Occluded Pedestrian Detection Based on Depth Vision Significance in Biomimetic Binocular

Wei Wei, Lidan Cheng, Yuxuan Xia, Pengcheng Zhang, Jihua Gu, and Xinyu Liu

***Abstract*— Pedestrian detection and tracking has become an important field in the field of computer vision research. However, the existing pedestrian detection algorithms have some problems, such as low accuracy and poor stability due to the similar background and overlapped occlusion interference. Therefore, an occluded pedestrian detection method based on binocular vision is proposed in this paper. We simulate the recognition of human brain and use the deep learning network MobileNet to detect and locate the initial pedestrians. Then, binocular depth is introduced as visual salience prior information, which solves the problem of identifying pedestrians with similar background and occlusion. The experimental results show that our pedestrian detection framework greatly improves the pedestrian error detection under similar background and occlusion conditions.**

***Index Terms*— Pedestrian detection, binocular vision, deep learning network, visual salience prior information.**

I. Introduction

PEDESTRIAN detection is one of the most important topics in target detection, it has been widely used in fields of intelligent video surveillance, autonomous robotic navigation, and automotive safety. Meanwhile, in the case of similar background and occlusion, it still has some problems, such as false recognition and missing recognition. Many scholars have carried out relevant research on pedestrian detection of occlusion. The methods can be roughly divided into two types: monocular-based and binocular-based. Under the traditional study of occlusion based on single-purpose processing, Gao *et al.* [1] proposed that each cell unit use a series of binary variables to correspond to whether it is a target object and separate pedestrians from the external occlusion. Ouyang and Wang [2] proposed for the first time to use occlusion likelihood map to predict occlusion information, divide gradient direction histogram (HOG) characteristic [3] carefully to separate pedestrians from occlusion, and use multi-person template to deal with occlusion problem [4]. Kaur and Bhattacharya [5] introduced a scene perception system trained using multicolumn CNN with edges, optical flow and scale space features to improve the performance of obstacle detection. Dollar *et al.* [6] estimate the occlusion degree of the target region by determining the visibility of each structural point of the target, and deal with different occlusion conditions. Hsiao and Hebert [7] effectively estimate the visibility of occlusion between objects, and predict the occlusion of occlusion in the scene of family life. Pepikj *et al.* [8] established a joint model of occlusion and occlusion to deal with occlusion problem. Yang and Ramanan [9] proposed a probability density mixture model to improve the score of visible parts to deal with partially occluded pedestrian detection. Zhang *et al.* [10] designed an attention network with self and external guidances, which can be added to the baseline FasterRCNN detector to improve the performance of occluded cases. Fernandez *et al.* [11] proposed a hybrid component model, which makes the component filter more compact hinged structure to make the detection more accurate. Sheng *et al.* [12] improved R-FCN proposed by Dai *et al.* [13] to strengthen the learning of occlusion and samples similar to the background, and solved the problem of missing and misclassification.

On the other hand, it is well known that the human binocular system can judge the three-dimensional information of the target object, but the monocular observation will lose the ability to judge the distance. Therefore, binocular camera can get more information than monocular camera, such as parallax and depth. The idea of using binocular information to help pedestrian detection has been proposed for a long time. In [14], a pedestrian detection system based on binocular vision is proposed. Extracting 3D Point Map based on canny feature and filtering according to neighborhood Standard. Candidates use subtraction clustering mechanism to locate and support vector machine (SVM) classifier to classify each candidate, then Kalman filter is used to improve the performance of pedestrian detection system. To solve the problem that features such as color, texture can be affected by illumination easily, Jia *et al.* [15] adopt Hu moment to detect human feature from human's disparity image. Mammeri *et al.* [16] proposed a binocular distance measurement of pedestrian detection system based on key points. HOG-SVM pedestrian detection method is used to detect pedestrians in left and right images. They also proposed a cross-redetection mechanism to improve the robustness of the detection method. All the above methods all needs considerable time because of complex algorithms,

