Computers, Materials & Continua
DOI: 10.32604/cmc.2022.027083
Article

Tech Science Press

Improved Lightweight Deep Learning Algorithm in 3D Reconstruction

Tao Zhang[1,*] and Yi Cao[2]

[1]School of Mechanical Engineering, North China University of Water Conservancy and Hydroelectric Power, Zhengzhou, 450045, China
[2]Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, N9B 3P4, Canada
*Corresponding Author: Tao Zhang. Email: ztncwu@126.com
Received: 10 January 2022; Accepted: 04 March 2022

Abstract: The three-dimensional (3D) reconstruction technology based on structured light has been widely used in the field of industrial measurement due to its many advantages. Aiming at the problems of high mismatch rate and poor real-time performance caused by factors such as system jitter and noise, a lightweight stripe image feature extraction algorithm based on You Only Look Once v4 (YOLOv4) network is proposed. First, Mobilenetv3 is used as the backbone network to effectively extract features, and then the Mish activation function and Complete Intersection over Union (CIoU) loss function are used to calculate the improved target frame regression loss, which effectively improves the accuracy and real-time performance of feature detection. Simulation experiment results show that the model size after the improved algorithm is only 52 MB, the mean average accuracy (mAP) of fringe image data reconstruction reaches 82.11%, and the 3D point cloud restoration rate reaches 90.1%. Compared with the existing model, it has obvious advantages and can satisfy the accuracy and real-time requirements of reconstruction tasks in resource-constrained equipment.

Keywords: 3D reconstruction; feature extraction; deep learning; lightweight; YOLOv4

1 Preface

Optical three-dimensional (3D) measurement technology [1] is one of the most important research fields and research directions in optical measurement. As an important method of three-dimensional measurement technology, striped structured light technology can quickly and accurately obtain 3D point cloud data on the surface of the measured object, and is widely used in quality inspection, cultural relic protection, human-computer interaction, biomedicine and other fields [2,3]. The basic process of the measurement algorithm is as follows: project one or a group of structural fringes onto the surface of the object, the camera captures the fringe image modulated by the height of the object, and the relevant algorithm is used to calculate the phase information carried in the fringe; according to the phase and height, world coordinates and image pixel coordinates The mapping relationship between to get the final 3D information. Techniques such as fringe analysis, phase extraction and

phase unwrapping have an important influence on the accuracy of 3D measurement. How to obtain high-precision depth information from the fringe image of the measured object is still the focus and difficulty of fringe projection 3D measurement technology.

Algorithms for obtaining depth information (or unfolding phase) from fringe images usually require two main steps: phase extraction represented by phase shifting and Fourier transform methods [4] and phase unfolding represented by spatial phase unwrapping and time phase unwrapping [5,6]. With the successful application of deep learning in the field of 3D measurement, 3D reconstruction technology [7–9] based on Convolutional Neural Network (CNN) has been continuously developed. The typical representative is Region-Convolutional Neural Network (R-CNN) series of algorithms based on region selection, but these methods take a long time to detect and cannot achieve the effect of real-time detection; Single Shot MultiBox Detector (SSD) [10] fusion multi-scale detection model has improved speed, but it detects small targets. Insufficient performance. The YOLO [11–14] series of algorithms are one of the most widely used algorithms in the field of deep learning. YOLOv1 has a fixed input size and has a poor detection effect on objects that occupy a relatively small area; YOLOv2 removes the fully connected layer and improves the detection speed; YOLOv3 obtained better detection performance, and can effectively detect small target objects, without a significant increase in speed. YOLOv4 is the fourth version of the YOLO series of algorithms, and its accuracy and speed have been significantly improved. With the increasing expansion of neural network model scale and increasing parameter scale, it needs to consume a lot of computing and storage resources, making it difficult to integrate into mobile terminals with limited resources, such as mobile phones and tablet computers.

Parameter compression on the constructed YOLOv4 model can well solve the contradiction between the huge network model and the limited storage space. The currently widely used model parameter compression method has weighted parameter quantization [15], Singular Value Decomposition (SVD) method [13] and so on.

