4D Light Field Superpixel and Segmentation

Hao Zhu, *Student Member, IEEE*, Qi Zhang, *Student Member, IEEE*, Qing Wang, *Senior Member, IEEE*, and Hongdong Li, *Member, IEEE*

***Abstract*— Superpixel segmentation of 2D images has been widely used in many computer vision tasks. Previous algorithms model the color, position, or higher spectral information for segmenting a 2D image. However, limited to the Gaussian imaging principle in a traditional camera, where each pixel is formed by summing lots of light rays from different angles, there is not a thorough segmentation solution to eliminate the ambiguity in defocus and occlusion boundary areas. In this paper, we consider the essential element of image pixel, i.e., rays in light space, and propose light field superpixel (LFSP) to eliminate the ambiguity. The LFSP is first defined mathematically and then two evaluation metrics, named LFSP self-similarity and effective label ratio, are proposed to evaluate the refocus-invariant and full-sliced properties of segmentation. By building a clique system containing 80 neighbors in light field, a robust refocus-invariant LFSP segmentation algorithm is developed. Experimental results on both synthetic and real light field datasets demonstrate the advantages over the current state of the art in terms of traditional evaluation metrics. Additionally, the LFSP self-similarity evaluations under different light field refocus levels show the refocus-invariance of the proposed algorithm. The full-sliced property of the proposed LFSP algorithm is verified by comparing it with the classical supervoxel algorithms. Finally, an LFSP-based application is demonstrated to show the effectiveness of LFSP in light field editing.**

***Index Terms*— Light field, superpixel segmentation, refocus-invariant, LFSP self-similarity, full-sliced, effective label ratio.**

Fig. 1. Light field superpixel segmentation on real scene light fields. The left image is a 2D slice of LFSP segmentation in central view. For each region in the right, the first row shows the close-up and the second and third rows are corresponding segmentations on horizontal and vertical EPIs respectively.

I. Introduction

SUPERPIXEL is the key fundamental to connect pixel-based low-level vision to object-based high-level understanding, which aims at grouping similar pixels into larger and more meaningful regions to increase the accuracy and speed of post processing [1]. To accomplish a good over-segmentation, previous works [2]–[8] have built various grouping methods to model the proximity, similarity and good continuation [1] in the classical Gestalt theory [9]. However, existing superpixel techniques are becoming more and more difficult to meet the requirements of modern applications. In computer vision, due to the Gaussian imaging model in traditional imaging system, there inevitably exist ambiguities in object boundaries where the light rays emitted from different objects are accumulated, including vignette, occlusions. These ambiguities may cause image degradation to disturb superpixel segmentation and further to decrease the accuracy of object segmentation and recognition. In computer graphics, previous techniques are designed for 2D image and cannot handle the recent 4D light field data [10]–[12] which has a 2D grid of 2D images. Each sub-aperture image in light field can only be segmented independently, so the full 4D light field cannot be edited simultaneously.

To overcome ambiguities and asynchrony in traditional superpixel segmentation, we introduce light field superpixel segmentation. It is known that light field [10], [13], [14] records scene information both in angular and spatial spaces, forming a 4D function named $L(u, v, x, y)$. The light field data can benefit superpixel segmentation on two aspects. First, since each ray is recorded in light field, the ambiguity in object boundaries can be well analyzed. Second, multi-view nature of light field enables the bottom-up grouping not only in the color and position but also in the structure.

Manuscript received October 25, 2018; revised April 26, 2019; accepted June 25, 2019. Date of publication July 15, 2019; date of current version September 12, 2019. This work was supported by the NSFC under Grant 61531014. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chia-Kai Liang. *(Corresponding author: Qing Wang.)*

H. Zhu, Q. Zhang, and Q. Wang are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: qwang@nwpu.edu.cn).

H. Li is with the College of Engineering and Computer Science, The Australian National University, Canberra, ACT 0200, Australia.

Digital Object Identifier 10.1109/TIP.2019.2927330

However, 4D light field segmentation is still a challenging task. As mentioned in [15], light field segmentation faces two major difficulties. First, each segmentation in light field ought to be propagated coherently to preserve the redundancy of 4D data. Second, although the depth is implicitly embedded in multi-view images, it is still unavailable, inconvenient and imperfect to segment the full 4D data.

In this paper, we explore superpixel segmentation on 4D light field. We show that the LFSP can represent proximate regions better, especially in object boundaries (in Section III). Traditional superpixel is just a 2D slice of LFSP by fixing angular dimensions. When fixing spatial dimensions, the angular segmentation in LFSP coincides with light field occlusion theory in [16]. Additionally, LFSP differs from supervoxel in definitions and assumptions.

In Section IV-A, we first propose a refocus-invariant LFSP segmentation algorithm by defining a clique system containing 80 neighbors in light field and introduce a 2D disparity map into the energy function. Then, two metrics, namely the LFSP self-similarity and effective label ratio, are proposed to evaluate refocus-invariant and full-sliced properties of segmentation. In Section V, extensive experiments are carried on synthetic data and real scene light fields captured by Lytro [11]. Quantitative and qualitative comparisons verify the effectiveness and robustness of our algorithm.

This is an extended version of the work at CVPR [17]. Compared with the conference paper, we analyze the differences between the proposed LFSP and the classical video supervoxel algorithms both in theory and experimental performance. Based on theoretical analysis, a new metric is designed for evaluating LFSP segmentation. Additionally, a new light field dataset is generated to show the differences better. Apart from the comparison with supervoxel, we also provide a deeper analysis on LFSP segmentation with more parameters and add more results both on synthetic and real light fields. Finally, a LFSP based application is demonstrated, showing the effectiveness of LFSP in light field editing.

In summary, our main contributions are,

1) The definition of light field superpixel.

2) Two evaluation metrics, namely the light field self-similarity and effective label ratio, are proposed for evaluating refocus-invariant and full-sliced properties of LFSP segmentation.

3) A robust refocus-invariant superpixel segmentation algorithm in the full 4D light field, which provides consistent segmentation for a same light field under different refocus levels.

4) A 4D light field segmentation dataset benchmark, which has more non-Lambertian object classes and a larger absolute disparity range.

II. Related Works

A. Light Field in Computer Vision

Unlike conventional imaging systems, light field cameras [11], [12] can record the appearance of objects in a higher 4D space, and have benefited many problems in computer vision, such as depth and scene flow estimation [16], [18], [19], saliency detection [20], super resolution [21] and material recognition [22]. Light field can generate depth map [16], [18], [23] from multiple cues such as epipolar lines, defocus and correspondence [23]. Compared with traditional multi-view stereo based matching methods, light field based methods can provide a high quality sub-pixel depth map, especially in occlusion boundaries. In this work, the algorithm developed by Zhu *et al.* [24] is utilized to generate depth map for LFSP segmentation.

For light field segmentation, only a few of approaches have been proposed in literatures, especially most of them are interactive. Wanner *et al.* [25] proposed GCMLA (globally consistent multi-label assignment) for light field segmentation, where the color and disparity cues of input seeds are used to train a random forest, which is used to predict the label of each pixel. Mihara *et al.* [26] improved the GCMLA by building a graph in 4D space. A '4-neighbouring system' in light field is defined and the 4D segmentation is optimized using the MRF. Hog *et al.* [27] exploited light field redundancy in ray space by defining free rays and ray bundles. A simplified graph-based light field is constructed, which greatly decreases computational complexity. Xu *et al.* [28] segmented 4D light field automatically. By defining the LF-linearity and occlusion detector in light field, a color and texture independent algorithm for transparent object segmentation is proposed. Compared with previous segmentation algorithms, our work focuses on a smaller unit – the superpixel in light field, which is the basis for many computer vision tasks [1], [20], [29]–[33]. More recently, Hog *et al.* [34] proposed super-rays for light field processing. The 2D grids of central view are projected to 4D light field according to the disparity. Then the whole 4D hyper volume is optimized in an iterative k-means clustering framework using the color and position (similar to [7]), which results in high computational complexity. In fact, it is not necessary to optimize the segmentation from 4D volume. Since 2D images in different views are correlated by the disparity, we only need to project 2D segmentation to 4D light field and optimize boundary pixels.

B. Superpixel Segmentation

Superpixel segmentation of 2D image has been researched for years and many excellent algorithms have been proposed. Shi and Malik [2] treated the image as a 2D graph using contour and texture cues. They proposed normalized cuts to globally optimize the cost function. Felzenszwalb and Huttenlocher [3] improved the efficiency of normalized cuts using an efficient graph cuts method. Liu *et al.* [35] introduced an entropy rate term and balance term into a clustering objective function to preserve jagged object boundaries. Achanta *et al.* [7] adapted a k-means clustering algorithm to seek cluster's center iteratively. Li and Chen [8] mapped traditional color and position features into a higher spectral space to produce more compact and uniform superpixels.

III. LFSP Definition

All previous superpixel approaches are built on traditional 2D image and are not suitable for 4D LFSP segmentation.

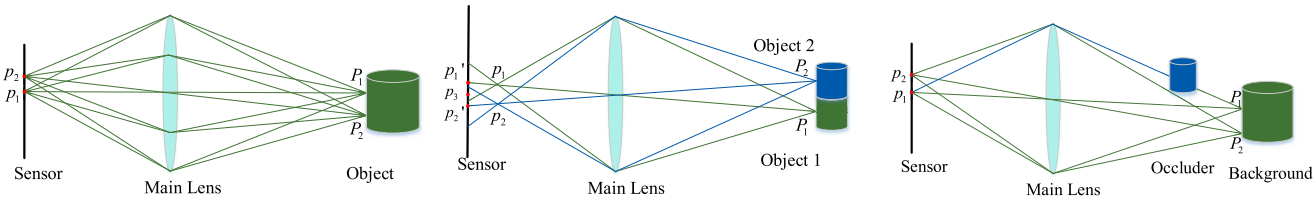


Fig. 2. (a) All rays emitted from P_1, P_2 to p_1, p_2 are contained in the LFSP. (b) The rays emitted from P_1, P_2 converge to p_1, p_2, forming two defocus areas centered at p_1', p_2' respectively. p_3 suffers from rays emitted from both P_1 and P_2. (c) There is an obstruction between the background and the main lens, and part of the rays emitted from the green point P_1 are occluded by the blur obstruction. There is an ambiguity in the segmentation for these mixed points here.

Although 4D light field can be treated as a serial of 2D images and each image can be segmented using these algorithms, ignoring the connection between these images not only cuts off segmentation consistency but also increases running time (Fig.17).

In contrast to previous superpixel segmentation algorithms, we treat 4D light field as a whole and improve the accuracy and running time of LFSP using angular coherence in light field. In this section, we first present the definition of light field superpixel (LFSP). Then the differences and characteristics of LFSP compared with traditional 2D superpixel and supervoxel are analyzed.

A. LFSP Definition

Superpixel algorithms model the proximity, similarity and continuation of the object in a 2D image. We ray-trace the pixels in the superpixel from a 2D image to the 3D space (see Fig.2a). In the propagation, each pixel spreads into multiple light rays and reaches the object in the real world. In this case, all rays are included in the LFSP.

The inverse propagation mentioned above can only model all-in-focus and non-occlusion situations, however the following two conditions are difficult to achieve actually. First, when the camera is focusing on a different depth (Fig.2b), defocus blurs occur on the sensor and original sharp boundary is blurred. Since the boundary pixel suffers rays emitted from different objects, it is ambiguous to segment it. Second, for the occlusion case (Fig.2c), when the camera is focusing on the background, a part of light rays emitted from the background point are occluded by foreground obstruction. As a result, the converged point on the imaging sensor is a mix of these rays – part from the background and part from the obstruction, which makes it difficult to segment the pixel. Fortunately, light field camera records all rays emitted from the physical world such that the defocus and occlusion cases can be well segmented in ray space using the LFSP.

