Journal of Artificial Intelligence and Technology, 2022, 2, 32-37
https://doi.org/10.37965/jait.2021.12005

ISTP
RESEARCH ARTICLE

Stud Pose Detection Based on Photometric Stereo and Lightweight YOLOv4

Xuan Zhang and Guohui Wang
School of Opto-Electronic Engineering, Xi'an Technological University, Xi'an 710021, China

(Received 13 December 2021; Revised 13 December 2021; Accepted 23 December 2021; Published online 27 December 2021)

Abstract: There are hundreds of welded studs in a car. The posture of a welded stud determines the quality of the body assembly, thus affecting the safety of cars. It is crucial to detect the posture of the welded studs. Considering the lack of accurate method in detecting the position of welded studs, this paper aims to detect the weld stud's pose based on photometric stereo and neural network. Firstly, a machine vision-based stud dataset collection system is built to achieve the stud dataset labelling automatically. Secondly, photometric stereo algorithm is applied to estimate the stud normal map which as input is fed to neural network. Finally, we improve a lightweight YOLOv4 neural network which is applied to achieve the detection of stud position, thus overcoming the shortcomings of traditional testing methods. The research and experimental results show that the stud pose detection system designed achieves rapid detection and high accuracy positioning of the stud. This research provides the foundation combining the photometric stereo and deep learning for object detection in industrial production.

Key words: Stud pose; photometric stereo; neural network; machine vision

I. INTRODUCTION

Stud is widely used in the modern machine building industry because of its high interchangeability [1]. There are hundreds of welded studs in a car and these studs are used for interior assembly in the car body. Whether the position of the welded studs meets the design requirements not only determines the subsequent assembly but also affects the performance of the vehicle directly. It is necessary to detect the poses of all studs in a car for quality control during the modern industrial automation production.

Coordinate-measuring machine (CMM) [2] cannot be adapted accordingly to different objects, and its material in probe damages the surface of the measured target easily. What's more, the speed of CMM is far from meeting the demand of more efficient measurement in higher precision. Recently, with the continuous development of computer technology, machine vision is widely used for 3D measurement of objects [3,4] due to its advantages of noncontact, fast speed, and high accuracy so that researchers prefer studying noncontact measurements for objects. There are three types of noncontact measurement methods: acoustic [5], optical [6], and electromagnetic methods [7], of which the optical 3D measurement is the most widely applied. Conventional optical measurement systems are laser scanner [8], laser radar [9], structure light scanner [10], monocular vision [11–13], multi-view stereo vision [14,15], and so on. Recently, neural networks have shown to superior performance in many object detection tasks due to its ability to learn from raw data automatically [16]. There are many kinds of networks in 3D object detection [17–20]. However, there are few studies in studs pose detection by machine vision and networks. In other words, the defects (e.g., large lens distortions, focal blur, heavy noise, and extreme poses) of the stud images limit the stud pose detection using only neural networks. Wu et al. [21] developed a novel method based on monocular vision for measuring the weld studs pose. Liu et al. [22] proposed a stud measurement system based on photometric stereo vision and Histogram of Oriented Normal (HON) feature extractor. Studies above have been limited in detecting stud poses due to the fact that there has a highly variant reflection property in studs.

Photometric stereo [23], an emerging technology estimating normal maps under different illuminations, has been extensively applied for precision improvement in object measurement combined with deep learning [24–27]. Photometric stereo uses normal maps to evaluate the 3D shape which contains more accurate information than 2D images and possesses lower cost. For this reason, more and more researchers dedicate to the combination with photometric stereo and deep learning for 3D reconstruction and 3D measurement; however, there are few studies for object detection. Liu et al. [28] implemented optical measurements of studs through normal vector map estimation and heat map training. On the basis of these studies, this paper proposes the method for stud pose detection based on photometric stereo and neural network. The main contributions in this work are threefold:

(1) The monocular vision is applied to calculate the coordinate parameters of the camera for calibration, which can achieve the stud dataset labelling automatically.

(2) Photometric stereo algorithm is applied to estimate the stud normal map which as input is fed to the neural network.

(3) The lightweight YOLOv4 network is improved to locate the stud by analysing the normal map images in studs, which directly processes normal maps and outputs prediction results with multi-prediction size.