The weight parameter quantization can achieve the purpose of reducing resource consumption by reducing the accuracy of the weight. For example, in common development frameworks [16–18], the activation and weight of neural networks are usually represented by floating-point data. Using low-level fixed-point data or even a small portion of training values to replace floating-point data helps reduce the bandwidth and storage requirements of the neural network processing system. The disadvantage is decreased data accuracy has caused a decrease in classification accuracy, and at the same time, the compression effect is difficult to improve. Peng [19] and others have greatly reduced the model parameters and resource occupation by adding the Ghost module and the Shuffle Conv module, but the accuracy is reduced by 0.2% compared with the original network. The SVD decomposition law achieves the purpose of reducing resource consumption by reducing the number of weights. Literature [20] proposed a global average pooling algorithm to replace the fully connected layer. Google Net uses this algorithm to reduce the scale of network training, and the removal of the fully connected layer does not affect the accuracy of image recognition. Google Net uses this algorithm to reduce the scale of network training, and the removal of the fully connected layer does not affect the accuracy of image recognition. The recognition accuracy of the algorithm in Image Net reaches 93.3%. At the same time, literature [21] proposed a $1*1$ convolution kernel, which was successfully applied to Google Net and Res Net, which played a role in reducing the amount of parameters.

This paper uses the YOLOv4 network model to extract the features of the striped structured light image. Considering that the features of the striped image are not obvious due to the influence of illumination and noise, the feature extraction network model is improved. The algorithm first uses

Mobilenetv3 structure to replace Cross-stage partial Darknet53 (CSPDarknet53) network of YOLOv4 to reduce the amount of backbone network parameters, and then introduces the Mish activation function and the CIoU loss function to calculate the improvement of the target frame regression loss, which effectively improves the generalization of feature extraction.

2 3D Reconstruction Algorithm

2.1 Stripe Structured Light 3D Reconstruction Algorithm

The principle of the fringe structured light 3D reconstruction algorithm is shown in Fig. 1. Assume that the light beam projected by the projection system intersects the reference plane at point B, which is imaged at point C on the camera image plane. When the object is placed, it suppose that another light beam intersects the object at point D, which is also imaged at point C in the camera image plane. For point C in the phase plane, there are two phase values before and after the object is placed. Therefore, the height h of point D can be derived from the phase difference.

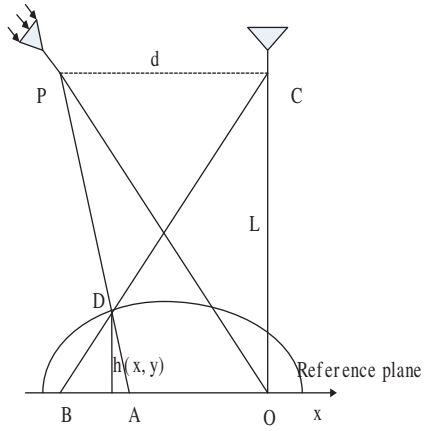

Figure 1: 3D measurement principle of structured light

The phase shift method is one of the commonly used methods of the fringe structured light 3D reconstruction technology. By projecting a series of fringe images with a phase shift of I_n (x, y) to the reconstruction target $2\pi/n$, the wrapped phase of the standard phase shift method is:

$$\varphi(x,y) = \arctan \frac{\sum_{n=0}^{N-1} I_n(x,y) \cdot \sin(2\pi n/N)}{\sum_{n=0}^{N-1} I_n(x,y) \cdot \cos(2\pi n/N)} \tag{1}$$

The wrapping phase is discontinuous, and the value range is between $[-\pi, \pi]$. The unfolding phase ϕ (x, y) required in the subsequent three-dimensional reconstruction work is obtained by phase unwrapping. Phase unwrapping aims to recover the continuous phase from φ (x, y), and reconstruct the physically continuous phase change by adding or subtracting an appropriate multiple k (x, y) of 2π, thereby eliminate phase jumps. Therefore, the relationship between the unfolding phase and the wrapping phase is as follows:

$$\phi(x, y) = \varphi(x, y) + 2\pi k(x, y) \tag{2}$$

Finally, the mapping expression between the unfolding phase and the height is determined and calibrated the mapping coefficients to realize the conversion of depth data and phase data of the measured object, and obtain the 3D topography information of the object surface.

