Blind-Depth Light Field Super-Resolution

Lei Zhang\textsuperscript{1}, Jianpeng Fan\textsuperscript{2}, \textsuperscript{*} and Jungang Yang\textsuperscript{2}

\textsuperscript{1}Beijing Institute of Tracking and Telecommunication Technology, Beijing, China. 
\textsuperscript{2}College of Electronic Science and Technology, National University of Defense Technology, Changsha, China.

Emails: zhanglei_bitt@163.com, jpfan@nudt.edu.cn (Corresponding Author), yangjungang@nudt.edu.cn

Abstract. Light field cameras have drawn much attention due to the ability of post-capture adjustments such as refocusing and 3D reconstruction. However, the low resolution has been the bottleneck of them and limits their further application. To solve this problem, various methods have been proposed to increase the resolution, namely, to recover details and reconstruct a high resolution (HR) image of the scene from several low resolution (LR) observations. Most of the previous super-resolution (SR) frameworks depend on prior knowledge of depth information, which is an indispensable part in their SR approaches. However, it may be difficult to obtain precise depth information in some practical situations. In this paper, we propose a novel light field super-resolution (LFSR) framework independent of the prior depth information. The framework combines a variational regularization-based SR approach with light field refocusing process, which can super-resolve the focused region while preserve the unfocused region and generate a series of multi-focus super-resolved images. We then employ a multi-focus image fusion algorithm based on stationary wavelet transform (SWT) and finally obtain an all-in-focus HR image. Synthetic and real-world datasets are utilized to demonstrate the effectiveness of the proposed framework both quantitatively and visually.

1. Introduction

Light field cameras have drawn more and more attention from both industry and academia recently. Compared to traditional 2D cameras, a light field camera not only catches the accumulated intensity at each point, but also separates intensity values for each ray direction. In other words, it captures the 3D scene geometry and reflectance property [1], possessing a strong capability of post-capture adjustments [31, 32]. It can be clearly seen that light field cameras are becoming increasingly popular and have a huge commercial potential.

To capture the light field signal, many camera architectures have been proposed such as microlens array based design [2], mask based design [3], coded aperture based design [4] and camera array based design [5]. Among these, the microlens array based design (also called plenoptic camera) is commercially available such as Lytro and Raytrix [6], which places a microlens array between main lens and CCD sensors to provide a sample of the light field. However, the design faces a trade-off between special and angular resolution [1]. Since the total sensor resolution is limited, one can either opt for a dense sampling in the angular (viewpoint) domain [7], or vice versa [2, 8, 9]. Relatively, camera array based design has shown the potential advantage in image resolution. The design sets several independent cameras in an array to get the parallax of the scene. The resolution of each sub-image caught by the camera can be higher and the baseline (the distance between two neighbouring camera) is adjustable. All these properties outperform the plenoptic camera.
Corporation in Columbia University has developed an ultra-thin high performance monolithic camera array called PiCam, which can capture light fields and synthesize high resolution (HR) images along with a range image (scene depth) through integrated parallax detection and superresolution (SR) \[10\]. Wang et al \[11\] set 8 low resolution (LR) side cameras around the central high-quality single lens reflex (SLR) lens and then formed a small camera array based light field imaging device. With the development of camera miniaturization, camera array based light field imaging devices can hopefully surpass other professional imaging devices such as SLR in the coming future.

Besides the preceding advantages, light field cameras can utilize the parallax obtained by different sub-cameras to super-resolve the composited image \[33, 34, 40, 41\]. SR is a technique for generating an HR image using one or several observed LR images \[35, 36, 37, 38, 39\]. Namely, the goal of SR is to increase the pixel density and use an appropriate reconstruction algorithm to estimate the true high-frequency components. Typical SR techniques utilize a maximum likelihood (ML) method \[12\], maximum a posterior (MAP) method \[13, 14\], iterative backward projection (IBP) method \[15\], frequency domain approach \[16\], or projection onto convex sets (POCS) method \[17, 18\] as a reconstruction algorithm. However, all the SR methods above assume that the sub-pixel shifts between each LR image are constants, which means that all pixels in an image correspond to the same shift. Nevertheless, in the sub-images caught by an array of cameras, shifts of pixels depend not only on the coordinate of the camera, but also on the depth in the scene. Due to the spatial-variant property of the light field, traditional SR techniques need to be modified. As a result, light field super-resolution (LFSR) techniques have been proposed. Bishop et al. \[19\] firstly designed a variational Bayesian framework to recover more information and super-resolved light field images. Wanner and Goldluecke \[1, 20\] optimized a continuous variational framework and generated super-resolved novel views of one scene. Nava et al. \[26\] introduced the focal stack transform and produced an all-in-focus image of a scene.

