IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 October 2022, accepted 2 November 2022, date of publication 5 December 2022, date of current version 9 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3226880

RESEARCH ARTICLE

A Multi-View Structured Light Measurement Method Based on Pose Estimation Using Deep Learning

TAO JIANG, KAITUO FANG, HAIFANG ZHAO, GUOBIN CHEN, (Graduate Student Member, IEEE), AND YANFENG WANG
School of Mechanical and Electrical Engineering, Suqian University, Suqian 223800, China

Corresponding author: Tao Jiang (TaoJiang@squ.edu.cn)

This work was supported in part by the General Program of Basic Science (Natural Science) Research in Universities of Jiangsu Province under Grant 22KJB460007 and Grant 22KJB460037; in part by the Research Foundation for Advanced Talents of Suqian University under Grant 2022XRC012; and in part by the Suqian Science and Technology Program under Grant K202206, Grant K202113, and Grant L202210.

ABSTRACT To enhance the effectiveness and flexibility of the data alignment in the multi-view measurement system, a measurement strategy based on pose estimation using deep learning is proposed. The object pose is estimated and established through a single-shot pose estimation network. Then, the coarse alignment of the data acquired from different views is performed using the estimated 6D pose. The ICP algorithm is utilized for global refinement. Different shapes are used to verify the effectiveness, robustness, and flexibility of the deep learning-based multi-view measurement strategy. Furthermore, error comparisons of data fusion using markers and deep learning are implemented. The translation error of pose estimation is 1.8-5 mm, and the angle error can reach 0.5-1 degree. The difference between the marker-based and proposed data alignment method is only 0.02 mm. The proposed method can achieve comparable data alignment accuracy with the marker-based method. Moreover, it increases the flexibility and convenience of the data alignment and provides an improved way for existing marker- and shape-based multi-view measurement systems.

INDEX TERMS Multi-view structured light measurement, pose estimation, deep learning, data alignment.

I. INTRODUCTION

With the increasing requirements for the entire shape measurement of complex objects, there is a growing research interest in multi-view structured light measurement (MSLM) technology [1], [2], [3]. Generally, an MSLM system, containing a projector and a single camera, can reconstruct a certain range of object surfaces, obtaining a high-density point cloud. However, due to the presence of cross occlusions, extended measurement depth, and limited imaging field of view (FOV), only a part of the object can be measured at a time. Therefore, to reconstruct the overall shape, MSLM followed by data registration affords a direct and effective approach [4], [5]. Typically, two sequential steps, including rough rigid transformation and global refinement, are considered a common strategy for such an alignment process. The Iterative Closest Point (ICP) algorithm is widely utilized in the global optimization process owing to its high precision and low time computation [6]. However, the effectiveness and accuracy of the ICP algorithm mainly rely on the rough registration result. Therefore, in the MSLM system, the determination of the initial transformation matrix between two measurement coordinate systems becomes the basic and crucial problem. This paper researches an MSLM system, proposing deep learning-based viewpoint estimation from a single image for initial point cloud registration.

The related work is divided into two categories: (1) coarse alignment mode in multi-view measurements. (2) deep learning-based point cloud registration. The first category introduces different modes of coarse alignment in the

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

existing multi-view measurement system. The second category reviews the deep learning-based point cloud registration for multi-view measurement.