2.2 YOLOv4 Network

YOLOv4 is mainly composed of Backbone, Neck and Head, as shown in Fig. 2. The Backbone part of YOLOv4 uses the CSPDarknet53 network, which is based on the Darknet53 network of YOLOv3 and formed by drawing on the ideas of CSPNet [22]. The Neck part is composed of the Spatial Pyramid Pooling Networks (SPPNet) structure and Path Aggregation Network (PANet). SPPNet is a spatial pyramid pooling network that can increase the receptive field of the network, and the PANet network is a path aggregation network that realizes the integration of deep features and shallow features of the Backbone network. In the head detection part, the YOLOv4 algorithm uses the YOLOv3 detection head to perform two convolution operations with a size of 3×3 and 1×1 to complete the detection.

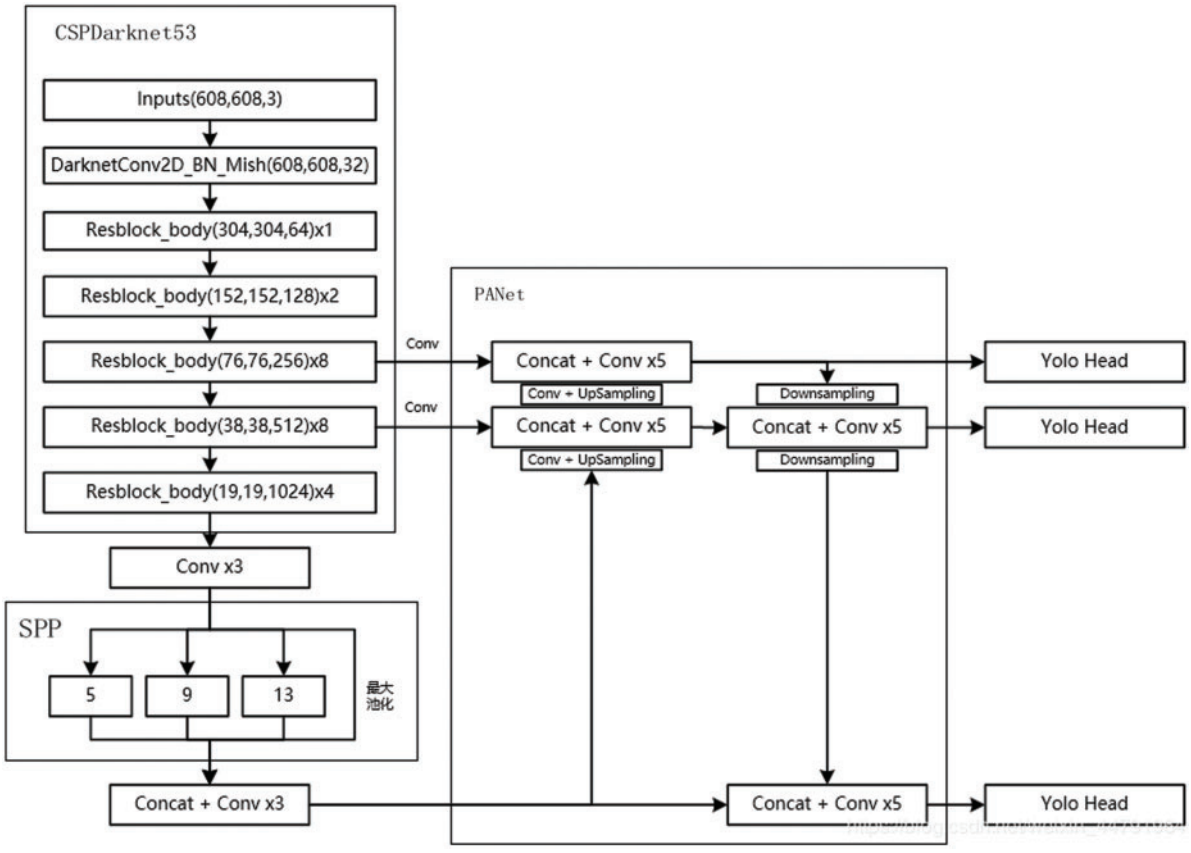

Figure 2: Stucture of YOLOv4

2.3 Network Model Compression

YOLOv4 network model is improved from two aspects: using the MobileNetV3 structure to replace the backbone feature extraction network of YOLOv4, and greatly reducing the number of backbone network parameters through the deep separable convolution in Mobilenetv3; introducing Mish activation function and CIoU loss function calculation to improve target frame regression loss, effectively improve the generalization of feature extraction.