Based on the above-mentioned analysis, we give the definition of LFSP as follows.

Definition 1: The LFSP is a light ray set which contains all rays emitted from a proximate, similar and continuous surface in 3D space.

Mathematically, supposing R is a proximate, similar and continuous surface in 3D space and the recorded light field is $L(u, v, x, y)$, the LFSP $s_R(u, v, x, y)$ is defined as,

$$s_R(u, v, x, y) = \bigcup_{i=1}^{|R|} L(u_{P_i}, v_{P_i}, x_{P_i}, y_{P_i}), \tag{1}$$

where $L(u_{P_i}, v_{P_i}, x_{P_i}, y_{P_i}) \subseteq L(u, v, x, y)$ is the recorded light field from i-th point P_i in the surface R. $|\cdot|$ denotes the number of elements in the set.

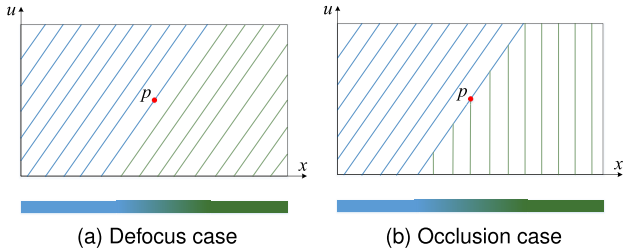


Fig. 3. The upper part shows light ray intensity distributions in defocus and occlusion cases in the EPI respectively. The lower part shows corresponding pixel intensity distributions in traditional 2D image.

B. Properties

1) Ambiguity Elimination: The LFSP eliminates the defocus and occlusion ambiguities essentially. In Fig.3, the object boundary is blurred in traditional 2D image (the bottom row) since all rays are accumulated in a same pixel. However, since all rays are recorded in light field, object boundaries are distinguishable in light ray space and can be well analyzed (the top row).

2) Limiting Cases: The definition above describes generic 4D LFSP and it can be reduced to 2D spatial or angular case by taking appropriate limits. On the one hand, considering fixing angular dimensions $(u, v) \rightarrow (u^*, v^*)$, the 4D LFSP reduces to a 2D superpixel segmentation s^{u^*,v^*} in the (u^*, v^*) view. On the other hand, if spatial dimensions (x, y) are fixed, the 4D LFSP reduces to an angular segmentation. When light field is refocused to a specific depth, the segmentation is a reference to determine the occlusion (see Fig.4). If all rays in $s_R(u, v, x^*, y^*)$ share a same label, there is no occlusion here and all views can be used to improve depth estimation. If $s_R(u, v, x^*, y^*)$ is segmented into two or more regions, the views sharing same label with central view are unoccluded views and others are occluded views. It coincides with light field occlusion theory [16], [24].

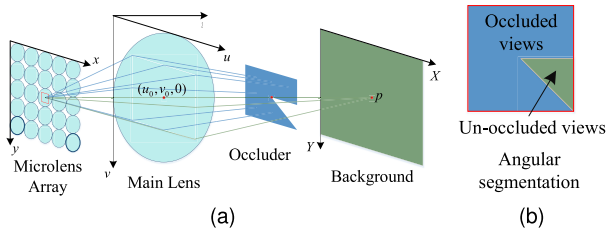


Fig. 4. The limiting case when fixing spatial dimensions. p is an occlusion boundary point. (a) Light field is refocused to the background, and only a few of views can observe p. The green rays belong to the background LFSP and blue rays belong to another LFSP. (b) The angular segmentation by fixing spatial dimensions of p. It can be seen that blue and green regions are occluded and unoccluded views respectively.

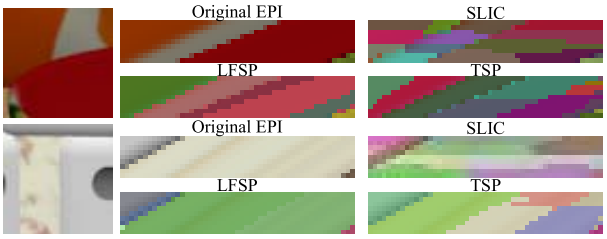


Fig. 5. Comparison of LFSP and supervoxel segmentation. Square boxes show the scaled patches of central view in our collected light fields. Right four rectangles are vertical EPI of light field, LFSP segmentation and the supervoxel segmentations of SLIC [7] and TSP [30], respectively. Different colors in segmentations represent different LFSP/supervoxel labels.

3) LFSP vs Supervoxel: There are two differences between LFSP and supervoxel. The first one is the 'full-sliced' property. In LFSP, the light rays emitted from a similar area should belong to a same LFSP. In other words, there are 2D slices of LFSP in all views of light field in free space (*i.e.* without occlusion). However, this 'full-sliced' property is not guaranteed in supervoxel which just describes a similar 3D volume in the video. Although the 2D slices of LFSP in different views describe a same object, they are unavoidably cut into multiple supervoxels when the number of views, *i.e.* the number of video frames is large enough. Fig.5 shows this difference. It is noticed that, a same LFSP is further cut into multiple supervoxels in different times. The second is different role of disparity or optical flow information. In supervoxel, optical flow can only be used to link neighboring frames. However, the disparity is also a basic attribute of pixel to distinguish different objects in LFSP segmentation.

All these two differences come from different definitions and assumptions of LFSP and supervoxel. Because light field describes a static scene, LFSP should contain all light rays emitted from a same area and the disparity plays a similar role as the textures in LFSP segmentation. Instead of static scene, the video describes dynamic scene at different time. Due to the motion assumption of objects with time, there is no guarantee that superpixel will appear in all frames and the disparity (or optical flow) should not be used to help each in-frame segmentation.

IV. Approaches

According to the definition of LFSP, each ray in the LFSP ought to be **refocus-invariant**, *i.e.*, the label of each ray should be unchangeable during the refocus operation, since the point in 3D space is unchangeable and light field itself is not changed.[1]. Apart from this, the LFSP segmentation should be **full-sliced**, *i.e.*, the LFSP should have 2D slices in all views of light field in free space. It is important to have a full-sliced LFSP segmentation in light field editing, where it is desirable to use few operations to propagate the editing from current view to full light field. To achieve these goals, we design the following refocus-invariant algorithm for LFSP segmentation.

A. Refocus-Invariant LFSP Algorithm

The nature of refocus is to shear pixels in each view [23], *i.e.*, only the disparity map of each image adds/subtracts with a constant value which is related to refocus level, while the content of each sub-aperture image does not change. To make the LFSP refocus-invariant, the disparity should be removed in the measurement of position distance.

Since it is difficult to obtain the disparity map for a full light field, in the proposed algorithm, a 2D disparity map d^{u_0,v_0} for central view (u_0, v_0) is obtained using occlusion-model guided anti-occlusion depth estimation algorithm [24]. To propagate the disparity from (u_0, v_0) to other views, the LFSP is modeled as a slanted plane in the disparity space. Suppose that $\pi_i = (A_i, B_i, C_i)$ assigns a plane function to the i-th LFSP s_i, the disparity of $p = (u, v, x, y) \in s_i$ can be computed as,

$$\hat{d}(p, \pi_i) = \frac{A_i x + B_i y + C_i}{1 + A_i(u - u_0) + B_i(v - v_0)}. \tag{2}$$

Please refer the Appendix A for detailed proof.

The full energy function is defined as,

$$\begin{aligned} &E(s, \pi, o) \\ &= \sum_{u,v}\sum_{p}\Big(E_c(p, s^{u,v}_{s(p)}) + \lambda_p E_p(p, s^{u,v}_{s(p)})\Big) + \lambda_d \sum_{p} E_d(p, \pi_{s(p)}) \\ &\quad + \lambda_s \sum_{(i,j)\in N_{seg}} E_s(\pi_i, \pi_j, o_{i,j}) + \lambda_b \sum_{(p,q)\in N_{80}} E_b(s(p), s(q)), \end{aligned} \tag{3}$$

where s is the segmentation in the full 4D light field and $s^{u,v}$ is the 2D slice of 4D LFSP in the view (u, v). $s(p)$ denotes the label that assigns to a pixel p. The o records the connection type between two neighboring LFSPs.

In Eqn.3, the terms E_c, E_p and E_d measure the color, position and disparity distance between the pixel p and superpixel center respectively. The term E_s measures the connectivity between two LFSPs in disparity space. Last but not least, the term E_b measures the 2D slice shape and **the connectivity between each 2D slice superpixel** $s^{u,v}$, which ensures that the LFSP is refocus-invariant.

[1]Noting that, because the recorded light field may be focused at different depth for a same scene, the refocus-invariance guarantees that the LFSP segmentation is always consistent with the view consistency no matter what the focus level is.

The color, position and disparity energy terms are defined as follows.

$$\begin{aligned} E_c(p, s^{u,v}_{s(p)}) &= \left\| L(p) - c_{s^{u,v}_{s(p)}} \right\|_2^2, \\ E_p(p, s^{u,v}_{s(p)}) &= \left\| p - \mu_{s^{u,v}_{s(p)}} \right\|_2^2, \\ E_d(p, \pi_{s(p)}) &= \left\| \frac{d^{u_0,v_0}(p) - \hat{d}(p, \pi_{s(p)})}{\max(d^{u_0,v_0}) - \min(d^{u_0,v_0})} \right\|_2^2, \end{aligned} \tag{4}$$

where $c_{s_i^{u,v}}$ and $\mu_{s_i^{u,v}}$ denote the color and position centers of the 2D slice $s_i^{u,v}$ respectively. $L(p)$ denotes the color of pixel p (the CIE-Lab color space is used here). The disparity term only works for central view image and it is normalized.

The smoothness term encourages the slanted planes of neighboring LFSPs (N_{seg}) to be similar. Like [36], it contains three types of LFSP boundaries, *i.e.*, the occlusion, hinge and co-planar, defined as,

$$E_s(\pi_i, \pi_j, o_{i,j}) = \begin{cases} 0 & o_{i,j} = occ \\ \dfrac{1}{|\mathcal{B}_{i,j}|} \displaystyle\sum_{p \in \mathcal{B}_{i,j}} (\hat{d}(p, \pi_i) - \hat{d}(p, \pi_j))^2 & o_{i,j} = hi \\ \dfrac{1}{|s_i \cup s_j|} \displaystyle\sum_{p \in s_i \cup s_j} (\hat{d}(p, \pi_i) - \hat{d}(p, \pi_j))^2 & o_{i,j} = co, \end{cases} \tag{5}$$

where $\mathcal{B}_{i,j}$ is the set of boundary pixels between s_i and s_j.