Manuscript received April 30, 2019; revised July 5, 2019; accepted July 5, 2019. Date of publication July 18, 2019; date of current version November 13, 2019. This work was supported in part by the Natural Science Research General Program of Higher Education of Jiangsu Province under Grant 16KJB510040 and in part by the Jiangsu Students' Platform for Innovation and Entrepreneurship Training Program under Grant 5731511816. The associate editor coordinating the review of this paper and approving it for publication was Prof. Kazuaki Sawada. *(Corresponding author: Wei Wei.)*

W. Wei, L. Cheng, Y. Xia, P. Zhang, and J. Gu are with the School of Optoelectronic Science and Engineering, Soochow University, Suzhou 215006, China (e-mail: weiwei0728@suda.edu.cn).

X. Liu is with the Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON M5S 2E8, Canada.

Digital Object Identifier 10.1109/JSEN.2019.2929527

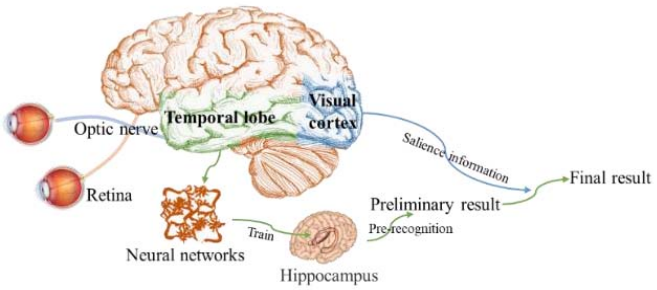

Fig. 1. The recognition process of human brain.

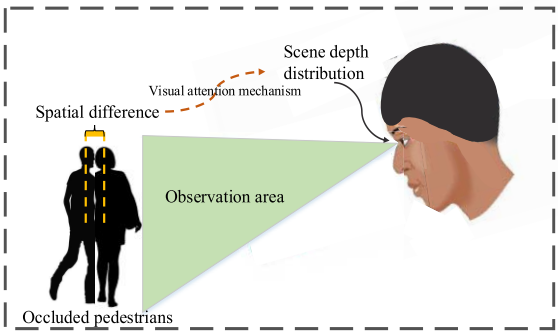

Fig. 2. Human eye attention mechanism based on depth information.

Fig. 3. Our binocular hardware system.

low computational efficiency, and are inapplicable to mobile platform.

In view of the above problems, we propose an occluded pedestrian detection based on binocular depth visual significance, which has the following contributions:

(1) Using a lightweight depth neural network MobileNet with low delay and high efficiency to detect and locate pedestrians. The delay, scale and accuracy of pedestrian detection model can be well balanced;
(2) A binocular hardware system was built to obtain the binocular gray parallax image, and the binocular visual salient mechanism was introduced to assist pedestrian detection;
(3) In visual salience mechanism, we improved the matching algorithm and emphasis targets. Besides, background suppression and the distinction between occluded and occluded targets are carried out on the detected area. Experiments show that the algorithm has good performance in the detection of occluded pedestrians.

The rest of this article is organized as follows: In Section 2, we briefly introduce the mechanism of visual attention, the related work of visual salience detection technology and the binocular hardware system that we built. Section 3 presents the network model we used for pedestrian detection. In Section 4, we introduce binocular depth as visual salience prior information based on the model in Section 3 and give the result of occluded pedestrian detection after target salience distinction. The comparison between the pedestrian detection result of MobileNet and our framework is given in Section 5. Sections 6 conclude the paper.

II. Related Work

Human visual system can deal with objects of interest selectively in complex scenes, and this adaptive selection ability is called visual attention mechanism [17]. As is shown in Fig. 1, images formed on the retina pass through the optic nerve to the brain and memories are stored in the hippocampus of temporal lobe after neural network training. When a target needs to be identified, preliminary result can be quickly identified in advance through the trained network. Then, according to the attention mechanism of human visual system, salience information can be obtained from visual cortex and the final result of recognition can be achieved.

