technique also manages to produce a high quality depth map by fusing the defocus cues with the detected occlusion edges. Our depth estimation maintains sharpness on occlusion edges while actively exploiting texture edges for reliable depth inference.

Fig.9 shows the disparity maps estimated by different methods. We normalize each disparity map to $[0, 1]$ via:

[1] We use the official codes/exes from [18], [19] to test the results. For [25], we were unable to use the exe file from the website and we have used our own implementation.

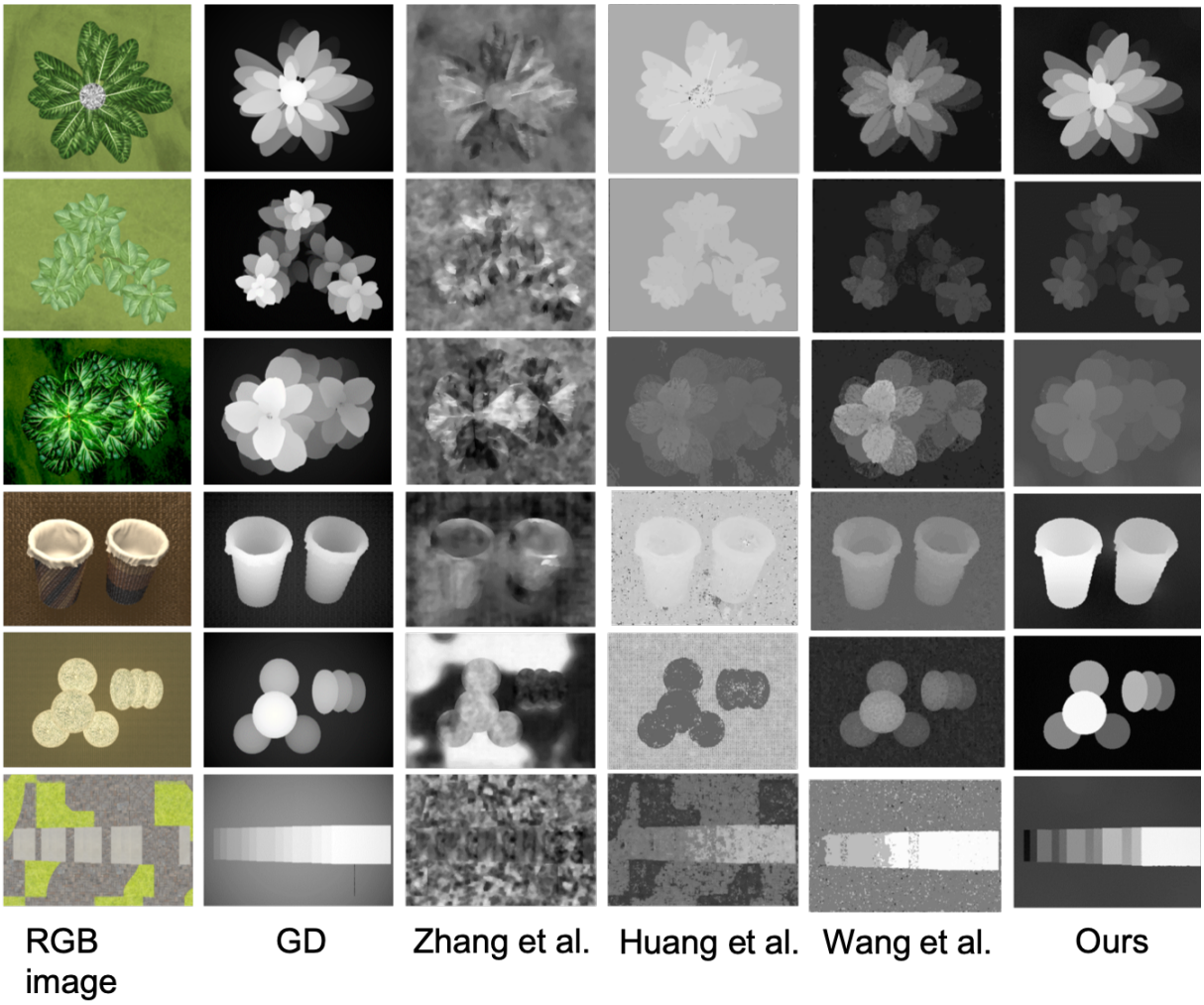


FIGURE 9. This figure shows comparison results on synthetic data using our algorithm and state-of-the-art methods. Our algorithm outperforms others and generates much better results with sharp depth boundaries. GD shows the ground truth disparity maps.

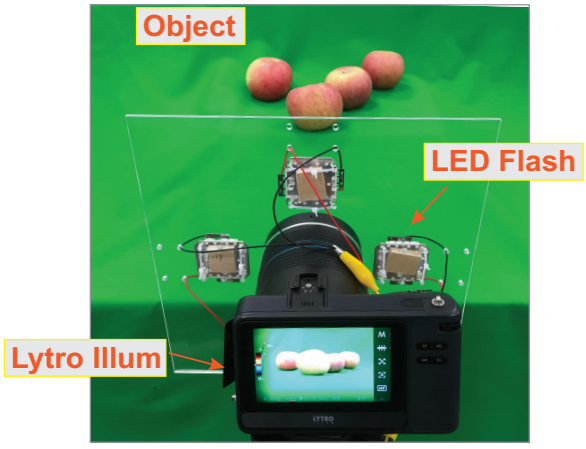


FIGURE 10. We couple a Lytro camera with a ring of four flash lights, and use a hardware trigger to control the camera and the lights.

$$D^* = \frac{D - min}{max - min} \tag{14}$$

where D is the original disparity map, and D^* is the normalized disparity map. min and max correspond to the minimum and maximum values in D.

The top three rows in Fig.9 show the disparity results on the plant scene. Each plant exhibits its unique occlusion pattern and shape while the leaves exhibit repeat-textures. [xxx] produces low quality reconstruction. This is because it relies on the linear color-disparity assumption that breaks under occlusions. The results from [xxx] exhibit deteriorated occlusion boundaries due to the use of sized windows. In contrast, [xxx] produces sharper edges by employing angular analysis similar to ours. However, the edges tend to be noisy and inconsistent. Benefiting from our multi-flash setting, we manage to produce sharp depth edges that shows advantages on depth estimation.

