IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 21, 2020, accepted May 29, 2020, date of publication June 3, 2020, date of current version June 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2999694

Detecting Small Scale Pedestrians and Anthropomorphic Negative Samples Based on Light-Field Imaging

YUFENG ZHAO[1], FAN SHI[1,2], MENG ZHAO[1], WENZHE ZHANG[1], AND SHENGYONG CHEN[1], (Senior Member, IEEE)

[1]Key Laboratory of Computer Vision and System, Ministry of Education, Tianjin University of Technology, Tianjin 300384, China
[2]MOE Key Laboratory of Weak-Light Nonlinear Photonics, Nankai University, Tianjin 300071, China

Corresponding authors: Fan Shi (shifan@email.tjut.edu.cn) and Meng Zhao (zh_m@tju.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, and in part by the National Natural Science Foundation under Grant 61906133, Grant 61703304, Grant 61906134, and Grant U1509207.

ABSTRACT Great progress has been made in the field of pedestrian detection, but the following two problems have not yet been well addressed. One problem is the missed detection of small scale pedestrain as false negative failure, and the other one is confusion with anthropomorphic negative samples like vertical structures as false positive failure. In this paper, to tackle the above two problems, we use the light-field camera to capture pedestrian images for the following reasons: (i) the light-field camera can obtain multi-depth refocused images in a single exposure using one sensor, (ii) compared with 2D images, these refocused images can provide different key representations for different parts of the image. We further establish a light-field pedestrian dataset with 1766 images for pedestrian detection. A multi-focus detection network proposed in this work consists of multiple-branch detection models and takes multiple refocused images as inputs. In order to select the appropriate candidate proposal bounding box as final detection results, we design a cumulative probability selection (CPS) layer to combine each refocused image branch and accumulate the probability of each candidate neighboring proposal. Experimental results demonstrate that the proposed method outperforms state-of-the-art methods on our light-field pedestrian dataset.

INDEX TERMS CNN, light-field imaging, pedestrian detection.

I. INTRODUCTION

Pedestrian detection is one of the key problems in the field of computer vision and multimedia, which has a number of applications, such as video surveillance, urban autonomous driving, and robotics. In recent years, a lot of work has been devoted to this study [1]–[10]. However, there are still some common problems existing in pedestrian detection task. Two of the main problems are as follows. One is the small scale pedestrian detection. Since small scale pedestrians are often far away from the camera lens, it is easy to blur and small scale pedestrians are easily detected as negative samples for the lack of strong feature representations. The other is confusion with anthropomorphic negative samples. These anthropomorphic negative samples are pedestrian-like objects such as vertical structures, dustbins, traffic lights, and trees [11]. Because the shape of these objects is very similar to that of pedestrians in some cases, they are usually incorrectly detected as positive samples.

At the same time, as the 2D images contain limited information in pedestrian detection tasks, there are many methods proposed utilizing RGB-D or multi-spectral sensors to detect pedestrians in various situations. Choi *et al.* [12] used Kinect to detect people in a living room combining an ensemble of detectors in a unified framework. Hsieh *et al.* [13] proposed a people counting system with Kinect achieving almost 100% bi-directional counting and real-time detecting. Chen *et al.* [14] detected people in crowded scenes by fusing RGB and depth images from Kinect. Premebida and Nunes [15] used the context-based multisensor for pedestrian detection in urban environment. Zhang and Tao [16] proposed a pedestrian codetection framework for detecting pedestrians in binocular stereo sequence. Park *et al.* [17] proposed a novel sensor fusion CNN framework for detecting pedestrians and

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaojie Ju.

improved the detection performance in the night. However, these methods are energy-consuming and inconvenient. For instance, the Kinect camera needs mobile power and bracket when working outdoors and the binocular vision systems need accurate calibration to work well.

Considering the unique imaging process of light-field camera [18], it records both light intensity and the direction of each ray simultaneously. Thus, it can get multi-depth refocused images in a single photographic exposure using one sensor. These refocused images provide different key representations for different parts of the image. Intuitively, the in-focus part transmits more information than the out-focus part. This inspires us to design reasonable strategies to enhance the performance of pedestrian detection according to the different emphases of the information expressed by different refocused images. However, the original data captured by the light-field camera is in 4D format which is quite different from 2D and stereo information. Until now, in the field of object detection, there is no ready-made solution that can easily read the original data and make good use of refocusing characteristics. To solve the above difficulties, we design a new method to read the original data and make full use of refocusing characteristics of light-field imaging. Reasons mentioned above lead to the motivation of our paper.

Therefore, in this paper, a multi-focus detection network (MFDN) which uses light-field refocused images as input is designed to boost pedestrian detection performance. For established light-field pedestrian dataset, we design two methods to obtain refocused images. These two methods are Lytoro software method and digital refocusing method, and the effects of two different methods on the results are also illustrated. In order to select the appropriate bounding boxes that can be retained from the results of multiple refocusing branches, a cumulative probability selection (CPS) layer is introduced to accumulate probabilities of different refocusing branches. Our method can effectively suppress the hard negatives (e.g., vertical structures) and improve the efficiency of the small scale pedestrian detection to some extent. Comparisons with state-of-the-art methods are also illustrated in Section IV. **Our key contributions are as follows**:

1) Dataset construction. Although there is a similar dataset of light-field pedestrian [19], it lacks the annotation information of pedestrian bounding box. We first establish a light-field pedestrian dataset with 1766 images with a complete annotation box. This dataset is obtained in complex scenes, which is totally different from [19] and will be public for researchers.
2) We devise two methods for extracting the multiple refocused images from the raw light-field images and analyze their advantages and disadvantages. These two methods are general and flexible enough to be applicable to any scenarios requiring speed or accuracy. Moreover, we also compare the effects of two refocusing methods on the detection results.
3) We propose a multi-focus pedestrian detection network with multiple refocused images as input to tackle the problems of small scale pedestrian detection and confusion with anthropomorphic negative samples, while introducing a cumulative probability selection (CPS) layer to combine the results of multiple detection branches. Compared with the baseline method [20], our method markedly reduces the log average missrate by 15.39% and achieves an overall missrate of 30.09% on our proposed light-field dataset.

The remainder of this paper is organized as follows. Section II elaborates related work on pedestrian detection and light-field imaging. We present the proposed method in Section III. Experimental results and discussion are introduced in Section IV, and we conclude the paper in Section V.