YOLOv4 algorithm uses the CSPDarknet53 network as the feature extraction network, which contains 5 residual blocks, which are respectively stacked by 1, 2, 8, 8, and 4 residual units. The algorithm has a total of 104 convolutional networks, including 72 convolutional layers, and uses a large number of standard 3×3 convolution operations. A large amount of computing resources are used in the calculation process, which makes it difficult to achieve real-time performance. With the transfer of multi-layer features, more convolutional layers will gradually reduce the ability of local refined feature extraction, which affects the detection performance of the algorithm for small features. Therefore, it is necessary to improve the YOLOv4 feature extraction network to meet the small target detection and real-time requirements.

The MobileNet network uses the depth separable convolution calculation to convert the traditional convolution into a deep convolution and a 1×1 dot convolution, and introduces a width multiplier and a resolution multiplier to control the amount of model parameters. Mobile NetV3 is the third generation of Mobile Net network development. It combines the deep separable convolution method in MobileNetV1, the Inverted Residuals, Linear Bottleneck and the Squeeze-and-Excitation (SE) attention mechanism in MobileNetV2. MobileNetV3 uses neural architecture search (NAS) to search for network configuration and parameters, while improving the swish activation function to reduce the amount of calculation for h-swish, which can achieve less calculation and higher accuracy. The Mobile Net network first uses three 3∗3 convolution kernels to convolve with each channel of the input feature map to obtain a feature map with an input channel equal to the output channel, and then uses N 1∗1 convolution kernels to convolve this feature map to obtain a new N-channel feature map. Compared with the CSPDarknet53 network, it not only maintains a relatively powerful feature extraction capability, but also reduces the size of the model to a large extent, making it more convenient to deploy in the mobile terminal of the industrial field. At the same time, it has less network depth than the CSPDarknet53 network, which can better extract local refined features and improve the feature detection performance of small targets.

The model is trained with a self-regular non-monotonic Mish activation function, which can ensure the effective return of training loss, and obtain better generalization ability and more accuracy while ensuring the convergence speed. The calculation formula is:

$$f(x) = x \cdot \tanh\left[\log\left(1 + e^{x}\right)\right] \tag{3}$$

where x is the input of the activation layer, and f (x) is the output of the activation layer.

In order to detect the target more accurately, the training loss is composed of the weighted sum of bounding box regression loss, confidence loss and classification loss, and calculates the return gradient. The calculation formula is:

$$\begin{aligned} L &= L_{box} + L_{obj} + L_{c} \\ L_{box} &= \lambda_{iou} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij,obj} L_{ciou} \end{aligned} \tag{4}$$

where L represents training loss; L_{box} represents bounding box regression loss; L_{obj} represents target confidence loss; λ_{iou} represents category classification loss; represents bounding box regression loss weight coefficient; S represents the number of grids; B represents anchor point candidates generated by each grid Box; $l_{ij,obj}$ indicates that there is a target; L_{ciou} indicates the boundary loss measured by CIoU.

λ_{iou} affects the proportion of the bounding box regression loss in the overall training loss, which can improve the detection accuracy. The calculation of confidence loss is as follows:

$$L_{obj} = \lambda_{cls} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij,obj} \lambda_C \left(C_i - \hat{C}\right)^2 + \lambda_{cls} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij,noobj} \lambda_C \left(C_i - \hat{C}\right)^2 \tag{5}$$

where λ_{cls} is the confidence loss weight coefficient; $l_{ij,noobj}$ indicates no target; λ_C is the loss weight coefficient corresponding to each category target; C_i is the confidence of the i-th grid; $\hat{C}$ is the target confidence.

By changing λ_{cls}, the influence weight of the confidence loss in the entire training loss can be adjusted; by changing λ_{C}, the influence weight of samples of different categories in the training loss can be set, so as to be compatible with categories with fewer training samples to solve complex problems.