The bottom three rows correspond to scenes composed simpler geometry with few occlusion. However, the fore-

ground objects have a similar appearance (color and texture) as the background. This is another violation of the color-depth assumption in [xxx] and their method produces low quality reconstruction. [xxx] also produces noisy results. The angular approach by Wang et al. [] produces reasonble results but their depth estimation is less accurate and the results are sitill noisy. In contrast, our technique produces much cleaner and sharper results that are close to the ground truth.

B. REAL DATA

For experiments on real data, we construct a multi-flash light field camera by surround a Lytro Illum camera with a ring of four flash lights as shown in Fig.10. Our flashes use LED lights and to synchronize the camera with the flash, we employ an Arduino board to control both the camera shutter and the flash trigger. To calibrate the Lytro camera, we adopt the Light Field Toolbox method [28], and extract $13 \times 13 \times 539 \times 359$ from each LF after calibration. For each real scene, Same as the synthetic data setting, we capture 5 LF images, one without flash and four with each individual flash on. Our SVM algorithm uses the polynomial order of 2.

Fig.11 shows the stereo matching results using ours vs. the state-of-the-art. Our MFLF outperforms the rest in both robustness and accuracy. For the plant scenes, such as the top two rows and the bottom row, our method produces very sharp and clean disparity maps. Instead, state-of-the-art methods miss these edges and produce overly blurred results. Wang et al. [xxx] produces better results but they are noisy near occlusion boundaries. For highly textured scenes (row 2 and 3), the results include large errors near the texture edges. In contrast, our approach produces uniformly high quality results on scenes with or without rich texture.

