IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 4, 2020, accepted September 4, 2020, date of publication September 8, 2020, date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022630

Scene Target 3D Point Cloud Reconstruction Technology Combining Monocular Focus Stack and Deep Learning

YANZHU HU, YINGJIAN WANG, AND SONG WANG, (Member, IEEE)
College of Modern Post, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Yingjian Wang (wangyingjian@bupt.edu.cn)

This work was supported in part by the Science and Technology Nova Plan of Beijing City under Grant Z201100006820122, in part by the Beijing Municipal Natural Science Foundation under Grant 4192042, in part by the Fundamental Research Funds for the Central Universities under Grant 2020RC14, and in part by the Beijing Science and Technology Plan Project under Grant Z191100001419001.

ABSTRACT In order to obtain the depth information of the target in the scene and realize three-dimensional (3D) reconstruction, in this paper, a target reconstruction method combining monocular focus stack image and deep neural network is proposed. This method makes full use of the advantages of light field imaging technology and can generate the all focus image. The method first collects multiple frames of continuous images at different focal lengths of the scene, using a divide and conquer algorithm strategy, uplink uses YOLO neural network to identify the target in 3D space and track the position information; the downlink reconstructs the four-dimensional (4D) light field data based on the focus stack image frequency domain back projection, and then uses light field imaging technology to invert the scene parallax; subsequently, achieve scene depth estimation and reconstruction of all focus image; finally, the uplink and downlink are merged to realize the reconstruction of the 3D point cloud of the space target. Experimental results on real scenes show the effectiveness of the proposed algorithm.

INDEX TERMS Focus stack image, deep learning, light field reconstruction, all focus image, 3D reconstruction.

I. INTRODUCTION

Scene target detection and three-dimensional reconstruction based on visual sensor terminals have received extensive attention in the field of mobile robot industrial detection and exploration. First, appropriate sensors are needed to capture the three-dimensional information of the world, binocular and RGB-D cameras are widely used, they are limited by the large size and lack of depth information; structured light obtains three dimensional scene information through the principle of triangle, which has short detection distance and low reconstruction accuracy; although the monocular camera is small in size, it has scale uncertainty, and can only record the two-dimensional (2D) space of the scene, with almost no angle information, as a result, conventional 3D vision is sensitive to the local structure of the scene.

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

As an emerging technology, light field imaging expands the field of computational imaging and computer vision, provides new methods for high-precision 3D vision sensing technology. Light field data can realize the simultaneous collection of light irradiation and direction information, there is a coupling relationship between light direction and scene depth information, which contains rich depth information and can reconstruct the scene depth image with higher accuracy than stereo vision [1], [2]. It can further realize the three-dimensional reconstruction of the scene. Light field imaging data can be obtained by direct methods of main lens-microlens array [3] and camera array [4], [5], indirect methods such as encoding mask [6], [7] or focus stack image [8] can also be used to reconstruct the light field. The focus stack is the compressed information of the light field data, which is achieved by keeping the imaging parameters constant by shifting the lens, achieving flexible acquisition of the real scene, and reconstructing the light field data at any angular resolution. Using light field data to reconstruct scene depth

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

information is dense and pixel-by-pixel, at the same time, it can generate all focused image of the scene, and project the all focus image into the scene to restore the 3D point cloud of rich colors and textures.

In the actual work of robots, 3D reconstruction is not applied to the whole scene but only at specific target areas, if the reconstruction of the whole scene is time-consuming and the effect is not obvious. In this case, machine learning will become the best choice for target recognition in the scene, in recent years, target detection and positioning based on deep neural networks have developed rapidly and have obvious advantages. Therefore, this paper proposes a fusion method to achieve 3D accurate reconstruction of the target area in the scene, the main contributions of the method mainly include the following three aspects:

- An algorithm framework is proposed, which fuses target detection of deep learning and scene reconstruction of light field imaging technology through up-down link parallel method, and finally restates 3D point cloud of specific target in the scene.
- Proposed a monocular passive 3D visual sensing technology suitable for small-dimensional robots, this technology is based on the back-projection of the focus stack image to form a 4D light field and performs scene depth estimation to achieve the reconstruction of the all focus image.
- Experimental studies are carried out in two different real scenes, and the influencing factors of target reconstruction are analyzed and discussed, finally reconstructed 3D point clouds with different set goals in the scene.

II. RELATED WORK

Passive and contactless measurement of visual information as feedback is highly valued for robot exploration and rescue, which can realize scene depth estimation and 3D reconstruction. Singh Mahesh Kr *et al.* used Kinect to obtain distance information and realized 3D reconstruction of scene by classifying the statistical thick and thin areas, but the TOF sensor was limited in resolution [9]. Yang *et al.* proposed a 3D reconstruction system combining binocular and TOF depth cameras to accurately identify the distance between objects and the camera, promote the scene acquisition resolution and stereo matching effect, and improve the reconstruction accuracy, however, the system adopts an active sensing method with large scale and high power consumption [10]. Compared with binocular and RGB-D camera, monocular camera has high adaptability and low power consumption. Newcombe *et al.* proposed a monocular vision pose estimation and sparse point cloud generation method, the method uses structure from motion (SFM) for scene basic grid prediction, and is distorted into a depth image to finally achieve scene model reconstruction, the algorithm needs to predict and update the optical flow of the scene [11].

Compared with the traditional imaging function that only preserves 2D information, light field imaging can obtain richer information during processing and reconstruction. The complete light field information of the scene is expressed by the seven-dimensional all-light function $L(V_x, V_y, V_z, \theta, \varphi, \lambda, t)$ proposed by Adelson in 1991. The two-plane parametric model of the light field proposed by Levoy and Gortler *et al.* simplifies seven-dimensional function approximately to the four-dimensional light field $L(x, y, u, v)$, plane (x, y) records spatial information, plane (u, v) records directional. The two-plane parametric model can be used to generate images with different parameters, which are widely used in actual imaging systems to achieve digital refocusing [12]–[14], depth reconstruction [15], [16], and high-precision 3D scene reconstruction [17], [18]. Ren N achieved 4D light field data through the lens-microlens array, further reconstructed the scene depth and arbitrary focus images, formed the Lytro handheld light field camera. The angular resolution was acquired at the expense of the spatial resolution by using a single sensor, resulting in a low final imaging resolution [19]. Raghavendra R *et al.* captured images with different focal points in the scene through the 4D light field camera, further rendered the attributes of multiple depth images, and improved the performance of 3d reconstruction and recognition of scene faces through the new resolution enhancement technology of discrete wavelet transform [20]. Camera array can realize light field data acquisition with rich visual angle by using multi-camera array or single camera combined with precision mobile platform, due to the large equivalent aperture, the resolution can be higher and the sampling density can be flexibly controlled artificially, but it has a large system and high cost. Disadvantages. Wang *et al.* used the camera array system to render the unfocused areas in the scene based on the anisotropy of the depth estimation, and then generated the refocused image through the reconstructed super-resolution method [5]. Xu *et al.* indirectly captured high-resolution optical field data by using two attenuation masks, and used a two-dimensional (2D) camera sensor to encode and sample a four-dimensional (4D) spectrum [7]. Based on the focal stack projection model, Liu *et al.* proposed the filtered-back-projection (FBP) algorithm for reconstructing the 4D light field, and get it according to the reconstruction formula [21]. Tao *et al.* used gradient detection to establish focus measure, and then fused matching consistency for global optimization of depth reconstruction, which could reconstruct high-precision scene depth [22]. Julia R. Alonso *et al.* proposed a refocusing method based on arbitrary shapes and sizes in the focus stack image, which realized the high-resolution refocusing of defocusing images by considering the visual information reorganization of the depth point transformation extension function [8].