Most of the previous work \[1, 6, 10, 11, 19, 20, 21\] depends on the depth map or disparity map to obtain pixels’ shearing shifts in LR images. However, it may be difficult to obtain precise depth information in some practical situations especially near edges or in richly-textured area. The increasing estimating error in depth map will mismatch the pixels and the shifts, which can reduce the sharpness of the SR result. Inspired by the light field refocusing process \[22, 31\], the depth information is actually not necessary in SR process. In this paper, we propose a novel LFSR framework which does not utilize estimated depth information. The framework combines a variational regularization-based SR method with light field refocusing process, generating an HR image with focused region super-resolved and unfocused region well-preserved. By varying the ‘slope’ of the refocusing process, a series of multi-focus and super-resolved images can be produced. We then employ a stationary wavelet transform (SWT) based fusion algorithm to obtain an all-in-focus HR image of the scene.

This paper is organized as follows: In Section 2 we formulate the light field degradation model and adapt the regularization-based SR model to light field situation. In Section 3 we present our Blind-Depth LFSR framework systematically. Extensive experiments are conducted using both synthetic and real-world datasets to test the effectiveness of the proposed framework in Section 4. Finally the conclusions are drawn in Section 5.

2. Problem Formulation

2.1. Lightfield Images Degradation Model

The HR image we need to reconstruct is a 2D projection of the 3D real-world scene. The light field sub-images caught by cameras suffer from shearing shifts (caused by parallax disparity), blurring (caused by optical distortions) and down sampling (caused by unideal CCD or CMOS density). Due to the 3D property of the light field, the refocused objects on various depth correspond to different shifts in each sub-image. For objects near the cameras, the parallax between each camera is obvious, which dues to the larger shearing shifts in sub-images. However, for objects in remoter area, the parallax between different cameras is tiny, which corresponds to smaller shearing shifts in sub-images. Since a
number of SR papers [12-18] neglect the depth property, we take this into consideration and formulate
the light field images degradation process as

\[ y_k = \sum_{i=1}^{p} D_k H_{i,k} F_{i,k} x_i + N_k \quad (k = 1, 2, \ldots, N) \]

\[ x = \sum_{i=1}^{p} x_i, \]  

where \( x \) is a lexicographically ordered vector, which represents the desired, ideal HR image captured by a selected reference camera (usually but not necessarily choose the centre camera). Since the refocused image on arbitrary depth has a depth of field, which means that there is a distance between the nearest and the furthest depth in focus. We divide the real-world scene into \( p \) depth layers. \( x_i \) is the \( i^{th} \) layer of \( x \), with pixels on the layer reserved and pixels out of the layer set to 0. \( y_k \) is the vectorized \( k^{th} \) observation caught by an sub-camera. Suppose that the observed sub-images are sized \( n \times n \) and \( m \) is the upsampling rate. The sizes of \( y_k, x_i \) and \( x \) are \( 2n \times 2n \), \( 2mn \times 2mn \) and \( 2mn \times 2mn \) respectively. As we have discussed above, pixels on different depth or caught by different cameras correspond to different shifts. \( F_{i,k} \) is the shifting matrix corresponding to \( k^{th} \) camera on \( i^{th} \) depth layer, which is sized \( 2mn \times 2mn \). \( H_{i,k} \) is also an \( 2mn \times 2mn \) matrix which represents the total blur. The \( 2n \times 2n \) matrix \( D_k \) reflects the down-sampling process from the HR real scene image by the \( k^{th} \) sub-camera’s sensor. \( N_i \) represents the imaging noise assumed to be i.i.d. Gaussian. Since \( H_{i,k} \) and \( D_k \) vary little with the depth, and each sub-camera has similar imaging configuration, we assume that each sub-image has the same blurring matrix \( H \) and down-sampling matrix \( D \). Then the degradation model becomes

\[ y_k = DH \sum_{i=1}^{p} F_{i,k} x_i + N_k \quad (k = 1, 2, \ldots, N) \]  

2.2. Variational Regularization-Based SR Model

After the degradation model being formulated, we need to estimate \( x \) through the observations \( y_k \). As the dimension of \( y_k \) is smaller than the dimension of \( x \), the problem becomes ill-posed and the solution is not unique. So we explore variational regularization-based SR method to solve the ill-posed LFSR problem. The objective function can be established as follows:

\[ \hat{x} = \arg \min_x \left\{ \sum_{k=1}^{p} \left\| y_k - DH \sum_{i=1}^{p} F_{i,k} x \right\|_2^2 + \lambda J(x) \right\} \]

\[ = \arg \min_x \left\{ E(x) + \lambda J(x) \right\} \]  

Subject to \( x = \sum_{i=1}^{p} x_i \).