II. RELATED WORK

A. LIGHT FIELD IMAGING

Adelson and Wang introduced the first light-field imaging technology in [21]. They built a plenoptic camera with an array of lenses. Ng *et al.* [22] invented the first hand-held device and introduced digital refocusing. Then, Ng founded Lytro company and successively launched two commercial hand-held light-field camera. Moreover, Veeraraghavan *et al.* [23] enriched the framework for obtaining the 4D information from light-field camera and widened the application scope of the light-field camera.

Recent years, many new light-field camera applications have been put forward. Shi *et al.* [24] introduced a novel technique that could simultaneously measure 3D model geometry and 3D surface pressure with a single light-field camera. Zhao *et al.* [25] believed that light-field camera could be applied to depth extraction of low-texture region. Fan and Yang [26] estimated the depth of real-world scenes containing an object semi-submerged in water using a light-field camera. The depth estimation of the light-field imaging had a certain develop [27], [28] and it made many applications possible, e.g., 3D scene reconstruction [29], saliency detection [30], and image super-resolution reconstruction [31]. The light-field imaging could be also used in biometric system [32] to improve face/iris recognition by its abundant information.

B. LIGHT FIELD DATASETS

The existing light-field datasets are based on the specific research purposes. For example, Raghavendra *et al.* [32] constructed two relatively large datasets to improve the performance of face and iris biometric recognition system. Li *et al.* [33] collected 100 scene images for saliency detection. Wang *et al.* [34] built a dataset which contained 1200 light-field images for material recognition. Jia *et al.* [19] established a dataset which contained about 1000 light-field pedestrian images. But the pedestrian dataset lacked bounding box (ground truth) and the scenes in the dataset was relatively simple. Therefore, it is necessary to build a dataset of light-field pedestrians with complete annotations in complex scenes.

C. PEDESTRIAN DETECTION

In recent years, there have been a lot of researches in the field of pedestrian detection, and a series of papers have been published. But two vital problems still exist in pedestrian detection task. The first is small scale of pedestrian detection. Song *et al.* [35] proposed a method which combined somatic topological line localization (TLL) and temporal feature aggregation for detecting small scale pedestrians. SAF R-CNN [36] utilized the divide-and-conquer philosophy to develop a Scale-Aware Fast R-CNN framework which contained multiple built-in subnetworks detecting various scale pedestrians. The second is confusion with hard negative samples. Zhang *et al.* [37] used boosted forests classifiers to effectively mine hard negatives based on the region proposal network (RPN) and high-resolution feature map. Wang *et al.* [38] proposed a novel loss function which improved the performance of the detector under occlusion cases. Liu *et al.* [39] proposed a asymptotic localization fitting module based on the single-stage detectors to improve the performance of pedestrian detection by continuously improving the accuracy of bounding box.

Overall, the above methods are more or less based on Fast/Faster R-CNN [20], [40] architecture with common 2D images as input. On the contrary, our proposed method relies on multi-focus image sequence as input, which can apply more information for detectors to suppress the hard negative samples and improve small scale pedestrian detection to some extent.

III. PROPOSED METHOD

In this section, we explain the principle of light-field imaging, and describe how to construct the pedestrian light-field dataset and how to label data. Furthermore, we describe the overall process of our framework in detail.

A. LIGHT FIELD IMAGING

A light-field camera consists of main lens, microlens array, and photosensor. Microlens array is a two-dimensional array composed of multiple microlens units, as indicated by Fig. 1(a). The pupil plane (uv plane) of the main lens and the photosensitive plane of the image sensor are conjugated with respect to the microlens array (st plane). That is to say, the light passing through each microlens unit will be projected onto the image sensor to form a small microlens image. Each microlens image contains several pixels. At the same time, the light intensity coming from the narrow beam is recorded by each pixel. As shown in Fig. 1(b), the narrow beam here is limited by a space composed of a microlens and a main lens and is also the discrete sampling form of the light-field. The position and direction of each narrow beam can be determined by the coordinate (s,t) of microlens unit and the coordinate (u,v) of sub-aperture. Furthermore, the light distribution $L(u, v, s, t)$ can be obtained.

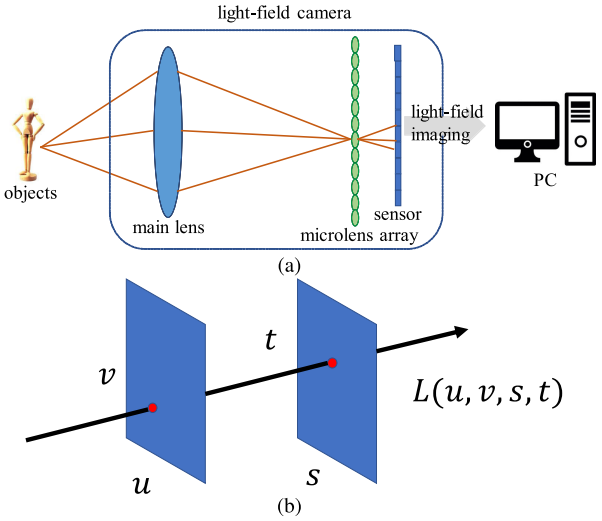

FIGURE 1. Schematic diagram of light-field camera imaging.

B. LIGHT FIELD DATASET CONSTRUCTION

1) LIGHT FIELD IMAGE ACQUISITION

In order to better illustrate the value of our experiment, we capture pedestrian images by different authors in relatively complex scenes, such as campus, busy street, and business district, using Lytro Illum camera [18] under different viewpoints and lighting conditions. The spatial resolution of the light field images is 376×541, and the angular resolution is 14×14. Considering that the imaging quality of the light field camera is sensitive to light, we limit the shooting time to 9am-4pm. Due to the small pixel size (1.4μm) of Lytro camera, the images are often too dark to use. All in all, we take 2000 images and keep 1766 images that are not so repetitive, dim, or blurred. The raw data exported from the Lytro camera is saved in LFR format. In order to refocus light-field images in subsequent experiments using the method in [41], we decode the raw data by using light-field MATLAB toolbox [42] and each size of decoded image is 118MB.