The light field imaging system based on the focus stack provides better applicability to the detection and rescue of the robot industry in terms of theory, environmental adaptability, and micro-grouping. Most of the existing application scenarios only deal with detection and rescue targets (such as target finding and location). The deep neural network identified candidate frames of multiple target regions in the

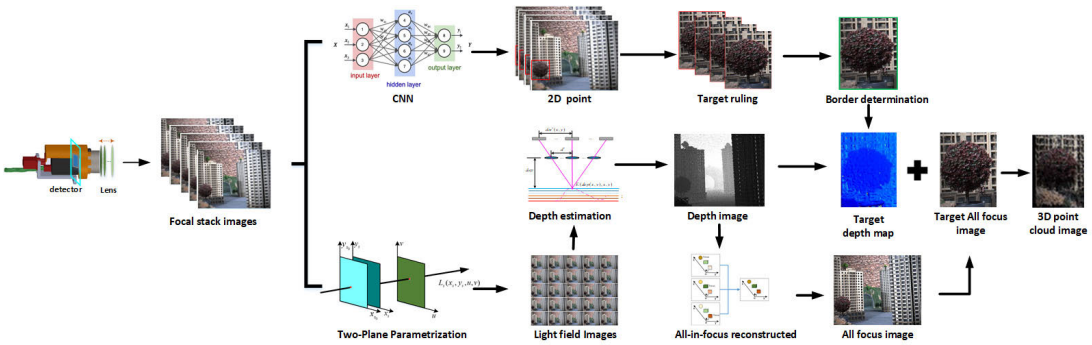


FIGURE 1. The algorithm framework of this paper.

scene, extracts features from the candidate frames and classifies information, finally returns to obtain the coordinate information of the target. With the continuous development of deep neural network, various types of network update and iteration, the target detection rate and positioning accuracy of scene image are constantly improved. Combining the target detection information of the deep neural network and the light field imaging of the focus stack, can implement a variety of imaging effect by flexible calculation, including scenes of various kinds of target depth estimation, refocus and scene reconstruction, have great potential in many fields, There are huge potentials in many fields, such as Virtual Reality (VR), robot navigation, and simultaneous localization and mapping (SLAM).

III. ALGORITHM

The frame diagram of the 3D reconstruction algorithm for specific targets in the scene proposed in this paper is shown in Fig. 1.

The algorithm of this paper first transmits the collected series of scene focus stack images to the deep neural network and scene reconstruction estimation model, the uplink is based on YOLO deep neural network to detect and locate the target area to be tested and obtain the optimal location information, which will be introduced in Section 3.1. Downlink generates scene depth estimation and all focused image based on focus stack image, which is executed in two steps, first, based on the projection model, the reconstructed light field data analysis algorithm (FBP) and iterative algorithm (Landweber) are formed, which are detailed in Section 3.2.1; then, light field imaging technology is used to reconstruct the scene parallax, depth information is estimated and all focus image of the scene is formed, this part is introduced in Section 3.2.2. Finally, Section 3.3 introduces the fusion of target detection and scene depth of up-downlink, and the reconstruction of the 3D point cloud of the target to be tested by perspective projection model.

A. SCENE TARGET DETECTION AND LOCATION OF UPLINK

The focus stack image within the visual range of the scene is realized by the precision servo controlling the micro movement of the monocular camera lens, the schematic

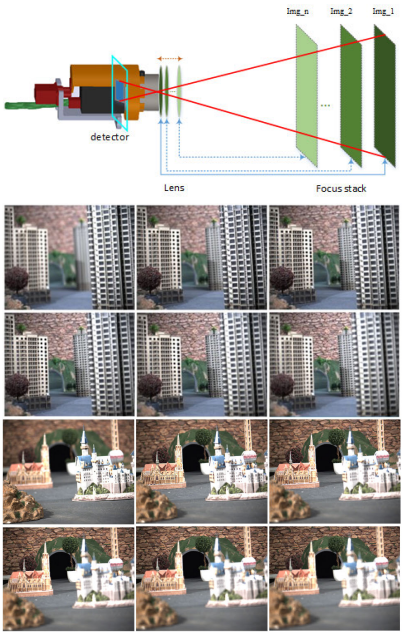


FIGURE 2. Diagram of acquisition system and partially focused images of scene.

diagram of the image acquisition system and the partially focused images of the scene are shown in Fig. 2.

Each frame of focused image is input into the deep learning neural network algorithm to realize target detection. The previous algorithms are two-step R-CNN algorithm (R-CNN [23], Fast R-CNN [24], Faster R-CNN [25]), which requires the use of the heuristic method or CNN network before classification and regression. This paper uses the classic YOLO algorithm, which uses a single CNN end-to-end process to handle object detection and frame positioning, which has higher computational efficiency than other methods. Algorithm mainly includes three aspects, first, the size of the image to be detected is adjusted to 448×448, and the image is divided into $S \times S$ grids, each grid is responsible for predicting whether the object center is detected, the grid containing the detected center of the object is taken as the initial condition to predict the border and confidence of the corresponding object, including five parameters, where (w, h) represents the border size, (x, y) represents the distance of

the center of the border offset from the corresponding grid, and confidence refers to the accuracy of the border and the possibility of the embedded object of the box. For the object classification, the probability value of prediction category is given for each cell, representing the probability that the object within the border predicted by the cell belongs to various categories.

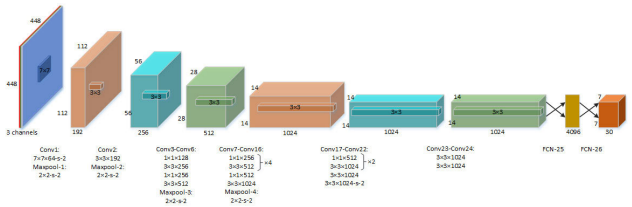


FIGURE 3. The framework of YOLO algorithm.

As shown in Fig. 3, the YOLO algorithm framework consists of 24 convolutional layers and 2 full-connection layers. The convolutional layer is composed of two different kernels, including 3×3 and 1×1 kernels, and the network outputs a vector of size $7 \times 7 \times 30$.

The target search and detection in the actual scene is based on the performance of the equipment carrying the monocular camera and the target data set. The target data set is composed of the actual image of the scene and the public data set. Based on the target detection result of the scene focus stack image obtained, erroneous detection and position information with low confidence are eliminated. In order to ensure that the position information covers the complete target, the maximum value of the frame of multiple sets of position information is taken as the final target position optimal information.

B. SCENE DEPTH ESTIMATION AND RECONSTRUCTION OF ALL FOCUSED IMAGE BASED ON FOCUS STACK IMAGE

This part is realized in two steps. First, the focus stack image is used to reconstruct the 4D light field, and a projection model is established to depict the relationship between the light field and the focus stack data space. Under this projection model, the filtered-back-projection method (FBP) of light field reconstruction is derived, and the Landweber method is used to achieve the optimal iteration of the light field data target functional. Second, reconstruct the scene parallax based on the above light field data and iteratively realize the minimization of the parallax mesh functional to complete high-precision scene depth estimation, and finally combine the focus stack image to form the scene all focus image.