A. COARSE ALIGNMENT MODE IN MULTI-VIEW MEASUREMENTS

The approaches of the coarse alignment vary from the tracking-based [7], [8], [9], [10], [11], [12], marker-based [13], [14] and shape-based [15], [16], [17] modes. Specifically, the tracking-based technique is the combination of the global tracking system such as monocular vision [9], stereovision [11], laser tracker [12], *etc*., and the scanning device such as the laser scanner, structured light system, *etc*. In such systems, the target with identifiable and measurable markers moves in the working volume of the tracking system, contributing to data transferring and global alignment. The tracking-based mode enhances the flexibility of multi-view reconstruction. However, the transformation between different perspectives is estimated by the position and location of the markers attached to the target. The data registration will be discontinuous if the markers on the target cannot be recognized from certain views. The marker-based multi-view measurement methods focus on the detection of the fiducial markers that are distributed on overlapping views. Typically, different marker types such as coded markers [17], circle markers [14], ArUco [18], *etc*. are designed to enhance the identifiability of different measurement ranges. Therefore, it is necessary to attach the markers to the object's surface and detect them at the early stage. As a result, the registration will be discontinuous or suspended when the common markers in different perspectives cannot be extracted correctly. Additionally, the marker density influences the accuracy and integrity of the 3-D reconstruction. The shape-based methods align the point clouds in different views based on the geometrical and topological features of the overlapping areas. The rich information and the invariant features under different imaging environments are the essential conditions of shape description and feature matching. However, weak textures and poor features are unable to provide matching information for data alignment. Especially for flat or uniform curvature geometries, such a shape-based technique is no longer practically appropriate. It can be concluded that the marker- and shape-based methods rely on the features, eighter designed markers or inherent features, on measurement objects, and the tracking-based methods need stable markers attached to the tracking targets. The data measured from different views are registered preliminarily with recognizable features in the overlapping areas. Additional optical elements, such as mirrors, can increase the one-shot measurement range [19], but the actual implementation is not flexible and limited by the object size [20], [21].

B. DEEP LEARNING-BASED POINT CLOUD REGISTRATION

Recent years have seen the development of the deep learning-based data registration method in optical measurement. The rapid development of deep learning technology has enabled new approaches in structured light measurement in terms of robust phase unwrapping, high-speed profilometry, sensor fusion, and others [1], [22]. The advantages of deep learning methods, including environment-independent, illumination immunity, and superior performance for high-dimensional, high-complexity problems, improve the stability and reliability of optical measurements that are certainly dependent on a stable measurement environment. The data from different perspectives can be easily fused by substituting the above-mentioned marker- and shape-based methods with deep learning where the pose of the acquisition system under each measurement perspective is estimated with a priori knowledge. Chang et al. proposed a registration architecture for relative pose estimating and 3-D point cloud registration based on Convolutional Neural Networks (CNN) [23]. However, the pose of the current viewpoint is estimated through the reconstructed point cloud, resulting in the increasing complexity and larger computation of network training. Furthermore, if 3D data is available, the measurement pose can be preliminarily obtained by registering with the CAD model, making the deep learning method unnecessary in such applications. Jack et al. [24] proposed a learn-based free-form deformation (FFD) method for 3D reconstruction from a single image. The network can be used to produce arbitrarily dense point clouds with fine-grained geometry. Different from estimating the pose with reconstruction, determining the object pose from one single image will significantly improve the accuracy and efficiency. Detecting 3D objects and estimating their 6D pose from one view image has been investigated in many works such as DPOD [26], YOLO6D [27], *etc*. The excellent performance in pose estimation of various shapes in an occlusion environment indicates its potential and valuable application in multi-view transformation establishment. Additionally, the learning-based approach has a better ability in managing geometric operations, such as calculating the camera pose [28] or the transformation parameters [29]. Therefore, it can be concluded that the learning-based method can improve the flexibility and efficiency of the data registration in multi-view measurement. Also, it is meaningful to develop a learning-based multi-view optical measurement system.

It can be inferred from the above research that the data registration in the multi-view structured light measurement system relies on the attached markers or the surface features, weakening the system's flexibility and efficiency. To establish and validate a general multi-view data registration approach without additional features, we propose pose estimation for measurement view determination and data alignment based on a deep learning network. The novelty and contribution are listed as follows. (a) We investigate a pose estimation network combined with the ICP algorithm. The additional markers or rich surface features are not required in the proposed method, contributing to the flexibility and easy implementation for multi-view measurement. (b) We provide experimental verification for the proposed measurement system, focusing on the accuracy and effectiveness of the pose estimation and

the global data registration. The major contribution is that we eliminate the dependence on the markers or features in the multi-view structured light measurement, replacing the traditional marker- or feature-based pose estimation method with a deep learning-based method and devoting to a more flexible data registration for the multi-view structured light measurement system.