2) ANNOTATION PROTOCOL

In order to annotate the light-field dataset, we need to get a 2D version of our dataset. First, we import all raw light-field images into the Lytro software and set the aperture to *f/16* (focus to infinity) in the Lytro software to get high-quality 2D images. Then we export all of the all-in-focus images into the disk and eventually get the 2D version of our dataset, denoted as 2D-light. In this paper, we do not consider the problem of occlusion. That is, if there is a pedestrian occlusion in the image, we will artificially determine the overall pedestrian proportion based on the area of the visible part and then annotate it. Meanwhile, similar to the format of VOC [43], we use the rectangle bounding box that covers the people from head to toe to annotate the image. This also applies to other types of people, such as sitting people and others. After annotating all images in the 2D-light, we divide them into

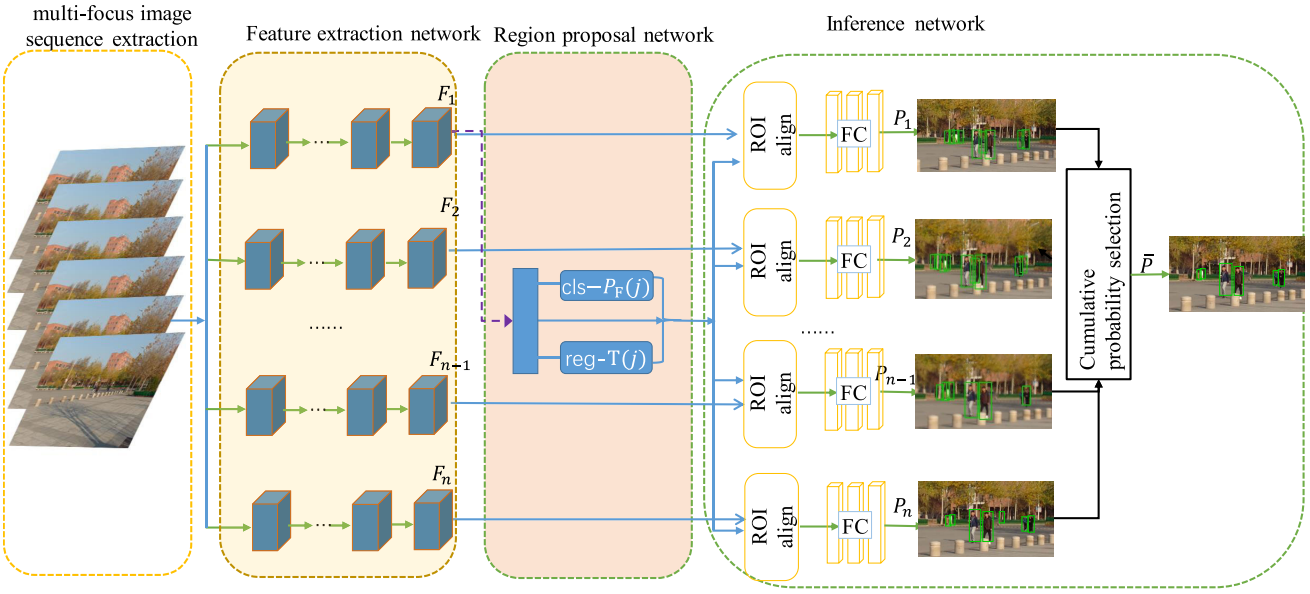

FIGURE 2. The illustration of our approach. Our proposed network consists of 4 modules, including multi-focus image sequence extraction, feature extraction network, region proposal network, and inference network.

TABLE 1. Statistics of our proposed dataset.

	Trainset	Testset	Subset
Images	1360	406	180
Persons	8599	2097	1331

the train set (1360) and test set (406). In order to study the performance of different pixel size of pedestrians in detector carefully, similar to Cityperson [44], we define the pedestrians which height between 30 pixels and 80 pixels as small scale pedestrian. Furthermore, we define a subset of images that contains at least five pedestrians and four of them are between 30 pixels and 80 pixels. The detailed information of our dataset is shown in Table 1.

C. THE PROPOSED SCHEME

Our approach consists of 4 parts, including multi-focus image sequence extraction, feature extraction network, region proposal network, and inference network, as indicated by Fig. 2. More concretely, the multi-focus image sequence extraction is to extract the multi-depth refocused images. Then, we feed these images into feature extraction network and get the convolutional features$(F_1, \cdots F_n)$. In the region proposal network, the proposals are generated on the feature F_1. In inference network, the cumulative probability selection (CPS) layer is proposed to select the candidate box with the optimal probability by accumulating the probability of the surrounding candidate proposals.

1) MULTI-FOCUS IMAGE SEQUENCE EXTRACTION

According to the requirements of our method, it is necessary to obtain the multi-focus image sequence of light-field

FIGURE 3. Considering the proportion of individual pedestrian pixels in the whole image, we divide rows and columns into 9 parts on average and get 81 regions, as shown by the dash lines in the figure.

images in advance when constructing the whole detection network. As far as we know, there are two common methods to refocus light-field images. These two methods are described in detail as following.

- Lytro software method. First, in order to get a wide scope of refocusing, the aperture is set to *f/1* in Lytro software [18]. Considering the proportion of individual pedestrian pixels in the whole image, we divide the image into 81 regions to cover as many areas as possible (Fig. 3). Depending on the functionality of the software, we click on the corresponding area (the region contained in the 81 regions) one by one and save the image to the disk. This repetitive process can be accomplished by automation tools like pyautogui library. We eventually get 81 refocused images (2450×1634) for one raw image. Considering the performance and time cost of image processing, all these refocused images are resized into 1225×817.

- Digital refocusing. Digital refocusing is proposed in Ng [22] and implemented by Tao *et al.* [41]. As far as we know, the code in Tao *et al.* [41] is the first and only open source. We rewrite the core module and accelerate it via GPU. Due to the limitation of number of microlens arrays, the resolution of the output refocused image is 541×434.

Comparison of two methods of the light-field image refocusing.

- Using Lytro software to refocus the image is time-consuming and it needs to save the temporary refocused images on the disk. However, the advantages of this method can generate high quality images and high resolution images.
- The advantages of digital refocusing is GPU acceleration and online processing, without storing refocused image to hard disk. So, it can fuse with CNNs better and achieve end-to-end training. However, due to the limitation of number of microlens arrays, the resolution of the refocused image is reduced and the height-width ratio is slightly deformed compared with the original image. This will more or less affect the final detection result.

2) FEATURE EXTRACTION NETWORK

To make full use of extracted refocused images, we feed each refocused image into a series of convolution layers (conv-layers). The initial parameters of conv-layers are from the ResNet-50 [45] pretrained model ('Conv1' to 'Conv5') on the ImageNet. Since original ResNet-50 has a large down-sampling rate at conv-layers to detect small pedestrians, we change it into a form by using a series of dilated-convolution and deconvolution resulted in the final feature map as 1/16 of input size [35]. The convolutional features from each branch are denoted as $(F_1, \cdots F_n)$, as shown in Fig. 2.