1) RECONSTRUCTION OF 4D LIGHT FIELD DATA BASED ON FOCUS STACK IMAGE

a: DESCRIBE THE POSITIVE PROCESS OF FORMING THE FOCUS STACK $E(s, x_s, y_s)$ BY THE 4D LIGHT FIELD $L_s(x_s, y_s, u, v)$ BASED ON THE PROJECTION OPERATOR P

The 2D focus image is a form of focus description of a 3D scene, and each image is a 2D projection of a 4D light field.

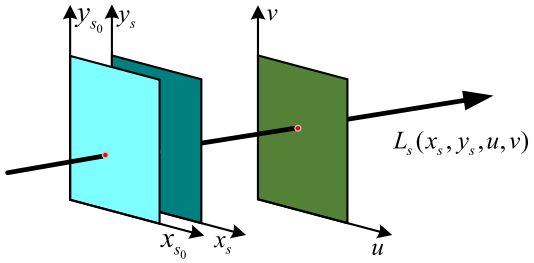


FIGURE 4. Two plane parametric representation of the light field.

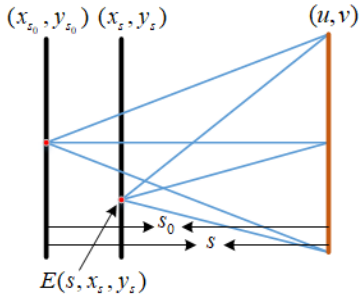


FIGURE 5. Focused imaging of different imaging planes.

The 4D light field $L_s(x_s, y_s, u, v)$ can be parametrically represented by the (x_s, y_s) and (u, v) two-plane [26], [27], as shown in Fig. 4, (x_{s_0}, y_{s_0}) and (x_s, y_s) denote the reference imaging plane and arbitrary imaging plane, respectively, (u, v) is the lens plane. The focusing imaging diagrams of different imaging planes are shown in Fig. 5, where $E(s, x_s, y_s)$ represents the focal stack imaging plane at depth s, s and s_0 are the distances from (u, v) to (x_s, y_s) and (x_{s_0}, y_{s_0}) planes, respectively.

When the distance between the two imaging planes is the same, $L(x, y, u, v)$ and $L_s(x_s, y_s, u, v)$ represent the same ray, the affine transformation expressions from (x, u) to (x_s, u) and (y, v) to (y_s, v) are:

$$\begin{pmatrix} x_s \\ u \end{pmatrix} = \begin{pmatrix} \frac{s}{s_0} & 1-\frac{s}{s_0} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \quad \begin{pmatrix} y_s \\ v \end{pmatrix} = \begin{pmatrix} \frac{s}{s_0} & 1-\frac{s}{s_0} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y \\ v \end{pmatrix} \tag{1}$$

Introduce Lemma: The positive process of the four-dimensional light field $L(x, y, u, v)$ forming the focus stack $E(s, x_s, y_s)$ is the focusing imaging process described by the projection operator.

$$\begin{aligned} E(s, x_s, y_s) &= P[L(x, y, u, v)] \\ &= \iint \iint L(x, y, u, v)\delta(\frac{s}{s_0}x + (1 - \frac{s}{s_0})u - x_s, \\ &\quad \frac{s}{s_0}y + (1 - \frac{s}{s_0})v - y_s)dudvdxdy \end{aligned} \tag{2}$$

Here: P denotes the focused imaging process and is a bounded linear projection operator, $x_s = (s/s_o)x + (1 - s/s_o)u$ and $y_s = (s/s_o)y + (1 - s/s_o)v$ is the projected integration path.

b: ANALYTICAL MODEL OF RECONSTRUCTING 4D LIGHT FIELD DATA BY FOCUS STACK IMAGE

The forward projection process establishes the projection model relationship from the 4D light field to the focus stack. Introducing CT ideas and techniques in the solution process of the reverse focus stack reconstruction of the 4D light field. On the one hand, based on the Fourier slice theorem, the FBP method of inversely deducing the focus stack to reconstruct the 4D light field; on the other hand, starting from solving the integral equation corresponding to the focus stack, the Landweber method is used for reconstruction and iterative optimization.

Fourier slicing theorem shows that the two-dimensional Fourier transform of the image $E(s, x_s, y_s)$ at depth s is a 2D slice of the four-dimensional Fourier transform of the light field $L(x, y, u, v)$ [14], [21].

$$\boldsymbol{F}[E(s, x_s, y_s)] = \boldsymbol{L}(\omega_u, \omega_v, \omega_x, \omega_y) \tag{3}$$

where $\boldsymbol{F}[E(s, x_s, y_s)]$ is the 2D Fourier transform of image $E(s, x_s, y_s)$, $\boldsymbol{L}(\omega_u, \omega_v, \omega_x, \omega_y)$ represents the 4D Fourier transform of the light field $L(x, y, u, v)$.

The corresponding frequency domain slice is selected as:

$$w_x = (s/s_o)w_1, \quad w_y = (s/s_o)w_2,$$
$$w_u = (1 - (s/s_o))w_1, \quad w_v = (1 - (s/s_o))w_2$$

Which is:

$$\boldsymbol{F}[E(s, x_s, y_s)] = \boldsymbol{L}((1 - \frac{s}{s_0})\omega_1, (1 - \frac{s}{s_0})\omega_2, \frac{s}{s_0}\omega_1, \frac{s}{s_0}\omega_2) \tag{4}$$

Integral variable substitution based on slice selection: $dw_u dw_x = J_1 dw_1 ds$ and $dw_v dw_y = J_2 dw_2 ds$, where J_1 and J_2 are the Jacobian determinant, the available variables are replaced by:

$$dw_u dw_x = \frac{1}{s_0}|w_1| dw_1 ds \quad dw_v dw_y = \frac{1}{s_0}|w_2| dw_2 ds \tag{5}$$

Based on the frequency-domain projection relationship between the 4D light field space and the focused stack space, the inversion analytical expression is calculated to form a filtered-back-projection (FBP) algorithm [21].

$$L(x, y, u, v) = (\frac{1}{s_0})^2 \int \boldsymbol{F}^{-1}(\boldsymbol{F}(E(s, x_s, y_s))|\omega_1||\omega_2|)ds \tag{6}$$

Here, $\boldsymbol{F}$ and $\boldsymbol{F}^{-1}$ represent the Fourier transform and inverse Fourier transform, respectively, and $|\omega_1||\omega_2|$ is the optimized filter function.

c: REALIZATION OF 4D LIGHT FIELD RECONSTRUCTION BASED ON LANDWEBER ITERATIVE OPTIMIZATION ALGORITHM

The problem of reconstructing the light field focusing on the stack image is to solve the integral equation $E(s, x_s, y_s) = P[L(x, y, u, v)]$, the Landweber method [28] adopted in this paper is the descent method for solving quadratic objective functional $\|E(s, x_s, y_s) - P[L(x, y, u, v)]\|^2$. In the actual calculation, it is converted to the approximate solution of discrete linear equations, the discrete expression is as follows: $AX = B$, $A = (a_{ij})_{M\times N}$ is the projection matrix and $B = (b_1, b_2 \cdots, b_M)^T \in R^M$ is the discrete focused image M-dimensional vector, b_i is the $i-th$ focused pixel value, $X = (x_1, x_2 \cdots, x_N)^T \in R^N$ is the N-dimensional finite vector of reconstructed light field, and x_j is the $j-th$ reconstructed pixel value. In the W-norm and V-norm, the discretized form of the equation is equivalent to the weighted least squares method to solve the optimization problem:

$$X^* = \arg\min\{\frac{1}{2}||B - AX||^2_{W,V}\} \tag{7}$$

The Landweber iterative expression for reconstructing the 4D light field is:

$$X^{(n+1)} = X^{(n)} + \alpha_n V^{-1} A^T W(B - AX^{(n)}) \tag{8}$$

Here, α_n denotes a relaxation factor, the smaller α_n is, the smaller the reconstruction artifact, and the larger α_n is, the faster the convergence rate, V and W represent two positive-definite diagonal matrices.