It is important to note that an important application of light field stereo matching is 3D reconstruction, i.e., the disparity map needs to be converted to point cloud and fused into 3D models. Fig.12 shows the final point cloud recovered by from our disparity map using MFLF. By changing the viewpoint, we observe that our approach preserves high fidelity in geometry (e.g., leaves are separated rather than adhesive to each other the plant scene).

VI. CONCLUSION

We have presented a multi-flash light field technique (MFLF) that employs the benefits from both multi-flash photography and light field imaging for depth estimation on scenes exhibiting complex occlusion patterns. In particular, we resort to multi-flash for occlusion boundary extraction and light field for reliable multi-view depth estimation. In particular, we adopt a novel SVM based technique to classify pixels and employ defocus anc occlusion cues for reliable disparity map generation. We have demonstrated MFLF is particularly useful for reconstructing scenes challenging to state-of-the-art approaches, e.g., plants and flowers that exhibit either no or extremely rich textures, along with complex occlusion boundaries.

Our current approach uses four flash images to extract depth edge in central subaperture view. For future work, we plan to explore the abundant ray geometry information embedded in light field for shadow analysis, i.e., the shadow geometry in 4D ray space. We also plan to explore how to fuse disparity maps obtained from MFLF from multiple viewpoints into the final mesh. Another potential extension is to use the mobile flash lights and cameras, e.g., latest Android phones are already equipped with multiple cameras (as many as 4). The work by Guo et al. [xxx] implemented a mobile multi-flash system that can be easily implemented on the multi-camera mobile phones for directly producing MFLF.

REFERENCES

[1] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm, "Patchmatch based joint view selection and depthmap estimation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1510–1517, 2014.
[2] 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.
[3] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," 2015 IEEE International Conference on Computer Vision (ICCV), pp. 873–881, 2015.
[4] J. Heinly, J. L. Schönberger, E. Dunn, and J. M. Frahm, "Reconstructing the world* in six days *(as captured by the yahoo 100 million image dataset)," 2015.
[5] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo - stereo matching with slanted support windows," in BMVC, 2011.
[6] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz, "Multi-view stereo for community photo collections," 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8, 2007.
[7] R. Feris, R. Raskar, L. Chen, K. Tan, and M. Turk, "Multiflash stereopsis: Depth-edge-preserving stereo with small baseline illumination," IEEE transactions on pattern analysis and machine intelligence, vol. 30, no. 1, pp. 147–159, 2008.
[8] P. Cavanagh and Y. G. Leclerc, "Shape from shadows." Journal of experimental psychology. Human perception and performance, vol. 15 1, pp. 3–27, 1989.
[9] R. T. Frankot and R. Chellappa, "A method for enforcing integrability in shape from shading algorithms," IEEE Trans. Pattern Anal. Mach. Intell., vol. 10, pp. 439–451, 1988.
[10] S. Savarese, "Shape reconstruction from shadows and reflections," 2005.
[11] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan et al., "Light field photography with a hand-held plenoptic camera," Computer Science Technical Report CSTR, vol. 2, no. 11, pp. 1–11, 2005.
[12] X. Guo, J. Sun, Z. Yu, H. Ling, and J. Yu, "Mobile multi-flash photography," in Digital Photography X, vol. 9023. International Society for Optics and Photonics, 2014, p. 902306.
[13] M. Levoy and P. Hanrahan, "Light field rendering," ser. SIGGRAPH '96. ACM, 1996.
[14] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," ser. SIGGRAPH '96. ACM, 1996.
[15] J. C. Yang, M. Everett, C. Buehler, and L. McMillan, "A real-time distributed light field camera."
[16] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4d light fields," in 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 41–48.
[17] C. Chen, H. Lin, Z. Yu, S. Bing Kang, and J. Yu, "Light field stereo matching using bilateral statistics of surface cameras," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1518–1525.
[18] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3487–3495.
[19] Y. Zhang, Z. Li, W. Yang, P. Yu, H. Lin, and J. Yu, "The light field 3d scanner," in 2017 IEEE International Conference on Computational Photography (ICCP). IEEE, 2017, pp. 1–9.