There are two major functions in the boundary term, corresponding to two different types of neighboring systems in light field, *i.e.*, spatial and angular neighboring systems. Additionally, these two types of neighboring systems are mixed to control full shape of 4D LFSP. For a 4D ray $p = (u, v, x, y)$ in light field, supposing its disparity is $\hat{d}(p)$, there are 8 pixels in its spatial and angular neighboring systems respectively,

$$N_{spa}(p) = \begin{cases} (u, v, x \pm 1, y + 1) \\ (u, v, x \pm 1, y - 1) \\ (u, v, x, y \pm 1) \\ (u, v, x \pm 1, y) \end{cases}$$

$$N_{ang}(p) = \begin{cases} (u \pm 1, v + 1, x \pm \hat{d}(p), y + \hat{d}(p)) \\ (u \pm 1, v - 1, x \pm \hat{d}(p), y - \hat{d}(p)) \\ (u \pm 1, v, x \pm \hat{d}(p), y) \\ (u, v \pm 1, x, y \pm \hat{d}(p)). \end{cases} \tag{6}$$

Apart from N_{spa} and N_{ang}, there is also a mixed neighboring system N_{mix} containing 64 rays in both spatial and angular domains simultaneously (see Appendix B). Fig.6 gives an illustration of these neighboring systems. (Noting that, all float values are rounded to integer in our implementation.)

In total, there are 80 rays (N_{80}) in p's neighboring system. Thus, the boundary term is defined as,

$$E_b(s(p), s(q)) = \begin{cases} 0 & s(p) = s(q) \\ E_{pen_s} & s(p) \neq s(q), N_{pq}\ is\ spatial \\ E_{pen_a} & s(p) \neq s(q), N_{pq}\ is\ angular \\ E_{pen_m} & s(p) \neq s(q), N_{pq}\ is\ mixed, \end{cases} \tag{7}$$

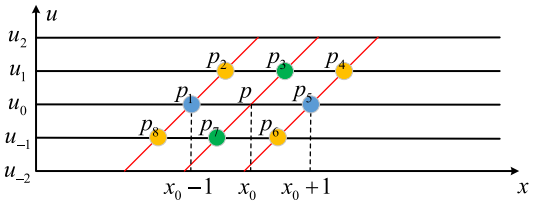


Fig. 6. An illustration of neighboring systems in light field. In the EPI space (the red lines are epipolar lines), for a pixel $p = (u_0, x_0)$, blue rays p_1 and p_5 are spatial neighbors, green rays p_3 and p_7 are angular neighbors, and orange rays p_2, p_4, p_6 and p_8 are mixed neighbors.

Algorithm 1 The LFSP Segmentation Algorithm

Input: The 4D light field $L(u, v, x, y)$
Output: The full 4D LFSP segmentation s

1: $d^{u_0,v_0} = DepthEstimation(L)$
2: $s^{u_0,v_0} = SLIC(L(u_0, v_0, x, y), d^{u_0,v_0})$
3: **for** $i = 1$ to $|s^{u_0,v_0}|$ **do**
4: $\quad \mu_{s_i^{u_0,v_0}} = \frac{1}{|s_i^{u_0,v_0}|} \sum_{p \in s_i^{u_0,v_0}} p$
5: $\quad \bar{d}_{s_i^{u_0,v_0}} = \frac{1}{|s_i^{u_0,v_0}|} \sum_{p \in s_i^{u_0,v_0}} d^{u_0,v_0}(p)$
6: $\quad$ **for** each view (u, v) **do**
7: $\quad\quad s_i^{u,v} = H(s_i^{u_0,v_0}, \bar{d}_{s_i^{u_0,v_0}}, u_0, v_0, u, v)$
8: $\quad$ **end for**
9: **end for**
10: **for** each view (u, v) **do**
11: $\quad$ **for** non-labeled pixel p **do**
12: $\quad\quad s(p) = \arg\min_{s(q)} \|p - q\|_2, q \in s^{u,v}$
13: $\quad$ **end for**
14: **end for**
15: $E(s, \pi, o) = \sum \sum (E_c + \lambda_p E_p) + \lambda_d \sum E_d + \lambda_s \sum E_s + \lambda_b \sum E_b$
16: $s = \arg\min_s E(s, \pi, o)$

where the penalty E_{pen_s} in spatial neighbouring system encourages 2D slice $s^{u,v}_{s(p)}$ to be regular, preferring straight boundaries. The penalty E_{pen_a} in angular neighboring system encourages 2D slice of LFSP to be 'regular' in epipolar plane, *i.e.* pixels in a same epipolar line share same LFSP label. It is the core to connect each 2D spatial slices of LFSP. Since the disparity is removed here, this term makes the LFSP to be refocus-invariant. The third penalty E_{pen_m} in mixed neighboring system encourages spatial 2D slice of LFSP to be regular in other views.

Remark: Reviewing the energy function, the refocus-invariance is guaranteed since (1) the 2D slices of LFSP in different views are segmented independently just using local 2D image information (E_c, E_p, E_d); and (2) angular penalty (E_{pen_a}) in the boundary term encourages similar slices to connect together according to the disparity. Additionally, the central view and boundary view is connected by a cascade angular and mix neighboring systems although the disparities are float values.

The full LFSP algorithm is summarized in the Algo.1. At first, a 2D depth map of central view d^{u_0,v_0} is calculated using [24] (line 1). Then an initial segmentation for 4D light field is obtained (lines 2-14). Finally, the LFSP result is optimized by minimizing the Eqn.3 (line 15-16) using the Block Coordinate Descent (BCD) algorithm [36]. Since the BCD algorithm only guarantees to converge to a local optima, a good initial value is in need. First, an initial superpixel

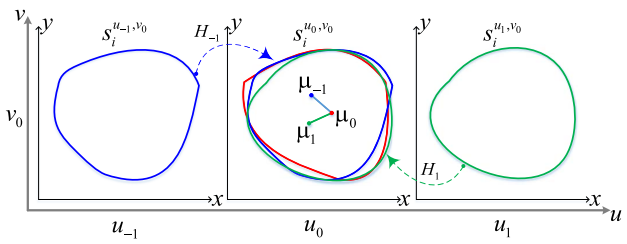


Fig. 7. An illustration of the self-similarity. The 2D slice of i-th LFSP in the view (u_{-1}, v_0), (u_0, v_0) and (u_1, v_0) are marked in blue, red, green respectively. Then $s_i^{u_{-1},v_0}$ and $s_i^{u_1,v_0}$ are projected to central view, and μ_{-1}, μ_1 are the projected centers. μ_0 is the center of $s_i^{u_0,v_0}$.

segmentation s^{u_0,v_0} of central view is obtained by embedding the disparity map d^{u_0,v_0} into the SLIC framework. Then the position and disparity centers of each superpixel $\mu_{s_i^{u_0,v_0}}$ and $\bar{d}_{s_i^{u_0,v_0}}$ are calculated and used to project s^{u_0,v_0} to 4D light field using Eqn.9 (lines 3-9). For each non-labeled pixel in other views, it is assigned as the nearest pixel's label (lines 10-14). Compared with [36], the added angular boundary term in Eqn.7 is indiscriminately calculated with spatial boundary term at the same time.

B. Evaluation Metrics

Existing evaluation metrics for superpixel segmentation concentrate on boundary adherence, such as under-segmentation error (UE), boundary recall (BR) and achievable segmentation accuracy (ASA) [35]. There is no proper metrics for the specific refocus-invariant and full-sliced properties. To measure these features, we propose the LFSP self-similarity and the effective label ratio.

1) Self-Similarity: The self-similarity SS_i of the i-th LFSP is defined as,

$$SS_i = \frac{1}{N_{uv} - 1} \sum_{u,v} \left\| \mu_{H(s_i^{u,v},d,u,v,u_0,v_0)} - \mu_{s_i^{u_0,v_0}} \right\|_2, \quad (8)$$

where N_{uv} is angular sampling number of light field. $s_i^{u,v}$ is a 2D slice of i-th LFSP in the (u, v) view and (u_0, v_0) is central view of light field. μ_s denotes the position center of superpixel s and $H(s_i^{u,v}, d, u, v, u_0, v_0)$ projects 2D superpixel $s_i^{u,v}$ from the (u, v) view to (u_0, v_0) according to ground truth disparity map d. For each pixel $p = (u, v, x, y)^\top \in s_i^{u,v}$, the projected coordinate $p' = (u_0, v_0, x', y')^\top$ is defined as (in homogeneous coordinate),

$$\begin{pmatrix} u_0 \\ v_0 \\ x' \\ y' \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 0 & 0 & 0 & u_0 \\ 0 & 0 & 0 & 0 & v_0 \\ -d(p) & 0 & 1 & 0 & u_0 d(p) \\ 0 & -d(p) & 0 & 1 & v_0 d(p) \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{H(s_i^{u,v},d,u,v,u_0,v_0)} \begin{pmatrix} u \\ v \\ x \\ y \\ 1 \end{pmatrix}. \quad (9)$$

We also give an intuitive explanation of above definition. For a light field (Fig.7) with 1×3 angular resolution, the slices $s_i^{u_{-1},v_0}$ and $s_i^{u_1,v_0}$ of i-th LFSP are projected to central view according to ground truth disparity. The new centers of the projected $s_i^{u_{-1},v_0}$ and $s_i^{u_1,v_0}$ are denoted as μ_{-1} and μ_1 respectively, and the center of $s_i^{u_0,v_0}$ is μ_0. The mean of $\|\mu_1 - \mu_0\|_2$ and $\|\mu_{-1} - \mu_0\|_2$ is the self-similarity of the i-th LFSP.

For a full segmentation in 4D light field, the LFSP self-similarity SS is defined as the mean of all SS_i,

$$SS = \frac{1}{K} \sum_{i=1}^{K} SS_i, \quad (10)$$

where K is the number of LFSP.

From the definition, the LFSP self-similarity is measured as pixel unit and a low SS value implies a high refocus-invariance. Apart from this, since the disparity changes with the refocus level, the LFSP self-similarity can measure the refocus-invariance of LFSP segmentation accurately.

2) Effective Label Ratio: As analyzed in Sec.III-B.3, one of the main differences between the LFSP and traditional supervoxel is the 'full-sliced' property, *i.e.*, the 4D LFSP should have the corresponding 2D slices in all views. To measure this property, the effective label ratio (ELR) is proposed and defined as,

$$ELR = \frac{K_{u_0,v_0}}{K}, \quad (11)$$

where K_{u_0,v_0} is the number of LFSP in the central view of light field. A larger ELR implies better full-sliced.

ELR is an important evaluation metric for light field editing. A larger ELR indicates that less operations are needed in order to propagate the edit from current view to full light field. On the contrary, a smaller ELR shows that there are missing regions in some views, resulting in more operations to be applied for light field editing.