In the field of computer vision, the attention mechanism of imitating human eyes is defined as the image salience detection technique, which is used to help the computer to quickly extract the target area of the image. As is shown in Fig. 2, human vision is constructed by binocular visual system and visual attention mechanism. Visual attention mechanism receives visual signals through the binocular vision system, and achieve objects based on the results of visual information rapidly [18]–[21]. The binocular vision system not only gives people the perception of colors, luminance and other essential information from images, but also obtains the depth distribution information of the scene according to the depth distribution feature to determine the shape, orientation, and distance of objects. This shows that the depth information distribution is also an important visual perception information, which can be used as a salient feature to deal with the regular pedestrian detection in cooperation, which can more accurately accomplish the target saliency detection under the similar background and occlusion conditions.

Inspired by the recognition process of human brain and visual properties mentioned above, we build a binocular hardware system and proposed a new framework for pedestrian occlusion detection based on depth vision salience in bionic binocular system. Our binocular hardware system is shown in Fig. 3, it is established by two same USB cameras.

Our framework is shown in Fig. 4. First, a small, low delay and low power consumption parameterized network is trained to get the preliminary detection and location of the pedestrian. Then, we obtain gray parallax depth map by binocular parallax method and enhance the area of depth maps with different colors. Then, we restrain the background influence of the target area with the weighted fusion of depth parallax salience map and global saliency graph obtained by FT (frequency-tuned salient region detection) method, and separately recognize

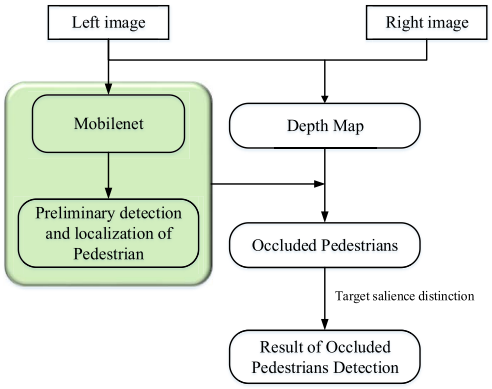

Fig. 4. Our framework for pedestrian occlusion detection.

front and rear occluded pedestrians with the salience information of depth map.

III. Pedestrian Detection

In the binocular vision system, we use the MobileNet convolution network to judge and locate the pedestrian. MobileNet is a lightweight deep neural network constructed from deep separable convolution [22]. It is a small, low delay and low power consumption parameterized model for mobile devices and embedded applications.

The input F dimension of the network is $D_F \times D_F \times M$ (M represents the channel of input), and the output G dimension is $D_F \times D_F \times N$(N represents the channel of output). Fig. 5 shows that MobileNet decomposes the standard convolution into a depthwise convolution (DWC) and a pointwise convolution (PWC). For standard convolution, the size of convolution kernel K is $D_K \times D_K \times M \times N$. After decomposition, the standard convolution filter is replaced by a set of M convolution kernels which size is $D_K \times D_K \times M$ and a set of N convolution kernels which size is $1 \times 1 \times M$. In order to build a network with smaller and less computation, a hyperparametric width multiplier α is introduced. The second hyperparametric resolution multiplier ρ is introduced to change the resolution of input layer. Now, the computational cost for the core layers of our network as depthwise separable convolutions can be expressed:

$$D_K \times D_K \times \alpha M \times \rho D_F \times \rho D_F + \alpha M \times \alpha N \times \rho D_F \times \rho D_F, \quad (1)$$

where $\alpha \in (0, 1]$, $\rho \in (0, 1]$. Resolution multiplier has the effect of reducing computational cost by ρ^2.

MobileNet network consists of 28 layers, the first layer is a standard convolution (SC), and a final average pooling (Avg Pool) reduces the spatial resolution to 1 before the fully connected (FC) layer. Counting depthwise convolutions (DWC) and pointwise convolutions (PWC) as separate layers.