The Landweber iterative algorithm firstly generates initial light field image $P_0[L(x, y, u, v)]$ with resolution angle $u \times u$ based on multi-frame focusing image $E(s, x_s, y_s)$ through backward projection matrix. Project the initial light field image forward to the focused image position to form the corresponding estimated focus image $E'(s, x_s, y_s)$, and calculate the error $E(s, x_s, y_s) - E'(s, x_s, y_s)$, it's the correction artifact. Finally, back-projection to the initial image of the light field to form a correction, and complete an iteration. The relaxation factor α_n affects the speed of iteration convergence and the size of artifacts, the optimal selection is based on the actual scene image reconstruction effect. The quality of the light field reconstructed image is not directly proportional to the number of iterations, after reaching a certain number of iterations, the image reconstruction quality will decline if the iteration continues.

2) SCENE DEPTH ESTIMATION AND ALL FOCUS IMAGE RECONSTRUCTION USING LIGHT FIELD DATA

The depth information of the scene is further retrieved by using the light field data. First, the scene parallax $dis^*(x, y)$ is reconstructed from the light field $L(x, y, u, v)$, and then the scene depth $dep(x, y)$ is reconstructed, finally, $dep(x, y)$ combined with the focus stack image to generate the all focus image.

In actual scene depth estimation, the estimation error is mainly composed of structural variable error and random error amount. As the real distance of the scene increases, the scene defocus becomes stronger and the focusing ability becomes weaker. At the same time, the lens rotation angle becomes smaller when the focused image is collected, and the micro lens structure variable error increases, resulting in lower depth estimation accuracy at long distances. On the other hand, light field vision measurement is affected by random errors such as image noise and lighting conditions,

and the presence of uniform areas such as weak textures in the scene causes the focus and defocus to be similar, which affects the accuracy of scene depth estimation.

a: RECONSTRUCTION OF SCENE PARALLAX BY LIGHT FIELD DATA

Based on the multi-view advantages of the 4D light field data, a pixel-by-pixel scene depth estimation can be obtained, that is, each pixel contains a depth value. In this paper, a target functional [29] with the matching term as the initial term, the gradient term and the classification term as the regular term is established to optimize the target functional to smooth the weak texture region on the basis of preserving the image edge, which is expressed as follows:

$$Func(dis(x, y)) = ||E(dis)||_{L2} + \alpha||Label(dis)||_{L1} + \beta||TV(dis)||_{L1} \quad (9)$$

Here, $||E(dis)||_{L2}$ is the matching term, $||TV(dis)||_{L1}$ is the gradient term, $||Label(dis)||_{L1}$is the classification item. α and β represent the adjustment scale.

The expression of classification item $||Label(dis)||_{L1}$ is:

$$Label(dis) \begin{cases} 1 & if\ conf(dis) \le \tau_1\ and\ E(dis) \le \tau_2 \\ -1 & if\ conf(dis) \le \tau_1\ and\ E(dis) > \tau_2 \\ 0 & others \end{cases} \quad (10)$$

Here, $conf(dis)$ is the confidence function, represents the number of pixels in the regional neighborhood $\Delta(W(x, y))$ meeting the matching conditions.

By setting threshold parameters τ_1 and τ_2, $Label(dis)$ is divided into exact match and mismatch, where 1 represents the smooth mismatching area, -1 represents the occlusion mismatching area, and 0 represents the accurate matching area.

Transform the depth of the reconstructed scene into the objective functional optimization iteration problem, and the scene parallax $dis^*(x, y)$ is obtained by solving the optimization:

$$dis^*(x, y) = \arg\min(Func(dis(x, y))) \quad (11)$$

The algorithm first obtains the initial parallax image by minimizing the block matching $||E(dis)||_{L2}$ of the reconstructed image in the light field. On this basis, the parallax image of the scene is obtained by iteratively minimizing the regular terms $||TV(dis)||_{L1}$ and $||Label(dis)||_{L1}$.

b: SCENE DEPTH ESTIMATION AND ALL FOCUS IMAGE RECONSTRUCTION

The scene depth $dep(x, y)$ can be calculated from the scene parallax image $dis^*(x, y)$ through the view point interval:

$$dep(x, y) = \frac{d'}{dis^*(x, y)} \quad (12)$$

Here, d' represents the viewpoint interval.

The schematic diagram of scene depth estimation is shown in figure 6.

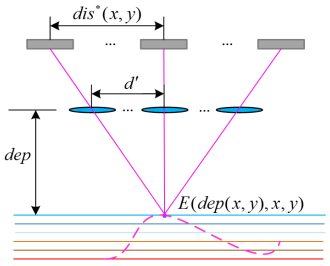


FIGURE 6. Schematic of scene depth estimation.

The scene depth image based on the reconstruction of the 4D light field data, although iterative optimization of the regular items is performed, due to the complexity of the actual scene, the pixels in the block area still have inaccurate depth estimation, which affects the accuracy of 3D reconstruction. In this paper, the non-local mean filtering (NL-means) algorithm [30] is used to perform the second optimization of the depth estimation image, the algorithm is based on the block area in the parallax reconstruction, and look for similar areas in the depth image to find the sum of the weighted average of pixels to filter and denoise, it has an optimized effect on the existence of structurally similar targets in the scene.

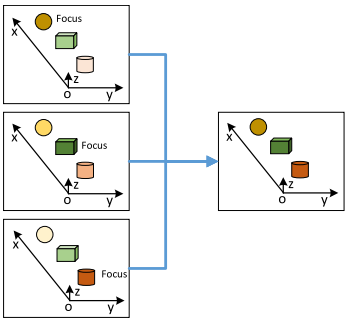


FIGURE 7. Diagram of all focus reconstruction.

The all focus image is obtained by merging the focus stack image with the depth information of each point in the depth image, the schematic diagram of the all focus reconstruction is shown in Fig. 7, the darker the color in the figure, the clearer the focus, each target in the all focus image is in focused.

The all focus model expression is:

$$I(x, y) = E(dep(x, y), x, y) \quad (13)$$

where $I(x, y)$ represents the all focus image, and the pixel value of (x, y) at the depth $dep(x, y)$ is the focal point. Assign the corresponding color value of $E(dep(x, y), x, y)$ to $I(x, y)$ to get the all focus image.