V. Experimental Results

We compare the proposed LFSP segmentation with state-of-the-art superpixel segmentation algorithms including SLIC [7] and LSC [8]. Noting that, the results of SLIC come from the vlfeat [37] library, and the code of LSC comes from the author's website. Apart from superpixel segmentation, two classical supervoxel algorithms, SLIC [7] and TSP [30], are also compared. The SLIC supervoxel algorithm only uses color cues while TSP algorithm also uses optical flow. All superpixel algorithms are evaluated both on synthetic data and real scene light fields while supervoxel algorithms are only tested on synthetic data.. For synthetic data, the HCI benchmark light field datasets [38] are used, which consist of four light fields with ground truth depth and segmentation. Each data includes a 9×9 (angular resolution) light field. We additionally generate other six light fields with ground truth disparity map and segmentation using the Blender [39], to better evaluate the performance of the proposed algorithm. Compared with previous data, our synthetic light fields contain more objects (4 to 6 objects in HCI data and 7 to 10 objects in our data) and introduce more non-Lambertian scenes (glass or mirror type objects). Apart from these, our light fields have a larger absolute disparity range (4 pixels in HCI data and 7 pixels in our data), which helps to compare different supervoxel algorithms better. The real scene light fields are captured

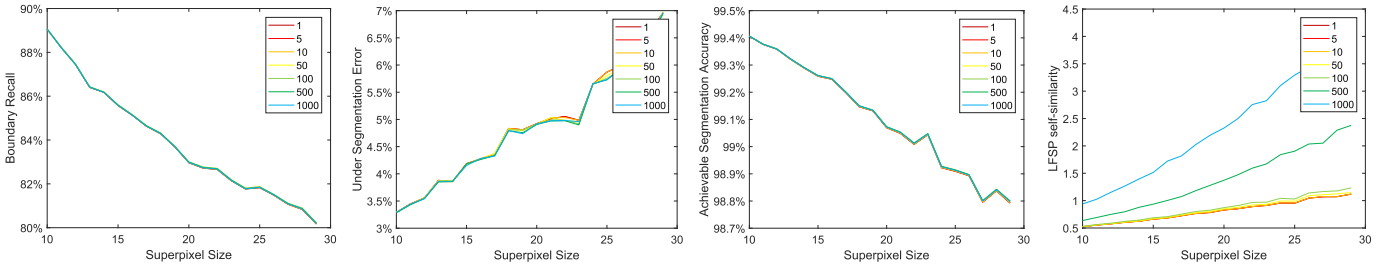


Fig. 8. Quantitative evaluation of LFSP with different depth weights λ_d.

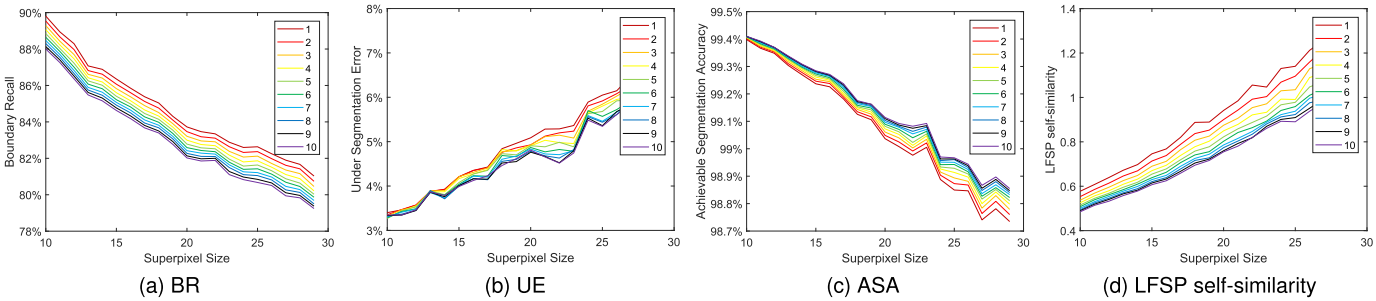


Fig. 9. Quantitative evaluation of LFSP with different E_{pen_a}.

by a consumer light field camera Lytro. The 4D light field data are extracted using the LFToolbox [40]. The quantitative evaluation contains the UE, BR, ASA [35], running time, LFSP SS and ELR. All evaluations are conducted on synthetic data since ground truth disparity map and segmentation are not available in real scene data, and so far, there is no light field segmentation benchmark in real scene data like classical Berkeley segmentation database [41]. The codes and synthetic light fields are available at [42] now.

A. Synthetic Scenes

1) Performance vs Parameters: The full 4D LFSP differs from traditional 2D image superpixel in two aspects: (1) The disparity map is always required instead of optional; and (2) The 2D slice superpixels in different views are tightly connected together instead of irrelevant. For the first point, we verify the LFSP with different depth weight λ_d. For the second one, the key parameter E_{pen_a}, which connects similar 2D slices, is evaluated with different values. Other parameters such as λ_p, λ_s and λ_b have been discussed in previous papers [7], [36] and will not be discussed here.

a) Depth weight λ_d: Fig.8 demonstrates evaluations of LFSP segmentation with different λ_d ranging from 1 to 1000. It can be seen that 7 lines are approximately coincident in the statistics of boundary adherence (Fig.8a,8b,8c). The λ_d has no effect on boundary adherence because each 2D slice of LFSP is segmented independently just using local 2D image information E_c,, E_p, E_d (in central view only). In Fig.8d, the LFSP SS decreases as λ_d decreases. In most cases, color and position terms can segment the image well. The disparity term not only fails to improve boundary adherence, but also increases the LFSP SS since only disparity map of the central view is utilized. However, this term cannot be ignored since previous color and position terms cannot segment object boundaries with similar textures (see Fig.11).

b) Angular neighbour penalty E_{pen_a}: Fig.9 shows comparisons on different penalty E_{pen_a} in angular neighboring system. It can be seen that both boundary recall and the LFSP SS decrease with the increase of E_{pen_a}. It is understandable that BR and LFSP SS are contradictory in the case of E_{pen_a}. Assuming there are only color and position terms in Eqn.3, all 2D slice superpixels in different views can fit boundaries tightly. When the boundary term is applied, especially the penalty in angular neighborhood, these 2D superpixels will "pull" boundary pixels in other views. The strength of this "pull" increases as E_{pen_a} increases, which makes boundary adherence decrease correspondingly.

2) Influence of Disparity Maps:

Disparity quality: Five estimated disparity maps from different algorithms [16], [24], [43]–[45] and ground truth are used to test our LFSP algorithm. Table I shows the RMS errors of state-of-the-art light field depth estimation algorithms, and Fig.10 demonstrates evaluations on different disparity maps. In general, the LFSP segmentation benefits from a good disparity map.

3) Ambiguity Elimination: Fig.12 demonstrates the comparison of the proposed LFSP with traditional SLIC and LSC algorithms on the defocused area. It is noticed that, the SLIC and LSC cannot accurately find the object boundaries in these areas due to the defocus blur, the superpixel boundaries are always cling with the circle of confusion, such as the red dot or the blue car frame. However, because the LFSP focuses on the segmentation of the light ray instead of the images cumulated by multiple light rays, it eliminates the ambiguities

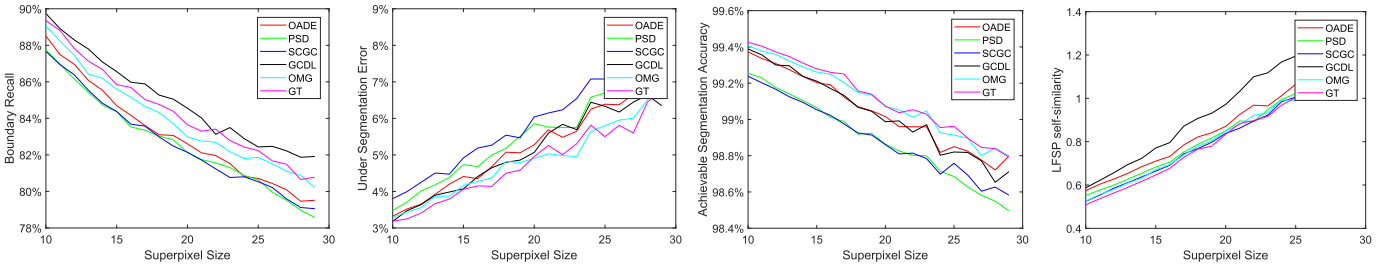


Fig. 10. Quantitative evaluation of LFSP with different disparity maps.

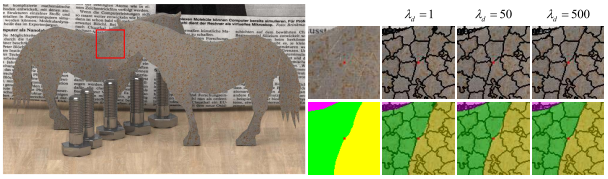


Fig. 11. The segmentation on Horses with the change of λ_d. There are two horses here, however it is hard to distinguish them due to similar textures. It can be seen that the boundary is preserved better with a larger depth weight λ_d.

TABLE I
RMS ERRORS OF DIFFERENT DEPTH ESTIMATION ALGORITHMS ON HCI DATA [38]

	Buddha2	Horses	Papillon	StillLife	Mean
OADE [16]	0.107	0.140	0.125	0.212	0.146
PSD [43]	0.070	0.129	0.237	0.135	0.143
SCGC [44]	0.079	0.205	**0.086**	0.111	0.120
GCDL [45]	0.094	0.163	0.158	0.184	0.150
OMG [24]	**0.051**	**0.074**	0.148	**0.110**	**0.096**

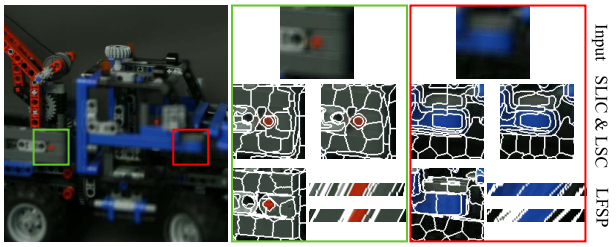


Fig. 12. Segmentation comparison on the defocused area.

TABLE II
PARAMETERS SETTING

Name	λ_p	λ_d	λ_s	λ_b	E_{pen_s}	E_{pen_m}	E_{pen_a}
Value	2	50	0.01	1	1	0.7	4

in defocus and occlusion boundaries. In Fig.12, the proposed LFSP segmentation algorithm segments boundaries of the red dot and blue car frame accurately.

4) Adherence to Boundaries: Unless otherwise stated, the pre-set parameters for all experiments are listed in Table II. Apart from λ_d and E_{pen_a}, λ_p balances the effect between the position and color distance and a larger λ_p leads to a more well-shaped superpixel. λ_s controls slanted plane function and it is mainly decided by initial disparity map. Since state-of-the-art depth estimation algorithms [16], [18], [23] always over-smooth occlusion boundaries, it is suggested to assign a small value to make the plane function more stable. For boundary terms E_{pen_s} and E_{pen_m}, small values are assigned, trying to encourage straight boundaries. λ_b balances the boundary adherence and shape. The boundary adherence decreases with the increase of λ_b.

Fig.13a-13c show quantitative results which are average values on the HCI segmentation datasets. It can be seen that the proposed LFSP algorithm obtains competitive results (red lines) over state-of-the-art algorithms (green and blues lines) in all three traditional metrics. Qualitative results are shown in Fig.15, from which we can see that the LFSP segmentation can produce more regular superpixels in occlusion boundary areas (the buddha in the first row and the butterfly in the second row). Fig.14 and 16 show quantitative and qualitative results on our synthetic light field data, respectively. Our algorithm also achieves competitive results compared with previous approaches. For example, in the first row of Fig.16, only the proposed algorithm preserves the boundaries of grey points in the dice well, while these areas are over-segmented or under-segmented in the results of SLIC and LSC. In the bottom row, the comparison is more obvious in the boundaries of the leaves.

Additionally, evaluations of initial value (the LFSP segmentation without BCD optimization) are also plotted in Fig.13 (black lines). It can be seen that the optimized segmentation is far superior to initial one, showing the effectiveness of optimization. In the fourth column of Fig.15, segmentation results in 4D space are partly exhibited. For each local region, the first row shows initial results and the second row shows the optimized results. Due to the occlusion, many pixels are assigned with wrong labels and segmentation boundaries do not agree with object boundaries in the EPI space at initial stage. After the optimization, these errors are amended and occlusion boundaries are preserved well.