IV. Get Salience Distinction of Fusion Binocular Disparity Map

The binocular can look at the same object from two different angles at the same time [23]. The different angle of view gives rise to the position difference in the direction, which

TABLE I
MobileNet Body Architecture

Type	Stride	Filter Shape	Input Size
SC	S=2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
DWC	S=1	$3 \times 3 \times 32$	$112 \times 112 \times 32$
PWC	S=1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
DWC	S=2	$3 \times 3 \times 64$	$112 \times 112 \times 64$
PWC	S=1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
DWC	S=1	$3 \times 3 \times 128$	$56 \times 56 \times 128$
PWC	S=1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
DWC	S=2	$3 \times 3 \times 128$	$56 \times 56 \times 128$
PWC	S=1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
DWC	S=1	$3 \times 3 \times 256$	$28 \times 28 \times 256$
PWC	S=1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
DWC	S=2	$3 \times 3 \times 256$	$28 \times 28 \times 256$
PWC	S=1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5 \times$ DWC / PWC	S=1	$3 \times 3 \times 512$ $1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$ $14 \times 14 \times 512$
DWC	S=2	$3 \times 3 \times 512$	$14 \times 14 \times 512$
PWC	S=1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
DWC	S=2	$3 \times 3 \times 1024$	$7 \times 7 \times 1024$
PWC	S=1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool	S=1	7×7	$7 \times 7 \times 1024$
FC	S=1	1024×1000	$1 \times 1 \times 1024$
Softmax	S=1	Classifier	$1 \times 1 \times 1000$

is called parallax. The parallax set of every pixel point is a disparity map. As shown in Fig. 6, disparity map is obtained after matching pair. Spatial relationship between objects can be obtained according to the size of parallax. Using spatial location information, target with similar background or front and rear occlusion can be distinguished.

In the process of constructing parallax map, coordinates of characteristic points and 64-bit descriptor of images are obtained from left and right camera with SURF method. Characteristic points of left and right images are defined as

$$\begin{cases} G^L = \left\{ \left(x'_1, y'_1\right), \left(x'_2, y'_2\right), \ldots, \left(x'_i, y'_i\right), \ldots, \left(x'_m, y'_m\right) \right\}, \\ \qquad 1 \le i \le m \\ G^R = \left\{ \left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_j, y_j\right), \ldots, \left(x_n, y_n\right) \right\}, \\ \qquad 1 \le j \le n, \end{cases} \quad (2)$$

where G^L represents feature point parameters of the left camera image, G^R represents feature point parameters of the right camera image, m and n represent the number of characteristic points of left and left image respectively, i and j represent the subscript of the left eye image and the right eye image feature point respectively.

Euclidean distance of all characteristic point parameters G^L in left image and characteristic point parameters G^R in right