In (3), \( \hat{x} \) is the estimation of the HR image. \( E(x) \) is called fidelity term which is an \( L_2 \) norm expression measuring the error between estimations and observations. \( J(x) \) represents regularization term which controls the smoothness of the solution. \( \lambda \) is a scalar to balance the weight between fidelity and smoothness, which depends on the kind of regularization terms and is generally set empirically.

We employ BTV regularization term as shown in (4), which is generally utilized and confirmed to have a better performance [23].

\[ J(X) = \sum_{l=1}^{p} \sum_{m=0}^{p} \alpha^l \| X - S_l^c S_m^c X \|_l \|, (l + m \geq 0) \]  

(4)
In (4), matrices $S_l^x$ and $S_m^y$ shift $X$ by $l$ and $m$ pixels in horizontal and vertical directions respectively. The scalar weight $\alpha$ ($0 < \alpha < 1$) is applied to give a spatially decaying effect to the summation of the regularization terms.

3. Blind-Depth LFSR Process

3.1. Blind-Depth LFSR Framework

Due to the special property of light fields, the optimization problem formulated in (3) cannot be solved directly. Since the shearing shifts are related to the depth of the scene, depth-relied LFSR techniques [1, 6, 10, 11, 19, 20, 21] first employ depth or disparity estimation to build a bridge between pixels’ locations and their corresponding shifts. However, as the refocusing techniques mentioned in [22], we do not need depth information while refocusing sub-images in various depth. Merely presenting the slope into a series of discrete values and traversaling most depth of the scene will be enough.

Inspired by the refocusing process, we propose a novel LFSR framework independent of the depth information. We first refocus the sub-images and project it onto a target HR buffer simultaneously. A bicubic interpolation method is used in the process. Then the refocused ‘HR’ image is adopted to a variational regularization-based SR algorithm as an initial value. After the iterations, the algorithm converges to an HR image with focused region super-resolved and unfocused region well-preserved. By traversaling the present slope parameter, we can get a series of multi-focus HR images. Finally we use SWT-based fusion method to generate an all-in-focus HR image of the scene. The process of the proposed framework is demonstrated in Figure 1.

![Blind-Depth LFSR framework](image)

Figure 1. Blind-Depth LFSR framework

3.2. Variational Regularization-Based LFSR Method

Since the initial input of the iteration is refocused on particular depth, which means objects in the depth of field are clear while objects out of the depth of field are blurred. So the LFSR method needs to be adapted. We propose a local optimal function replacing the global optimal function represented in (3), as

$$\hat{X}_i = \arg \min_{x_i} \left\{ \sum_{k=1}^{N} \| \nu_k - DHF_{i,k} X_i \|^2 + \lambda J (X_i) \right\}$$

(5)

Where $X_i$ is the desired, vectorized HR image focused on the $i^{th}$ depth layer, with the region in focus clear and out of focus blurred. Then we can represent $X_i$ as

$$X_i = x_i + P_i$$

(6)

$x_i$ has the same implication as it has in (1). $P_i$ corresponds to the unfocused region while focusing on $i^{th}$ depth layer. It is obvious that the Hadamard product of $x_i$ and $P_i$ is zero vector. Then we derive the variational degradation model as
Y_{i,k} = DHF_{i,k}X_i + N_k = DHF_{i,k}X_i + DHF_{i,k}P_i + N_k \quad (y_{i,k} \in Q_{i,k} = 0) \quad (7)

Y_{i,k} is the projection of X_i in LR spaces with focused region aligned and unfocused region blurred. Correspondingly, \( y_{i,k} \) represents the focused region in the \( k^{th} \) sub-image and on the \( i^{th} \) depth layer while \( Q_{i,k} \) represents the corresponding unfocused region. Assume that there is no occlusion, the Hadamard product of \( y_{i,k} \) and \( Q_{i,k} \) is still a zero vector. Based on this property, we establish a local SR model and separate the minimization into two parts.

\[
\arg\min_{X_i} \left\{ \sum_{k=1}^{N} \left[ y_{i,k} - DHF_{i,k}X_i \right] + \lambda J(X_i) \right\} = \arg\min_{T(X_i)} + \arg\min_{T(P_i)}
\]