5) LFSP Self-Similarity: Apart from these traditional evaluation metrics (*i.e.*, BR, UE and ASA), we also evaluate the LFSP using the LFSP SS. Since there is no previous work on light field superpixel segmentation, it is unfair to directly compare it with traditional 2D superpixel segmentation. We refocus light field for 7 times (the refocus level $1 - \frac{1}{\alpha}$ varying in $-1.5, -1, -0.5, 0, 0.5, 1, 1.5$) and segment

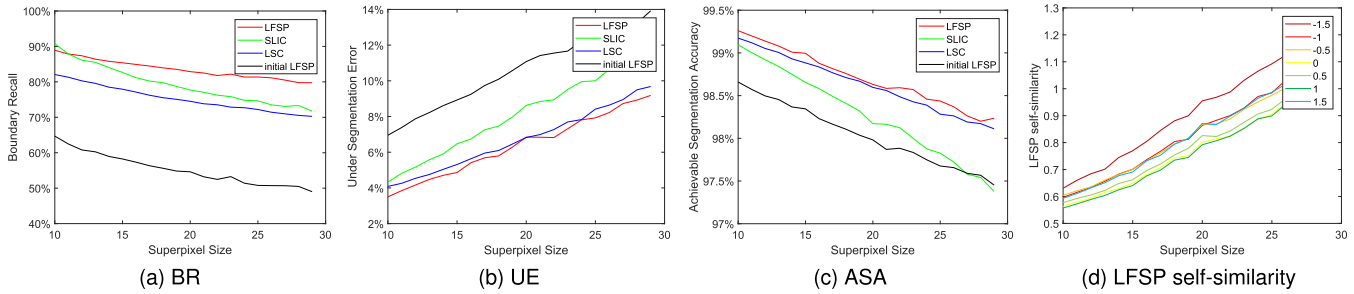


Fig. 13. Quantitative evaluation of different superpixel segmentation algorithms on HCI synthetic light field data [38].

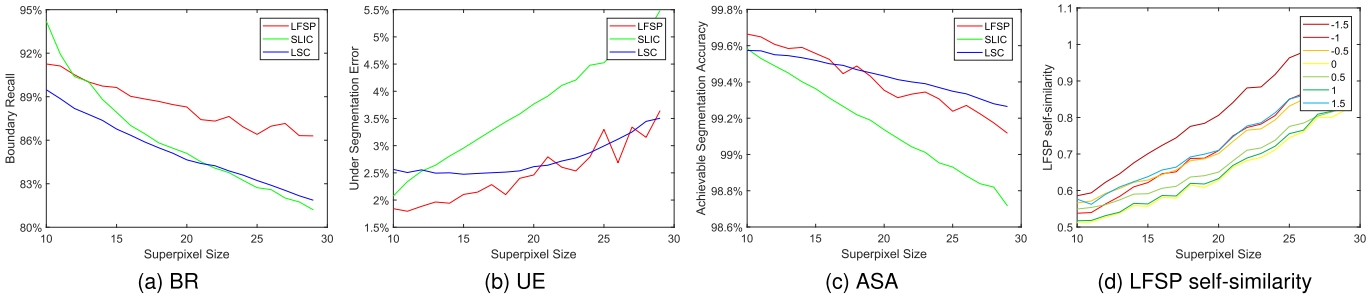


Fig. 14. Quantitative evaluation of different superpixel segmentation algorithms on our synthetic light field data.

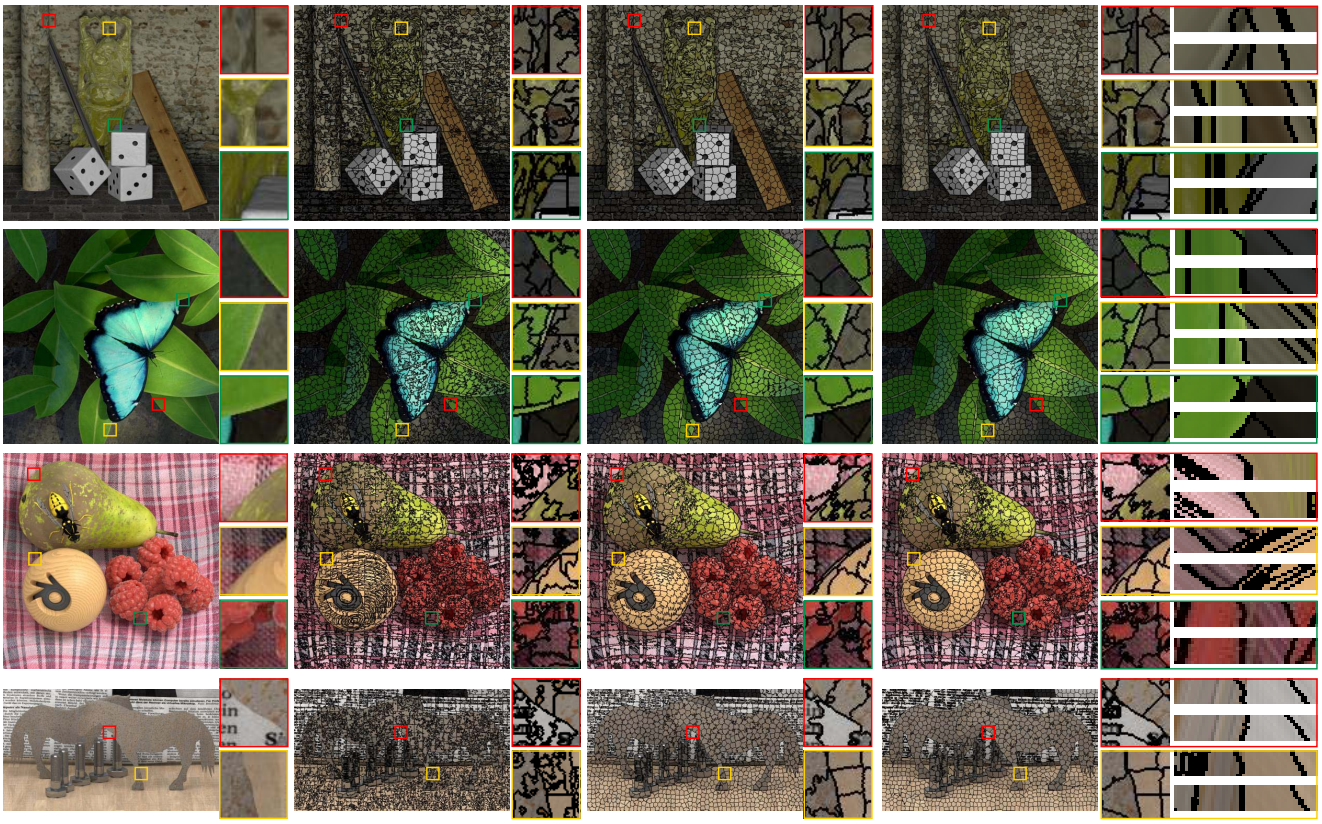


Fig. 15. Segmentation results on HCI synthetic light field data [38] (the superpixel size is 20). The first column shows central view of input light field. The second to fourth columns show the results of SLIC, LSC and the proposed algorithm respectively. For each region in our results (the right-most column), the upper row shows initial segmentation in the EPI space, and the lower row shows the optimized segmentation result.

them. Then the LFSP SSs on each segmentation are plotted in Fig.13d and 14d. It can be seen that the curves always maintain at a low level and all values are smaller than 1 pixel, which shows good refocus-invariance of the proposed LFSP algorithm. Furthermore, the curves are very close to each other, which indicates the stability of our LFSP algorithm.

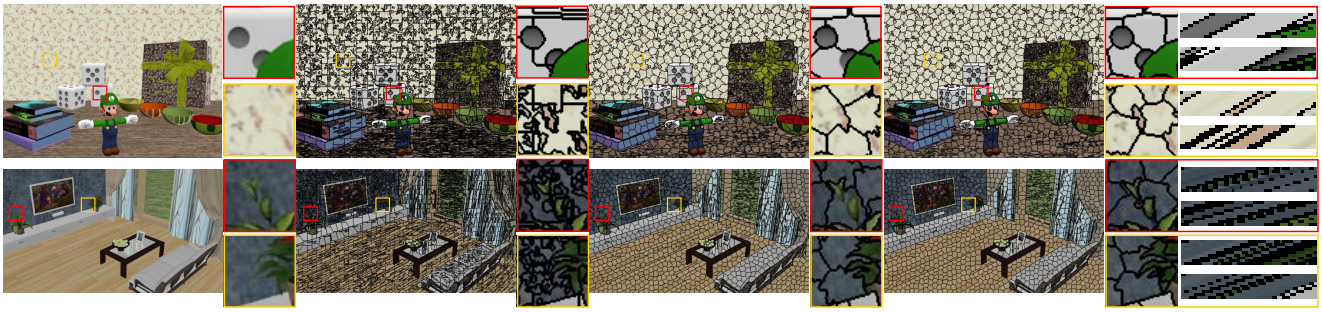


Fig. 16. Segmentation results on our synthetic light field dataset (the superpixel size is 20). For each region in our results (the right-most column), the first and second rows show the segmentation in the horizontal and vertical EPIs respectively.

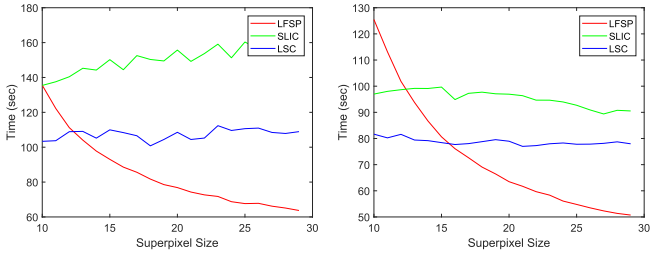


Fig. 17. Running times of different algorithms on HCI and our synthetic light field data.

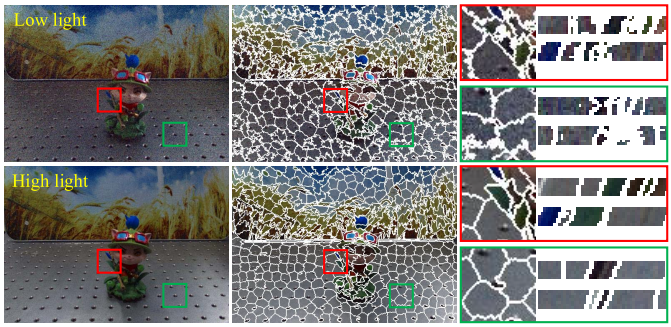


Fig. 18. Comparison of LFSP segmentation under different illuminations.

6) Running Time: Fig.17 shows the running time of different algorithms on HCI and our light field data respectively. Noting that, because the disparity map is considered as a basic property of the input light field for post editing tasks [15] and can be pre-computed, its running time is not counted in our results in Fig.17. All algorithms are evaluated on the same desktop computer with a 3.4 GHz i7 CPU. For HCI data, there are 768×768 pixels in each sub-aperture image. Our data contains 760×760 or 720×1024 pixels in each sub-aperture image. The time of SLIC and LSC are the sum of time costs that are taken on each view image of light field by the two algorithms respectively. It can be seen that our un-optimized Matlab/C implementation shows great advantages over previous works with the increasing of superpixel size since the 4D light field is treated as a whole instead of multiple independent images. Besides, the BCD algorithm just iteratively optimizes boundary pixels in the LFSP segmentation, which decreases the complexity a lot.