3) REGION PROPOSAL NETWORK

We adopt the region proposal network (RPN) [20] to generate the candidate proposals. These proposals can be parameterized to a tuple $L(i) = [x_i, y_i, w_i, h_i]$, where (x_i, y_i) is the center point and (w_i, h_i) are weight and height diameters of the bounding box. Because the refocusing operation often affects the local features of the image and the RPN will convolute the overall image to generate a mass of proposals, the slight change of these local features will not have a great impact on the generation of the proposals. So, in this paper, our method only utilizes the convolutional feature F_1 to generate proposals for simply processing.

4) INFERENCE NETWORK

The inference network is to get the pedestrian probability and position coordinate of each pedestrian proposal according to convolution features of each pedestrian proposal. To improve the performance of pedestrian detection, we adopt the region-of-interest (ROI) align [46] rather than ROI pooling [40] scheme compared with [20]. To select the appropriate candidates as outputs from each refocused image branch, we propose a cumulative probability selection (CPS) layer to select the appropriate candidate bounding box as the final bounding box.

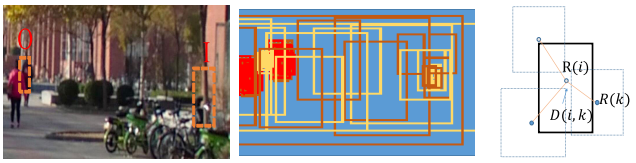

FIGURE 4. Schematic diagram of CPS layer. (a) An example image may produce false positives. (b) Proposals corresponding to (a), the red color rectangles represent proposal with high pedestrian probabilities and yellow color rectangles represent proposal with low pedestrian probabilities. (c) A bounding box neighborhood system.

a: REGION-OF-INTEREST (ROI) ALIGN

As the fully-connected layer requires the input to be a fixed size vector, similar to [46], we transform the features in any effective ROI into a small feature map with a fixed size of 7×7. The pedestrian probability of each ROI in each refocused image branch can be expressed as $P_1(i), P_2(i) \ldots P_n(i)$.

b: CUMULATIVE PROBABILITY SELECTION LAYER

In pedestrian detection task, the false positives usually occur in two situations. First, because overlapping areas (such as the region "O") contain the part of pedestrians, the probability of these areas is often higher than that of other candidate bounding box (Fig. 4(a)). The region similar to "O" is easily detected as false positives. It is worth noting that the non-maximum suppression (NMS) process cannot suppress false positives well in this case. Second, the false positives usually appear at some region (such as the region "I") which is usually called pedestrian-like vertical structure (Fig. 4(a)). These two kinds of problems can be alleviated by considering the relevance of neighborhood proposals. Most of the existing detection methods [20], [47], [48] deem that each pedestrian candidate proposal is independent and these proposals are sensitive to false positives. To overcome this, we propose a cumulative probability selection (CPS) layer which accumulates the estimated probabilities at each pedestrian proposal with its surrounding proposals. To describe the surrounding candidate bounding box, we introduce the concept of neighborhood space, denoted as $N(i)$. Moreover, we use $R(i)$ and $R(k)$ to represent neighboring pedestrian proposals where $k \in N(i)$. The l2-norm distance can be expressed as $D(i, k) = \|R(i) - R(k)\|^2$. The comprehensive pedestrian probability at each location can be expressed as

$$\bar{P}(i) = \frac{1}{N_p} \sum_{k \in N(i)} \sum_{j \in \{1, \cdots, n\}} \exp(-D(i, k)) L_j(k) P_j(k). \quad (1)$$

where $P_j(k)$ is the pedestrian probability of neighboring pedestrian proposals $R(k)$. $L_j(k)$ is a $N \times n$ weight

matrix which represents the contribution of each candidate proposal to the final pedestrian probability. The $L_j(k)$ can be initialized by normal distribution and updated by iterative training. The normalization factor is $N_p = \sum_{k\in N(i)} \sum_{j\in\{1,\cdots,n\}} \exp(-D(i,k))L_j(k)$.

This design not only combines proposals of each branch, but utilizes the method of accumulating the probability of neighbor bounding box without hard negative mining [49]. To update $P_j(k)$ in iterative training, the CPS layer must be derivable. The derivative of the final loss L with respect to $P_j(k)$ can be expressed as

$$\frac{\partial L}{\partial P_j(k)} = \frac{\partial L}{\partial \bar{P}(i)} \frac{\partial \bar{P}(i)}{\partial P_j(k)} = \frac{\partial L}{\partial \bar{P}(i)} \frac{1}{N_p} \exp(-D(i,k))L_j(k). \quad (2)$$

D. NETWORK TRAINING

We use two loss functions L_{RPN}, L_{CPS} for the whole network. The total loss of the whole network can be expressed as

$$L_{\text{total}} = L_{RPN}\left(T, P_F, T^*, P_F^*\right) + L_{CPS}\left(T', \bar{P}, T^*, P^*\right). \quad (3)$$

where T^* is the ground truth pedestrian bounding box.

1) REGION PROPOSAL NETWORK LOSS

According to the positive and negative judgment method of anchor in [20], we assign an anchor class label to each proposal candidate. If an anchor is positive, set $P_F^*(j) = 1$, and if an anchor is negative, set $P_F^*(j) = 0$. According to these definitions, the L_{RPN} can be expressed as

$$L_{RPN}\left(T, P_F, T^*, P_F^*\right) = \lambda_1 \sum_j L_{RPN}^{cls}\left(P_F(j), P_F^*(j)\right) + \lambda_2 \sum_j P_F^*(\mathrm{j}) L_{RPN}^{reg}(T(j), T^*(j)). \quad (4)$$

where L_{RPN}^{cls} is the classification loss and L_{RPN}^{reg} is the regression loss. λ_1 and λ_2 are balancing factors between L_{RPN}^{cls} and L_{RPN}^{reg}, respectively. For the regression loss, we use $L_{RPN}^{reg}(T(j), T^*(j)) = \beta(T(j) - T^*(j))$ where β is the robust loss function (smooth L_1) defined in [40]. The term $P_F^*(\mathrm{j})L_{RPN}^{reg}$ means the regression loss which is activated only for positive anchors $P_F^*(j) = 1$ and is disabled otherwise $P_F^*(j) = 0$. The $T(j)$ and $T^*(j)$ can be computed according to [20].