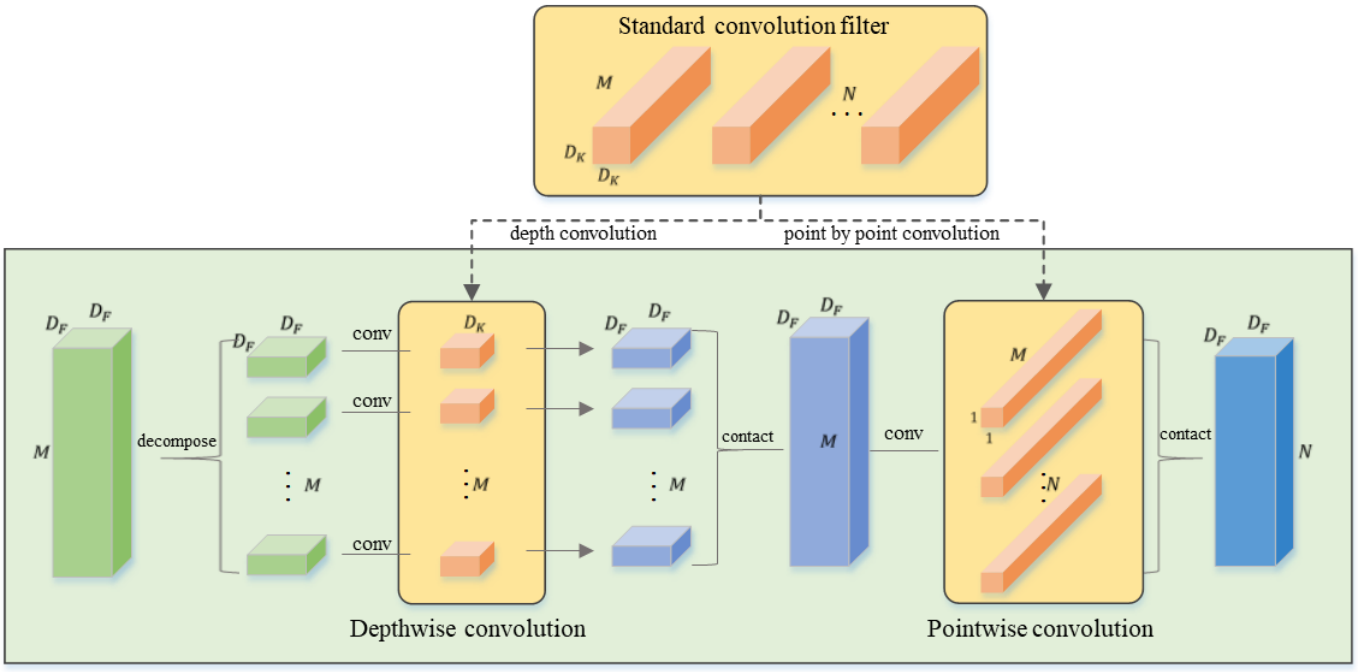

Fig. 5. The standard convolution filter is replaced by two layers: depthwise convolution (dwc) and pointwise convolution (pwc) to establish a depth detachable filter.

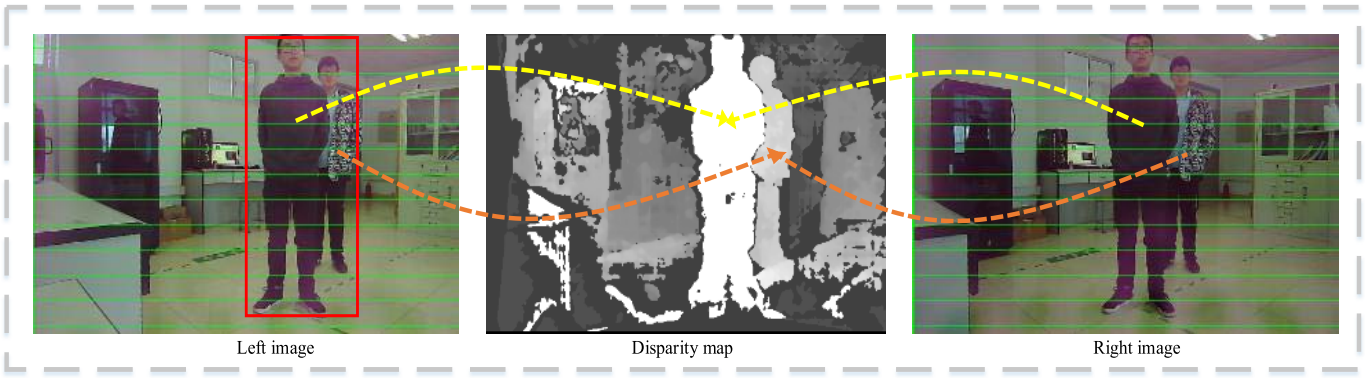

Fig. 6. Disparity map is obtained after matching pair.

image. The point with the smallest Euclidean distance is selected as the rough matching point. Rough matching points are ordered in ascending order of Euclidean metric and outliers are delete. We define the first K matching points as:

$$G^K = \left\{ \begin{matrix} \left\{ \left(x_1', y_1'\right), \left(x_1, y_1\right) \right\}, \ldots, \left\{ \left(x_i', y_i'\right), \left(x_j, y_j\right) \right\}, \\ \ldots, \left\{ \left(x_K', y_K'\right), \left(x_K, y_K\right) \right\} \end{matrix} \right\}, \quad 1 \le i, j \le K. \tag{3}$$