We adopt gradient descent method in [10] to obtain the approximate solution, as

\[
\nabla T(X_i) = \nabla T(x_i) + \nabla T(P_i) = \sum_{k=1}^{N} F_{i,k}^TH^TD^T(DHF_{i,k}X_i - y_{i,k}) + \lambda \nabla J(x_i) + \sum_{k=1}^{N} F_{i,k}^TH^TD^T(DHF_{i,k}P_i - (y_i - y_{i,k})) + \lambda \nabla J(P_i) \quad (9)
\]

We employ the projection method proposed in [24]. Fully taking the sparseness of the degradation matrix into consideration, we can transfer matrix multiplication into operations upon the image matrix, which can greatly reduce the computational complexity.

\[
\hat{X}_i^{(n+1)} = \hat{X}_i^{(n)} - \beta \times \nabla T\left( \hat{X}_i^{(n)} \right) \quad (10)
\]

During the iterations, we first project the estimated HR image into LR space and subtract the estimation from LR images. Then we project the residuals onto HR grid and correct the estimated HR image. \( \beta \) is the step size to balance the weight of correction.

Figure 2. Illustration of ‘slope’ while refocusing light field sub-images
(a) Objects on different depth layers correspond to different slopes. (b) Refocusing the sub-images by slope 1 will cause objects on other depth layers unaligned

3.3. Focused Region Recognition

After the iteration in (10), both focused and unfocused region are ‘super-resolved’. However, as Figure. 2 shows, the unfocused region does not match to a correct shift in the refocusing process. In other words, they
are not aligned by $F_{i,k}$. As illustrated in Figure 5 ‘D3’ and ‘D5’, it will cause an ill result in unfocused regions, which results in the poor SR quality and thus damages the final fusion result. Consequently, we have to employ a method to recognize the focused region and prevent the unfocused region from being damaged.

Inspired by the epipolar image (EPI) proposed in [1], [6], we represent a novel approach. Consider that the focused region in LR space is aligned by the shearing shifts. Thus the undulation among the residuals in LR space is tiny. However, due to the unalignment of the unfocused region, there is a much larger undulation, which causes a larger variance. Therefore, we judge the region in focus exactly by the variance of the pixel residuals and match the variance to the weight of correction matrix in the iteration.

![Figure 3. Focused Region Recognition](image)

(a) The ‘Lego Knights’ scene refocused on the two side knights; (b) The variance map of the residual matrix; (c) The plot of transform proposed in (11); (d) The transformed weight map

As Figure 3 (a) shows, we use the camera array light field datasets from Stanford University [27] and refocus them on the depth of two side knights. Then we calculate the residuals and draw the variance map. From Figure. 3 (b) we can see that the region in focus corresponds to smaller variance. We provide a mapping relationship to match the variance to the weight of the correction matrix, which is demonstrated in (11) and Figure 3 (c).

$$V = -0.1 \exp{\theta}$$  \hspace{1cm} (11)

Where $V$ represents the variance in certain pixel and $\theta$ is the weight of the corresponding element in the correction matrix. And the correction matrix can be obtained by (12).

$$C(X_i) = \sum_{k=1}^{N} F_{i,k}^T H^T D^T (\Theta e(DHF_{i,k} X_i - y_i)) + \lambda \nabla J(X_i)$$  \hspace{1cm} (12)

Where $\Theta$ is the weight vector matching to the corresponding pixels in the residual vector in LR space. By (11), the variance map can be transformed into weight map, which is depicted in Figure 3 (d).

### 3.4 Multi-Focus Image Fusion

Multi-focus image fusion techniques aim at the problem that a clear image with all objects in focus cannot be obtained within one snapshot due to the camera’s limited depth of field. And the goal of the techniques is to create an all-in-focus image from multi-focus images [25]. Since the camera array has a large equivalent aperture, the algorithm is required to deal with a large number of images with an extremely shallow depth of field. To solve the problem, we employ an SWT-based multi-focus image
fusion algorithm. We first use SWT to decompose multi-focus source images, then integrate and select the wavelet coefficients under different scales. Finally we fuse the chosen coefficients with inverse SWT to generate a composited all-in-focus image. The fusion process is demonstrated in Figure. 4.

We assume that the multi-focus source images named \( I_n \) \((n = 1, 2, ..., N)\) are decomposed into \( k \) levels, which contain approximate components \( A \), horizontal details \( H \), vertical details \( V \) and diagonal details \( D \) in each level. Suppose the size of \( I_n \) is \( P \times Q \), due to the avoidance of downsampling process \[26\], the size of all SWT coefficients \( A, H, V, D \) in each level is \( P \times Q \) as well. Then different fusion rules are adopted to select the coefficients.