2) CUMULATIVE PROBABILITY SELECTION LOSS

Compared with the L_{RPN}, the CPS layer loss L_{CPS} can produce more accurate pedestrian probability. The L_{CPS} also includes classification loss and regression loss functions and can be expressed as

$$L_{CPS}\left(T', P', T^*, P^*\right) = \lambda_3 L_{CPS}^{cls}\left(\bar{P}, P^*\right) + \lambda_4 P^* L_{CPS}^{reg}\left(T', T^*\right). \quad (5)$$

where L_{CPS}^{cls} is the classification loss and L_{CPS}^{reg} is the regression loss. λ_3 and λ_4 are balancing factors between L_{CPS}^{cls} and L_{CPS}^{reg}, respectively. According to the rules of anchor positive and negative judgments, we also make positive and negative judgments on candidate proposal. If candidate proposal is positive, set $P^* = 1$, otherwise, set $P^* = 0$. The L_{CPS}^{reg} can also be computed similar to the L_{PRN}^{reg}. In the process of network training, minimizing the total loss function of L_{RPN} and L_{CPS} can make the network provide the best performance of pedestrian detection.

TABLE 2. Comparison results of different *num_refocus*.

num_refocus	8	16	32	64	128	256
MR	40.35%	37.56%	35.34%	**32.36%**	32.32%	32.32%
Time(s.)	0.75	0.98	1.15	**1.75**	2.75	5.45

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present the settings and details of our experiments. Then, we compare our method with state-of-the-art methods and analyze the effects of some parameters on the results. In order to explicitly illustrate the role of our method, we make a qualitative analysis in the end.

A. EXPERIMENT SETTINGS

Our framework is implemented using the Pytorch deep learning library. We train the network using a NVidia GeForce GTX 1080Ti GPU and adjust the learning rate from 0.01 to 0.001 after 60k iterations (in total 120k iterations). The SGD solver is adopted to optimize the network. In our experiments, our network is implemented with the fixed parameter for all datasets:$\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{1, 10, 1, 10\}$ and the number of anchors is set to 9 similar to [20]. The size of anchors is defined as 16, 32, and 64. The number of neighboring proposals is set to 7. The parameters worth discussing in this experiment are $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and the number of neighboring proposals. We conduct several experiments with these parameters, and obtain the balance of speed and accuracy. The standard evaluation metric [50] is adopted for our experiment: log missrate is averaged over the false positive per image (FPPI) in (10^{-2},100), denoted as MR.

B. ANALYSIS OF THE DEPTH-RESOLUTION IN DIGITAL REFOCUSING

In order to analyze the depth-resolution of the refocused image on the detection performance, we choose the digital refocusing method to illustrate. From [22] and the open-source code in [41], we know that the α_{step} determines resolution of the refocused light-field images and the number of refocused images, as in (6). We use *num_refocus* to denote the number of refocused images.

$$num\_refocus = \frac{\alpha_{\max} - \alpha_{\min}}{\alpha_{\text{step}}}. \quad (6)$$

The default value of α_{max} and α_{min} are 2 and 0.2, respectively. In order to analyze the effect of different α_{step} on detection result, we compare the effects of different *num_refocus* on the results.

Table 2 shows that the MR corresponding to the number of different refocused image and the time required to

TABLE 3. Comparison results of different refocusing method.

Testset \ Method	Lytro Software method	Digital Refocusing
All	30.09%	33.56%
Subset	33.34%	38.64%

TABLE 4. Comparison results of different backbone.

Backbone	Vgg16	ResNet-50	ResNet-101	MobileNetV1
MR	32.23%	**30.09%**	31.16%	**38.36%**
Test Time	1.82s/img	**1.62s/img**	2.02s/img	**1.45s/img**

detect one light-field image. We find that if *num_refocus* $\geq$ 128, MR remains unchanged. If *num_refocus* equals to 64 or 128, MR does not change much, but the time is shortened obviously. So, if using the digital refocusing method and the *num_refocus* equals to 64, a speed-accuracy trade-off result can be achieved.

C. COMPARISON STUDY OF TWO DIFFERENT REFOCUSING METHODS

As mentioned in chapter III, different refocusing methods generate refocused images with different resolutions. So, in order to study the effect of refocusing methods on the final results, we present the detection results on the all dataset and subset.

Table 3 shows that the missrate of the light-field images refocusing algorithm using Lytro software is lower than that of Tao *et al.* [41] on both test dataset and subset. This confirms that using the Lytro software methods to generate high resolution images will retain more useful image information and provide more detection performance. It is worth noting that as Lytro company dose not provide the open source implementation code of Lytro software, we do not know the specific principle of obtaining refocused images by clicking the corresponding area with the mouse. However, the refocused images obtained from the Lytro software can be used for theoretical verification.

D. COMPARISON STUDY OF DIFFERENT BACKBONES

To illustrate the influence of backbone on detection performance, we use several common backbones including Vgg16 [51], ResNet-101, and MobileNetV1 to replace the ResNet-50 in feature extraction network and evaluate the results of experiments, respectively. Four groups of experiments adopt the Lytro software method to refocus light-field images.

Table 4 shows that using the ResNet-50 as the backbone can get the best detection performance with less time cost. Using the MobileNetV1 as the backbone can achieve the fastest detection speed, but the performance of the detector is greatly reduced. Overall, the ResNet-50 is the best feature extractor backbone at present. Many recent pedestrian detection methods [38], [39], [52] all adopt it as the feature extractor. Because it takes a lot of time to process the light-field image data, it is also a good choice to adopt a lighter feature extractor such as MobileNetV1.

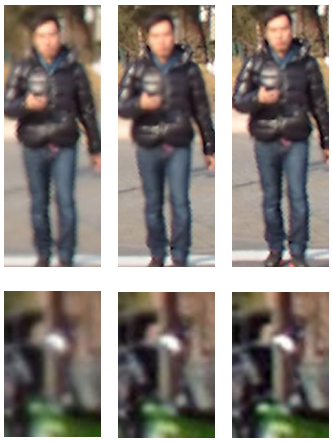

FIGURE 5. Some examples of pedestrians and hard negatives from different refocused images.

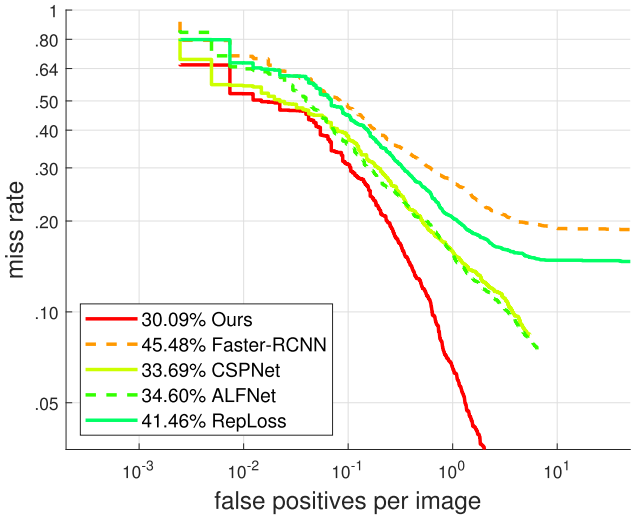

FIGURE 6. Comparisons of detection results with state-of-the-art methods.