The structure of the rest of this paper is as follows: Section II provides basic methods in automatic labelling of stud datasets, estimating normal maps and building neural network; Section III presents the detailed experiments; in Section IV, data and the

Corresponding author: Xuan Zhang (e-mail: spritexuan7@163.com)

experimental results are presented; and Section V draws the conclusions of this work.

II BASIC METHOD

Combining photometric stereo and deep learning, as shown in Fig. 1, we first build a photometric stereo vision system and a machine vision measurement system to capture images of studs under eight different light sources (LED lights). We calculate the closed solution in camera calibration to obtain the internal and external parameters of the camera. Then, we derive the image pixel coordinates of studs in the images by Harris corner point detection algorithm [29] for automatically labelling the studs. Secondly, all stud images are processed by the light vector pseudo-inverse matrix to obtain the normal maps of the studs, which as the training images are input to the neural network. Finally, all the training images and the corresponding labels (ground truth) are input to the neural localisation network for iterative training and testing to achieve the pose detection of studs. As long as the nominal position of the stud is accurate, the pixel coordinates of the top and bottom centre points of the studs agree with the nominal position of the suds.

A. MONOCULAR VISION-BASED DATASET CONSTRUCTION

Figure 2 illustrates the interrelationship between the point P in 3D space and its corresponding point p in the image, which contains coordinate transformation in four coordinate systems. These four coordinate systems are the world coordinate system, camera coordinate system, image coordinate system, and pixel coordinate system, respectively. As shown in Fig. 2, the 3D right-angle coordinate $O_w - X_w Y_w Z_w$ is the world coordinate, which can be set artificially. The 3D right-angle coordinate $O_c - X_c Y_c Z_c$ is the camera coordinate, O_c is located at the projection centre of the lens; the Z_c axis is perpendicular to the image sensor and coincides with the optical axis of the lens; the X_c axis is parallel to the long side of the image sensor image array; and the Y_c axis is determined by the right-hand rule. The 2D right-angle coordinate $o - xy$ is the image coordinate, and $o_0 - uv$ is the pixel coordinate [30].

The relationship of the transformance between the world coordinates and the pixel coordinates is expressed as:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R_{3\times3} & T_{3\times1} \\ 0_{1\times3} & 1_{1\times1} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \quad (1)$$

where $f_x = f/dx$, $f_y = f/dy$. f is the focal length of the lens, and dx,dy are the physical dimensions of a pixel in x-axis and y-axis, respectively. As the external parameters of the camera, R and T are the rotation matrix and the translation vector, respectively. By equation (1), the parameters of the camera are obtained for camera calibration. On the basis of which, we construct the stud datasets. The details in dataset construction are as follows:

a. Calculating the internal and external parameters of the camera for the camera calibration.

b. Calculating the pixel values of the top and bottom centre points of the stud in the image coordinate from the 3D coordinate of the stud.

c. Labelling the stud by the image coordinate and defining the bottom centre point of the stud as *studb*, the top centre point as *studt*.

d. Feeding the pixel coordinates of the studs as ground truth to the neural network.

B. PHOTOMETRIC STEREO SYSTEM

Photometric stereo is a method to obtain local normal maps in several images under different illuminations. This paper applies eight LED lights with different orientations for improving the accuracy and robustness of the result. The complexity of the threads on the stud surface and the soot from welding leads to a more pronounced diffuse reflection of the stud itself, so the photometric stereo vision system is established based on the Lambertian reflection.

According to the Lambertian reflection, the intensity of any pixel $p(x,y)$ in the image can be expressed as:

$$I_i(x,y) = \rho_{(x,y)}(S_i \cdot n_{(x,y)}) \quad (2)$$

where $I_i(x,y)$ is the pixel intensity under i^{th} illumination in x^{th} row and y^{th} column, and ρ is the albedo at the corresponding point of