B. Real Scenes

Fig.1 and 19 show experimental results on real scene light fields, captured by a Lytro camera (the superpixel size is set as 20 here). Due to low signal-to-noise ratio of Lytro camera, the SLIC and LSC algorithms cannot produce reliable results from single image of central view. However, due to the introduction of angular neighboring system, the proposed LFSP algorithm can produce more convincing results. Compared with the SLIC and LSC, our algorithm can generate superpixels with more regular shapes (the stone bench in the first row and the leaves in the second row) and more similar size (the trash can in the third row and the plant in the fourth row). Apart from this, occlusion boundaries in the EPI space are also preserved well. The segmentation boundaries can always cling occlusion boundaries or remain the same direction with EPI lines, which verifies good LFSP self-similarity of the proposed algorithm.

Although the proposed LFSP also outperforms SLIC and LSC in real light fields, the qualitative results are not as good as the results of synthetic data shown in Fig.15. This is due the low signal-to-noise ratio of the Lytro camera. In Fig.18, the segmentation comparisons under both the low and high illuminations are provided. In low light environment, the EPI consistency is broken due to the noise, so the results are worse than those of high light environment. It is suggested to record light fields in high illumination environment to obtain good LFSP segmentations.

In Fig.20, segmentation results under different refocus levels $(-0.5, 0, 0.5)$ are demonstrated. It can be seen that although the direction of EPI lines changes with different refocus levels, segmentation boundaries always agree with object boundaries, which further validates good refocus-invariance of the proposed algorithm.

C. LFSP vs Supervoxel

As the 4D light field can also be treated as 3D videos and following be segmented using the supervoxel algorithms, to fairly compare the performance of the proposed LFSP algorithm and the traditional supervoxel algorithms, we organize the 3D video from the 4D light field using two different orders, namely the row first (RF) and column first (CF), respectively.

[1]Noting that it is a mean ratio, the ELR on the Buddha is 0.999 while others are all 1.000.

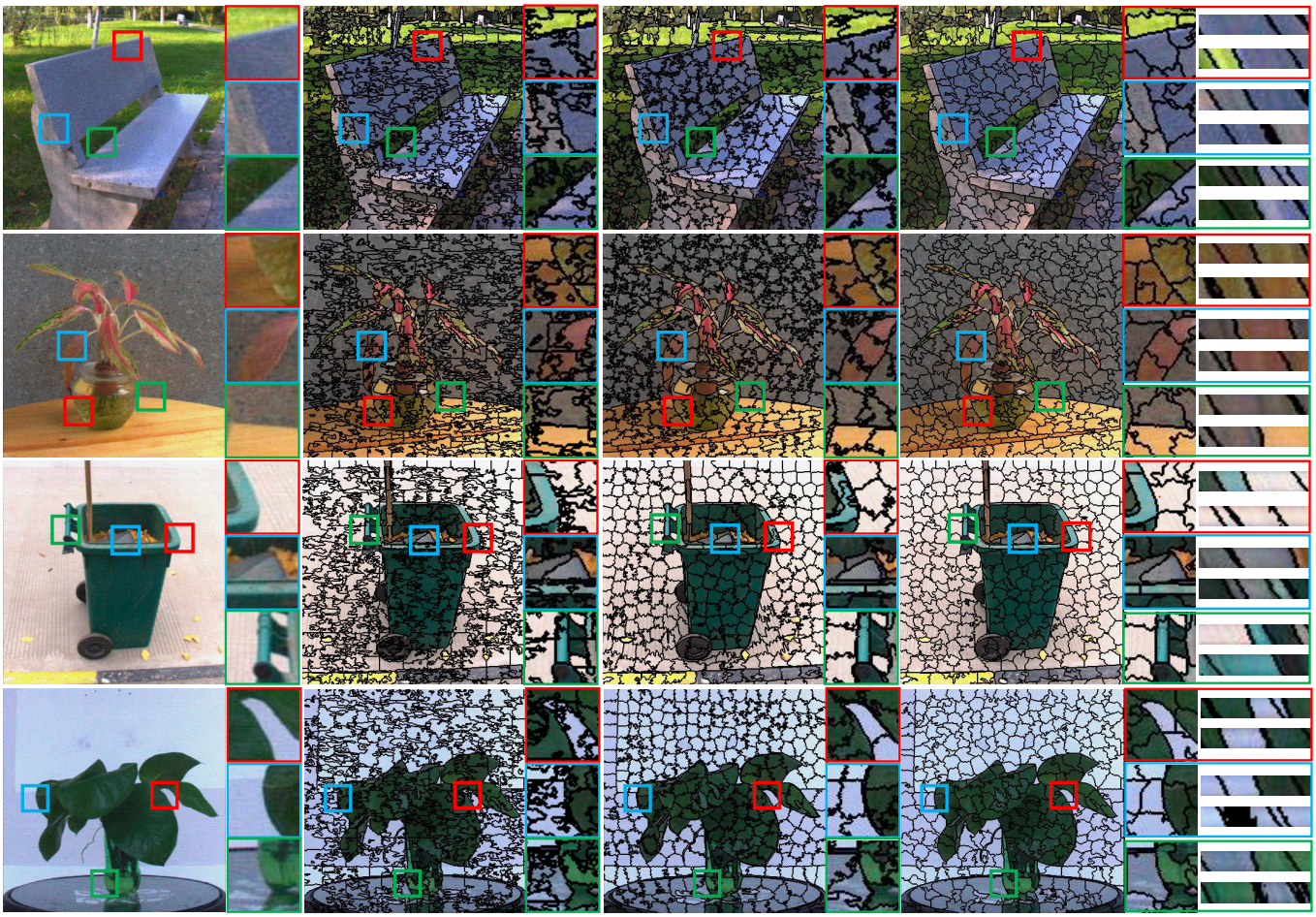


Fig. 19. Segmentation results on real scene light fields. For each region in our results (the right-most column), the upper and lower rows show the segmentation in the horizontal and vertical EPIs respectively.

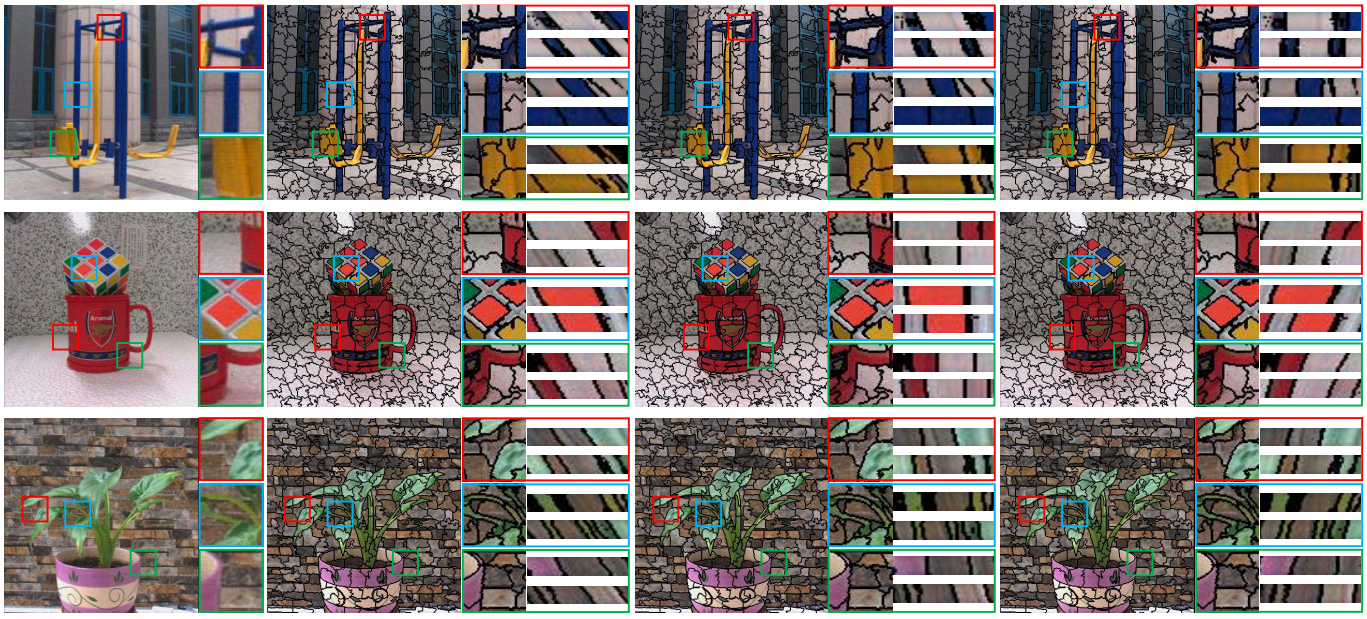


Fig. 20. Segmentation results of real scene light fields under different refocus levels. The first column shows central view of input light field. The second to fourth columns show the results under different refocus levels $(-0.5, 0, 0.5)$ respectively.

Tab.III shows the quantitative comparison of LFSP and supervoxel algorithms on HCI and our light fields. The terms 'HCI', 'Ours' and 'Both' refer to the results on HCI light fields, our light fields and mean of two datasets. Since there is little difference in BR, UE, ASA and SS between the segmentation results on HCI and our light fields, the mean results

