IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 3, 2019, accepted October 20, 2019, date of publication October 29, 2019, date of current version December 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2950141

Blurring-Effect-Free CNN Network of Structural Edge for Focus Stacking

GUIJIN WANG[1,2], (Senior Member, IEEE), WENTAO LI[1], XINGHAO CHEN[1], XUANWU YIN[3], XIAOWEI HU[1], CHENBO SHI[4], AND LONG MENG[4]

[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]Sichuan Energy Internet Research Institute, Tsinghua University, Beijing 100084, China
[3]Department of Kirin Chipset and Technology Development, Hisilicon, Beijing, China
[4]Shandong Mingjia Technology Company, Ltd.

Corresponding author: Guijin Wang (wangguijin@tsinghua.edu.cn)

ABSTRACT Focus stacking is a promising computational technique to extend depth of field by fusing images focused at different focal planes. However, existing focus stacking methods could not cope with the blurring-effect problem produced in structural edges, where depth values change abruptly. In this work, we firstly extract structural edges robustly by designing Des(depthmap-based extraction of structural edges)-ResNet. Then we propose a novel convolutional neural network (BEF-CNN) to restore blurring-effect-free image patches in order to enhance all-in-focus performance. To the best of our knowledge, it is the first work to utilize CNN to generate all-in-focus image directly instead of pixel-to-pixel correspondence with depthmap. Experimental results validate that our proposed algorithm has achieved best all-in-focus image while keeping the accuracy of depthmap.

INDEX TERMS All-in-focus, focal stack, Des-ResNet, BEF-CNN, structural edge.

I. INTRODUCTION

In general photography, optical lenses focus on a specific depth plane and leave other regions blurred by various scales. With the development of digital imaging technique, focus stacking, as an extended depth of field (EDOF) manner, has drawn more and more attentions of researchers [1]–[4]. It captures focal stack composed of a group of images focused at various depth planes and fuses them into an all-in-focus image.

Focus stacking could be divided into 2 different categories: transform-based method and depth-estimation-based method. In the first category, source RGB images are converted to certain feature domains, then the final all-in-focus image is reconstructed by the inverse transformation of the fused corresponding coefficients [5]–[11]. However, these methods are unstable and sensitive to fluctuation of transform coefficients. In depth-estimated-based methods [12], [13], [13], [14], the fusion is done in spatial domain. Researchers extract depth values of image edges by comparing various sharpness measurements. Then they propagate depth values from sparse edge positions to all pixels in the image and construct a dense depthmap. At last, all-in-focus image is fused by extracting pixel intensities from focal stack pixel-by-pixel correspondence to depthmap. So most researchers considered improving accuracy of depthmap to refine the all-in-focus image. Some researchers designed robust gradient measurements to extract depth values of edge points accurately and optimized propagation methods of final depthmap. Others treated the depthmap estimation as an image segmentation problem and proposed solutions based on deep learning (eg. CNN). However, all the methods above overlooked the blurring-effect problem produced in structural edges where depth values change abruptly. Even if depthmap is estimated accurately, the pixel-to-pixel correspondence between all-in-focus image and depthmap would reserve blurring effects on structural edges and degrade the performance of all-in-focus image.

In this paper, a novel depth-estimation-based method is proposed for focus stacking with two main contributions. Firstly, we design a residual network to extract structural edges robustly and call it Des-ResNet (depth-based extraction of structural edges). Here structural edge is defined as edge point at boundary of different depth planes, same as the definition in [15]. These structural edges are then utilized to propagate the accurate entire depthmap. Secondly, we propose

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Di Martino.

BEF-CNN to remove blurring-effects of all structural edges and refine all-in-focus image. To the best of our knowledge, it is the first work to utilize CNN to generate all-in-focus RGB images directly instead of pixel-to-pixel correspondence with depthmap. Extensive experiments show that our method has achieved higher all-in-focus performance than state-of-the-art methods.

A preliminary version of this paper is presented in [16]. This paper extends the preliminary work in several aspects: (1) A more extensive survey on related work is provided, including focus stacking and CNN for depth segmentation. (2) we design Des-ResNet in order to extract structural edges robustly. (3) We optimize the extraction of patch as input of BEF-CNN in order to guarantee sharpness of extracted patches as much as possible. (4) The visual performance of our method is compared with other state-of-the-art methods. Advantages of our modules for all-in-focus fusion are also discussed in detail.

The rest of the paper is organized as follows. Related work is introduced in detail in Section II. The blurring-effect problem is produced detailedly in Section III. Section IV depicts the approach of our work in detail, including Des-ResNet and BEF-CNN. In Section V, extensive experiments certify the effectiveness and robustness of our proposed method. Finally, we conclude our work in Section VI.

II. RELATED WORK

In this section we review the related work, which includes focus stacking and CNN for depth segmentation.

A. FOCUS STACKING

Focus stacking is the technique of fusing sharpest pixels into a single all-in-focus image and could be divided into 2 categories: transform- domain-based methods and depth-estimation methods.

In transform-domain-based methods, Forster *et al.* [5] proposed a complex wavelet method to extend DOF of microscopy images. Haghighat *et al.* [6] presented an approach for fusion of multi-focus images based on variance calculated in discrete cosine transform (DCT) domain. Sroubek, Redondo *et al.* [7], [8] fused the decomposed discrete wavelet transform(DWT) coefficients to get the all-in-focus image. Dense scale invariant feature transform (DSIFT) [9] was utilized for the activity level measurement to fuse multi-focus images. Kuthirummal *et al.* [10] presented Focal Sweep Imaging (FSI) to extend the DOF with 2D deconvolution, where the sensor moved along the optical axis during one exposure. Llavador *et al.* [11] extended the FSI to generate large depth-of-field integral microscopic images with liquid lens. These transform-domain-based methods are usually unstable, complicated and even sensitive to tiny perturbation of transform coefficients.

In depth-estimation-based methods, most researchers focused on improving accuracy of depthmap. Aguet *et al.* [12] estimated the all-in-focus image with a model based 2.5D deconvolution method. Suwajanakorn *et al.* [13] regarded the depthmap fusing problem as a multi-label MRF optimization problem on a regular 4-connect grid given a sharpness measurement, and defined the pairwise energy as total variation of gradients of neighboring pixels. M. Seitz. introduced the first depth from focus (DFF) method capable of computing depth and all-in-focus from mobile phones and other hand-held cameras [13]. Wang *et al.* [14] proposed directional-max-gradient flow and iterative-labeled Laplacian depth propagation method to extract true depth values for edge points to refine depthmap as well as all-in-focus image. However, all these methods fused the all-in-focus image by pixel-to-pixel correspondence with the depthmap and could not remove blurring effect problems occurred on structural edges.