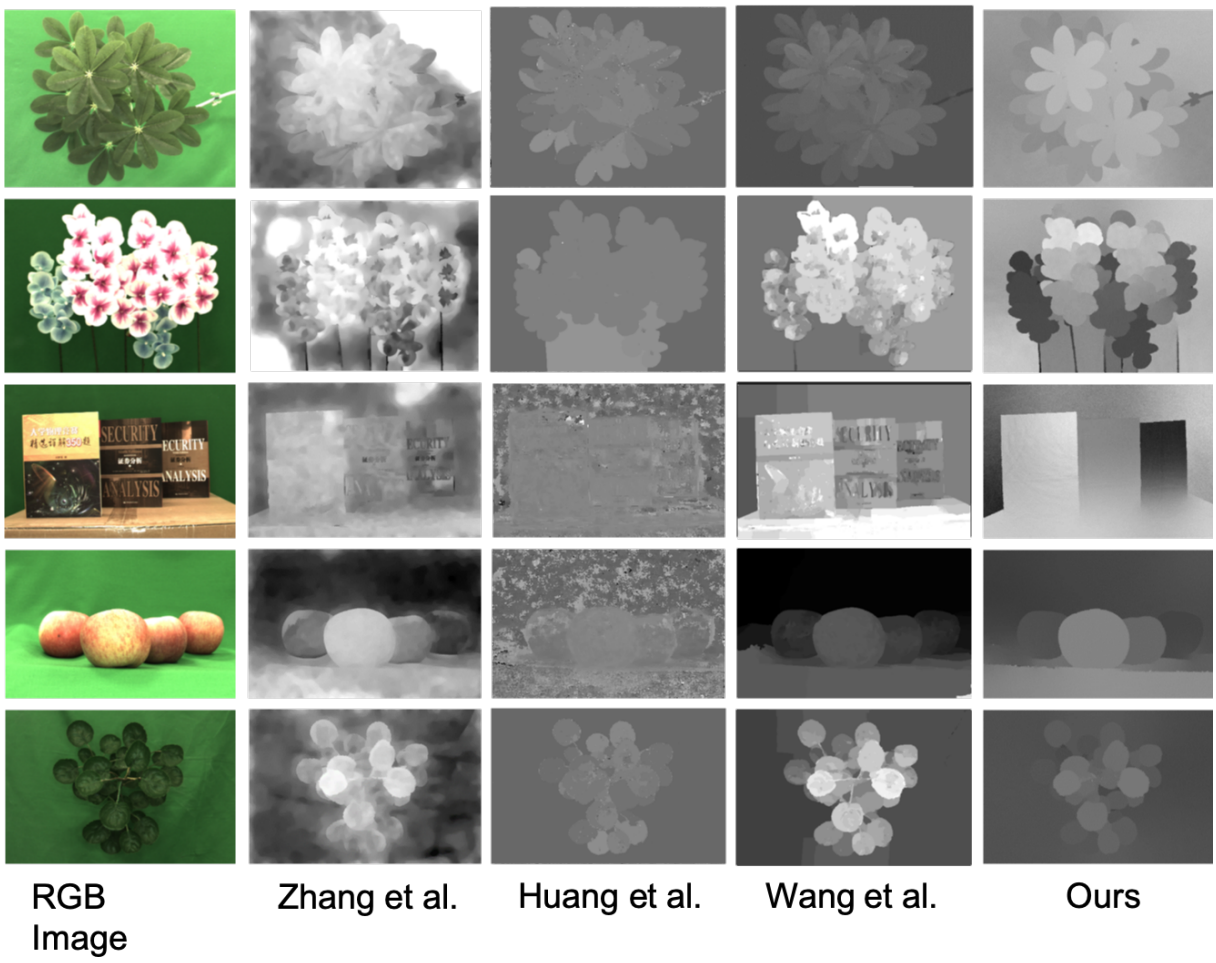


FIGURE 11. This figure shows comparison results on real data using our algorithm and state-of-the-art methods. Our algorithm is the most robust and accurate compared to the others. Due to the accurate depth edges from multi-flash setting, our disparity maps keep precise depth boundaries.