The remainder of the paper is organized as follows. Section II states the principle of the proposed strategy, including the network preparation and training details and the data alignment procedure. Section III presents the experiments and discussions about pose estimation and data registration. The accuracy analysis is provided. Section IV ends the paper with conclusions.

II. PRINCIPLE AND METHOD

A. PROPOSED LEARNING-BASED STRATEGY

Figure. 1 indicates the principle of the proposed MSLM strategy. As shown in Figure.1, the Single Shot Pose network based on the YOLO model and PnP algorithm is utilized for pose estimation. Then, the object pose relative to the measurement coordinate system is established via the learning-based pose estimation. Lastly, the corresponding 3-D measured data from different perspectives are aligned with the estimated pose. The precise fusion is achieved through the ICP algorithm.

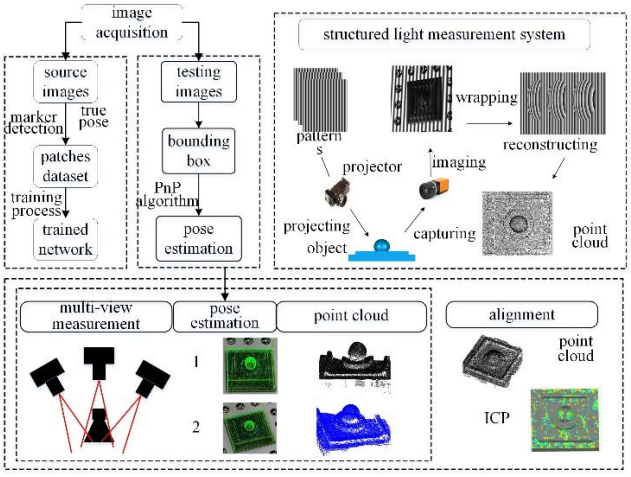


FIGURE 1. The proposed strategy of multi-view structured light measurement.

To improve the measurement efficiency of the MSLM system, the single-shot 2D object detector YOLO was employed to predict the projections of the 3D bounding box corners in the image. YOLO is end-to-end trainable and precise without any *a posteriori* refinement. It also has an impressive performance of fast detection. In addition, since the object pose is estimated with the image coordinates and the corresponding spatial coordinates, the symmetry of the object shape is not restricted strictly. Based on the above advantages, we integrated the YOLO network program into the MSLM system, achieving efficient and fast data alignment. Specifically, the YOLOv3 mode is utilized and improved to estimate the 8 projected corners of the 3D bounding box of the measured object.

B. POSE ESTIMATION NETWORK

The basic idea of pose estimation using deep learning is as follows. (1) The imaging points of the three-dimensional bounding box corners of the measured object are identified by the YOLO network. (2) The pose of the object is solved by the PNP algorithm. This method can quickly solve the pose from a single image with high accuracy. Figure. 2 is the overall diagram of pose estimation.

As shown in Figure. 2, the network directly recognizes the pixel coordinates of the bounding box corners. The loss function L is defined as

$$L = \lambda_p \sum_{i=1}^{9} ||p_i - \tilde{p}_i||^2 + \lambda_{conf} L_{conf} \tag{1}$$

where p and $\tilde{p}$ denote the true corners and predicted corners, respectively. $\|\cdot\|^2$ is the L-2 norm operator whose physical meaning is to calculate the Euclidean distance between two vectors. $\lambda_p = 1$ denotes the weight of corner errors. λ_{conf} is the credibility weight. L_{conf} is the loss function of credibility. The number of the 3D bounding box and the centroid is 9 in total. The Euclidean distance is also used to solve the deviation between the projection of the centroid of the target object on the image and the predicted projection coordinates. If the deviation distance is D and the distance threshold is defined as d, the credibility function is defined as.

$$L_{conf} = \begin{cases} \exp\left(1 - \dfrac{D}{d_{th}}\right), & D < d_{th} \\ 0, & otherwise \end{cases} \tag{2}$$

According to the network, the image is divided into cells during training. If the centroid projection point locates in the cell, $\lambda_{conf} = 5$, otherwise, $\lambda_{conf} = 0.1$.