B. CNN FOR DEPTH SEGMENTATION

Convolutional Neural Network(CNN) is a typical deep learning model, which attempts to learn a hierarchical representation of a single image with different levels of abstraction. It is generally used in multi-focus image fusion to estimate the decision map or defocus map. Liu *et al.* [17] proposed a deep CNN network to learn a direct mapping between source images and focus map, whose edges are calculated with pixel-wise weighted-average rule. Tang *et al.* [18] proposed a pixel-wise convolutional neural network (p-CNN) to recognize focused and defocused pixels in source images from neighbourhood information. It could be thought of as a learned focus measure(FM) and provided more efficiency than conventional handcrafted FMs. Du and Gao [19] achieved depthmap segmentation through a multi-scale CNN. They performed a multi-scale analysis on each input image to derive the respective feature maps on region boundaries between focused and defocused regions. However, in all these methods, although depthmap is aimed to be estimated accurately, structural edges where depth values change sharply would reserve image blurring effects and degrade the performance of all-in-focus image.

III. BLURRING-EFFECT PROBLEM

In this section, we formulate blurring-effect problem.

Structural edge is defined as boundary of two different depth planes in this paper. So based on the fact that nearer objects shelter farther objects, when the farther plane is focused, the object at the nearer depth plane is defocused and its blur kernel would propagate and interfere the sharp texture at the farther plane. This phenomenon is called as blurring-effect, which blurs farther plane's sharp texture near structural edges.

Fig. 1 explains the production of blurring-effect. Fig. 1(a) and Fig. 1(b) present pixel intensities of same structural edge from two different images of focal stack while Fig. 1(c) is the depth groundtruth. Fig. 1(a) and Fig. 1(b) are respectively focused on farther depth plane (blue) and nearer depth plane (red). We use black bounding box to highlight the region of farther plane near structural edge, where blurring-effect happens. When the nearer plane is focused, the sharp texture in the bounding box is blurred due to defocusing shown

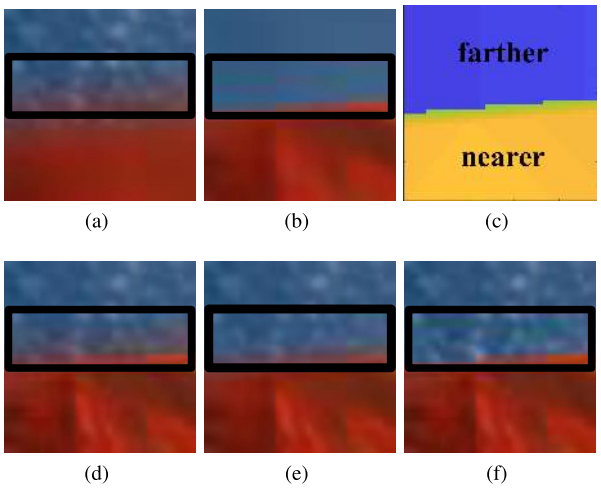


FIGURE 1. Blurring-effect is produced in farther plane near structural edge and is reserved by pixel-by-pixel all-in-focus fusion based on depthmap. Our method removes blurring-effect to improve all-in-focus performance (a) stack 1. (b) stack 2. (c)depthmap groundtruth (d) pixel-by-pixel fused patch. (e)blurring-effect-free patch generated with our method (f) all-in-focus groundtruth.

as Fig. 1(b). When the farther plane is focused, however, the blurred pixels from red nearer plane interfere the clear texture and make the region of the black box blurred purple. Therefore, structural edges' surrounding farther pixels have no sharp and clear texture in the entire focal stack.

In traditional depth-estimation-based methods, the all-in-focus image FI is reconstructed based on depthmap following the equation below.

$$FI(x, y) = I_{d(x,y)}(x, y), \tag{1}$$

where x and y are coordinates of pixel, $d(x, y)$ is depthmap of pixel (x, y) and $I_{d(x,y)}$ represents the $d(x, y)$-th image in the focal stack.

Since the all-in-focus image is reconstructed by pixel-to-pixel correspondence to depthmap, the blurred pixels near structural edges are all reserved and they degrade the performance of all-in-focus image. Fig. 1(d) shows the fused patch, where purple artifacts near the structural edge are all reserved. The blurring-effect-free patch generated by our method is displayed in Fig. 1(e), which removes purple noises in the black bounding box and gets better all-in-focus performance. Fig. 1(f) displays the all-in-focus groundtruth image, whose obtaining process would be described in Section V-A.

IV. APPROACH

As shown in Fig. 2, our hierarchical framework consists of two modules: depthmap estimation based on structural edges extracted by Des-ResNet, and all-in-focus image fusion by BEF-CNN.

Firstly, we utilize max-gradient flow (MGF) to extract depth values of source points in the focal stack and estimate initial dense depthmap d' with traditional Laplacian optimization. We propose Des-ResNet utilizing this depthmap to classify source points as structural edges and texture edges. Here structural edge is source point at boundary of different depth planes and texture edge is source point nearly on the same depth plane. Then the dense depthmap d is refined with labeled-Laplacian optimization and behaves sharp at structural edges while smooth at texture edges. Initialized all-in-focus image suffering from blurring effects is also generated by pixel-to-pixel correspondence with depthmap d. Secondly, we extract image patches around each structural edge and propose BEF-CNN to fuse blurring-effect-free patches, whose details are described further in Section IV-B. Finally, we utilize output of our BEF-CNN as replacement of image patches around corresponding regions from initialized all-in-focus image to remove blurring effects. In this way, our whole framework gets accurate depthmap and high-quality all-in-focus image at the same time.

A. DEPTHMAP GENERATION WITH RESNET

In this section, we introduce how to generate accurate depthmap and extract structural edges robustly with Des-ResNet.

1) MAX-GRADIENT FLOW

We utilize max-gradient flow referring to [20] to extract valid depth values of image edges. Firstly a max-gradient image MG whose pixels record maximum gradient values across the

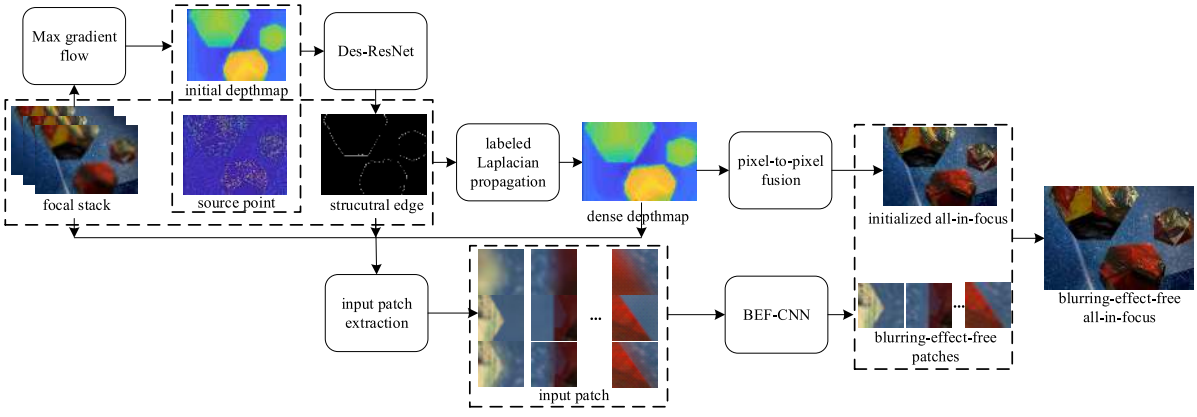


FIGURE 2. The framework of our work.