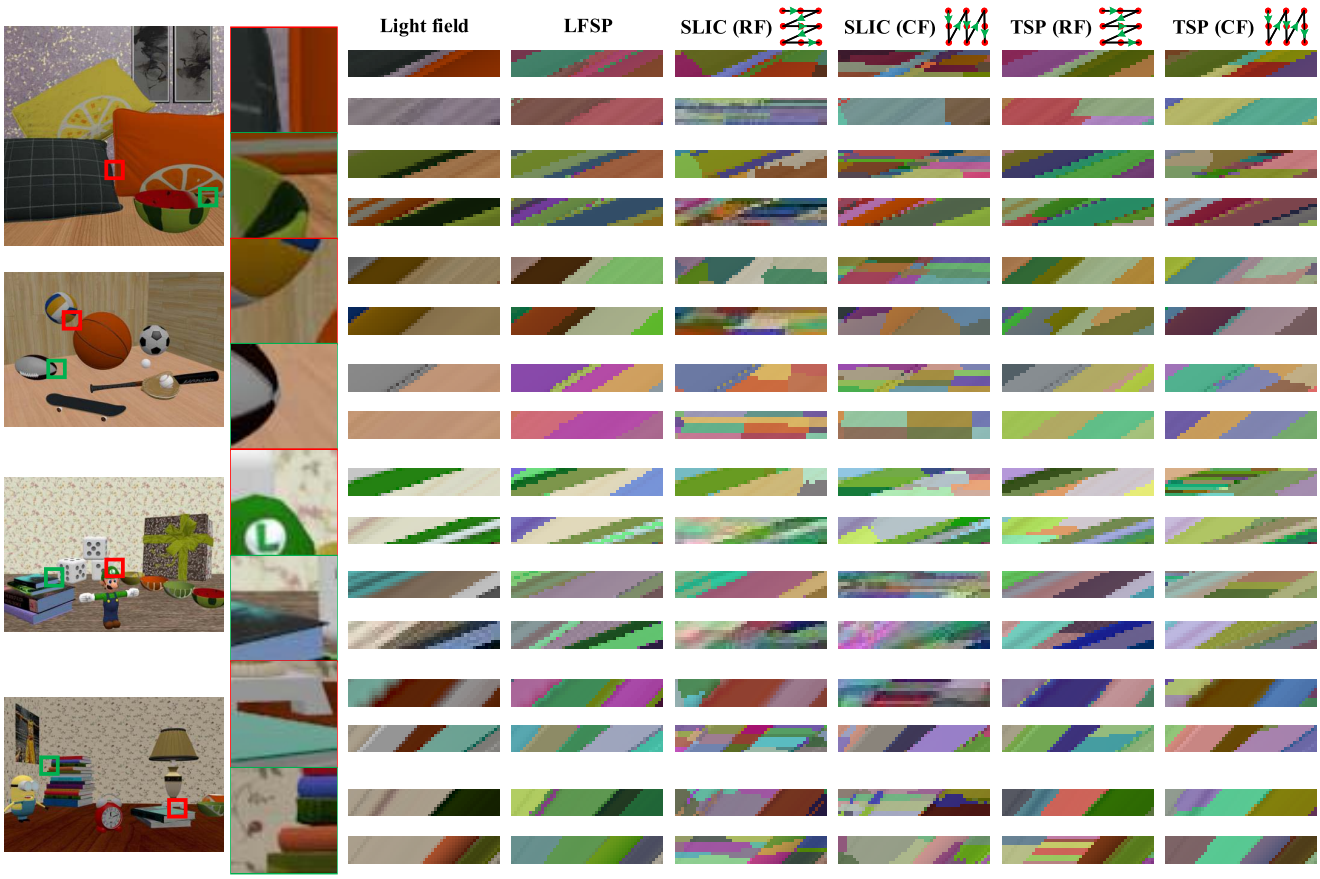


Fig. 21. LFSP segmentation vs supervoxel segmentation. For each scene, the right square patches are zoom-in of red/green boxes. To the right of each scaled patch, the first and second rows are the horizontal and vertical EPIs, respectively. From left to right, original EPIs, LFSP segmentations, SLIC and TSP supervoxel segmentations. The terms 'RF' and 'CF' refer to two types of order of different views in video, namely the row first and column first, respectively.

TABLE III

RESULTS COUNTING ON SYNTHETIC LIGHT FIELDS (LFSP VS SUPERVOXEL)

	SLIC [7]			TSP [30]			Ours
	RF	CF	Mean	RF	CF	Mean	
BR (Both)	0.832	0.828	0.830	0.647	0.645	0.646	0.824
UE (Both)	0.060	0.067	0.064	0.071	0.074	0.072	0.043
ASA (Both)	0.988	0.987	0.987	0.984	0.983	0.984	0.990
SS (Both)	2.602	2.290	2.446	1.317	1.386	1.352	0.773
ELR (HCI)	0.501	0.493	0.497	0.929	0.909	0.919	1.000^2
ELR (Ours)	0.428	0.397	0.412	0.783	0.689	0.736	1.000

are listed. On classical boundary adherence metrics (BR, UE, ASA), LFSP achieves slight advantages. The performance of LFSP improves in SS counting. Since there is no optical flow constraint, SLIC has a high SS value. TSP has a smaller SS than SLIC due to the use of optical flow. In the ELR counting, LFSP achieves 100% on both datasets while SLIC and TSP have small values. SLIC failed in ELR since only color and position are utilized. Because of the view by view optical flow calculation which is prone to noise, it is difficult to build the angular connection between the first and last views in light field, and TSP has a low ELR value. In addition, it is noticed that the ELR values of SLIC and TSP decrease a lot from HCI to our light fields. The main reason for this phenomenon is that our light fields have a larger absolute disparity, which increases the difficulty to build angular connections.

Considering the 'full-sliced' property cannot be well demonstrated in the HCI light fields due to the small absolute disparity, we only plot the qualitative comparisons on our light fields (see Fig.21). Since only color and position of video are utilized, a same LFSP is cut into multiple LFSP fragments in time axis and the EPI consistency is severely damaged. TSP provides good segmentation when the orientation of EPI agrees with the view order in video organization, in other words, it achieves good results in horizontal (vertical) EPI when the video is organized using row (column) first rule. However, the EPI consistency is undetermined on another orientation EPI. The main reason is that optical flows have a same direction and small values in a same angular row (column) in light field, while the direction changes and optical flow values increase a lot between the end of current angular row (column) and the first of next angular row (column) in light field. Compared with SLIC and TSP, the angular connections in both horizontal and vertical EPIs are built at the same time in LFSP, so the EPI consistencies in both horizontal and vertical EPIs are protected.

D. Limitations

In experiments, we find that our algorithm cannot handle specular (*i.e.*, non-Lambertian) or cluttered regions well (see Fig.22). For specular areas, since the captured scene

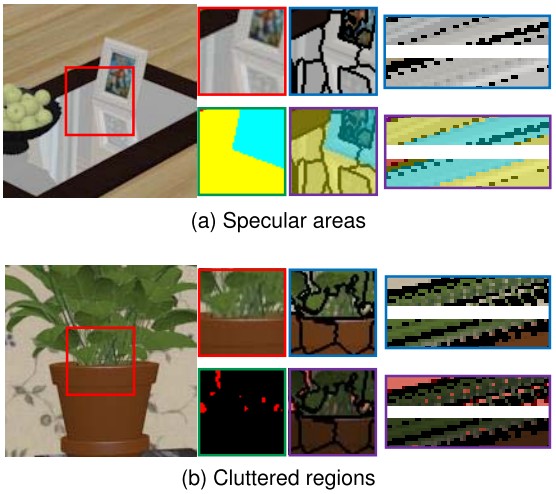


Fig. 22. Limitations. These two results demonstrate the LFSP segmentations on specular and cluttered areas respectively. For each rectangle in the subfigure, red is the input data, green is ground truth, blue is the LFSP segmentation and purple is the comparison of the LFSP and ground truth. The most-right column refers to the segmentations/comparisons in horizontal or vertical EPI spaces.

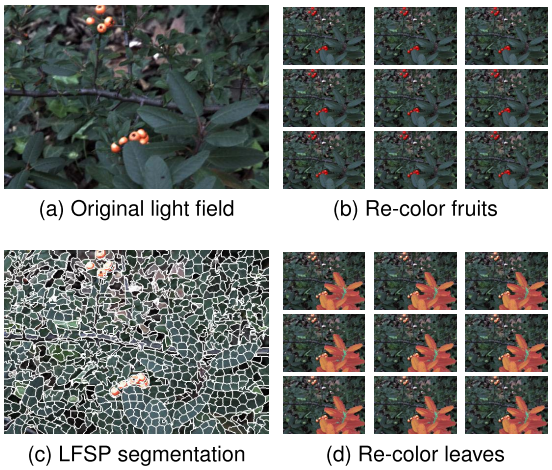


Fig. 23. LFSP based re-color. (a) The central view of input light field. (c) The 2D slice of LFSP segmentation in central view. (b), (d) The light fields after re-coloring the fruits and leaves.

may be affected by the light (*e.g.*, metal materials) or neighboring scenes (*e.g.*, mirror type objects) (see Fig.22a), the color or depth cues are not reliable and boundaries of these objects cannot be segmented well. For cluttered regions, because there are too many small objects which are often much smaller than the given superpixel size (see Fig.22b), it is unavoidable to include many small objects in the segmented superpixel. These two challenge issues are also not solved well by traditional algorithms.

E. Extended Application

With the help of LFSP, full light field can be edited with only one operation. Fig.23 demonstrates the LFSP based re-color, where the light field is captured using the Lytro Illum camera. The color of fruits and leaves in all views are adjusted by simply changing the HSV of related LFSPs.

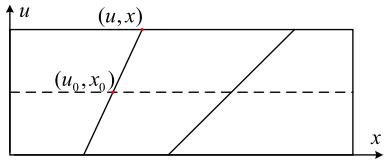


Fig. 24. By fixing (v, y) as (v_0, y_0), the epipolar plane image appears. (u_0, x_0) is a ray in central view and (u, x) is the correspondence in the (u, v_0) view. (u_0, x_0) and (u, x) share a same epipolar line.

VI. Conclusions and Future Work

In the paper, we have defined the light field superpixel (LFSP). The LFSP is defined in 4D space and can essentially eliminate the defocus and occlusion ambiguities in traditional 2D superpixel segmentation. We have proposed a refocus-invariant LFSP segmentation algorithm. By embedding 2D disparity map into superpixel segmentation and defining a clique system with 80 (spatial, angular and mixed) neighbors in the full 4D light field, the proposed algorithm not only outperforms the state-of-the-arts in term of traditional evaluation metrics but also achieves good refocus-invariant and full-sliced properties. In the future, we will not only explore the LFSP on more challenging non-Lambertian surfaces and clutter scenes but also try to segment and edit light field using LFSP.

Appendix A
Proof of the Plane Equation 2

Suppose that $\pi_i = (A_i, B_i, C_i)^\top$ assigns a planar equation to the i-th LFSP s_i. The disparity of pixel $p = (u_0, v_0, x_0, y_0) \in s_i^{u_0, v_0}$ in central view can be obtained by,

$$\hat{d}(p, \pi_i) = A_i x_0 + B_i y_0 + C_i. \tag{A.1}$$

We employ the EPI to derive the disparity of $p = (u, v, x, y) \in s_i$ (see Fig.24) in other views. Suppose that (u_0, x_0) is a ray in central view and (u, x) is the correspondence in the (u, v) view. It is known that (u, x) and (u_0, x_0) are in a same epipolar line. The disparity of (u, x) or (u_0, x_0) is defined as,

$$\hat{d} = \frac{x - x_0}{u - u_0}. \tag{A.2}$$

In other words, if we know the disparity $\hat{d}$ of ray (u, x), the corresponding (u_0, x_0) in central view can be computed as,

$$x_0 = x - \hat{d} \cdot (u - u_0) \tag{A.3}$$

Similarly, if the disparity $\hat{d}$ of (v, y) is known, the corresponding (v_0, y_0) in central view can be computed as,

$$y_0 = y - \hat{d} \cdot (v - v_0) \tag{A.4}$$

Substituting Eqns. A.3 and A.4 into Eqn.A.1, we have

$$\hat{d}(p, \pi_i) = A_i \cdot (x - \hat{d} \cdot (u - u_0)) + B_i \cdot (y - \hat{d} \cdot (v - v_0)) + C_i. \tag{A.5}$$

Revisiting Eqn.A.5, the disparity of $p = (u, v, x, y) \in s_i$ is obtained,

$$\hat{d}(p, \pi_i) = \frac{A_i x_0 + B_i y_0 + C_i}{1 + A_i (u - u_0) + B_i (v - v_0)}. \tag{A.6}$$

APPENDIX B
THE 80-NEIGHBOURING SYSTEM

For a ray $p = (u, v, x, y)$ in 4D light field, supposing its disparity is $\hat{d}(p)$, its spatial and angular neighbors are,

$$N_{spa}(p) = \begin{cases} (u, v, x \pm 1, y + 1) \\ (u, v, x \pm 1, y - 1) \\ (u, v, x, y \pm 1) \\ (u, v, x \pm 1, y) \end{cases}$$

$$N_{ang}(p) = \begin{cases} (u \pm 1, v + 1, x \pm \hat{d}(p), y + \hat{d}(p)) \\ (u \pm 1, v - 1, x \pm \hat{d}(p), y - \hat{d}(p)) \\ (u \pm 1, v, x \pm \hat{d}(p), y) \\ (u, v \pm 1, x, y \pm \hat{d}(p)) \end{cases} \tag{B.1}$$

To demonstrate better, 64 mixed neighbors are divided into 4 parts (each part contains 16 neighbors).