The input of the network is the original image and the output is the target pose. At the same time, the corner image coordinates of the target object bounding box, the length and width of the 2D bounding box, and the target category can be additionally obtained. If there is only one tested object, the category value is set to 1. According to Figure. 2, the truth input and output data should be prepared before training. The details are as follows.

Firstly, a large number of images containing the coded markers and objects are captured from different views. The mask images can be obtained simultaneously by setting the background of the simulation scene to be empty. In practice, the mask images are processed manually. Secondly, for each image, the target markers are detected and a global coordinate system is established with the image coordinates and the corresponding spatial coordinates. If the image points of the coded markers are noted as $m_i(u_i, v_i)$ and the corresponding spatial coordinates are noted as $M_i(X_i, Y_i, Z_i)$. According to

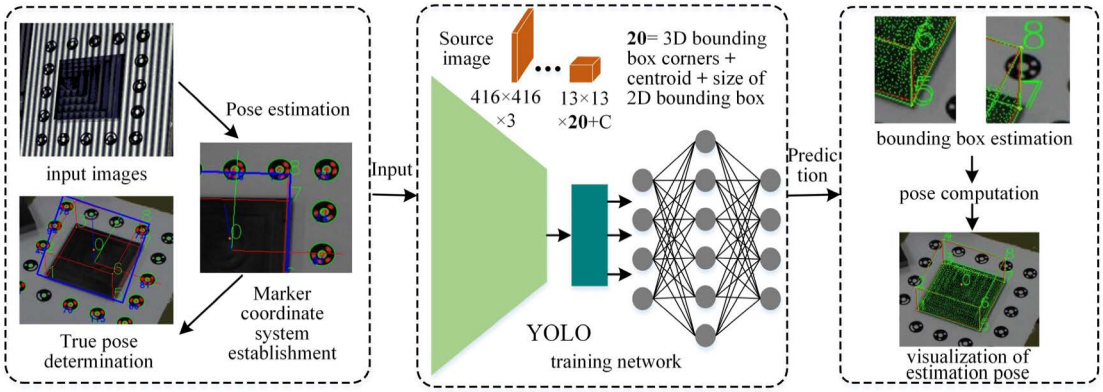


FIGURE 2. Pose estimation based on YOLO network.

the camera imaging model,

$$s\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} \tag{3}$$

where $R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}, t = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}$.

Expand Eq.(3) and eliminate s, $u_i = \frac{r_{11}X_i+r_{12}Y_i+r_{13}Z_i+t_1}{r_{31}X_i+r_{32}Y_i+r_{33}Z_i+t_3}$ and $v_i = \frac{r_{21}X_i+r_{22}Y_i+r_{23}Z_i+t_2}{r_{31}X_i+r_{32}Y_i+r_{33}Z_i+t_3}$ can be obtained. If the pose matrix is $T(R,t)_{3\times4} = [r_1, r_2, r_3]^T$ where r_j denotes the transposed vector of jth row ($j = 1, 2, 3$) of T, (3) can be expressed as,

$$\begin{cases} r_1^T M_i - r_3^T M_i u_i = 0 \\ r_2^T M_i - r_3^T M_i v_i = 0 \end{cases} \tag{4}$$

Then, the linear equations are established by using the image coordinates recognized by coding points and their corresponding spatial three-dimensional coordinates.

$$\begin{bmatrix} M_0^T & 0 & -u_0 M_0^T \\ 0 & M_0^T & -v_0 M_0^T \\ M & M & M \\ M_8^T & 0 & -u_8 M_8^T \\ 0 & M_8^T & -v_8 M_8^T \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} = 0 \tag{5}$$

The coefficient matrix is decomposed by the singular value decomposition method to obtain the least square solution of T. A series of the true pose T_i (R_i, t_i), where R_i, t_i denote the rotation matrix and translation vector at i-th view, are obtained at this step, respectively.