focal stack is constructed following Eq. (2).

$$MG(x, y) = \max_i G_i(x, y), \tag{2}$$

where $G_i(x, y)$ is the gradient of pixel $I_i(x, y)$ captured at depth position i. Then the max-gradient flow is constructed by calculating gradients of the max-gradient image along x-axis and y-axis following Eq. (3):

$$MGF(x, y) = \begin{bmatrix} \dfrac{MG(x + \Delta x, y) - MG(x, y)}{\Delta x} \\ \dfrac{MG(x, y + \Delta y) - MG(x, y)}{\Delta y} \end{bmatrix}, \tag{3}$$

where Δx and Δy denote the increment of pixel coordinate along x-axis and y-axis respectively. This flow models propagation of image edges' gradients and chooses edge points which satisfy Eq. (4) as source points whose depth values are valid. Then the depth values $\tilde{d}$ of source points are calculated as Eq. (5), where n is the image number focal stack contains.

$$\nabla \cdot MGF(x, y) > 0. \tag{4}$$

$$\tilde{d}(x, y) = \arg\max_{j \in [1,n]} G_j(x, y). \tag{5}$$

Readers are referred to our previous work [20] for more detailed definition and its satisfying performance of estimating depth values of edge points.

2) LABELED-LAPLACIAN DEPTHMAP GENERATION

After calculating depth values of source points, we describe how to propagate depth values to all pixels and construct dense depthmap d.

Like the method in our previous work [14], the depth propagation problem is formulated as minimizing the following cost energy

$$E(d) = d^T L d + \lambda (d - \hat{d})^T D (d - \hat{d}), \tag{6}$$

where D is diagonal matrix whose element $D(i, j) = 1$ if pixel i has valid depth value $\hat{d}(i, j) > 0$. The scalar λ controls the balance between smoothness of depth propagation and the fidelity of source points. d is the depthmap we want and L is the labeled-Laplacian matrix defined as follows:

$$L(i, j) = \sum_{k|(i,j)\in\omega_k} (\delta_{ij} - \frac{1}{|\omega_k|}(1 + (I_i - \chi(i, k))^T \times (\Sigma_k + \frac{\varepsilon}{|\omega_k|} U_3)^{-1} (I_j - \chi(j, k)))), \tag{7}$$

where

$$\chi(i, k) = (1 - \Pi_i) I_i + \Pi_i \mu_k. \tag{8}$$

Here δ_{ij} is the Kroecker delta, U_3 is identity matrix and ω_k is a small window covering pixels i and j. μ_k and Σ_k are mean vector and covariance matrix of guided image I in ω_k.

We classify source points into structural edges ($\Pi_i = 1$) and texture edges ($\Pi_i = 0$), then propagate the depth values of these two kinds of edges differently. Here the classification is realized by Des-ResNet, which would be explained in detail in following Section IV-A3. If pixel i belongs to structural edge, the depth boundary should be aligned with intensity edge, and the similarity $L(i, j)$ between i and its neighbouring pixel j is calculated from mean and covariance matrix of colors in ω_k in order to reserve sharp depth boundary at structural edges. If pixel i is texture edge, we have $\chi(i, k) = I_i$ to force pixels in the patch ω_k have same color. In this way, we guarantee accuracy of depthmap by smoothing depth values at texture edges and reserving depth sharpness at structural edges. Specially, if we regard all source points as structural edges, the depth propagation is strongly dependent of colored texture of guided image and Eq. (7) degenerate into traditional Laplacian optimization [21]. Even though it guarantees depth sharpness of structural edges, it also reserves depth variation noises at texture edges whose depth values should keep constant. So we should refine the depthmap by extracting structural edges, which directly influences accuracy of depthmap propagation.

3) DES-RESNET BASED STRUCTURAL EDGES EXTRACTION

Here we introduce how to extract structural edges from source points with Des-ResNet.

Based on the definition that structural edge is the boundary of two different depth planes while texture edges does not influence depth values, depthmap around edges could be seen as features to distinguish these two kinds of edges. Here we utilize depthmap d' propagated from source points with traditional Laplacian optimization in order to reserve sharpness of structural edges' depthmap and increase representativeness of our network.

We formulate $(2N+1) \times (2N+1)$ sized patch region p centered with source point i. We then extract initial depthmap d' of patch p as d'_p and subtract it with initial depth value of source point i to construct the $(2N + 1) \times (2N + 1)$ sized feature ξ as standardization following Eq. 9.

$$\xi(l) = d'_p(l) - d'(i), \tag{9}$$

where l is local pixel in the extracted patch region p. Since the extracted patch feature ξ depicts depth variation around source point, it should be utilized as inputs of the network.

The structure of our proposed Des-ResNet is shown in Fig. 3, Our network is composed of 1 convolutional layer, 2 resnet layers and 1 full-connected layer. The two residual layers includes 3 convolutional layers with 3×3 filters. Max-pooling is executed in order to decrease the number of parameters. At last, the network outputs probability of classifying as structural edges and texture edges. The patch with red bounding box is labeled as structural edge while the black patch is classified as texture edge. The experiment in Section V shows the result of our structural edge extraction and its advantage over other method in [14].

Finally, with the classification result of structural edges and texture edges, we utilize Eq. (7) to propagate the final accurate depthmap d. Initialized all-in-focus image FI' is also reconstructed by pixel-to-pixel correspondence with d following Eq. (1).

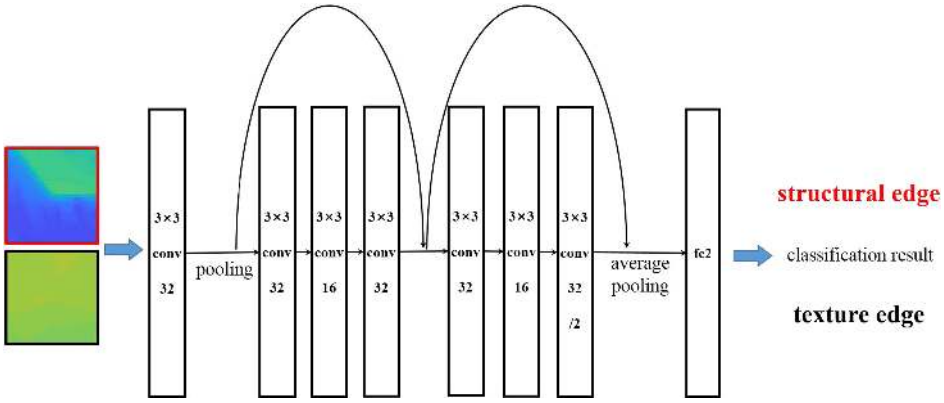


FIGURE 3. The framework of our Des-ResNet.