3 Experiment and Result Analysis

In order to verify the reliability of the algorithm and the effect in the actual measurement, a set of grating three-dimensional projection measurement system composed of a projector and a camera was built, as shown in the Fig. 3. The resolution of the camera (Hikvision MV-CA060-10GC) is 3072∗2048, the resolution of the projector (BenQ es6299) is 1920∗1200, the high-speed vision processor (CPU i9-10900X, 3.7 GHz, 4.5 GHz Turbo, memory 64 GB DDR4, 32-bit Windows operating system).

Figure 3: Experimental system

The experimental steps are as follows:

(1) Generate sine grating fringes, where a four-step phase shift fringe pattern is used.
(2) Project the sine grating fringe pattern to the homogeneous whiteboard, and collect the grating fringe modulated by the surface of the object.

(3) Use the training data to train the YOLOv4 network model to obtain the mapping between the fringe image and the depth image.
(4) Use the trained network to obtain the depth data of the fringe image.

For deep learning network training, the training rounds are uniformly set to 100, the batch size is 16, the initial learning rate is 1e-3, and the initial weights are all set to 1. There are a total of 5012 photos in the training set. In each round of training, 90% of the photos are used for training, and the other 10% of the photos are used for real-time detection of the training effect. This experiment will select a set of weight files with the lowest loss in each round to compare mAP size, model size, and real-time detection Frames Per Second (FPS).

As shown in Tab. 1, the model size of standard YOLOv4 is about 220 M, and FPS is 6.33. After replacing CSPDarknet53 with Mobilenetv3, the model size further decreased to only 50 M, FPS increased to 14.35, but mAP also dropped to 77.48%. It can be concluded that although Mobilenetv3 can greatly simplify the network structure, mAP will also be greatly reduced. After the improved model is used in the algorithm in this paper, mAP increases to 82.11% and the model size becomes 52 M, and the FPS is 13.67. Although the algorithm causes the model to become slightly larger and the FPS to drop slightly, it ensures a higher mAP.

Table 1: Comparison of deep learning model

Model	mAP	Model size	FPS
YOLOv4	83.64%	220 M	6.33
YOLOv4 + Mobilenetv3	77.48%	50 M	14.35
Improved algorithm in this paper	82.11%	52 M	13.67

According to the built experiment system and trained deep learning model, 3D reconstruction calculations are performed on objects with a simple shape and a complex shape respectively. The experiment uses a high-speed visual processor for training, and uses pre-training weights to train the original YOLOv4 network and the improved YOLOv4 model in this article. Finally, the results of the above three models are compared. Fig. 4 is the simple shape image of the pony spoon inputted by the test, respectively taking 4 fringe images with different phases. Fig. 5 is the final optimal depth image. Tab. 2 is the 3D reconstruction effect of the three models.

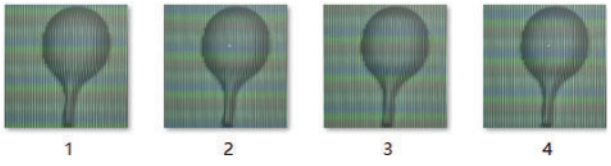

Figure 4: Test fringe image

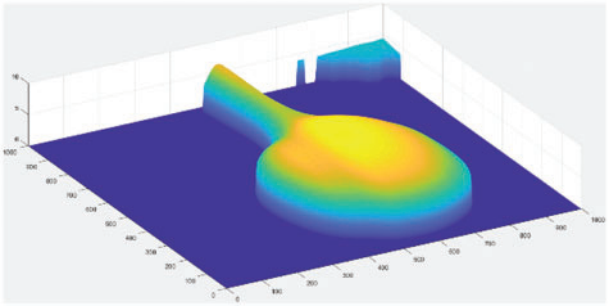

Figure 5: Depth image

Table 2: Comparison of 3D reconstruction results

Model	Average phase error	Point cloud restoration rate	Operation time
YOLOv4	0.12	73.4%	7.23
YOLOv4 + Mobilenetv3	0.057	80.4%	4.32
Improved algorithm in this paper	0.034	90.1%	3.64

Fig. 6 is the complicated shape image of the human face inputted by the test, respectively taking 4 fringe images with different phases. Fig. 7 is the final optimal depth image. Tab. 3 is the 3D reconstruction effect of the three models.