These points are matched to obtain discrete visual handicap map and compute visual handicap projection. Regional visual difference $\mathcal{L}_{Rj}$ can be figured. The greater regional parallax is, the higher significantly depth of the region is

$$L_{Rj} = \begin{cases} median\left(\{p_1, p_2, \ldots, p_n\}\right), & n \ge 3 \\ \min\left(\{p_1, p_2\}\right), & 0 < n < 3 \\ 0, & n = 0, \end{cases} \quad j \in [1, m], \quad i \in [1, n], \tag{4}$$

where j represents the label value of the merged region, m represents the number of merged regions, n represents the total number of parallax points falling into a region. p_i represents the visual difference of the i visual handicap in the region.

Target area obtained by preliminary pedestrian detection includes occluded pedestrian and background. The FT (frequency-tuned salient region detection) algorithm was proposed by Achanta *et al.* [24]. It is simple to calculate and can detect significant target regions effectively. Therefore, the fusion salient map of target area can be obtained with the weighted fusion of depth parallax salience map and global saliency graph obtained by FT method.

$$\begin{cases} H_{fuse}(x, y) = w_1 \cdot H_{parallax}(x, y) + w_2 \cdots H'_{FT}(x, y) \\ H_{FT}(x, y) = \left\| \left[\bar{L}\bar{a}\bar{b}\right] - I_G(x, y) \right\| \\ H'_{FT}(i) = \dfrac{1}{n} \sum_{j=1}^{n} H_{FT}(x, y), \end{cases} \tag{5}$$

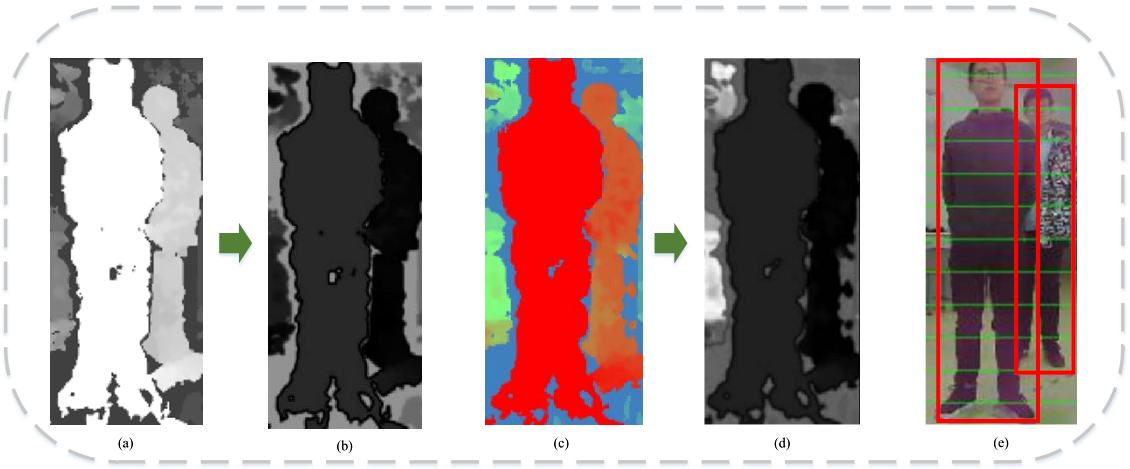

Fig. 7. (a) target region disparity map; (b) the final fusing saliency map of (a) and FT map; (c) target region color disparity map; (d) the final fusing saliency map of (c) and FT map; (e) the result of occluded pedestrian detection after target salience distinction.