B. BEF-CNN

After structural edges are extracted robustly, in this section, we introduce structure of our BEF-CNN to remove blurring-effects produced on structural edges of initialized all-in-focus image in detail.

1) INPUT EXTRACTION

Based on the assumption that small patch around each structural edges contains only two different depth planes, we extract depthmap of patch region p around structural edge as d_p and classify it into 2 categories s_A and s_B with K-means (K=2). Here we use clustering centers A and B as respective depth values of these two planes.

Although pixel-to-pixel correspondence between depthmap and all-in-focus produces noises, information about sharp texture of dual planes should be concluded from images of the focal stack. So it is reasonable to inspect pixel intensities in patch region around each structural edge point when it is focused in respective depth planes. For structural edge point i, for example, we obtain $(2N+1)\times(2N+1)$ sized image patches p_A and p_B, which records intensity values of patch region p in focal stack images I_A and I_B.

These two patches depict sharp objects focused on two depth planes, so they choose all sharp texture when patch p is strictly dual-classified. However, when depth values of the patch is diverse and complicated, some pixels belonging to s_A might have large depth deviation from A and might produce many defocused blurs on patch p_A. Since inputs of our BEF-CNN aim to reserve sharp texture of both depth planes as much as possible, we extract image patches as network inputs following:

$$\begin{aligned} p'_{AB}(i) &= p_{d(i)}(i), \\ p'_A(i) &= p_{d(i)}(i)\delta(i \in s_A) + p_A(i)\delta(i \in s_B), \\ p'_B(i) &= p_{d(i)}(i)\delta(i \in s_B) + p_B(i)\delta(i \in s_A) \end{aligned} \tag{10}$$

where $d(i)$ is the depthmap value of pixel i. Here p'_A and p'_B record sharp pixels belonging to s_A and s_B respectively, and p'_{AB} is pixel-to-pixel fused based on depthmap.

We choose p'_A as an example to explain the process of input patch combination. When pixel i belongs to s_A, patch p'_A chooses sharp pixels from whole focal stack instead of only two images. In this way, pixels belonging to s_A are all focused in p'_A. On the other hand, when pixel i belongs to s_B, $p'_A(i)$ should be blurred. So we choose intensity value from p_A because tiny depth perturbation would not affect intensities of originally defocused pixels.

There are two advantages of our input patches: Firstly, these patches are fused according to depthmap, so they reflect edge shape and depth differences, the two main factors of blurring-effects in structural edges. Secondly, since p'_A and p'_B include sharpest pixels of s_A and s_B respectively, our network is also suitable for occasions where structural edges are not strictly dual-classified. Fig. 4 shows the performance of our patch extraction. Even though some pixels (orange in Fig. 4(a)) belonging to s_A have slight depth differences and behave blurred in p_A(Fig. 4(b)), they still keep sharp and clear in our input patch p'_A. Therefore, our fusion method guarantees sharpness in focused regions of extracted patches as much as possible.

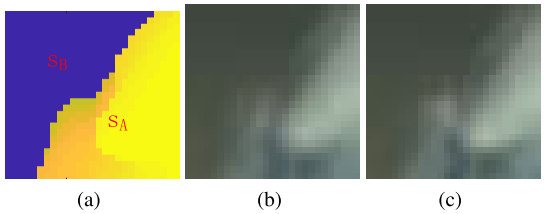


FIGURE 4. our patch combination process reserve more sharp texture in BEF-CNN's input patches (a) depthmap. (b) p_A. (c)p'_A.

2) NETWORK STRUCTURE DESIGN

Overall architecture of our proposed BEF-CNN is shown in Fig. 5. The proposed BEF-CNN has three convolutional layers, and the generation process of each convolutional layer H_i could be described as follows:

$$H_i = f(H_{i-1} \otimes W_i + b_i), \tag{11}$$

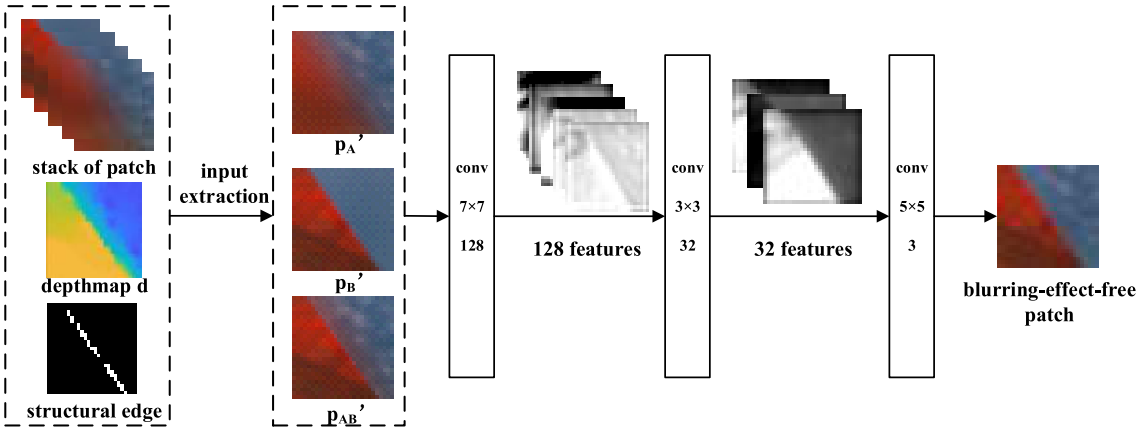


FIGURE 5. Network architecture of proposed BEF-CNN.

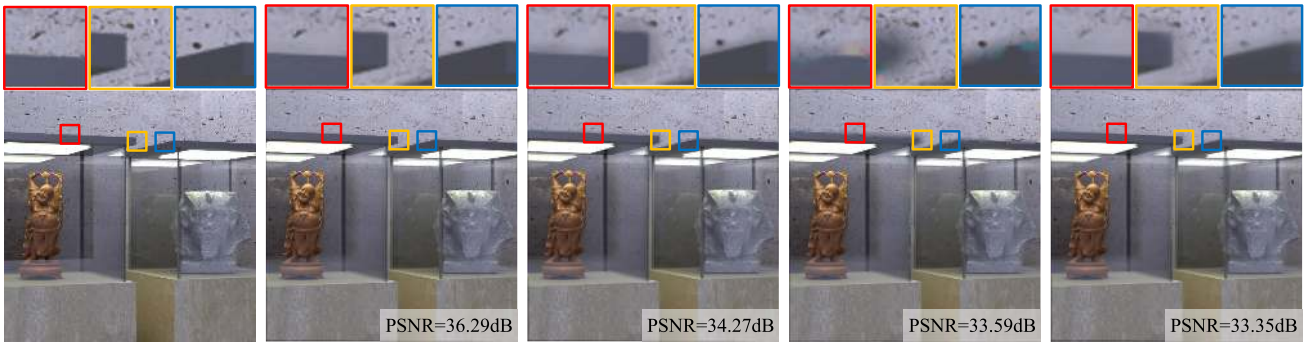


FIGURE 6. All-in-focus performance of 'museum' dataset. Our method removes blurring-effects at structural edges and reserves more sharp texture near the edge than other methods (a) groundtruth. (b) our method. (c) DMGF-based. (d) DWT-based. (e) DCNN.

where $\otimes$ denotes the convolution operation, W_i is the convolutional kernel and b_i is the bias. In this work, $f()$ is the non-linear ReLU activation function.