Usually, the ‘max absolute value (Max Abs)’ rule is employed to select the detail coefficients because the focused area always corresponds to more edges and textures, whose detail coefficients’ absolute values are large. Max Abs rule can be represented as

\[
H^k(i, j) = \arg \max_{n} \|H^k_n(i, j)\| \quad (1 \leq n \leq N)
\]

(13)

Where \( N \) is the total number of the source multi-focus images, \( n \) is the frame of the source images ranged 1 to \( N \), \((i, j)\) represents the coordinate in decomposed coefficient images. \( k \) represents the decomposition level.

\( H^k_n \) represents the horizontal detail’s coefficients of image \( n \) in \( k^\text{th} \) level and \( H^k \) represents the selected horizontal detail’s coefficients in \( k^\text{th} \) level. The same selecting process can be employed to the coefficients \( V \) and \( D \).

As for approximate components which contain some large-scale information, ‘average (Avg)’ rule is commonly adopted in many algorithm, which is defined as

\[
A^k(i, j) = \frac{1}{N} \sum_{n=1}^{N} A^k_n(i, j)
\]

(14)

Where \( A^k \) represents the approximate components’ coefficients of image \( n \) in \( k^\text{th} \) level and \( A^k \) represents the integrated approximate components’ coefficients in \( k^\text{th} \) level. Both rules above are confirmed efficient by practice. Finally, the integrated coefficients need to be fused to obtain the output image \( I_f \).

![Figure 4. SWT-based Multi-focus Image Fusion Methodology](image)

4. Blind-Depth LFSR Process

4.1. Blind-Depth LFSR Framework

In this section, both synthetic and real-world scenes are employed to test the effectiveness of the proposed framework. As for synthetic scenes, we use the datasets produced by the light field laboratory in Stanford University \[27\]. We first simulate the degradation process (shifting, blurring and downsampling) and then employ the simulated LR sub-images to the algorithms. We set the original center-view sub-image as a reference image. By comparing the results with the reference image, we can get precise quantitative metrics of the methods, which is significant for the evaluation.
Average Gradient (AG), Spatial Frequency (SF), Root Mean Square Error (RMSE), Peak Signal to Noise Ratio (PSNR) and Correlation Coefficient (Corr) are employed to make a quantitative assessment [28], [29].

The Stanford University’s light field datasets are produced by a 17*17 array of cameras. We choose the central 5*5 sub-images to run the experiment. We arbitrarily select the scenes named ‘Lego Knights’, ‘Stanford Bunny’ and ‘Tarot Cards’, with each sub-images’ resolution 1024*1024. As for all the datasets, we down-sample the source sub-images into the resolution 256*256, then employ a 3*3 Gaussian blurring kernel with the variance 0.5 to simulate the blurring effect. Finally, we add a zero mean standard Gaussian white noise to each sub-image. In the LFSR process, we set the correction rate $\beta$ to 0.05 and the regularization weight $\lambda$ to 0.001. And the number of iteration is set to 10. In the fusion process, the SWT decomposition level is set to 7. All of the parameters are optimized to balance the effectiveness and the computational complexity. Bicubic interpolation method, LFSR approach without focused region recognition (rough LFSR method) and the proposed LFSR method are compared. The visual results are shown in Fig. 5 and the quantitative results are listed in Table 1.

![Figure 5. Reconstruction results of the synthetic scenes](image)

From left to right: A1–A5 are from the central-view sub-images, which are utilized as reference HR images; B1–B5 are the central-view degraded LR images, which are employed as inputs of the framework; C1–C5 are the bicubic interpolation results; D1–D5 are the results of rough LFSR approach; E1–E5 are the results of our proposed LFSR framework. And from up to down: Images in Row 1 are refocused on one certain depth. Row 2 and the Row 3 show the magnificent details of Row 1 in focus and out of focus respectively. Row 4 shows the final all-in-focus fusion results and Row 5 shows the magnificent details of Row 4.
Table 1. Formatting sections, subsections and subsubsections