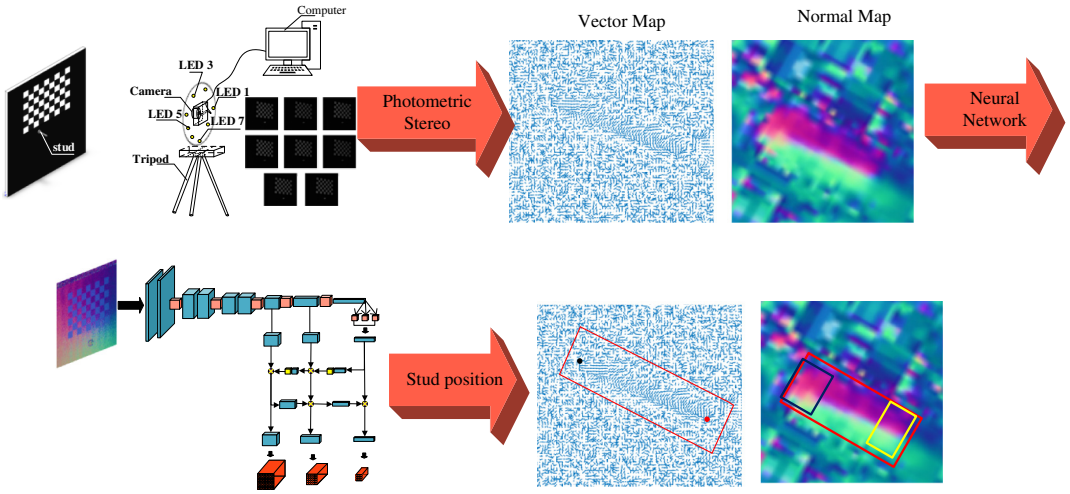


Fig. 1. Stud pose detection system.

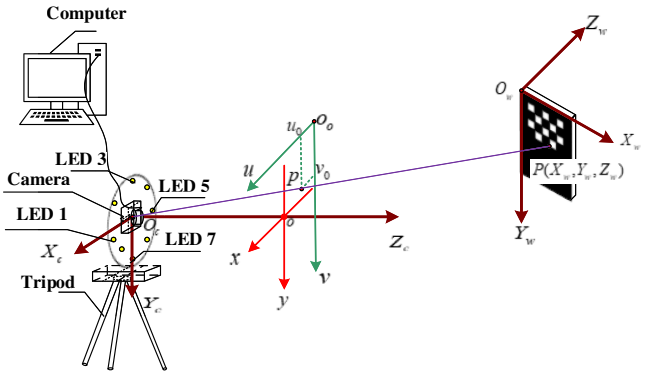


Fig. 2. Dataset collection system based on monocular vision.

pixel $p(x,y)$, $S_i = [s_{x_i}, s_{y_i}, s_{zi}]^T$ denotes the direction of light source projection. The equation (2) can be formulated as:

$$I(x,y) = \rho(x,y)\frac{1 + p_s p(x,y) + q_s q(x,y)}{\sqrt{1 + p_s^2 + q_s^2}\sqrt{1 + p^2(x,y) + q^2(x,y)}}$$
$$= \rho(x,y)\vec{n}(x,y) \cdot \vec{s} = \mathrm{g}(x,y) \cdot \vec{s} \quad (3)$$

$$\mathrm{g}(x,y) = \rho(x,y)\vec{n}(x,y) \quad (4)$$

where $\vec{n}$ is the surface unit normal vector, which is estimated by applying eight LEDs and calculating the pseudo-inverse matrix of the light source vectors in this research. On the basis of which, the equation (3) can be described as:

$$S^T \cdot I = \rho \cdot S^T \cdot S \cdot n \quad (5)$$

The surface normal $\vec{n}$ of pixel $p(x,y)$ can be estimated:

$$\vec{n}(x,y) = \frac{g(x,y)}{\rho(x,y)} = \frac{g(x,y)}{\|g(x,y)\|} \quad (6)$$

Finally, calculating every pixel through equation (6) repeatedly for normal map.

C. DEEP NETWORK-BASED NORMAL MAP OF STUDS IN LOCALISATION

In this paper, a lightweight YOLOv4 network based on YOLOv4 [31] is proposed to locate the stud by analysing the normal map images in studs. As shown in Fig. 3, the size fed to the network is $608 \times 608 \times 3$, where 3 indicates the three channels. The lightweight YOLOv4 network applies convolution layers, upsampling, downsampling, and deep concatenation layers to directly process normal maps and output prediction results with multi-prediction size. The multi-size output contains three kinds of sizes: $76 \times 76 \times 24$,$38 \times 38 \times 24$, and $19 \times 19 \times 24$, which can get better network performance in extracting important features from the training data. Root mean squared error (RMSE) is used as the regression loss function during the training process:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad (7)$$

where y_i is the ground-truth value, $\hat{y}_i$ is the predicted value, and N is the number of the testing samples of stud normal maps.