E. ABLATION STUDY ON CPS LAYER

The CPS layer has two functions: one is to use the cumulative probability of surrounding proposals to suppress hard negative samples, and the other is to select more appropriate proposals as our final detection results from the multiple refocusing branches. In order to evaluate the role of $N(i)$ neighborhood system, we remove the l2-norm term from the CPS layer. The pedestrian probability can be expressed as

$$\bar{P}(i) = \frac{1}{N_p} \sum_{j \in \{1,\cdots,n\}} \exp L_j P_j. \tag{7}$$

where normalization factor $N_p = \sum_{j\in\{1,\cdots,n\}} \exp L_j$. The experiment results on all test dataset are as follows.

Table 5 shows that the missrate of the method without $N(i)$ system is higher than that of the method with $N(i)$ neighborhood system. The $N(i)$ neighborhood system in CPS

FIGURE 7. Comparisons with state-of-the-art methods. (From left to right) ALFNet, CSPNet, Faster-RCNN, Reploss, and Ours. The green rectangle represents the detection results and the red rectangle represents missed detection and error detection.

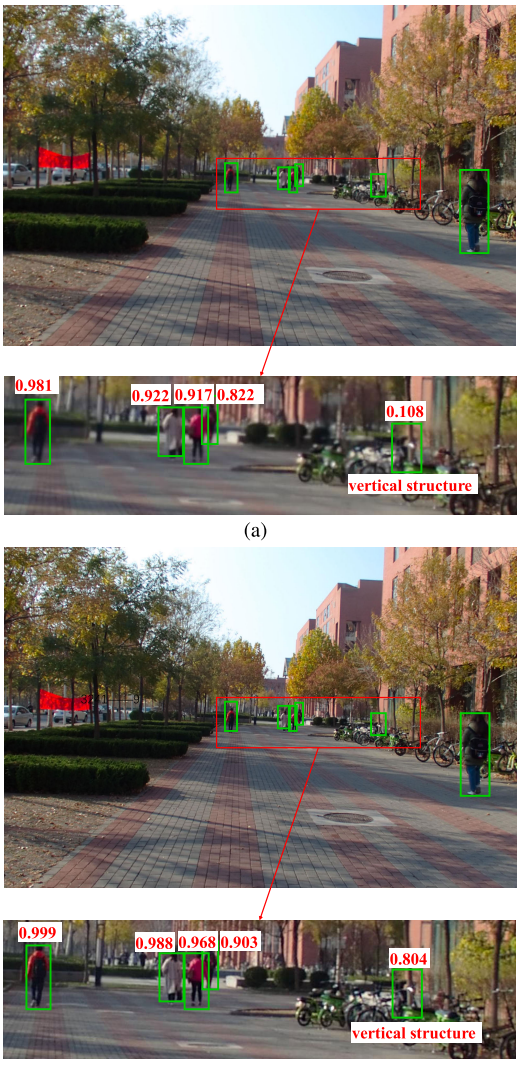

FIGURE 8. Analysis of the confidence score at same position of different refocused images. (a) Detection results of refocusing on the region 9. (b) Detection results of refocusing on the region 11.

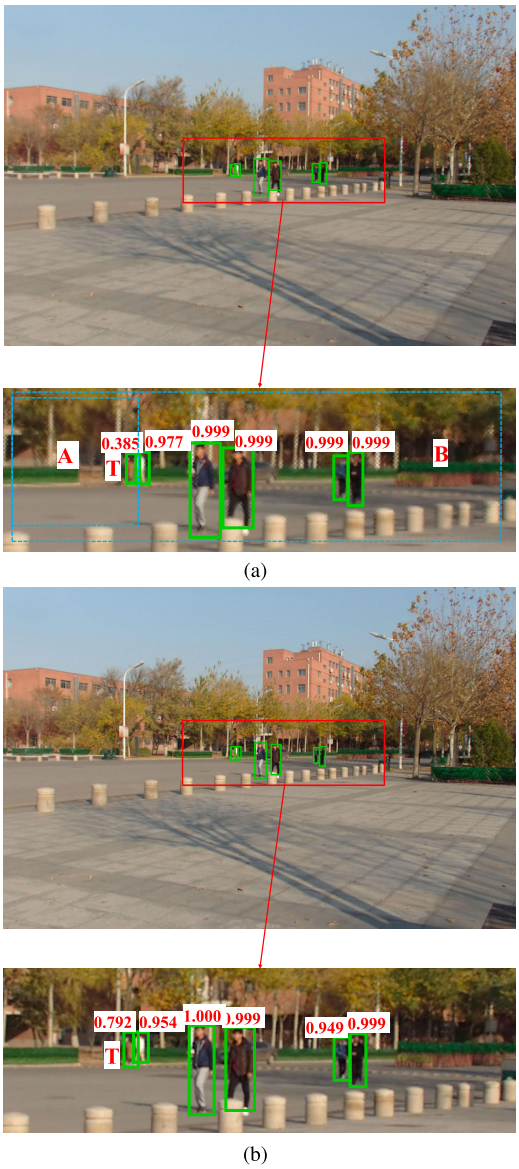

FIGURE 9. Analysis of the confidence score at same position of different refocused images. (a) Detection results of refocusing on the region 33. (b) Detection results of refocusing on the region 36.

layer indeed improves the detection performance. As mentioned in chapter III, 81 refocused images will be obtained after refocusing operation (using Lytro software method). To show the effect of pedestrians and hard negatives in different refocused images more specifically, we cut out examples from some refocused images and find that the sharpness is different for the same location of each image (Fig. 5). These parts have different feature representations and roles in the whole detection network and the effect of different refocused parts on the detection results is illustrated in the *Qualitative Analysis* section. Moreover, due to the randomness of the location of the pedestrians and hard negative samples in each of refocused images, it is necessary to input all the refocused images into the network to ensure that the feature representation of each refocused image is not missed. To evaluate the effect of different refocused branches on the detection results, we use odd, even, and interval two sampling method (noted as a, b, and c) to sample 81 refocused branches, and the number of reserved branches finally included in the study is 41, 40, and 27, respectively. The comparative results show that if some branches are missing, the performance of the whole system will decline to some extent (Table 6).