|                | Bicubic Interpolation | Rough LFSR | Proposed LFSR |
|----------------|-----------------------|------------|---------------|
| Lego Knights   |                       |            |               |
| AG             | 2.4259                | 4.9836     | 2.8800        |
| SF             | 5.9292                | 13.0703    | 6.7474        |
| Rmmse          | 8.9717                | 14.2384    | 8.5524        |
| PSNR           | 29.0734               | 25.0616    | 29.4891       |
| Corr           | 0.9898                | 0.9746     | 0.9907        |
| AG             | 1.4016                | 2.0827     | 1.7697        |
| SF             | 3.8286                | 5.3735     | 4.3809        |
| Stanford Bunny |                       |            |               |
| Rmmse          | 3.6939                | 4.5420     | 3.5649        |
| PSNR           | 36.5592               | 34.7609    | 36.8649       |
| Corr           | 0.9989                | 0.9983     | 0.9990        |
| AG             | 4.0493                | 8.2750     | 4.9619        |
| SF             | 8.4964                | 17.6346    | 10.1711       |
| Tarot Cards    |                       |            |               |
| Rmmse          | 15.3030               | 19.3742    | 13.8524       |
| PSNR           | 24.2478               | 22.1988    | 25.1130       |
| Corr           | 0.9614                | 0.9395     | 0.9685        |

From Figure 5 we can see that the bicubic interpolation method has little contributions in the superresolution but causes no distortion in the fusion results. In contrast, the rough LFSR method can produce highly super-resolved results in focused region but causes severe distortion in unfocused region, which destroys the fusion results obviously. By comparison, although the proposed LFSR framework loses some sharpness in focused region, it can preserve the unfocused region and generate a satisfactory fusion result without distortion.

By analysing the quantitative results from Table 1, we can also infer that the distortion caused by the rough LFSR method in unfocused region makes a huge difference in fusion results. It not only brings in the noise, which results in a high AG and SF, but also reduces the fidelity, which reduces the PSNR, Corr and increases the Rmmse. In contrast, our proposed method can deal with the detail-riched scenes and generate the results with best quantitative metrics.

The experiments with synthetic scenes preliminarily demonstrate the outperformance of our proposed framework. However, the simulation cannot be equivalent to the real degradation process. In order to test the practical value of the framework, experiments with real-world scenes are undoubtedly necessary. Since the single camera scanning approach is equivalent to the camera array approach when dealing with statistic scenes. We employ the self-developed scanning gantry with cameras on it to capture the real-world scene in our own laboratory. The devices and the scenes is shown in Figure 6.

![Figure 6. Self-developed light field experimental devices](image)

(a) The scanning gantry with the Leika Q camera on it. (b) The scanning gantry with the Meizu smart phone on it. (c) The controller of the scanning gantry. (d) The real-world scene employed in the experiments

In the experiments, we transfer the camera into 25 different positions to form a 5*5 array. With the Leika Q camera, the exposure time is set to 0.1 second and the size of camera’s aperture is set to 2 mm. Since the Meizu (Meilan 5) smart phone camera has a certain aperture (2.2mm) which is unadjustable, we set the exposure time to 0.05 second which can achieve a proper brightness. After the camera
calibration [30] and sub-images registration, the resolutions of rectified sub-images are 2148*2048 for Leika Q camera and 2560*1440 for phone camera respectively. To reduce the computational complexity, we cut the central 1024*1024 pixels and super-resolve them into a 4096*4096 resolution. The level of the SWT decomposition is set to 10 and other parameters are the same as those in former synthetic scene experiments. Figure. 7 and Figure. 8 show the visual results of the Leika Q camera and the Meizu smart phone camera respectively.

**Figure 7.** Reconstruction results using real-world scene datasets captured by Leica Q camera
*From left to right: A1~A5 are the central-view sub-images; B1~B5 are the bicubic interpolation results; C1~C5 are the results of the rough LFSR method; D1~D5 are the results of our proposed LFSR framework. And from up to down: Images in Row 1 are refocused on the calendar. Row 2 and the Row 3 show the magnificent details of Row 1 in focus and out of focus respectively. Row 4 shows the all-in-focus fusion results and Row 5 shows the magnificent details of Row 4.*
Figure 8. Reconstruction results using real-world scene datasets captured by the Meizu smart phone camera. From left to right: A1~A5 are the central-view sub-image; B1~B5 are the bicubic interpolation results; C1~C5 are the results of the rough LFSR method; D1~D5 are the results of our proposed LFSR framework. And from up to down: Images in Row 1 are refocused on the right calendar. Row 2 and the Row 3 show the magnificent details of Row 1 in focus and out of focus respectively. Row 4 shows the all-in-focus fusion results and Row 5 shows the magnificent details of Row 4.