III EXPERIMENTS

A. DATASET AND EXPERIMENTAL PLATFORM

In this study, a total of 5000 groups of samples for studs are constructed. We apply the software MTLAB to program the microcontroller program Arduino to ensure that the LEDs are lit in the clockwise from the number 1 in Fig. 2 for capturing stud images. Every group of the stud sample contains eight stud images from different illuminations. These images are calculated by

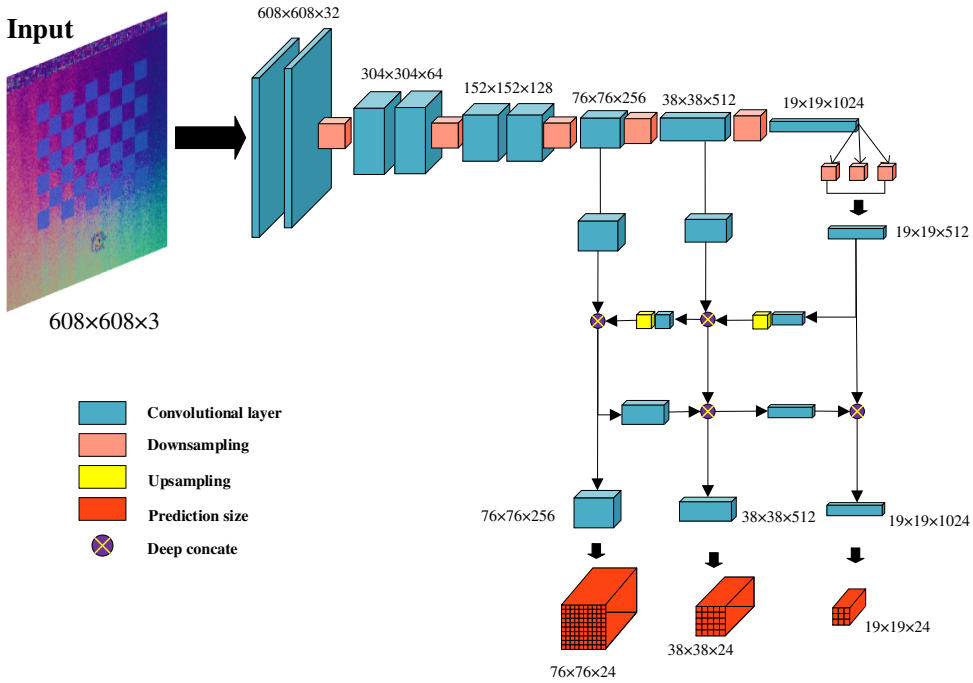


Fig. 3. Lightweight YOLOv4 network.

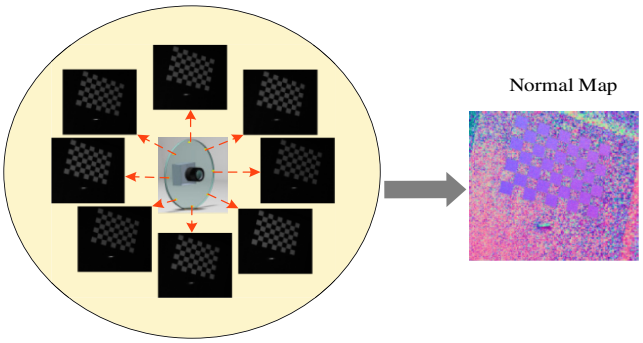


Fig. 4. Normal map diagram for a stud pose.

photometric stereo to obtain the normal maps of studs, which as the input are fed to neural network for training.

The hardware server configuration for the experiments is: Intel(R) Core (TM) i5-9600KF processor, NVIDA GeForce GTX 2080Ti graphics card. The software environments are Ubuntu 18.04, Python 3.7.7, TensorFlow-gpu-2.1.0, and PyCharm 2020. 1. Proposed method in this paper utilises several libraries such as NumPy, Pillow, and OpenCV.