C. SCENE TARGET 3D POINT CLOUD RECONSTRUCTION WITH UPLINK AND DOWNLINK FUSION

Fusion of uplink target detection information with downlink scene depth estimation information, the all focus and depth information of the target in the scene was obtained, and the

3D point cloud image of the target was reconstructed by perspective projection model. The expression of 3D point cloud coordinate (x_w, y_w, z_w) generated from depth image $dep(x, y)$ and all focus image $I(x, y)$ [31]:

$$\begin{cases} x_w(x, y) = -(dep(x, y)^* x)/f \\ y_w(x, y) = -(dep(x, y)^* y)/f \\ z_w(x, y) = dep(x, y) \end{cases} \tag{14}$$

where f is the focal length of the camera, each pixel value of the target all focus image is rendered to the target 3D point cloud coordinate to form the 3D point cloud image $f(x_w, y_w, z_w)$ of the target:

$$f(x_w, y_w, z_w) = I(x, y) \tag{15}$$

IV. TEST VERIFICATION

Experiment uses a high-precision servo motor to drive the lens to rotate to capture the focus stack image and perform scene target detection, scene depth estimation and target 3D point cloud reconstruction. The focus stack image acquisition system includes a camera module, a highly mobile servo module and a power supply module, the camera module consists of a Point Grey industrial camera (GS3-U3-60S6M-C) and a Kowa fixed focus lens (LM35JC10M) with a focal length of 25mm, the servo module contains a 32-bit ARM embedded chip and lens drive circuit. The exposure time of each image captured by the camera is 15ms, the aperture value F takes a value of 1.6, and the gain value is set to 0. The depth range of the first two actual scenes selected in the experiment was 5 meters, and a frame of focus images were collected at an interval of 300mm, respectively. A total of 17 focus images were collected in each scene to form the scene focus stack image. In order to analyze the influence of strongly defocused stack image on the algorithm of this paper, collects 20 focused images in a scene with a depth of field of 13 meters to conduct depth estimation and all focus comparison experiments.

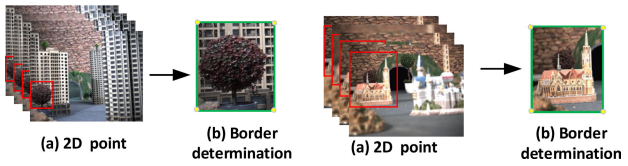


FIGURE 8. Target detection and location.

The target detection experiment of uplink focus stack image was completed with Dell PowerEdge R740 server and independent video card, the YOLO deep network learning rate parameter is set to 0.005, batch size is set to 16 according to server performance, the optimizer parameter is set to 0.95, and the number of iterations is set to 80 according to the actual training target size of the scene set. The red part in Fig. 8a shows the target detection and location of multiple focusing images in two actual scenes, respectively. Fig. 8b shows the optimal target area determined for multiple sets of target positioning information, the four corner information is obtained after target selection and frame maximization.

The light field image of the downlink focus stack reconstruction is realized by (7)-(8), in the experiment, the angle of view resolution of the light field back projection reconstruction in the experiment are usually 3×3, 5×5 and 7×7, and the angle of view resolution shown in Fig. 9 is 5×5 and 3×3's schematic diagram of light field reconstruction.

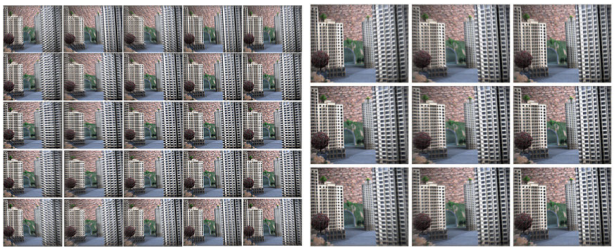


FIGURE 9. Reconstructed light fields with different resolution angles.

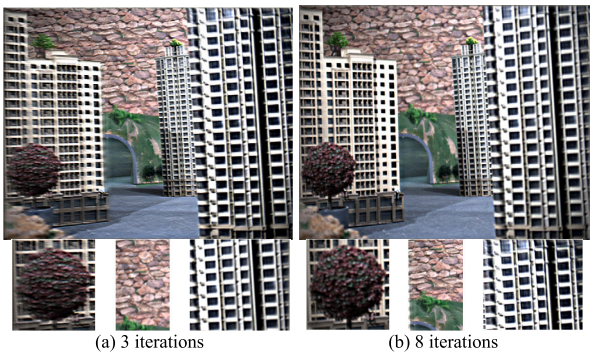


FIGURE 10. Light field reconstruction iteration result.

Fig. 10 shows the reconstruction iteration results of 3 and 8 resolving angles at 5×5 and enlarged views of some targets.

According to the above figure, as the number of iterations increases, it can be seen from the partially enlarged view that the trees, walls, and buildings are clearer in 8 iterations than in 3 iterations. In this iterative algorithm, sinusoidal window function is selected as the filter to improve the detail description of the original image and reduce the fuzziness. Fig. 11 shows the light field reconstruction diagram after sinusoidal window filtering under the iteration times of Fig. 10.

According to the figure above, the iteration results under the sinusoidal window filter function have obvious improvement in clarity and detail compared with the original. This paper selects the average gradient value of the reconstructed image as the evaluation index of the light field reconstruction clarity, Table 1 and Fig. 12 and show the average gradient of the light field image under different iteration times.

The analysis results show that the average gradient of the image increases with the increase of iteration times in the first 8 times, and decreases slightly after the 8 times, although the

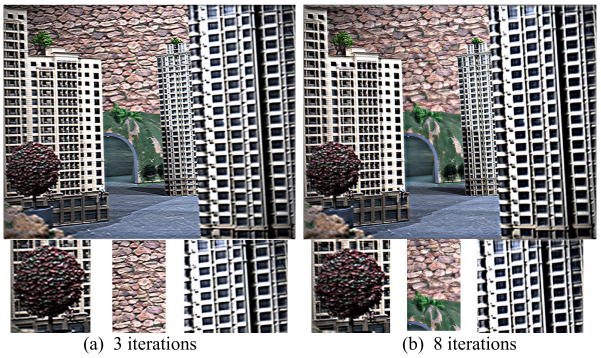


FIGURE 11. Iterative reconstruction results under sinusoidal window filtering.

TABLE 1. Average gradient of light field images with different iteration times.

Number of iterations	2	3	5	8	9	10	12	13
Average gradient	25.452	26.020	26.183	26.233	26.214	26.208	26.203	26.210

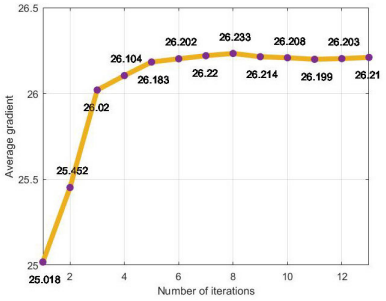


FIGURE 12. Relationship between the number of iterations and the average gradient.

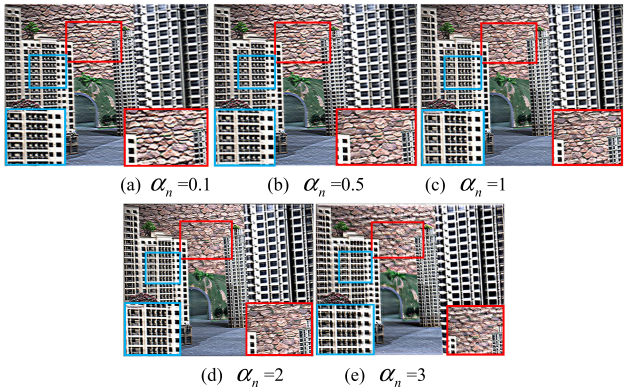


FIGURE 13. Reconstruction results under different values of relaxation factor.