Firstly, we extract three RGB patches p_A, p_B and p_{AB} following Eq. (10) in the last section and concate them into a 9-channel patch. The first and second convolutional layers obtain 128 and 32 feature maps respectively by 7×7 and 3×3 sized kernel filter. The kernel sizes are set to cover the propagation regions of structural edges entirely. The last layer obtains the final blurring-effect-free RGB patches with the kernel size of 5×5. It's possible to add more convolutional layers to increase the non-linearity. But this would increase the complexity of the model and demand more training time.

Finally, we utilize outputs of BEF-CNN as replacement of image patches around corresponding regions from initialized all-in-focus image to remove blurring effects. Since we extract patch region centered with each structural edge, this replacement improves the quality of structural edges while maintaining the all-in-focus performance of texture edges. Further, the blurring-effect occurs in the center portion of extracted patch, and it barely influences the pixel intensities at patch's borders. In this way, the discontinuity in fusion process has little influence on our fusion method and could be neglected.

It should be noted that our method is based on the assumption of focal stack's densely-capturing. Section V-E would discuss it in detail.

V. EXPERIMENT

In this section, we present the experimental performance of our proposed method. In Section V-A, we introduce the datasets used in our experiments. We explain our training process in detail in Section V-B. Section V-C evaluates the overall performance of our proposed method with state-of-the-art methods on synthesized data. In section V-D, we analyze the advantages of two proposed modules: Des-ResNet and BEF-CNN in detail respectively. In the last Section V-E, we discuss the limitation of our method and expect the future research work.

A. SETUP

In this work, we use synthesized focal stack to evaluate our performance. This focal stack is constructed by light field data taken from Training set of 4D Light Field Benchmark [22]. The dataset has light field data sampled with 9×9 angular resolution and depthmap groundtruth of 20 different scenes. Following the discrete projection relationship between focal stack imaging and sub-aperture

image of 4D light field based on discrete refocusing equation formulated in [23], each focal stack dataset is constructed by light field dataset and is composed of 49 512 × 512 sized images while central view of light field data in each scene is utilized as corresponding all-in-focus groundtruth. Since black boundaries around images are produced in the focal stack construction process, we exclude the black boundaries (5 to 10 pixel-widths) during all-in-focus performance measurement.

B. TRAINING

There are two different networks in our work, so we introduce their training process respectively.

For Des-ResNet, We firstly extract source edge points with max-gradient flow. Then we do edge detection of groundtruth depthmap with Sobel operator. Source points with large Sobel gradients are labeled as structural points. The other source points are labeled as texture edges. Also, the input patches of Des-ResNet is extracted from groundtruth depthmap.

For BEF-CNN, we mainly extract RGB values of structural edges' surrounding patches as network inputs following the Eq. (1) and Eq. (10) with groundtruth depthmap. Meanwhile, the network output is composed of image patches extracted from groundtruth all-in-focus image.

Since extracted patch region should cover depth variation in Des-ResNet and propagation of blurring effect at structural edges in BEF-CNN, we choose $N = 13$ and make the patch size 27 × 27 in this paper. There are lots of overlapping regions between adjacent patches, so we extract image patches every 5 structural edge points for training to decrease the repetition of training set. To avoid the repetition between training set and testing set, we apply n-fold ($n = 20$) crossing validation by choosing structural patches from one group of focal stack for testing while others for training.

C. OVERALL PERFORMANCE

In this section, we compare our method(BEF-CNN-v2) with DWT-based method [8], DCNN-based method [17], DMGF-based method [14] and our previous BEF-CNN-based method [16] (BEF-CNN-v1) as baseline work. The DCNN-based method is designed mainly for two-images-fusion problem. To deal with focus stacking which contains more than two multi-focus images, we fuse them one by one in series following the author's advice.

Table 1 shows the peak signal-to-noise ratio (PSNR) values of reconstructed all-in-focus image with different methods. It also shows the area ratio that extracted patches of structural edges occupy in an entire image as *ratio*(%) displayed in Table 1. Since blurring-effect occurs at structural edges, the area ratio determines the extent of removed blurring-effect and influences the final all-in-focus performance. Generally, the more accurate the structural edges are extracted, the more blurring-effects are removed and the better all-in-focus images are obtained.

From Table 1, our two versions of BEF-CNN method get higher all-in-focus performance than other methods since BEF-CNN is effective to remove blurring-effects occurred at structural edges. Although patches of structural edges cover 16% and 23.9% entire area of a single image respectively, our two BEF-CNN methods improve averagely 0.72dB and 0.95dB PSNR than other state-of-the-art methods. Compared with our BEF-CNN-v1 method, the updated BEF-CNN-v2 designs Des-ResNet to improve extraction of structural edges and it finds 7.9% more are of true structural edges' patches and gets 0.23dB higher PSNR of all-in-focus image.

Table 2 shows the Feature Similarity(FSIM) values of reconstructed all-in-focus image with different methods. Although the FSIM scores are similar, our method still behaves better all-in-focus performance than state-of-the-art methods.

TABLE 1. PSNR (dB) of different methods on synthesized data.

	BEF-CNN-v2 (*ratio*(%))	BEF-CNN-v1 (*ratio*(%))	DMGF	DWT	DCNN
antinous	43.82(*12.8*)	43.85(*10.6*)	43.30	**45.14**	42.58
boardgame	**41.07**(*5.1*)	40.80(*0*)	40.80	40.68	39.80
boxes	**35.60**(*34.6*)	35.27(*16*)	35.30	33.67	33.78
cottons	47.76(*6.5*)	47.82(*5.8*)	47.75	47.96	**48.94**
dino	41.87(*24.9*)	**41.90**(*16.2*)	41.20	39.44	39.69
dishes	**33.44**(*21.5*)	32.88(*0*)	32.88	31.38	32.47
greek	**40.52**(*24.7*)	39.31(*9.8*)	37.77	37.29	37.95
kitchen	**37.06**(*31.1*)	36.72(*28.9*)	35.04	33.13	34.61
medieval2	**37.88**(*17.1*)	37.79(*8.4*)	37.20	33.99	37.01
museum	**36.29**(*35.9*)	35.28(*20.4*)	34.27	33.59	33.35
pens	**39.15**(*31.0*)	38.75(*23.6*)	37.35	35.08	36.73
pillows	37.46(*10.7*)	**37.55**(*10.2*)	37.23	35.87	35.44
platonic	38.49(*23.1*)	38.41(*25.7*)	36.40	38.26	**39.04**
rosemary	**35.60**(*35.3*)	35.39(*24.4*)	35.00	32.01	33.10
sideboard	32.01(*45.2*)	**32.09**(*34.6*)	30.73	29.22	30.58
table	**36.60**(*39.8*)	36.22(*13.5*)	35.46	35.19	35.23
tomb	40.78(*10.8*)	40.82(*12.7*)	40.74	**40.91**	40.51
tower	**36.66**(*30.0*)	36.51(*21.1*)	35.79	35.92	35.88
town	**38.58**(*28.7*)	38.47(*24.4*)	36.91	36.84	36.01
vinyl	42.05(*19.7*)	**42.20**(*14.3*)	41.08	40.99	40.37
avg	**38.63**(*23.9*)	38.40(*16.0*)	37.68	37.15	37.08

TABLE 2. FSIM of different methods on synthesized data.