B. NORMAL MAPS OF STUDS

We estimate the vector maps of studs by the least square algorithm based on photometric stereo. Eight stud images of the same stud pose with different illuminations are integrated into a stud vector map. The normal maps of studs are obtained by converting the channels of the stud vector map. This paper displays the normal map of one stud pose in Fig. 4.

C. EVALUATION METRICS

In this paper, the RMSE and mAP (mean Average Precision) are used to evaluate the model:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

RMSE suggests the precision of the measurement, which indicates the overall difference between the predictions and the ground truth for all testing samples. TP stands for true-positive, FP for false-positive, and mAP as an important evaluation metrics is used to evaluate the accuracy of object detection.

D. NETWORK TRAINING

The network is trained on Adaptive moment (Adam) estimation method, which possesses a very fast convergence rate and powerful generalisation ability with optimisation. Mosaic and Image augmentation (Imgaug) are applied to expand the stud dataset with a total of 30,000 data samples. All the labelled data (corresponded with ground truth) are randomly divided into the training and testing datasets with the ratio of 4 : 1. During the network training, the epoch and batch size of the training data are set to 60 and 4, respectively. The weights of the Pascal VOC (Pascal Visual Object Classes) are used as the initial weight input. The learning rate given an initial value with 0·001 is updated every 2500 iterations.

IV RESULTS

The training loss curves are shown in Fig. 5. The numbers on x-axis and y-axis represent the training epochs and the loss values, respectively. Figure 5(a) shows the trend of loss values, Figs. 5(b) and 5(c) show the distribution of loss values in the locally enlarged region of Fig. 5(a) respectively. Figure 5 indicates that the loss function converges rapidly at the beginning of the training with oscillating decrease in the followed training. After 60 epochs of the network training, the loss value is 13·9392 (unnormalised) and the prediction result performs best.

The weights perform best trained in network are used to predict the stud pose. Figures 6(a) to 6(d) illustrate the prediction result in stud normal map images and stud images captured by the camera directly. It is obvious that the neural network provides a good performance in detecting stud normal maps with its key points on the top and bottom of the stud in Figs. 6(a) and 6(b). However, the raw image of the stud is detected incorrectly under the complex background shown in Figs. 6(c) and 6(d). Figures 6(c) and 6(d) show that the top and bottom key points of the stud are not accurately recognised, or are arbitrarily recognised as other key points, or are not recognised. RMSE and mAP in proposed network are 0·074% and 99·65%, respectively, a low error and high precision. In terms of detecting the speed for every stud image, our method requires less average computation time of 0·002584 s, which indicates that the proposed method can be applied in a real production environment for stud real-time detection.

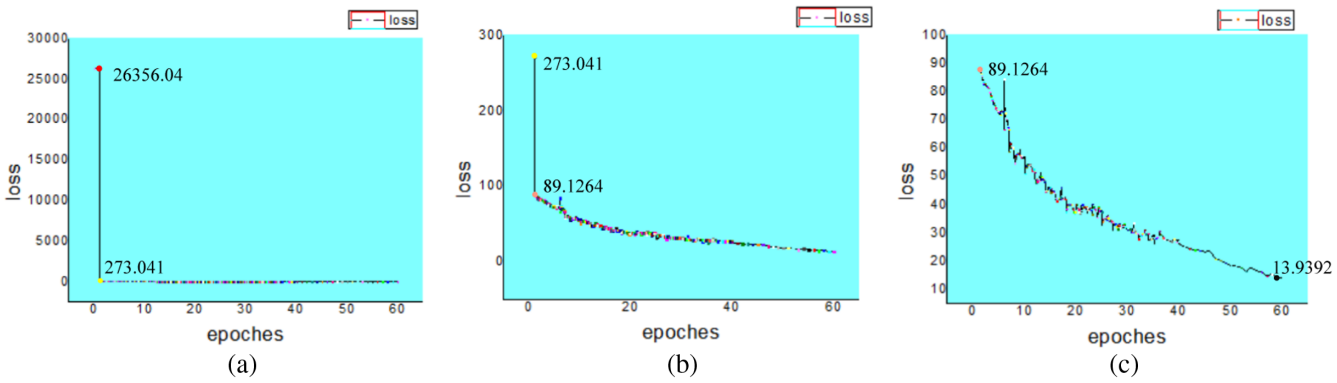
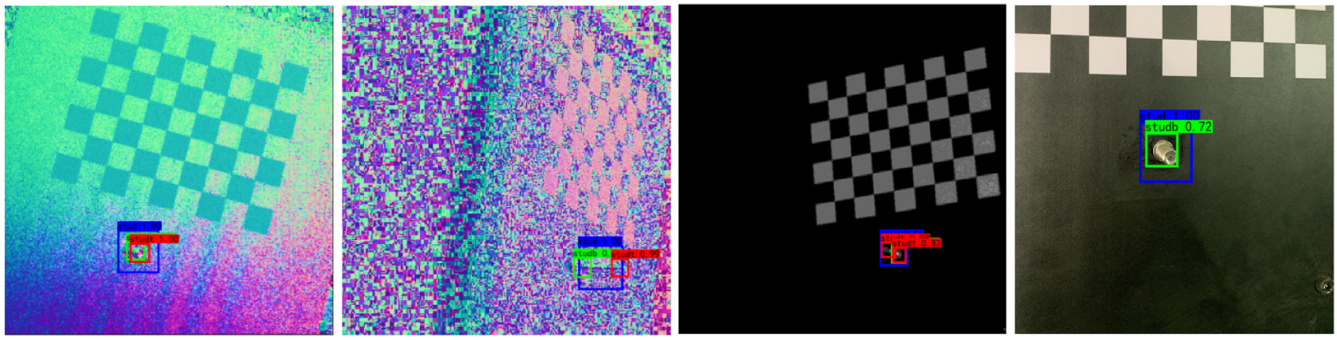