FIGURE 12. We generate the point cloud from our estimated depth map, and show two views of the point cloud.

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

[21] R. Raskar, K.-H. Tan, R. Feris, J. Yu, and M. Turk, "Non-photorealistic camera: depth edge detection and stylized rendering using multi-flash imaging," in ACM transactions on graphics (TOG), vol. 23, no. 3. ACM, 2004, pp. 679–688.

[22] R. Feris, R. Raskar, L. Chen, K.-H. Tan, and M. Turk, "Discontinuity preserving stereo with small baseline multi-flash illumination," vol. 1, 11 2005, pp. 412– 419 Vol. 1.

[23] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 673–680.

[24] S. F. Johnston, "Lumo: illumination for cel animation," in Proceedings of the 2nd international symposium on Non-photorealistic animation and rendering. ACM, 2002, pp. 45–ff.

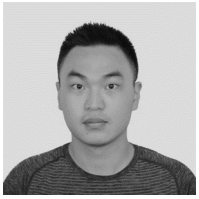
[25] C.-T. Huang, "Robust pseudo random fields for light-field stereo matching," in The IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[26] S. Wanner, S. Meister, and B. Goldlücke, "Datasets and benchmarks for densely sampled 4d light fields," in Vision, Modeling & Visualization, 2013, pp. 225–226.

[27] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4dÂălight fields," in Computer Vision – ACCV 2016, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham: Springer International Publishing, 2017, pp. 19–34.

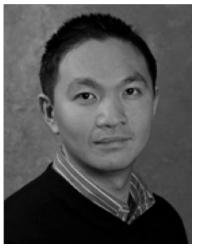
[28] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 1027–1034.

YINGLIANG ZHANG received the BE degree of communication engineering from Ningbo University, China, in 2014. He is pursuing the Ph.D degree of computer science from ShanghaiTech University, China. His research interests include image-based 3D reconstruction, light field rendering, and light field reconstruction.



CHENG SHEN received the BS degree in 2018 at ShanghaiTech University, School of Information Science and Technology. He was advised by Jingyi Yu. His research interest is in computer vision, including light-field technologies.He is currently pursuing his MS degree in Computer Science at Columbia University.

WEI YANG received the BEng and MS degrees from the Huazhong University of Science and Technology and Harbin Institute of Technology respectively, and the PhD degree from the University of Delaware (UDel) in Dec 2017. He joined the DGene. Corp (Prev. Plex-VR) as a research scientist in Mar 2018. His research interests include computer vision and computer graphics, with special focus in computational photography and 3D reconstruction.



JINGYI YU received the B.S. degree from the California Institute of Technology, Pasadena, CA, USA, in 2000, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2005. He is currently an Professor with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China, and an Associate Professor with the Department of Computer and Information Sciences and the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA. His current research interests include computer vision and computer graphics, in particular, computational cameras and displays. Prof. Yu has served as the Program Chair of the 2011 Workshop on Omnidirectional Vision and Camera Networks, General Chair of the 2008 International Workshop on Projector-Camera Systems, and Area and Session Chair of the 2011 International Conference on Computer Vision. He was a recipient of the NSF CAREER Award and the AFOSR YIP Award. He is an Editorial Board Member of the IEEE Transactions on Pattern Analysis and Machine Intelligence, The Visual Computer Journal, and Machine Vision and Application.