	BEF-CNN-v2	BEF-CNN-v1	DMGF	DWT	DCNN
antinous	0.9829	0.9841	0.9838	0.9908	**0.9966**
boardgame	**0.9980**	0.9976	0.9976	0.9947	0.9977
boxes	**0.9861**	0.9845	0.9825	0.9767	0.9755
cottons	0.9982	0.9982	0.9982	**0.9989**	0.9985
dino	0.9972	**0.9976**	0.9970	0.9917	0.9951
dishes	**0.9876**	0.9850	0.985	0.9744	0.9785
greek	**0.9937**	0.9916	0.9886	0.9868	0.9889
kitchen	**0.9937**	0.9932	0.9927	0.9872	0.9889
medieval2	**0.9932**	0.9931	0.9930	0.9907	0.9914
museum	**0.9817**	0.9780	0.9747	0.9571	0.9657
pens	**0.9929**	0.9922	0.9893	0.9792	0.9671
pillows	**0.9966**	0.9966	0.9963	0.9891	0.9918
platonic	0.9868	0.9872	**0.9879**	0.9874	0.9793
rosemary	**0.9880**	0.9878	0.9874	0.9771	0.9707
sideboard	0.9858	**0.987**	0.9792	0.9787	0.9749
table	**0.9923**	0.9920	0.9908	0.9873	0.9888
tomb	0.9951	0.9952	0.9951	0.9939	**0.9954**
tower	**0.9868**	0.9868	0.9830	0.9814	0.9826
town	**0.9965**	0.9964	0.9943	0.9903	0.9936
vinyl	**0.9976**	0.9976	0.9973	0.9930	0.9965
average	**0.9915**	0.9911	0.9897	0.9853	0.9859

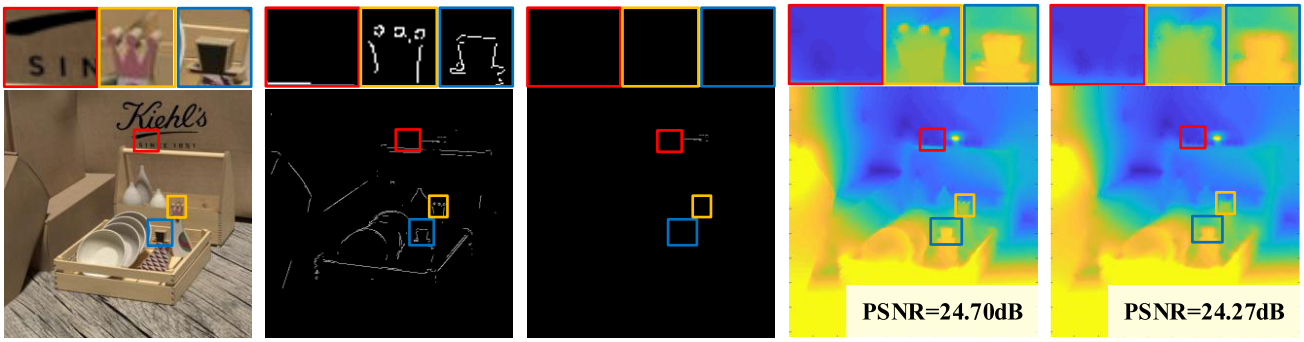


FIGURE 7. Depthmap comparison with other methods. Our Des-ResNet extracts more structural edges and better depthmap (a) scene (b) our structural edge (c) DMGF-based's structural edge (d) our depthmap (e) DMGF-based's depthmap.

Fig. 6 shows the visual performances of all-in-focus images of dataset 'museum' and its three different patches at structural edges. Our method has the best visual performance since it reserves most sharp textures and removes blurring effects at structural edges. For DWT-based method, some objects are mistakenly blurred and some false colors are produced in Fig. 6(d) due to its unstability. DCNN-based method leaves out sharp texture near structural edges because its depthmap generated from CNN network has unclear boundaries between different depth planes. Method of DMGF reserves sharp textures, but pixels near structural edges are noised by propagated blurred edges due to blurring-effect problems.

D. MODULE ANALYSIS

In this module, we mainly analyze the advantage of our main two modules. Firstly, we show how proposed Des-ResNet extract structural edges and compare depthmap with other state-of-the-art depth-estimation method. Secondly, we discuss how BEF-CNN removes blurring effects and refine all-in-focus image. Finally, we do self-comparison to evaluate effects of modules including Des-ResNet and Input Extraction for BEF-CNN on final all-in-focus performance.

1) DEPTHMAP COMPARISON

In this section, we compare our depthmap generated by Des-ResNet with that of [14] and our previous conference paper [16], whose structural edges are extracted with K-means.

The comparisons of depthmap as well as structural edge extraction on dataset 'dishes' are shown in Fig. 7. Fig. 7(b) and Fig. 7(c) show structural edges extracted from the same input of source points with our proposed Des-ResNet and K-means method respectively. K-means method iteratively updates clustering centers based on depth-edge feature operator for each source point and the classification result is influenced by depth distribution in one single image. For example, 'dishes' has some structural edges with large depth discontinuity in estimated depthmap, other structural edges with slight depth discontinuity would not be classified correctly. That is why K-means shown in Fig. 7(c) misses many true structural edges. The learning-based method in this paper, whose result is shown in 7(b), is more robust to intensity of depth discontinuity at structural edges and finds more true structural edges. So the depthmap generated in this paper shown in Fig. 7(d) is also better than K-means-based method shown in Fig. 1 because the depthmap keeps sharpness at structural edges while behaving smoothness at texture edges at the same time.

2) BLURRING-EFFECT-FREE WITH BEF-CNN

In this section, we show the advantage of our BEF-CNN to remove blurring-effects of structural edges.