Fig. 5. Loss trend during training.



(a) (b) (c) (d)

Fig. 6. Prediction of stud normal map image and stud raw images. (a, b) prediction of stud normal map images; (c, d) prediction of stud raw images. The blue box, red box, and green box indicate the position, the top point, and the bottom point of stud, respectively.

V. CONCLUSIONS

In this paper, a dataset system for automatically collecting and labelling studs was built. The photometric stereo with eight light sources was applied to estimate stud normal maps as input to improved neural network with good experimental results in stud poisoning. After the prediction, RMSE and mAP were used as the evaluation metrics to validate the prediction performance. A comparison of stud normal maps with stud raw images fed in network was made and suggested that proposed method indicated superior prediction performance. The conclusions in this paper are also applicable to multi-stud identification and detection. This research provides the foundation combining the photometric stereo and deep learning for object detection in industrial production. In future, the combination of deep learning and photometric stereo will be studied more intensively to improve the accuracy and speed of object detection.

ACKNOWLEDGEMENTS

The work is partly supported by the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2016JM6041).

References

[1] X. Chen, "Research on non-contact obstacle detection system of urban rail train," Master Dissertation, Shanxi Univ, 2018.
[2] R. J. Hocken, *Coordinate Measuring Machines and Systems*, Second Edition, CRC Press, Boca Raton, Florida, US, 2017.
[3] Z. Liu and B. Qu, "Machine vision based online detection of PCB defect," *Microprocess. Microsyst.*, vol. 82, no. 9, p. 103807, 2021.
[4] B. Kostov and V. Hristov, "Implementation of 3D measuring sensor for calibrating robot coordinate systems," *2021 5th Int. Symp. Multidiscip. Stud. Innov. Technol. (ISMSIT)*, Ankara, Turkey, pp. 795–798, Oct. 21–23, 2021.
[5] M. Szczodrak et al., "A system for acoustic field measurement employing Cartesian robot," *Metrol. Meas. Syst.*, vol. 23, no. 3, 2016.
[6] C. Duan et al., "Improving the performance of 3D shape measurement of moving objects by Fringe projection and data fusion," *IEEE Access*, vol. 9, pp. 34682–34691, 2021.
[7] W. Skierucha et al., "Estimation of electromagnetic sensor measurement volume using combined 3D EM simulation and electronic design software," *12th Int. Conf. Electromagn. Wave Interact. Water Moist Subst., ISEMA*, Lublin, Poland, pp. 1–9, June 4–7, 2018.
[8] Y. M. Zhao et al., "Laser scanner for 3D reconstruction of a wound's edge and topology," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, pp. 1761–1773, 2021.
[9] Y. Fan, L. Zheng, and Y. Liu, "3D environment measurement and reconstruction based on LiDAR," *IEEE Int. Instrum. Meas Technol. Conf. (I2MTC 2018)*, Houston, TX, USA, pp. 1–4, May 14–17, 2018.
[10] Budianto, W.Law, and D. P. K. Lun, "Deep learning based period order detection in structured light three-dimensional scanning," *IEEE Int. Symp. Circuits Syst.*, IEEE, Sapporo, Japan, pp. 1–5, May 6–29, 2019.
[11] X. Shi, Z. Chen, and T. K. Kim. "Distance-normalized unified representation for monocular 3D object detection," *Eur. Conf. Comput. Vision*, Springer, Cham, pp. 91–107, 2020.