TABLE 5. Performance of our methods with different CPS layers.

Method	Backbone	MR
Ours (CPS)	**ResNet-50**	**30.09%**
Ours(CPS without $N(i)$system)	ResNet-50	31.69%

F. COMPARISONS WITH STATE-OF-THE-ART METHODS

Recent open source methods are trained and evaluated on the 2D-light dataset for a reasonable comparison with ours.

As shown in Table 7 and Fig. 6, we compare our approach with some state-of-the-art methods including Faster R-CNN

TABLE 6. Performance of our methods with different branches input.

Sampling method	Number of branches	Backbone	MR
a	41	ResNet-50	32.29%
b	40	ResNet-50	32.32%
c	27	ResNet-50	34.89%
Full	**81**	**ResNet-50**	**30.09%**

TABLE 7. Comparisons with state-of-the-art methods on our dataset.

Method	Backbone	All	Subset
Faster-RCNN [20]	VGG-16	45.48%	43.38%
RepLoss [38]	ResNet-50	41.46%	43.26%
ALFNet [39]	ResNet-50	34.60%	38.34%
CSPNet [52]	ResNet-50	33.69%	39.45%
Ours	**ResNet-50**	**30.09%**	**33.34%**

[20], CSPNet [52], ALFNet [39], and RepLoss [38]. These methods use default parameters in the training process. Our approach achieves the MR of 30.09% and 33.34% on all test set and subset, respectively, and reduces 15.39% and 10.04% compared with the baseline method (Faster R-CNN). Our method also works better than other 3 approaches on both dataset and subset. The MR of all methods is between 30%-50%. Compared with the above four methods on Cityperson and Caltech datasets, the value of MR is in a reasonable range. This shows that the dataset we build is reasonable and has certain difficulty. It can properly reflect the differences of the algorithm and can be used to evaluate the performance of pedestrian detection algorithm.

As shown in Fig. 7, the first row of images is captured in ice arena. The second to the fourth rows are campus, commercial street, and busy street. The green rectangle in the image represents the detection results and the red rectangle represents missed detection and error detection. The Faster R-CNN method dose not optimize the pedestrian situation, such as increasing the size of feature map and mining the hard negatives, resulting more false positive in detection results. The ALFNet, CSPNet, and Reploss all have different levels of missed detection and error detection. The missed detection usually occurs in case of small scale pedestrian or occlusion, and the error detection usually occurs in case of anthropomorphic negatives. As shown in first row, our method still work in a crowded scene. The second row shows that our method can still achieve a better result in case of bad light. All of these show that our proposed method is robust in real world.

G. QUALITATIVE ANALYSIS

To explore the effect of different refocused images on the detection results, we output the confidence scores of the detection results of different refocused images. From Fig. 3, one light-field image is divided into 81 regions for refocusing operation. Then, we select a light-field image focused on the region 9 (Fig. 8(a)) and region 11 (Fig. 8(b)) respectively for specific explanation. As shown in Fig. 8, the hard negative (e.g., vertical structure) is detected with the confidence score of 0.108 and 0.804 in (a) and (b), repectively. From the perspective of intuition, the sharpness of the vertical structure part seems to be better in (b) than in (a). If the threshold is set to 0.9, it can really suppress the vertical structure in Fig. 8(a), but it also suppress positive samples with a confidence score less than 0.9. In our experiment, the threshold is usually set to 0.5. Therefore, the detection branch of different refocused images processed by CPS layer can suppress some hard negative (e.g., vertical structure) which are usually detected as false positive. If the hard negative samples is as close to the light-field camera as the positive samples, that is to say, they are in the same imaging plane, and they will have the same sharpness with the change of focusing distance. It may affect the detection results more or less, but this situation is very rare in the whole dataset after all.

For small scale pedestrian detection, we also output the confidence scores of the detection results of different refocused images. As shown in Fig. 9(a), we define target T as the small scale pedestrian target which needs to be studied. Then, we feed the region A and B into the detection network, and the confidence score of T is 0.865 and 0.387, respectively. This phenomenon shows that different feature maps classified by fully-connected layer will interact with each other. Namely, some feature maps are activated and some are suppressed. Whereas the confidence score of target T in Fig. 9(b) is 0.792. This phenomenon shows that the representation of feature map around the target T is enhanced to some extent compared with which in Fig. 9(a) in some cases. So, the target T in Fig. 9(b) is considered to be a true positives. Overall, multiple refocused images will increase the representation ability of feature maps and can extract more useful features than a single 2D image.

V. CONCLUSION

In this paper, we design two methods to extract multiple refocused images from raw light-field images and propose a multi-focus detection network (MFDN) based on Faster R-CNN architecture. MFDN uses multiple refocused images as input and makes full use of the information of light-field images. To select the appropriate pedestrian candidate proposal from detection branch of each refocused image, we propose cumulative probability selection (CPS) layer. The evaluation experiments demonstrate that our approach improves the performance of pedestrian detection. Overall, our method provides a new idea for detecting pedestrians with light-field camera and promotes the application of light-field imaging in object detection.

REFERENCES

[1] Y. Zhang, P. Yi, D. Zhou, X. Yang, D. Yang, Q. Zhang, and X. Wei, "CSANet: Channel and spatial mixed attention CNN for pedestrian detection," *IEEE Access*, vol. 8, pp. 76243–76252, 2020.

[2] P. Tumas, A. Nowosielski, and A. Serackis, "Pedestrian detection in severe weather conditions," *IEEE Access*, vol. 8, pp. 62775–62784, 2020.

[3] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for robust pedestrian detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3820–3834, Jan. 2020.

[4] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "WiderPerson: A diverse dataset for dense pedestrian detection in the wild," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 380–393, Feb. 2020.