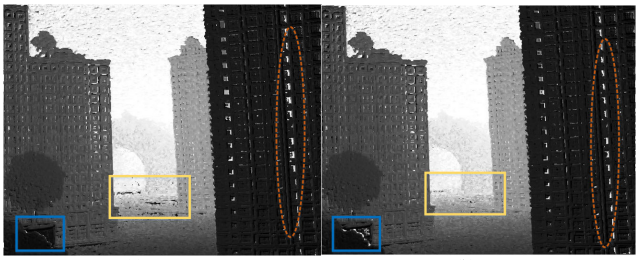


FIGURE 14. Scene depth estimation for different size block areas.

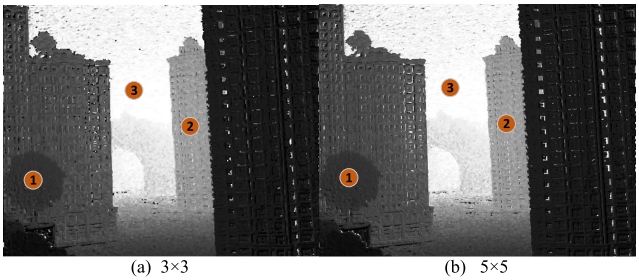


FIGURE 15. NL-means optimization of different block regions.

average gradient increases slightly in the 13th iteration, it is still lower than the result of the 8 iterations.

The increase of the relaxation factor α_n makes the iteration convergence speed faster, but the use of a larger α_n increases the correction error in the iterative correction, making the image divergence easy to produce artifacts. The experiment iterated for 5 times under the sinusoidal window filtering function, α_n was used to reconstruct the light field with 0.1, 0.5, 1, 2 and 3 respectively, and the results were shown in Fig. 13.

It can be seen from the figure that when α_n increases to 3, the artifact in the reconstructed figure has been very obvious, α_n is set as 2 to reduce the operation time while ensuring that the influence of the artifact is small.

The above 4D light field image are used to achieve the depth estimation of the scene by (9)-(12). Fig. 14 shows that the viewpoint interval is 4 mm, $\lambda = 1.5$, and the image block area size are 3×3 and 5×5 depth estimation, respectively.

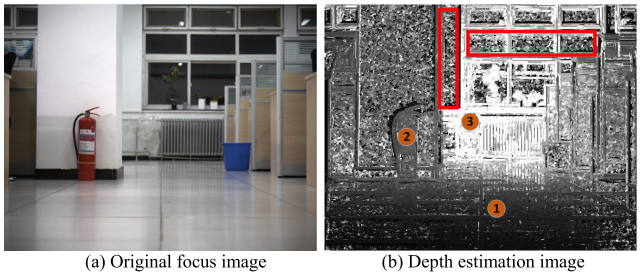
The two figures can resolve the depth well in the edge region, and the depth estimation can be reconstructed at the weak texture of the image. There are still error depth estimation areas in the scene (pixels in the rectangle and ellipse boxes), NL-means can be used for secondary depth optimization, Fig. 15 shows the block area parameters optimization results.

The optimization of NL-means can improve the consistency and estimation accuracy of depth information. Depth estimation and actual measurement of the three marked points in the depth image. Table 2 shows the results of two depth estimations for three target points.

The table above shows that the depth information of the scene obtained by 3D vision sensor is highly consistent and can achieve millimeter accuracy. In order to analyze the influence of strong defocusing on depth estimation under deeper depth of field, this article uses the method of scene 3 to estimate the depth. The original focus image and depth estimation image of scene 3 are shown in Figure 16.

TABLE 2. 3D visual depth estimation of scene.

Marks	Actual target distance（cm）	3D visual sensing（cm）	
1	160	161.8	161.3
2	275	277.4	276.8
3	500	497.6	498.3



(a) Original focus image (b) Depth estimation image

FIGURE 16. Original focus image and depth estimation of scene 3.

It can be seen from the figure that although this article sets iterative optimization to reduce the defocusing effect of weak texture areas, but at the same time, due to the increase in field depth and uneven illumination, the depth estimation accuracy of scene 3 is lower than that of scene 1. Especially in the long-distance uniform area depth estimation error is large, and some pixel depth estimation information is lost, as shown in the red box in the figure. The depth estimation information of the three target points in the selected figure is compared with the actual measurement. Table 3 shows the results of two depth estimations for the three target points.

TABLE 3. 3D visual depth estimation of scene.

Marks	Actual target distance（cm）	3D visual sensing（cm）
1	225	223.3
2	510	514.8
3	1100	1108.7

The reconstruction of the all focus image of the scene is realized by (13)-(14), Fig. 17 shows the all focus image of the three scenes.

The figure above show that the four enlarged sections in each figure have high resolution and are all focused. In the experiment, the all focus image of scene 1 and scene 3 respectively and the initial focus images were selected to evaluate the fuzziness. The gradient values of the focus images originally collected in the two scenes are shown in Fig. 18a and Fig. 18c. Fig. 18b and Fig. 18d are the average gradient values of the all focus images after the two scenes are reconstructed 5 times.

The gradient values of the original focus images in Fig. 18a are in the range of 3.18∼4.16, and the average value is 3.55,

(a) Scene 1

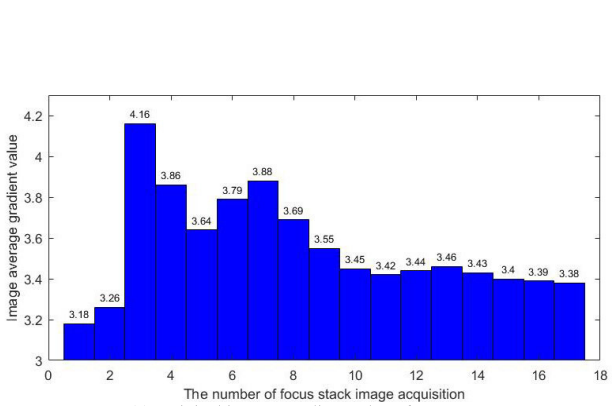
(b) Scene 2

(c) Scene 3

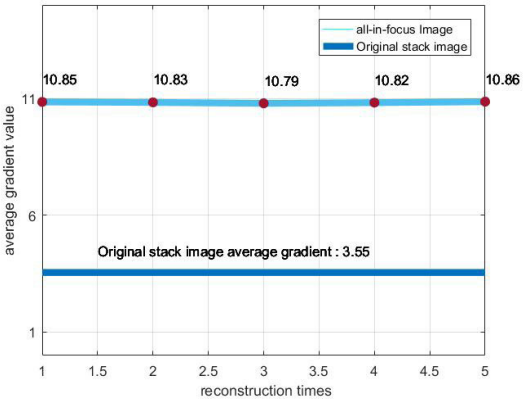
FIGURE 17. All focus image and enlarged image of three scenes.

Fig. 18b is the gradient values of the five times all focused reconstruction images, the sizes are 10.85, 10.83, 10.79, 10.82 and 10.86. The gradient values of the original focus images in Fig. 18c are in the range of 1.09∼1.80, and the average value is 1.67, Fig. 18d is the gradient values of the five times all focused reconstruction images, the sizes are 3.76, 3.72, 3.83, 3.65and 3.84. It is found that the all focus image has better global resolution and small detail presentation than the original stack diagram. Affected by conditions such as the range of depth of field and on-site care, the gradient value of the original focus image and the full focus image collected when the depth of field is deeper is smaller than the value of shallow depth of field, and the reconstruction accuracy and details are lower.