TABLE II
THE ACCURACY AND RUNTIME OF DIFFERENT METHODS

Authors	Accuracy	Runtime(fps)
Andrew G. Howard [22]	85.4%	75fps
Our Framework	87.7%	34fps

where H_{fuse} is the fusion salient map, $\left[\bar{L}\bar{a}\bar{b}\right]$ is the arithmetic mean of the three components L, a, b of an image. $I_G\,(x, y)$ is a smooth image used Gaussian blur. $H_{parallax}\,(x, y)$ represents sparse parallax graph, H'_{FT} represents the mean saliency of the global salience map detected by FT method in each partition region. i represents the i combined region, j represents the j pixel point in the region, and n represents the number of pixel points in the region. w_1 and w_2 are weights.

We sort the salience values of all merged regions and take 80% of the value as the threshold. If the significant value is below the threshold, the region is considered to be the background area and set the significant value of the region to 0 to reduce the impact of background areas. We get depth significant map after regional background suppression H_{final}:

$$H_{final} = \begin{cases} H_{fuse} & H_{fuse} > 0.8 \times median\left(H_{fuse}\right) \\ 0 & otherwise. \end{cases} \tag{6}$$

Then, we transform grey disparity map into color disparity map to enhance significance of occluded targets and recognize separately.

V. EXPERIMENTAL RESULT

We evaluate our approach in actual scene. The contrast of disparity map is shown in Fig. 6. 500 occluded Pedestrian images are obtained to prove our framework have advantages in detecting occluded Pedestrian. Experimental results indicate that the accuracy of our framework is increased by 2.3% than MobileNet. Although the runtime of our framework is dropped by 41 fps (frames per second), it is fully applicable to mobile platforms and embedded vision applications.

VI. CONCLUSIONS

A new framework for occluded pedestrian detection based on depth visual significance in bionic binocular vision is proposed in this paper:

First, we preliminary judge and locate pedestrians with a low delay and low power convolutional network MobileNet based on the left camera. Second, we obtain depth parallax salience map by binocular matching and weight fuse with global salient graph obtained by FT method to restrain the background influence of the target area. Then, we distinguish the occluded pedestrians separately by the significance of the depth map. The experimental results show that:

(1) Compared with the pure network, our pedestrian detection framework where visual saliency information is introduced reduces the pedestrian error detection under similar background and occlusion, and improves the accuracy of pedestrian detection by 2.3%.
(2) The performances of low delay, fast response and high-precision indicate that our framework is suitable for mobile and embedded vision applications to a certain extent.

VII. CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] T. Gao, B. Packer, and D. Koller, "A segmentation-aware object detection model with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1361–1368.

[2] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3258–3265.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 886–893.

[4] W. Ouyang and X. Wang, "Single-pedestrian detection aided by multi-pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3198–3205.

[5] B. Kaur and J. Bhattacharya, "Scene perception system for visually impaired based on object detection and classification using multimodal deep convolutional neural network," *Proc. SPIE*, vol. 28, no. 1, Feb. 2019, Art. no. 013031.

[6] 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.

[7] E. Hsiao and M. Hebert, "Occlusion reasoning for object detectionunder arbitrary viewpoint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3146–3153.

[8] B. Pepikj, M. Stark, P. Gehler, and P. Gehler, "Occlusion patterns for object class detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3286–3293.

[9] W. U. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[10] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6995–7003.

[11] D. Fernandez *et al.*, "Pedestrian recognition for intelligent transportation systems," in *Proc. ICINCO*, Sep. 2005, pp. 292–297.

[12] J. Sheng *et al.*, "Pedestrian detection method based on R-FCN," *Comput. Eng. Appl.*, vol. 54, no. 18, pp. 180–183, Nov. 2017.

[13] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region based fully convolutional networks," in *Proc. IEEE Conf. Appl. Comput. Vis.*, Dec. 2016, pp. 379–387.

[14] X. Liu and Z. Jin, "A pedestrian detection system based on binocular stereo," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2012, pp. 1–6.

[15] S. Jia, L. Zhao, J. Sheng, X. Li, and W. Cui, "Performance improvement of human detecting and tracking based on stereo vision," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2010, pp. 1762–1767.