In Fig. 8, we compare our method with the pixel-to-pixel all-in-focus fusion based on groundtruth depthmap. We also show the comparisons of visual performance of dataset 'platonic' in Fig. 9. Even though the groundtruth depthmap has accurate depth values, the sharp depth boundaries cause serious blurring effects at structural edges, which is shown in Fig. 9(c). Our BEF-CNN produces blurring-effect-free patches and has higher fusion accuracy in most

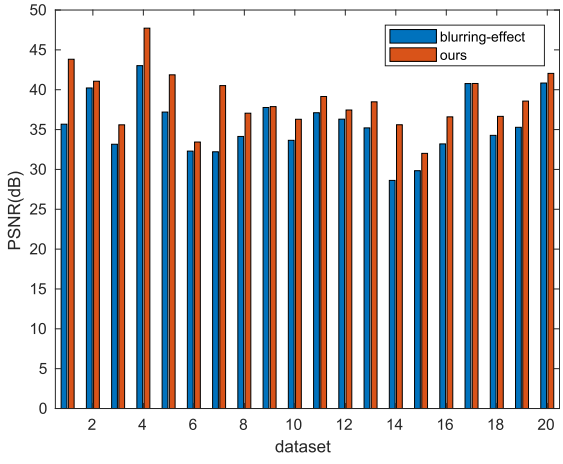
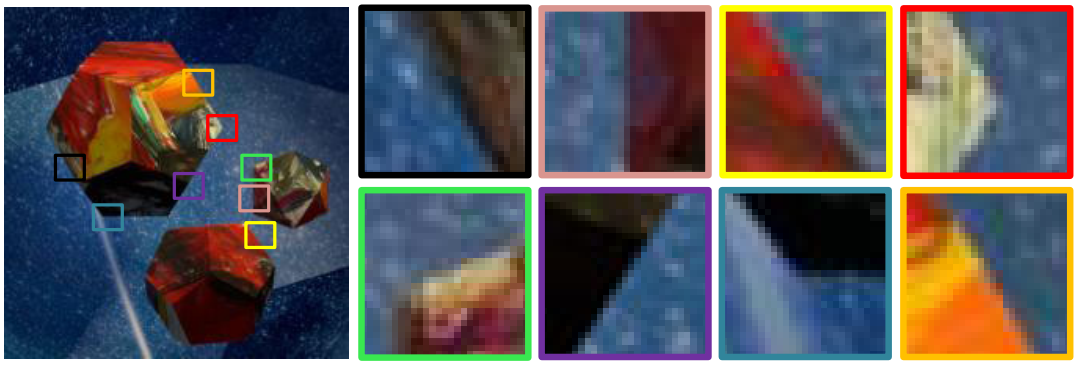
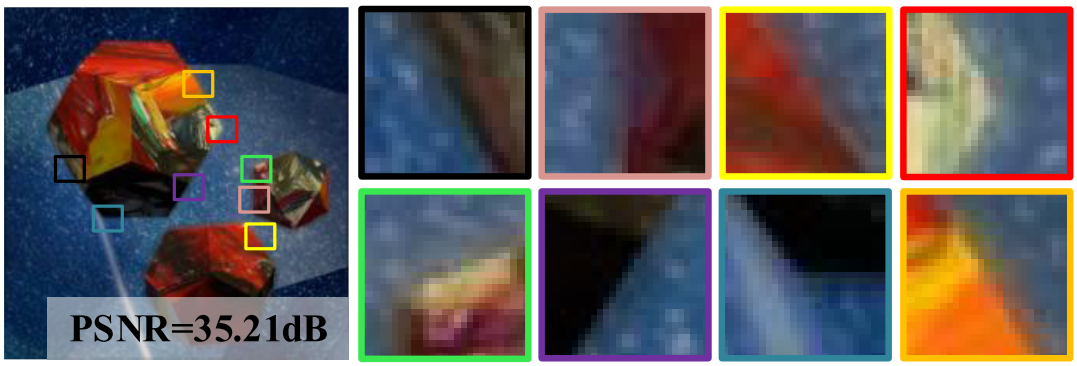


FIGURE 8. Advantage of our BEF-CNN: all-in-focus performance compared with pixel-by-pixel fusion based on depthmap groundtruth.



(a)

(b)



(c)

FIGURE 9. All-in-focus image comparison with pixel-to-pixel fusion with groundtruth depthmap. Our method removes blurring-effects produced near structural edges (a) groundtruth (b) our blurring-effect-free image (c) pixel-to-pixel fusion with depthmap groundtruth.

of the datasets. Therefore, our BEF-CNN validly remove blurring effects and improve all-in-focus performance.

3) SELF-COMPARISON

In this section, we evaluate the effects of modules including Des-ResNet and Input Extraction for BEF-CNN in Section IV-B1 on the final all-in-focus performance. We show averaged PSNR as well as area ratio of structural edges with different modules in Table 3. From this table, Des-ResNet extracts more structural edges and improves the averaged PSNR by nearly 0.2dB. It proves our proposed Des-ResNet's validity of structural edges extraction and all-in-focus fusion. Input Extraction of Section IV-B1 is designed to reserve sharp texture in BEF-CNN's input patches as many as possible. Although it does not affect structural edge extraction, it also

TABLE 3. Area ratio and PSNR (dB) of methods with different modules.

methods	ratio	PSNR(dB)
BEF-CNN	16%	38.40
BEF-CNN+Des-ResNet	23.9%	38.59
BEF-CNN+Des-ResNet+Input Extraction	23.9%	38.63

improves the all-in-focus performance and raises the PSNR up to 38.63dB. This proves that Input Extraction indeed refines the performance of BEF-CNN.

E. DISCUSSION

Our method is based on the assumption that the focal stack is densely-captured and has plenty of images. The density is optimal if union depth of fields of all images cover the depth range of scene's objects. When the scene's depth range is F and FOV of each image is F_0, the minimum optimal number of images in focal stack is $\frac{F}{F_0}$. If focal stack does not contain enough images, some objects are not focused in all images and all-in-focus performance would degrade. Therefore, in the future, we would focus on how to combine all-in-focus fusion with focal stack capturing to maintain all-in-focus performance when focal stack does not have enough images.

VI. CONCLUSION

In this work, we propose a novel all-in-focus fusion method based on Des-ResNet and BEF-CNN. Firstly, we utilize a two-layer Des-ResNet to extract structural edges and estimate accurate depthmap. Secondly, we propose BEF-CNN to remove blurring-effects on structural edges and improve all-in-focus image fusion accuracy. Experimental presents that our method behaves much better than state-of-the-art methods.

REFERENCES

[1] E. W. Allen and S. Triantaphillidou, *The Manual of Photography and Digital Imaging*. Waltham, MA, USA: Focal Press, 2012.

[2] N. T. Goldsmith, "Deep focus; a digital image processing technique to produce improved focal depth in light microscopy," *Image Anal. Stereol.*, vol. 19, no. 3, pp. 163–167, 2011.

[3] I.-H. Lee, M. T. Mahmood, and T.-S. Choi, "Robust depth estimation and image fusion based on optimal area selection," *Sensors*, vol. 13, no. 9, pp. 11636–11652, 2013.