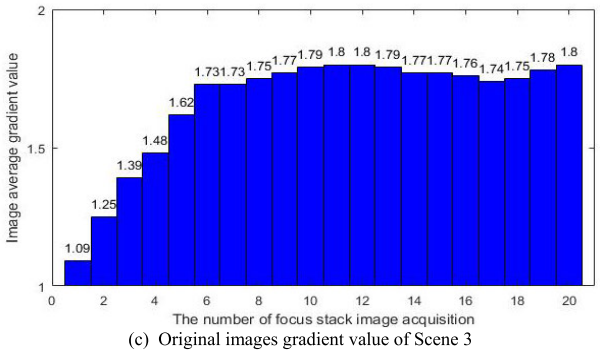
Based on the above analysis, the algorithm framework proposed in this paper is more suitable for scene depth estimation and target point cloud reconstruction in the small scene range.



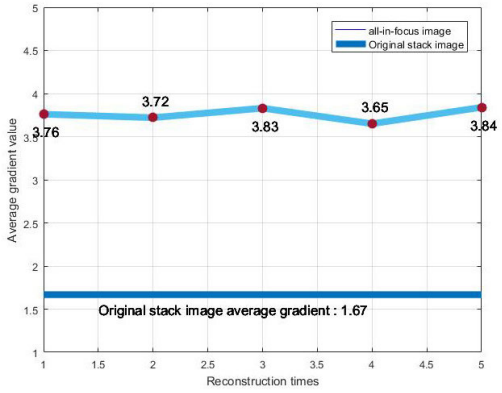
(a) Original images gradient value of Scene 1



(b) All focus image gradient value of Scene 1

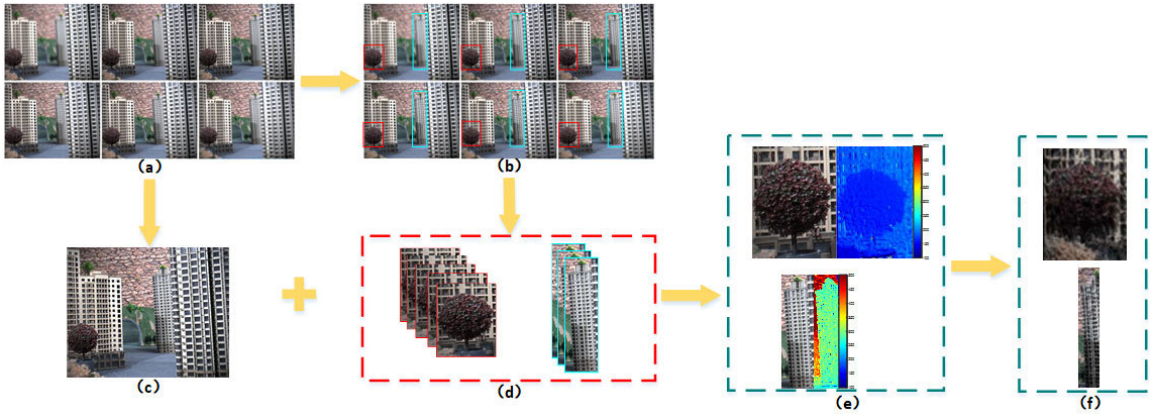


(c) Original images gradient value of Scene 3



(d) All focus image gradient value of Scene 3

FIGURE 18. Average gradient of original images and all focus image.

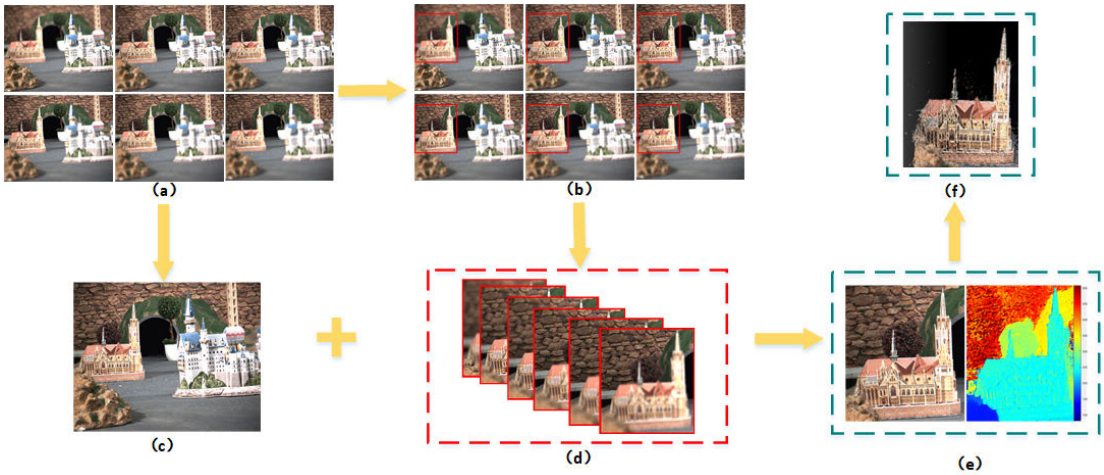


(a) Multi-frame focus stack image. (b) Target recognition in the scene. (c) All focus image. (d) Multi-frame image target positioning. (e) Target depth images and all focus images. (f) Target 3D reconstruction.

FIGURE 19. Scene 1 target reconstruction experiment diagram.

The experiment carries out the target 3D point cloud reconstruction experiment for the actual scene 1 and 2, the experiment flow chart is shown in Fig. 19 and Fig. 20 respectively.

The 3D reconstruction of the target of the two scenes mainly includes the following steps. Fig. 19a and Fig. 20a capture the focus stack image of the two scenes, and the comparison shows that the focus in the image is from near to far. Multi-target detection and positioning of the image (a) are performed through the YOLO neural network, scene 1 detects two targets in each frame of the image, including a building and a tree, and is marked with a light blue and red rectangular frame, as shown in Fig. 19b; Scene 2 detects a house in



(a) Multi-frame focus stack image. (b) Target recognition in the scene. (c) All focus image. (d) Multi-frame image target positioning. (e) Target depth images and all focus images. (f) Target 3D reconstruction.

FIGURE 20. **Scene 2 target reconstruction experiment diagram.**

each frame of the image and marks it with a red rectangular frame, as shown in Fig. 20b. Fig. 19 d and Fig. 20d show images of multi-frame target detection and optimal position information. Fig. 19c and Fig. 20c are depth estimation image and all focus image reconstructed based on image Fig. 19a and Fig. 20a. Fig. 19e and Fig. 20e are the target all focus image and depth information obtained after fusion of c and d. Finally, the 3D point cloud images of the two detection targets were reconstructed, as shown in Fig. 19f and Fig. 20f.