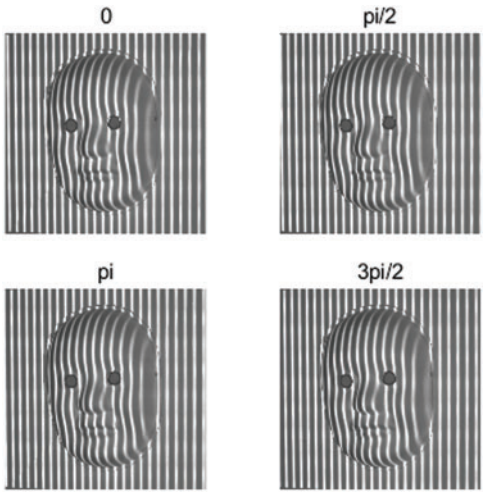

Figure 6: Test fringe image

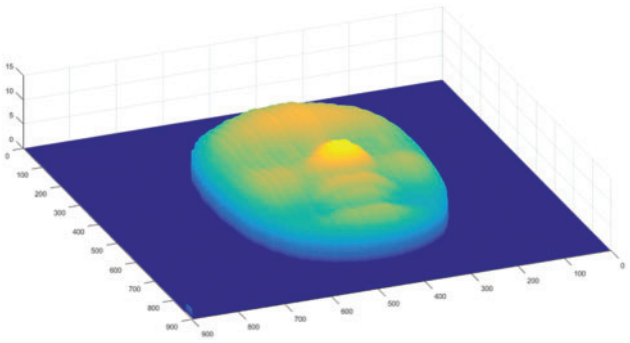

Figure 7: Depth image

Table 3: Comparison of 3D reconstruction results

Model	Average phase error	Point cloud restoration rate	Operation time
YOLOv4	0.054	81.2%	8.21
YOLOv4 + Mobilenetv3	0.034	84.4%	5.01
Improved algorithm in this paper	0.012	92.3%	3.13

By simulating the 3D reconstruction process of two different objects, compared with the simple model of first example, the model of second example is more complex, has more abundant fringe features, and be convenient to obtain the phase change, so it is better than the first example in reconstruction accuracy and speed. At the same time, it can be concluded from the simulation results of the three algorithms: the lightweight YOLOv4 model in this paper is superior to the other two models in terms of average phase error, point cloud restoration rate and running time, but it still needs further research at the sub-pixel level in the detail reconstruction.

4 Conclusions

Based on the 3D model of striped structured light construction, this paper proposes a stripe image feature extraction algorithm based on lightweight YOLOv4. The advantage of this model is that it uses a lightweight Mobile Net network to replace the CSPDarknet backbone network in YOLOv4, which simplifies the network structure and improves the real-time performance of model detection; uses the Mish activation function and the CIoU loss function to calculate and improve the target frame regression loss, which is effective Improved feature detection accuracy and real-time performance. The experimental results show that, compared with the existing 3D reconstruction methods, the depth information calculated by the proposed method has higher accuracy and improves the accuracy of the 3D measurement results of fringe images. Therefore, it can be effectively used in the field of fringe projection 3D measurement and is better to meet the needs of 3D shape measurement of objects in scientific research and practical applications. The next step will continue to study the effectiveness of the proposed method in other more experimental scenarios, such as the effectiveness and accuracy of the fringe image depth estimation in the case of colored objects, high-light objects, and projection out-of-focus conditions. On the other hand, the generalization ability of the model is a common problem

in deep learning, and it is also a key issue that needs to be paid attention to in the next work to improve the proposed method.

Acknowledgement: The authors thank Dr. Jinxing Niu for his suggestions. The authors thank the anonymous reviewers and the editor for the instructive suggestions that significantly improved the quality of this paper.

Funding Statement: This work is funded by the Training Plan for Young Backbone Teachers in Colleges and Universities in Henan Province under Grant No. 2021GGJS077.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. R. Zhang, W. F. Zhang, W. Sun, X. M. Sun and S. K. Jha, "A robust 3-D medical watermarking based on wavelet transform for data protection," *Computer Systems Science & Engineering*, vol. 41, no. 3, pp. 1043–1056, 2022.

[2] T. Y. Tao, Q. Chen and J. Da, "Real-time 3D shape measurement with composite phase-shifting fringes and multi-view system," *Optics Express*, vol. 24, no. 18, pp. 20253–20269, 2016.