[5] J. Cao, Y. Pang, J. Han, B. Gao, and X. Li, "Taking a look at small-scale pedestrians and occluded pedestrians," *IEEE Trans. Image Process.*, vol. 29, pp. 3143–3152, Dec. 2020.
[6] Y. Pang, J. Cao, J. Wang, and J. Han, "JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 12, pp. 3322–3331, Dec. 2019.
[7] Y. Zhao, Z. Yuan, and B. Chen, "Accurate pedestrian detection by human pose regression," *IEEE Trans. Image Process.*, vol. 29, pp. 1591–1605, Sep. 2020.
[8] X.-M. Xia, Z.-L. Jiang, and P.-F. Xu, "A detection algorithm of spatter on welding plate surface based on machine vision," *Optoelectron. Lett.*, vol. 15, no. 1, pp. 52–56, Jan. 2019.
[9] X. Zhang, D. Wang, Z. Zhou, and Y. Ma, "Robust low-rank tensor recovery with rectification and alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 16, 2019, doi: 10.1109/TPAMI.2019.2929043.
[10] W. Liu, X. Chang, L. Chen, D. Phung, X. Zhang, Y. Yang, and A. G. Hauptmann, "Pair-based uncertainty and diversity promoting early active learning for person re-identification," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 2, pp. 1–15, Mar. 2020.
[11] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving occlusion and hard negative handling for single-stage pedestrian detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 966–974.
[12] W. Choi, C. Pantofaru, and S. Savarese, "Detecting and tracking people using an RGB-D camera via multiple detector fusion," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 1076–1083.
[13] C.-T. Hsieh, H.-C. Wang, Y.-K. Wu, L.-C. Chang, and T.-K. Kuo, "A Kinect-based people-flow counting system," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst.*, Nov. 2012, pp. 146–150.
[14] X. Chen, K. Henrickson, and Y. Wang, "Kinect-based pedestrian detection for crowded scenes," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 31, no. 3, pp. 229–240, Mar. 2016.
[15] C. Premebida and U. Nunes, "Fusing LIDAR, camera and semantic information: A context-based approach for pedestrian detection," *Int. J. Robot. Res.*, vol. 32, no. 3, pp. 371–384, Mar. 2013.
[16] Z. Zhang and W. Tao, "Pedestrian detection in binocular stereo sequence based on appearance consistency," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1772–1785, Sep. 2016.
[17] K. Park, S. Kim, and K. Sohn, "Unified multi-spectral pedestrian detection based on probabilistic fusion networks," *Pattern Recognit.*, vol. 80, pp. 143–155, Aug. 2018.
[18] R. Ng. *Lytro Redefines Photography With Light Field Cameras*. Accessed: Oct. 18, 2018. [Online]. Available: http://www.lytro.com/
[19] C. Jia, F. Shi, Y. Zhao, M. Zhao, Z. Wang, and S. Chen, "Identification of pedestrians from confused planar objects using light field imaging," *IEEE Access*, vol. 6, pp. 39375–39384, 2018.
[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[21] E. H. Adelson and J. Y. A. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992.
[22] R. Ng, M. Levoy, M. Brédif, G. Duval, and M. Horowitz, "Light field photography with a hand-held plenoptic camera," Ph.D. dissertation, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep. 2005-02, 2005.
[23] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 1–12, Jul. 2007.
[24] S. Shi, S. Xu, Z. Zhao, X. Niu, and M. K. Quinn, "3D surface pressure measurement with single light-field camera and pressure-sensitive paint," *Exp. Fluids*, vol. 59, no. 5, p. 79, Apr. 2018.
[25] H. Zhao, Z. Miao, Y. Peng, and S. Zhao, "Depth extraction of low-texture region using a light field camera," *Opt. Eng*, vol. 56, no. 7, pp. 65–75, Jul. 2017.
[26] J. Fan and Y.-H. Yang, "Depth estimation of semi-submerged objects using a light-field camera," in *Proc. 14th Conf. Comput. Robot Vis. (CRV)*, May 2017, pp. 80–86.
[27] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon, "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1547–1555.
[28] W. Zhou, E. Zhou, G. Liu, L. Lin, and A. Lumsdaine, "Unsupervised monocular depth estimation from light field image," *IEEE Trans. Image Process.*, vol. 29, pp. 1606–1617, Oct. 2020.
[29] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–12, Jul. 2013.
[30] Y. Piao, X. Li, M. Zhang, J. Yu, and H. Lu, "Saliency detection via depth-induced cellular automata on light field," *IEEE Trans. Image Process.*, vol. 29, pp. 1879–1889, Oct. 2020.
[31] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 57–65.
[32] R. Raghavendra, K. B. Raja, and C. Busch, "Exploring the usefulness of light field cameras for biometrics: An empirical study on face and iris recognition," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 5, pp. 922–936, May 2016.
[33] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1605–1616, Aug. 2017.
[34] T. Wang, J. Zhu, E. Hiroaki, M. Chandraker, A. Efros, and R. Ramamoorthi, "A 4D light-field dataset and CNN architectures for material recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 121–138.
[35] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 554–569.
[36] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.
[37] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 443–457.
[38] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7774–7783.
[39] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 643–659.
[40] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[41] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 673–680.
[42] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1027–1034.
[43] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[44] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4457–4465.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[46] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[47] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. 14th Eur. Conf. Comput. Vis. ECCV*, Amsterdam, The Netherlands, Oct. 2016, pp. 354–370.
[48] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May 2017.
[49] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.
[50] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–14.
[52] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5182–5191.

YUFENG ZHAO received the bachelor's degree from the Tianjin University of Technology and Education, Tianjin, China, in 2015. He is currently pursuing the Ph.D. degree with the Key Laboratory of Computer Vision and System, Ministry of Education, Tianjin University of Technology. His research interests include object detection and computer vision.

FAN SHI received the Ph.D. degree from Nankai University, Tianjin, China, in 2012. He is currently an Associate Professor with the Tianjin University of Technology, China. His research interests include machine vision, pattern recognition, and optics.

MENG ZHAO received the Ph.D. degree from Tianjin University, Tianjin, China, in 2016. She is currently a Lecturer with the Tianjin University of Technology, China. She is a part of the Chinese National Natural Science Foundation Young Project. Her research interests include medical images and machine learning.

WENZHE ZHANG received the bachelor's degree from the Weifang Institute, Shandong, China, in 2018. He is currently pursuing the master's degree with the Key Laboratory of Computer Vision and System, Ministry of Education, Tianjin University of Technology. His research interests include light field reconstruction and machine learning.

SHENGYONG CHEN (Senior Member, IEEE) received the Ph.D. degree in computer vision from the City University of Hong Kong, Hong Kong, in 2003. In 2006 and 2007, he was with the University of Hamburg. He is currently a Professor with the Tianjin University of Technology, China. He has authored over 100 scientific articles in international journals. His research interests include computer vision, robotics, and image analysis. He is a Fellow of IET and a Senior Member of CCF. He was a recipient of the National Outstanding Youth Foundation Award of China, in 2013. He received the Fellowship from the Alexander von Humboldt Foundation, Germany.