From the results of the real-world scene experiments, it can be seen that the proposed framework works well and produces results with the highest visual quality. In Fig. 7 we observe the details ‘date’, ‘remarks’ and ‘days of the week’ and in Fig. 8 we observe the numbers on the calendar. The proposed LFSR framework produces more clear results than the original central-view sub-images and the results of bicubic interpolation. Compared with the results of rough LFSR method, our proposed framework have a better performance in the reduction of the block distortion and generates much cleaner images. The two scenes with different cameras reasonably demonstrate the effectiveness and the practical value of our framework.

5. Conclusion
In this paper, we propose a novel LFSR framework free of prior estimated depth information. Refocused image is used as an initial value of the iteration and a regularization-based SR algorithm is adapted to produce an HR image with focused region super-resolved. By traversing the preset slope parameters, we can obtain a series of multi-focus images. Finally we use the SWT-based fusion
approach to generate an all-in-focus HR image of the scene. Extended experiments have demonstrated the effectiveness of the framework. How to reduce the computational complexity and optimize the parameters are worthy to be further researched.

6. References

[1] Wanner, S.; Goldluecke, B. Variational light field analysis for disparity estimation and super-resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence 2014, 36, 606–619.

[2] Ng, R.; Levoy, M.; Bredif, M.; Duval, G.; Horowitz, M.; Hanrahan, P. Light field photography with a hand-held plenoptic camera. Stanford University CSTR 2005, 2.

[3] Veeraraghavan, A.; Raskar, R.; Agrawal, A.; Mohan, A.; Tumblin, J. Dappled photography: Mask enhanced cameras for heterodyned lightfields and coded aperture refocusing. ACM Transactions on Graphics 2007, 26.

[4] Liang, C.-K.; Lin, T.-H.; Wong, B.-Y.; Liu, C.; Chen, H. H. Program-mable aperture photography: Multiplexed light field acquisition. ACM Transactions on Graphics 2008, 27, 55:1–55:10.

[5] Wilburn, B.; Joshi, N.; Vaish, V.; Talvala, E.-V.; Antunez, E.; Barth, A.; Adams, A.; Horowitz, M.; Levoy, M. High performance imaging us-ing large camera arrays. ACM Transactions on Graphics 2005, 24, 765–776.

[6] Mitra, K.; Veeraraghavan, A. Light field denoising, light field superreso-lution and stereo camera based refocussing using a GMM light field patch prior. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops; 2012; pp. 22–28.

[7] Perwass, C.; Wietzke, L. The next generation of photography. White Paper,http://www. raytrix. de 2010, 24, 765–776.

[8] Bishop, T.; Favaro, P. Full-resolution depth map estimation from an aliased plenoptic light field. In Asian Conference on Computer Vision; 2010; pp. 186–200.

[9] Georgiev T.; Lumdsaine, A. Focused plenoptic camera and rendering. Journal of Electronic Imaging 2010, 19, 021106-021106.

[10] Venkataraman, K.; Lelescu, D.; Duparre, J.; McMahon, A.; Molina, G.; Chatterjee, P.; Mullis, R.; Nayar, S. Picam: An ultra-thin high perfor-mance monolithic camera array. ACM Transactions on Graphics 2013, 32, 166-166.

[11] Wang, Y.; 222 Liu, Y.; Heidrich, W.; Dai, Q. The light field attachment: turning a dslr into a light field camereausing a low budget camera ring. IEEE Transactions on Visualization and Computer Graphics 2016, 99, 1-1.

[12] Tom, B.; Katsaggelos, A. Reconstruction of a high-resolution image by simultaneous registration, restoration, and interpolation of low-resolution images. In International Conference on Image Processing; 1995; pp. 539–542.

[13] Schultz, R.; Stevenson, R. A bayesian approach to image expansion for improved definition. IEEE Transactions on Image Processing 1994, 3, 233–242.

[14] Schultz, R.; Stevenson, R. Extraction of high-resolution frames from video sequences. IEEE Transactions on Image Processing 1996, 5, 996–1011.

[15] Irani, M.; Peleg, S. Improving resolution by image registration. Graphical Models and Image Processing 1991, 53, 231–239.

[16] Tsai, R.; Huang, T. Multiframe image restoration and registration. In International Conference on Computer Vision and Image Processing; 1984; pp. 317–339.

[17] Eren, P.; Sezan, M.; Tekalp, A. Robust object-based high resolution image reconstruction from low-resolution video. IEEE Transactions on Image Processing 1997, 6, 1446–1451.

[18] Patti, A.; Altunbasak, Y. Artifact reduction for set theoretic super resolution image reconstruction with edge adaptive constraints and higher-order interpolants. IEEE Transactions on Image Processing 2001, 10, 179–186.