V. CONCLUSION

To effectively solve the problem of target detection, depth estimation and 3D reconstruction in a specific area of the scene, a fusion method based on deep learning for target detection and focus stack image reconstruction is proposed. In the algorithm framework, 3D visual perception acquires focus stack image through monocular lens, combines the advantages of deep learning and light field imaging, detects specific target area of stack image, reconstructs 4D light field and scene depth estimation, and fuses the target location and depth information of scene to achieve 3D reconstruction of the target. The algorithm in this paper is affected by the depth of field of the scene and the on-site lighting environment. In the long-distance scene, the adjustment accuracy of the micro lens is limited, and the information obtained has a large error. It can obtain more accurate depth estimation and reconstruction effects in the small scale scene. This method is not only applied in the field of robot industrial detection and rescue, but also suitable for 3D face reconstruction and recognition in the scene. The previous related work has introduced the high-resolution reconstruction and recognition of human faces based on the images collected by the 4D light field camera. The theory and algorithm flow of face recognition and 3D reconstruction based on the scene focus stack image are consistent with this article. In the actual operation, a reasonable number of focused images are selected according to the actual scene size to reconstruct the depth estimation image in the scene. Combine the YOLO algorithm to identify different human facial features in the scene, and detect facial location information, finally achieve the point cloud reconstruction of the face target. The future work includes extending the algorithm framework fusion of multiple perspectives of the scene, conducting more extensive research on target detection and reconstruction based on the background of 3D point cloud confidence assessment, and completing the reconstruction of the whole scene, provide theoretical basis and related experimental verification for the follow-up robot visual navigation and SLAM research work.

REFERENCES

[1] V. Adhikarla, J. Sodnik, P. Szolgay, and G. Jakus, "Exploring direct 3D interaction for full horizontal parallax light field displays using leap motion controller," *Sensors*, vol. 15, no. 4, pp. 8642–8663, Apr. 2015.

[2] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3487–3495.

[3] Z. Xin, D. Wei, X. Xie, M. Chen, X. Zhang, J. Liao, H. Wang, and C. Xie, "Dual-polarized light-field imaging micro-system via a liquid-crystal microlens array for direct three-dimensional observation," *Opt. Express*, vol. 26, no. 4, p. 4035, Feb. 2018.

[4] M. U. Mukati and B. K. Gunturk, "Light field super resolution through controlled micro-shifts of light field sensor," *Signal Process., Image Commun.*, vol. 67, pp. 71–78, Sep. 2018.

[5] Y. Wang, J. Yang, Y. Guo, C. Xiao, and W. An, "Selective light field refocusing for camera arrays using bokeh rendering and superresolution," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 204–208, Jan. 2019.

[6] H. Nagahara, "Programmable aperture camera using LCoS," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 1–8.

[7] Z. Xu, J. Ke, and E. Y. Lam, "High-resolution lightfield photography using two masks," *Optics Express*, vol. 20, no. 10, p. 10971, 2012.

[8] J. R. Alonso, A. Fernández, and A. José Ferrari, "Reconstruction of perspective shifts and refocusing of a three-dimensional scene from a multi-focus image stack," *Appl. Opt.*, vol. 55, no. 9, p. 2380, 2016.

[9] M. K. Singh, K. S. Venkatesh, and A. Dutta, "Accurate rough Terrain modeling from fused 3D point cloud data," in *Proc. Annu. IEEE India Conf. (INDICON)*, Dec. 2015, pp. 1–6.

[10] Y. Yang, X. Meng, and M. Gao, "Vision system of mobile robot combining binocular and depth cameras," *J. Sensors*, vol. 2017, pp. 1–11, Oct. 2017.
[11] R. A. Newcombe and A. J. Davison, "Live dense reconstruction with a single moving camera," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1498–1505, doi: 10.1109/CVPR.2010.5539794.
[12] C. Li and X. Zhang, "High dynamic range and all-focus image from light field," in *Proc. IEEE 7th Int. Conf. Cybern. Intell. Syst. (CIS)*, Jul. 2015, pp. 7–12.
[13] Y. Yuan, B. Liu, S. Li, and H.-P. Tan, "Light-field-camera imaging simulation of participatory media using Monte Carlo method," *Int. J. Heat Mass Transf.*, vol. 102, pp. 518–527, Nov. 2016.
[14] N. Ren, "Fourier slice photography," *ACM Trans. Graph.*, vol. 24, no. 3, p. 735, 2005.
[15] V. Boominathan, K. Mitra, and A. Veeraraghavan, "Improving resolution and depth-of-field of light field cameras using a hybrid imaging system," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, May 2014, pp. 1–6.
[16] 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.
[17] X. Lei, "Multiframe super-resolution reconstruction algorithm based on total variation regularization," *Electron. Meas. Technol.*, no. 1, pp. 76–79, 2012.
[18] C. Kim *et al.*, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, p. 1, 2013.
[19] N. Ren, "Light field photography with a hand held plenopic camera," Stanford Tech. Rep. CTSR 2005-02.
[20] R. Raghavendra, K. B. Raja, B. Yang, and C. Busch, "Comparative evaluation of super-resolution techniques for multi-face recognition using light-field camera," in *Proc. 18th Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2013, pp. 1–6.
[21] C. Liu, J. Qiu, and M. Jiang, "Light field reconstruction from projection modeling of focal stack," *Opt. Express*, vol. 25, no. 10, p. 11377, May 2017.
[22] 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. 1–4.
[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[24] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[25] 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, Oct. 2015.
[26] M. Levoy, "Light field rendering," in *Proc. Conf.*, Aug. 1996, p. 45.
[27] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 1996, vol. 96, no. 30, pp. 43–54.
[28] L. Landweber, "An iteration formula for fredholm integral equations of the first kind," *Amer. J. Math.*, vol. 73, no. 3, pp. 615–624, Jul. 1951.
[29] C. Liu, J. Qiu, and S. Zhao, "Iterative reconstruction of scene depth with fidelity based on light field data," *Appl. Opt.*, vol. 56, no. 11, pp. 3185–3192, 2017.
[30] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in *Proc. CVPR*, Jun. 2005, pp. 60–65.
[31] X. Zhao, C. Liu, L. Dou, J. Qiu, and Z. Su, "3D visual sensing technique based on focal stack for snake robotic applications," *Results Phys.*, vol. 12, pp. 1520–1528, Mar. 2019.

YANZHU HU was born in Beijing, China, in 1970. He received the B.S. degree in control science and engineering from the Beijing University of Aeronautics and Astronautics, in 1991, the M.S. degree in economic management from the Chinese Academy of Social Sciences, in 1997, and the Ph.D. degree in systems engineering from Beijing Jiaotong University, in 2005. From 2006 to 2007, he did his Postdoctoral Research at Beijing Jiao Tong University. Since 2008, he has been a Professor with the Beijing University of Posts and Telecommunications. He is the author of more than 15 articles, His research interests include intelligent monitoring and image processing.

YINGJIAN WANG received the B.S. degree in applied physics from Anhui Jianzhu University, Hefei, in 2014, and the M.S. degree in navigation guidance and control from the Beijing University of Science and Technology Information, Beijing, in 2018. He is currently pursuing the Ph.D. degree in control science and engineering with the Beijing University of Posts and Telecommunications, Beijing. His research interests include machine vision depth estimation and 3D scene reconstruction.

SONG WANG (Member, IEEE) received the B.S. degree in electronic information engineering from the Beijing Electronic Science and Technology Institute, Beijing, in 2010, the M.S. degree in control science and engineering from Beijing Technology and Business University, Beijing, in 2015, and the Ph.D. degree in control science and engineering from the Beijing University of Posts and Telecommunications, Beijing. His research interests include safety monitoring and intelligent signal processing.