$$N_{mix,1}(p) = \begin{cases} (u + 1, v + 1, x + \hat{d}(p) \pm 1, y + \hat{d}(p) + 1) \\ (u + 1, v + 1, x + \hat{d}(p) \pm 1, y + \hat{d}(p) - 1) \\ (u + 1, v + 1, x + \hat{d}(p), y + \hat{d}(p) \pm 1) \\ (u + 1, v + 1, x + \hat{d}(p) \pm 1, y + \hat{d}(p)) \\ (u - 1, v + 1, x - \hat{d}(p) \pm 1, y + \hat{d}(p) + 1) \\ (u - 1, v + 1, x - \hat{d}(p) \pm 1, y + \hat{d}(p) - 1) \\ (u - 1, v + 1, x - \hat{d}(p), y + \hat{d}(p) \pm 1) \\ (u - 1, v + 1, x - \hat{d}(p) \pm 1, y + \hat{d}(p)) \end{cases} \tag{B.2}$$

$$N_{mix,2}(p) = \begin{cases} (u + 1, v - 1, x + \hat{d}(p) \pm 1, y - \hat{d}(p) + 1) \\ (u + 1, v - 1, x + \hat{d}(p) \pm 1, y - \hat{d}(p) - 1) \\ (u + 1, v - 1, x + \hat{d}(p), y - \hat{d}(p) \pm 1) \\ (u + 1, v - 1, x + \hat{d}(p) \pm 1, y - \hat{d}(p)) \\ (u - 1, v - 1, x - \hat{d}(p) \pm 1, y - \hat{d}(p) + 1) \\ (u - 1, v - 1, x - \hat{d}(p) \pm 1, y - \hat{d}(p) - 1) \\ (u - 1, v - 1, x - \hat{d}(p), y - \hat{d}(p) \pm 1) \\ (u - 1, v - 1, x - \hat{d}(p) \pm 1, y - \hat{d}(p)) \end{cases} \tag{B.3}$$

$$N_{mix,3}(p) = \begin{cases} (u + 1, v, x + \hat{d}(p) \pm 1, y + 1) \\ (u + 1, v, x + \hat{d}(p) \pm 1, y - 1) \\ (u + 1, v, x + \hat{d}(p), y \pm 1) \\ (u + 1, v, x + \hat{d}(p) \pm 1, y) \\ (u - 1, v, x - \hat{d}(p) \pm 1, y + 1) \\ (u - 1, v, x - \hat{d}(p) \pm 1, y - 1) \\ (u - 1, v, x - \hat{d}(p), y \pm 1) \\ (u - 1, v, x - \hat{d}(p) \pm 1, y) \end{cases} \tag{B.4}$$

$$N_{mix,4}(p) = \begin{cases} (u, v + 1, x \pm 1, y + \hat{d}(p) + 1) \\ (u, v + 1, x \pm 1, y + \hat{d}(p) - 1) \\ (u, v + 1, x, y + \hat{d}(p) \pm 1) \\ (u, v + 1, x \pm 1, y + \hat{d}(p)) \\ (u, v - 1, x \pm 1, y - \hat{d}(p) + 1) \\ (u, v - 1, x \pm 1, y - \hat{d}(p) - 1) \\ (u, v - 1, x, y - \hat{d}(p) \pm 1) \\ (u, v - 1, x \pm 1, y - \hat{d}(p)) \end{cases} \tag{B.5}$$

REFERENCES

[1] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. IEEE ICCV*, Oct. 2003, pp. 10–17.

[2] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[3] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.

[4] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Proc. ECCV*, 2008, pp. 705–718.

[5] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "TurboPixels: Fast superpixels using geometric flows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2290–2297, Dec. 2009.

[6] O. Veksler, Y. Boykov, and P. Mehrani, "Superpixels and supervoxels in an energy optimization framework," in *Proc. ECCV*, 2010, pp. 211–224.

[7] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[8] Z. Li and J. Chen, "Superpixel segmentation using linear spectral clustering," in *Proc. IEEE CVPR*, Jun. 2015, pp. 1356–1363.

[9] Gestalt. (2018). *Gestalt Principles*. [Online]. Available: http://facweb.cs.depaul.edu/sgrais/gestalt_principles.htm

[10] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. ACM SIGGRAPH*, Aug. 1996, pp. 31–42.

[11] Lytro. (2011). *Lytro Redefines Photography With Light Field Cameras*. [Online]. Available: http://www.lytro.com

[12] Raytrix. (2012). ∞ *raytrix*. [Online]. Available: http://www.raytrix.de

[13] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," *Comput. Models Vis. Process.*, vol. 1, no. 2, pp. 3–20, 1991.

[14] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. ACM SIGGRAPH*, Aug. 1996, pp. 43–54.

[15] A. Jarabo, B. Masia, A. Bousseau, F. Pellacini, and D. Gutierrez, "How do people edit light fields?" *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–146, 2014.

[16] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Depth estimation with occlusion modeling using light-field cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2170–2181, Nov. 2016.

[17] H. Zhu, Q. Zhang, and Q. Wang, "4D light field superpixel and segmentation," in *Proc. IEEE CVPR*, Jul. 2017, pp. 6709–6717.

[18] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[19] P. P. Srinivasan, M. W. Tao, R. Ng, and R. Ramamoorthi, "Oriented light-field windows for scene flow," in *Proc. IEEE ICCV*, Dec. 2015, pp. 3496–3504.

[20] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE CVPR*, Jul. 2014, pp. 2806–2813.

[21] T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 972–986, May 2012.

[22] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4D light-field dataset and CNN architectures for material recognition," in *Proc. ECCV*, 2016, pp. 121–138.

[23] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE ICCV*, Dec. 2013, pp. 673–680.

[24] H. Zhu, Q. Wang, and J. Yu, "Occlusion-model guided antiocclusion depth estimation in light field," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 965–978, Oct. 2017.

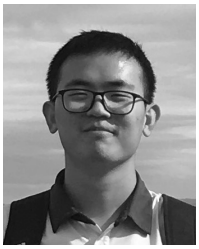
[25] S. Wanner, C. Straehle, and B. Goldluecke, "Globally consistent multi-label assignment on the ray space of 4D light fields," in *Proc. IEEE CVPR*, Jun. 2013, pp. 1011–1018.

[26] H. Mihara, T. Funatomi, K. Tanaka, H. Kubo, Y. Mukaigawa, and H. Nagahara, "4D light field segmentation with spatial and angular consistencies," in *Proc. IEEE ICCP*, May 2016, pp. 1–8.

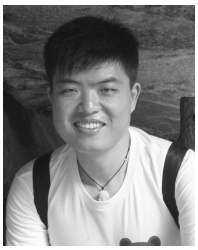
[27] M. Hog, N. Sabater, and C. Guillemot, "Light field segmentation using a ray-based graph structure," in *Proc. ECCV*, 2016, pp. 35–50.

[28] Y. Xu, H. Nagahara, A. Shimada, and R.-I. Taniguchi, "Transcut: Transparent object segmentation from a light-field image," in *Proc. IEEE ICCV*, May 2015, pp. 3442–3450.

[29] J. Yang and H. Li, "Dense, accurate optical flow estimation with piecewise parametric model," in *Proc. IEEE CVPR*, Jun. 2015, pp. 1019–1027.
[30] J. Chang, D. Wei, and J. W. Fisher, III, "A video representation using temporal superpixels," in *Proc. IEEE CVPR*, Jun. 2013, pp. 2051–2058.
[31] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara, "Gesture recognition in ego-centric videos using dense trajectories and hand segmentation," in *Proc. IEEE CVPR Workshops*, Jun. 2014, pp. 688–693.
[32] J. Chen, J. Hou, Y. Ni, and L.-P. Chau, "Accurate light field depth estimation with superpixel regularization over partially occluded regions," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4889–4900, Oct. 2018.
[33] H. Zhu *et al.*, "Full view optical flow estimation leveraged from light field superpixel," *IEEE Trans. Comput. Imag.*, to be published.
[34] M. Hog, N. Sabater, and C. Guillemot, "Superrays for efficient light field processing," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1187–1199, 2017.
[35] M. Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *Proc. IEEE CVPR*, Jun. 2011, pp. 2097–2104.
[36] K. Yamaguchi, D. McAllester, and R. Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in *Proc. ECCV*, 2014, pp. 756–771.
[37] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 1469–1472.
[38] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *Proc. VMV*, 2013, pp. 225–226.
[39] (2014). *Blender*. [Online]. Available: https://www.blender.org/
[40] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE CVPR*, Jun. 2013, pp. 1027–1034.
[41] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2. Jul. 2001, pp. 416–423.
[42] CVPGNWPU. (2015). *Computer Vision and Computational Photography Group*. [Online]. Available: http://www.npu-cvpg.org/opensource
[43] H.-G. Jeon *et al.*, "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE CVPR*, May 2015, pp. 1547–1555.
[44] L. Si and Q. Wang, "Dense depth-map estimation and geometry inference from light fields via global optimization," in *Proc. ACCV*, 2016, pp. 83–98.
[45] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *Proc. IEEE CVPR*, Jul. 2012, pp. 41–48.



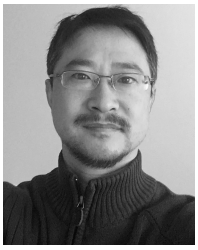
Hao Zhu received the B.E. degree from the School of Computer Science, Northwestern Polytechnical University in 2014, where he is currently pursuing the Ph.D. degree with the School of Computer Science. His research interests include computational photography, and light field computing theory and application.



Qi Zhang received the B.E. degree in electronic and information engineering from the Xi'an University of Architecture and Technology in 2013, and the M.S. degree in electrical engineering from Northwestern Polytechnical University in 2015, where he is currently pursuing the Ph.D. degree with the School of Computer Science. His research interests include computational photography, and light field computing theory and application.

Qing Wang (M'05–SM'19) graduated from the Department of Mathematics, Peking University, in 1991, and received the master's and Ph.D. degrees from the Department of Computer Science and Engineering, Northwestern Polytechnical University, in 1997 and 2000, respectively. He is currently a Professor with the School of Computer Science, Northwestern Polytechnical University. He was a Research Assistant and Research Scientist with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University from 1999 to 2002. He was also as a Visiting Scholar with 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 over 100 papers in the international journals and conferences. His research interests include computer vision and computational photography, such as 3D structure and shape reconstruction, object detection, tracking and recognition in dynamic environment, and light field imaging and processing.

Dr. Wang is a Senior Member of the China Computer Federation (CCF) and a member of the ACM. He was a recipient of the Outstanding Talent Program of New Century by the Ministry of Education, China, in 2006.



Hongdong Li received the Ph.D. degree from Zhejiang University. He is currently a Reader with the Computer Vision Group, The Australian National University (ANU). He is also a Chief Investigator of the Australia ARC Centre of Excellence for Robotic Vision (ACRV). He started as a Post-Doctoral Fellow with ANU in 2004. His research interests include 3D vision reconstruction, structure from motion, multi-view geometry, and the applications of optimization methods in computer vision. Prior to 2010, he was with NICTA Canberra Labs working on the Australia Bionic Eyes Project. He was a recipient of a number of best paper awards in computer vision and pattern recognition and the CVPR Best Paper Award in 2012 and the Marr Prize Honorable Mention in 2017. He served as the Area Chair in recent year ICCV, ECCV, and CVPR. He was the Program Chair of the Australia Conference on Robotics and Automation (ACRA) 2015 and the Program Co-Chair of the Asian Conference on Computer Vision (ACCV) 2018. He is an Associate Editor of the IEEE T-PAMI.