[3] X. R. Zhang, X. Sun, X. M. Sun, W. Sun and S. K. Jha, "Robust reversible audio watermarking scheme for telemedicine and privacy protection," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 3035–3050, 2022.

[4] F. Q. Li and W. J. Chen, "Phase error analysis and correction for phase-shift profilometry using crossed grating," *Acta Optica Sinica*, vol. 41, no. 14, pp. 95–106, 2021.

[5] L. Cheng, Y. J. Pan and D. D. Xu, "Phase unwrapping correction method for dual-frequency fringe projection pofilometry," *Laser & Optoelectronics Progress*, vol. 58, no. 12, pp. 216–222, 2021.

[6] S. Zhang, "Absolute phase retrieval methods for digital fringe projection profilometry: A review," *Optics and Lasers in Engineering*, vol. 107, no. 8, pp. 28–37, 2018.

[7] B. Wang and B. L. Wang, "Detection method of catenary insulator defects based on convolutional neural network," *Urban Mass Transit*, vol. 23, no. 12, pp. 90–94+112, 2020.

[8] W. M. Guo, K. Liu and H. F. Qu, "Welding detection of X-ray images based on faster R-CNN model," *Journal of Beijing University of Posts and Telecommunications*, vol. 42, no. 6, pp. 20–28, 2019.

[9] S. Q. Ren, K. M. He and R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[10] Y. Han and H. Hahn, "Visual tracking of a moving target using active contour based SSD algorithm," *Robotics & Autonomous Systems*, vol. 53, no. 3, pp. 265–281, 2005.

[11] A. Baccouche, B. Garcia-Zapirain, C. C. Olea and A. S. Elmaghraby, "Breast lesions detection and classification via yolo-based fusion models," *Computers, Materials & Continua*, vol. 69, no. 1, pp. 1407–1425, 2021.

[12] J. Liu, X. Zhu and M. M. Song, "End-to-end Chinese character detection in natural scene based on improved YOLOv2," *Control and Decision*, vol. 36, no. 10, pp. 2483–2489, 2021.

[13] Y. Ding and Z. Fu, "Multi-UAV cooperative GPS spoofing based on YOLO nano," *Journal of Cyber Security*, vol. 3, no. 2, pp. 69–78, 2021.

[14] W. J. Li, G. W. Xu and W. G. Kong, "Research on target detection of plant leaf-stem intersection based on improved YOLOv4," *Computer Engineering and Applications*, vol. 31, no. 4, pp. 1–8, 2021.

[15] Y. Feng and J. Z. Li, "Improved convolutional neural network prdestrian detection method," *Computer Engineering and Design*, vol. 41, no. 5, pp. 1452–1457, 2020.

[16] R. Dubey and J. Agrawal, "An improved genetic algorithm for automated convolutional neural network design," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 747–763, 2022.
[17] Y. J. Ren, Y. Leng, J. Qi, K. S. Pradip, J. Wang *et al.,* "Multiple cloud storage mechanism based on blockchain in smart homes," *Future Generation Computer Systems*, vol. 115, no. 2, pp. 304–313, 2021.
[18] Y. J. Ren, F. J. Zhu, K. S. Pradip, T. Wang, J. Wang *et al.,* "Data query mechanism based on hash computing power of blockchain in internet of things," *Sensors*, vol. 20, no. 1, pp. 1–22, 2020.
[19] X. X. Li and L. Yang, "Real-time detection algorithm for small-scale pedestrians in complex road scenes," *Computer Engineering and Applications*, vol. 56, no. 22, pp. 124–131, 2020.
[20] H. D. Han, Y. R. Xu and B. Sun, "Active infrared detection of aerospace electronic solder joint defects based on improved tiny-YOLOv3 network," *Chinese Journal of Scientific Instrument*, vol. 37, no. 11, pp. 42–49, 2020.
[21] C. Li and J. X. Li, "Orthogonality feature extraction method and its application in convolutional neural network," *Journal of Shanghai Jiaotong University*, vol. 55, no. 10, pp. 1320–1329, 2021.
[22] K. M. He, X. Y. Zhang and S. Q. Ren, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2021.