[12] S. Wang and X. Li, "A real-time monocular vision-based obstacle detection," *ICCAR*, Singapore, pp. 695–699, April. 20–23, 2020.
[13] A. Simonelli et al., "Disentangling monocular 3D object detection: from single to multi-class recognition," in *IEEE TPAMI*, pp. 8969–8979, 2021.
[14] Q. Wang et al., "Deep learning and binocular stereovision to achieve fast detection and location of target,". *CISC*, Springer, Singapore, vol. 593, pp. 306–313, 2019.
[15] T. Schöps et al., "A multi-view stereo benchmark with high-resolution images and multi-camera videos," *IEEE CVPR*, Honolulu, HI, USA, pp. 2538–2547, July 21–26, 2017.
[16] P. N. Druzhkov and V. D. Kustikova, "A survey of deep learning methods and software tools for image classification and object detection," *Pattern Recognit. Image Anal.*, vol. 26, no. 1, pp. 9–15, 2016.
[17] W. Kehl et al., "Deep learning of local RGB-D patches for 3D object detection and 6D pose estimation," *ECCV*, Springer, Cham, vol. 9907, pp. 205–220, 2016.
[18] N. Lu et al., "Deep learning for fall detection: three-dimensional CNN combined with LSTM on video kinematic data," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 1, pp. 314–323, Jan. 2019.
[19] Y. Guo et al., "An integrated framework for 3-D modeling, object detection, and pose estimation from point-clouds," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 3, pp. 683–693, 2014.
[20] F. Yi et al., "Deep learning integral imaging for three-dimensional visualization, object detection, and segmentation," *Opt. Lasers Eng.*, vol. 146, p. 106695, 2021.
[21] B. Wu, F. Zhang, and T. Xue, "Monocular-vision based method for online measurement of pose parameters of weld stud," *Measurement*, vol. 61, pp. 263–269, 2015.
[22] H. Liu et al., "Optical challenging feature inline measurement system based on photometric stereo and HON feature extractor," *Opt. Micro Nanometrol.*, vol. 10678, p. 1067812, 2018.

[23] R. J. Woodham, "Determining surface curvature with photometric stereo," *Proceedings*, ICRA, Scottsdale, AZ, USA, pp. 36–42, May 14–19, 1989.

[24] Y. Ju, J. Dong, and S. Chen, "Recovering surface normal and arbitrary images: a dual regression network for photometric stereo," *IEEE Trans. Image Process.*, vol. 30, pp. 3676–3690, 2021.

[25] E. Song and M. Chang, "Photometric stereo using CNN-based feature-merging network," *20th Int. Conf. Control Autom. Syst. (ICCAS)*, Busan, Korea (South), pp. 865–868, Oct. 13–16, 2020.

[26] H. Santo et al., "Deep photometric stereo networks for determining surface normal and reflectances," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 114–128, 2022.

[27] B. Shi et al., "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo," *IEEE CVPR*, Las Vegas, NV, USA, pp. 3707–3716, Jun. 7–30, 2016.

[28] H. Liu et al., "Efficient optical measurement of welding studs with normal maps and convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 70, Art no. 5000614, pp. 1–14, 2021.

[29] C. Guo et al., *A Fast and Accurate Corner Detector Based on Harris Algorithm*," Nanchang, China: IITA, pp. 49–52, 2009.

[30] G. H. Wang and K. M. Qian, "Review on line-scan camera calibration methods," *Acta Opt. Sin.*, vol. 40, no 1, pp.181–193, 2020.

[31] A. Bochkovskiy, C. Y Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," [OL]. https://arxiv.org/abs/2004.10934.