[16] A. Mammeri, A. Boukerche, and M. Zhao, "Keypoint-based binocular distance measurement for pedestrian detection system," in *Proc. ACM Int. Symp. ACM*, Sep. 2014, pp. 9–15.

[17] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[18] S. Yantis, "To see is to attend," *Science*, vol. 290, no. 5603, pp. 54–56, Jan. 2003.

[19] N. Kanwisher and E. Wojciulik, "Visual attention: Insights from brain imaging," *Nature Rev. Neurosci.*, vol. 1, no. 2, pp. 91–100, Nov. 2000.

[20] H.-C. Nothdurft, "Attention shifts to salient targets," *Vis. Res.*, vol. 42, no. 10, pp. 1287–1306, May 2002.

[21] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it," *Nature Rev. Neurosci.*, vol. 5, no. 6, pp. 495–501, Jun. 2004.

[22] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861

[23] Z. Zhang, W. Tao, K. Sun, W. Hu, and L. Yao, "Pedestrian detection aided by fusion of binocular information," *Pattern Recognit.*, vol. 60, pp. 227–238, Dec. 2016.

[24] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.

Wei Wei received the Ph.D. degree from Harbin Engineering University, Heilongjiang, China. From 2009 to 2010, he was a Visiting Scholar with Kagawa University, Kagawa, Takamatsu, Japan. He is currently an Associate Professor with the Department of Optoelectronic Science and Engineering, Soochow University.

His research interests include the development of exoskeleton robot, optical sensor, pattern recognition, and intelligent systems. His awards and honors include the Best Paper Award at IEEE ICME 2013 and best paper nomination awards at IEEE ICMA 2018.

Lidan Cheng received the B.E. degree from the School of Optoelectronic Science and Engineering, Soochow University, Suzhou, Jiangsu, China, in 2018, where she is currently pursuing the M.E. degree in detection technology and automatic equipment.

Her research interests include pattern recognition and intelligent systems, machine vision, and image processing.

Yuxuan Xia received the B.E. degree from the School of Optoelectronic Science and Engineering, Soochow University, Suzhou, Jiangsu, China, in 2018, where he is currently pursuing the M.E. degree in detection technology and automatic equipment.

His research interests include the development of exoskeleton robot and intelligent robot.

Pengcheng Zhang received the B.E. degree from the School of Optoelectronic Science and Engineering, Soochow University, Suzhou, Jiangsu, China, in 2018, where he is currently pursuing the M.E. degree in detection technology and automatic equipment.

His research interests include the development of exoskeleton robot and intelligent robot.

Jihua Gu received the M.E. degree from Tongji University, Shanghai, China, in 1984, and the Ph.D. degree from Ruhr-Universität Bochum, Germany, in 1993.

He is currently a Professor with the Department of Physics and Technology, Soochow University. His research interests include the study on photothermic nondestructive characterization of physical properties of materials, digital image, and speech processing, and the application of photoelectric technology.

Xinyu Liu is an Associate Professor with the Department of Mechanical and Industrial Engineering and the Department of Biomedical Engineering, University of Toronto, Canada. He is also the Research Chair of the National Microfluidic and Bio-MEMS Technology Research and a member of the national institutes of the Health Funding Board for Biomedical Engineering. Dr. Liu was a recipient of the International Summit on Microsystems and Nano-Engineering Young Scientist Award for Excellence in 2018, the McGill University Christophe Pierre Award for Outstanding in 2017, and the Canadian Grand Challenges Foundation Global Health Star Award in 2012. His research achievements have won five best paper awards and eight best paper nomination awards at IEEE and ASME international conferences. He was an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, the IEEE ROBOTICS AND AUTOMATION LETTERS, the *International Journal of Advanced Robotic Systems*, and the *IET Cyber-Systems and Robotics*. He also holds the post of editorial committee of ICRA, IROS, and CASE IEEE international conferences on robotics and automation.