[19] Bishop, T.; Zanetti, S.; Favaro, P. Light field superresolution. In International Conference on Intelligent Computer Communication and Processing; 2009; pp. 1–9.

[20] Wanner, S.; Goldluecke, B. Spatial and angular variational super-resolution of 4d light fields. In European Conference on Computer Vision; 2012; pp. 608–621.
[21] Wang, Y.; Hou, G.; Sun, Z. A simple and robust super resolution method for light field images. In IEEE International Conference on Image Processing; 2016; pp. 1459–1463.

[22] Vaish, V.; Wilburn, B.; Joshi, N. Using plane + parallax for calibrated dense camera arrays. In IEEE Conference on Computer Vision and Pattern Recognition; 2004; pp. 1–1.

[23] Rudin, I.; Osher S.; Fatemi, E. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 1992, 60, 259–268. Physica D

[24] Farsiu, S.; Robinson, D.; Elad, M. Fast and robust multiframe super resolution. IEEE Transactions on Image Processing 2004, 13, 1327–1344.

[25] Li, S.; Kang, X.; Fang, L. Pixel-level image fusion: A survey of the state of the art. Information Fusion 2017, 33, 100–112.

[26] Zhou, Z.; Lin, J.; Jin, W. Image fusion by combining SWT and variational model. In International Congress on Image and Signal Processing; 2011; pp. 1907–1910.

[27] Stanford. The (new) stanford light field archive, http://lightfield.stanford.edu. 2008.

[28] Xydeas, S.; Petrovic, V. Objective image fusion performance measure. Electronics letters 2000, 36, 308–309.

[29] Piella, G.; Heijmans, H. A new quality metric for image fusion. In IEEE International Conference on Image Processing; 2003; pp. III–173.

[30] Zhang, Z. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000, 22, 1330–1334.

[31] Wang, Y.; Yang, J.; Guo, Y.; Xiao, C.; An, W. Selective light field refocusing for camera arrays using bokeh rendering and superresolution. IEEE Signal Processing Letters 2019, 26, 204–208.

[32] Wang, Y.; Wu, T.; Yang, J.; Wang, L.; An, W.; Guo, Y. DeOccNet: Learning to See Through Foreground Occlusions in Light Fields. In International Workshop on Applications of Computer Vision; 2020; pp. 118–127.

[33] Wang, L.; Wang, Y.; Liang, Z.; Lin, Z.; Yang, J.; An, W.; Guo, Y. Learning Parallax Attention for Stereo Image Super-Resolution. In IEEE Conference on Computer Vision and Pattern Recognition; 2019; pp. 12250–12259.

[34] Ying, X.; Wang, Y.; Wang, L.; Sheng, W.; An, W.; Guo, Y. A Stereo Attention Module for Stereo Image Super-Resolution. IEEE Signal Processing Letters 2020, 27, 496–500.

[35] Wang, L.; Guo, Y.; Lin, Z.; Deng, X.; An, W. Learning for video super-resolution through HR optical flow estimation. In Asian Conference on Computer Vision; 2018; pp. 514–529.

[36] Wang, L.; Guo, Y.; Liu, L.; Lin, Z.; Deng, X.; An, W. Deep Video Super-Resolution using HR Optical Flow Estimation. IEEE Transactions on Image Processing 2020, 29, 4323 - 4336.

[37] Wang, Y.; Wang, L.; Yang, L.; An, W.; Yu, J.; Guo, Y. Spatial-Angular Interaction for Light Field Image Super-Resolution. arXiv preprint arXiv:1912.07849 2019.

[38] Ying, X.; Wang, 274 L.; Wang, Y.; Sheng, W.; An, W.; Guo, Y. Deformable 3D Convolution for Video Super-Resolution. arXiv preprint arXiv:2004.02803 2020.

[39] Wang, L.; Wang, Y.; Lin, Z.; Yang, J.; An, W.; Guo, Y. Learning for Scale-Arbitrary Super-Resolution from Scale-Specific Networks. arXiv preprint arXiv:2004.03791 2020.

[40] Wang, Y.; Wang, L.; Yang, J.; An, W.; Guo, Y. Flickr1024: A large-scale dataset for stereo image super-resolution. In IEEE International Conference on Computer Vision Workshop; 2019; pp. 3852-3857.

[41] Wang, Y.; Wang, L.; Yang, J.; An, W.; Guo, Y. Flickr1024: A dataset for stereo image super-resolution. arXiv preprint arXiv:1903.06332, 2019 2019.