[4] H. Liu, H. Li, J. Luo, S. Xie, and Y. Sun, "Construction of all-in-focus images assisted by depth sensing," *Sensors*, vol. 19, no. 6, p. 1409, 2019.

[5] B. Forster, D. Van De Ville, J. Berent, D. Sage, and M. Unser, "Complex wavelets for extended depth-of-field: A new method for the fusion of multichannel microscopy images," *Microsc. Res. Techn.*, vol. 65, nos. 1–2, pp. 33–42, 2004.

[6] M. B. A. Haghighat, A. Aghagolzadeh, and H. Seyedarabi, "Real-time fusion of multi-focus images for visual sensor networks," in *Proc. 6th Iranian Conf. Mach. Vis. Image Process. (MVIP)*, Oct. 2010, pp. 1–6.

[7] R. Redondo, F. Sroubek, S. Fischer, and G. Cristóbal, "Multifocus fusion with multisize windows," *Proc. SPIE*, vol. 5909, Sep. 2005, Art. no. 59091B.

[8] R. Redondo, F. Šroubek, S. Fischer, and G. Cristóbal, "Multifocus image fusion using the log-Gabor transform and a multisize windows technique," *Inf. Fusion*, vol. 10, no. 2, pp. 163–171, 2009.

[9] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense SIFT," *Inf. Fusion*, vol. 23, pp. 139–155, May 2015.

[10] S. Kuthirummal, H. Nagahara, C. Zhou, and S. K. Nayar, "Flexible depth of field photography," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 58–71, Jan. 2011.

[11] A. Llavador, G. Scrofani, G. Saavedra, and M. Martinez-Corral, "Large depth-of-field integral microscopy by use of a liquid lens," *Sensors*, vol. 18, no. 10, p. 3383, 2018.

[12] F. Aguet, D. Van De Ville, and M. Unser, "Model-based 2.5-D deconvolution for extended depth of field in brightfield microscopy," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1144–1153, Jul. 2008.

[13] S. Suwajanakorn, C. Hernandez, and S. M. Seitz, "Depth from focus with your mobile phone," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3497–3506.

[14] G. Wang, W. Li, X. Yin, and H. Yang, "All-in-focus with directional-max-gradient flow and labeled iterative depth propagation," *Pattern Recognit.*, vol. 77, pp. 173–187, May 2018.

[15] S. Liu, F. Zhou, and Q. Liao, "Defocus map estimation from a single image based on two-parameter defocus model," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5943–5956, Dec. 2016.

[16] W. Li, G. Wang, X. Chen, X. Yin, and X. Hu, "Blurring-effect-free CNN for optimization of structural edges in focus stacking," in *Proc. IEEE Int. Conf. Image Process.*, Taipei, Taiwan, Sep. 2019, pp. 4634–4638.

[17] Y. Liu, X. Chen, H. Peng, and Z. F. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36 pp. 191–207, Jul. 2017.

[18] H. Tang, B. Xiao, W. Li, and G. Wang, "Pixel convolutional neural network for multi-focus image fusion," *Inf. Sci.*, vol. 433, pp. 125–141, Apr. 2018.

[19] C. Du and S. Gao, "Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network," *IEEE Access*, vol. 5, pp. 15750–15761, 2017.

[20] X. Yin, G. Wang, W. Li, and Q. Liao, "Large aperture focus stacking with max-gradient flow by anchored rolling filtering," *Appl. Opt.*, vol. 55, no. 20, pp. 5304–5309, 2016.

[21] S. Zhuo and T. Sim, "Defocus map estimation from a single image," *Pattern Recognit.*, vol. 44, no. 9, pp. 1852–1858, 2011.

[22] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 19–34.

[23] X. Yin, G. Wang, W. Li, and Q. Liao, "Iteratively reconstructing 4D light fields from focal stacks," *Appl. Opt.*, vol. 55, no. 30, pp. 8457–8463, 2016.

GUIJIN WANG received the B.S. and Ph.D. degrees (Hons.) from the Department of Electronics Engineering, Tsinghua University, China, in 1998 and 2003, respectively, all in signal and information processing. From 2003 to 2006, he was with Sony Information Technologies Laboratories as a Researcher. Since 2006, he has been with the Department of Electronics Engineering, Tsinghua University, China, as an Associate Professor. He published over 100 International journals and conference papers, holds tens of patents with numerous pending. His current research interests include computational imaging, pose recognition, intelligent human–machine UI, intelligent surveillance, industry inspection, and AI for big medical data. He was the TPC member of ICIP2017 and the Track Chair of ChinaSIP 2015. He received the reward (the first prize) of Science and Technology Award from the Chinese Association for Artificial Intelligence, in 2014, and the reward (the first prize) of Electronic Society Science and Technology, in 2018. He was an Associate Editor of the *IEEE Signal Processing Magazine* and the Guest Editor of *NeuroComputing*.

WENTAO LI received the B.S. degree from Tsinghua University, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree in electronic engineering. From 2018 to 2019, he was a Visiting Ph.D. Student with Northwestern University, USA. His current research interests include depth sensing technique, 3D imaging, and computational imaging.

XINGHAO CHEN received the B.S. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2013, where he is currently pursuing the Ph.D. degree. From 2016 to 2017, he was a Visiting Ph.D. Student with Imperial College London, U.K. His current research interests include deep learning, hand pose estimation, and gesture recognition.

XUANWU YIN received the B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2011 and 2017, respectively. Since 2017, he has been with the Department of Kirin Chipset and Technology Development, Hisilicon, focusing on image signal processing algorithms. His current research interests include passive active depth sensing technique, 3D reconstruction, and computational imaging.

XIAOWEI HU received the B.S. degree (Hons.) from the University of Electronic Science and Technology of China, Chengdu, China, in 2015. He is currently pursuing the Ph.D. degree in electronic engineering with Tsinghua University, Beijing, China. His current research interests include passive/active depth sensing technique, 3-D reconstruction, and computational imaging.

CHENBO SHI received the B.S. and Ph.D. degrees from the Department of Electronics Engineering, Tsinghua University, China, in 2005 and 2012, respectively. From 2012 to 2016, he was a Postdoctoral Researcher with the EE Department, Tsinghua University. Since 2016, he has been the CTO and the CEO of Shandong Mingjia Technology Company, Ltd., which is focused on applications of industrial computer vision. He published over 20 international journal and conference papers, and holds tens of patents. His current research interests include image stitching, stereo matching, matting, and object detection and tracking.

LONG MENG received the B.S. degree from the Beijing University of Posts and Telecommunications, and the M.S. degree in signal and information processing from the Department of Electronics Engineering, Tsinghua University, China, in 2004. From 2004 to 2006, she was with Sony Japan, as an Algorithm Engineer for digital camera. From 2007 to 2011, she was with the Sony China Research Laboratory, as a Researcher. In 2011, she joined a start-up company, Shandong Mingjia Technology Company, Ltd., which is focused on industry inspection. She received the reward (the first prize) of Science and Technology Award from the Chinese Association for Artificial Intelligence, in 2014.