Thirdly, for i-th view, 8 corners, noting $P_{ij}(x_{ij}, y_{ij}, z_{ij})$, $j = 1, 2, \ldots, 8$, of the 3-D bounding box and the centroid $P_{i0}(x_0, y_0, z_0)$ are projected onto the image with the camera matrix K using the equation 6.

$$s_{ij}p_{ij} = KT_iP_{ij}, \quad j = 1, \ldots, 8 \tag{6}$$

where s_{ij} is the scale factor, p_{ij} is the image point. The 2-D bounding box is obtained by fitting the 8 projected image points p_i to a minimum bounding rectangle. The normalized image points, the length, and the width of the 2-D bounding box are saved as the label files for the training. In total, 21 parameters are obtained for each capturing view. According to the network structure, the input is the original image and the outputs are the 21 parameters. Finally, the training is performed using the prepared label files.

Each iteration of the network generates eight corner coordinates of the bounding box. The pose is estimated by using equations (3) ∼ (5). Then, the corner of the theoretical spatial bounding box is re-projected onto the image to calculate the value of the loss function and adjust the weight of the network. The estimation accuracy of the pose is determined by translation accuracy and rotation accuracy which are defined as follows.

$$\begin{aligned} E_T &= \frac{1}{N}\sum_{i=1}^{N} \|T_{ie} - T_{i0}\|^2 \\ E_A &= \frac{1}{N}\sum_{i=1}^{N} \|A_{ie} - A_{i0}\|^2 \end{aligned} \tag{7}$$

where: T_{ie} and T_{i0} represent the estimated translation vector and the corresponding real translation vector after the i-th iteration, respectively. A_{ie}, A_{i0} represents the estimated Euler angle and the corresponding real Euler angle after the i-th iteration respectively; N is the number of test sets.

C. MULTI-VIEW DATA REGISTRATION

In the i-th view measurement, the 8 corners' image points p_i and the object pose T_i relative to the camera coordinate system are estimated from a given image with the trained network. If the measurement data at i-th and j-th view are M_i and M_j, respectively, M_j can be transformed to the coordinate system of M_i through

$$M_j' = M_j T_i^{-1} T_j \tag{8}$$

TABLE 1. Average 2D re-projection error comparison (Pixels).

Method Object	Brachmann [31]	SSD [32]	BB8 [33]	YOLO-Based
Ape	6.25	5.54	5.88	5.26
Benchvise	3.08	2.89	3.12	2.45
Cam	5.71	4.59	5.77	4.85
Can	4.24	4.13	4.16	3.76
Cat	4.56	4.97	4.95	4.31
Driller	4.06	4.12	4.25	3.72
Duck	3.48	2.86	2.94	2.16
Eggbox	3.87	4.87	4.91	4.26
Glue	3.06	3.34	3.35	2.48
Holepuncher	4.94	4.75	4.95	4.28
Iron	3.18	3.05	3.44	2.89
Lamp	3.89	3.46	3.68	3.04
Phone	5.07	4.56	4.85	4.15
Average	4.26	4.09	4.33	3.66

Since M_i and M_j' are under the same coordinate system, the initial registration and the global point cloud fusion can be achieved by putting them together, i.e. $M = \{M_i, M_j'\}$. Typically, to align the multi-view point clouds, M_0 is considered as the initial point cloud, the global point cloud can be obtained through $M = \{M_0, M_1', \ldots, M_j', \ldots\}$, where M_j' is the transformed point cloud of M_j at j-th measurement perspective, calculated with Eq.(8). Then the ICP algorithm is employed to refine the global registration result.

III. EXPERIMENT AND DISCUSSION

A. PERFORMANCE OF POSE ESTIMATION NETWORK

The public data set LineMod [30] is utilized to test the overall performance of the network. LineMod is a *de facto* standard benchmark for 6D pose estimation of textureless objects. 13 objects with more than 15000 images are assigned with ground-truth rotation and translation parameters. In addition, the 3D mesh files are also available in LineMod. To verify the performance of the YOLO-based pose estimation network, several existing pose estimation networks [错误!未找到引用源。-错误!未找到引用源。] are compared. The 2D re-projection error and the visualization of the pose estimation are shown in Table 1 and Figure.3.

The re-projection error is calculated with the ground-truth corners and the re-projected corners. According to the imaging process (Eq.3), the reprojected image points are based on the camera parameters and the pose parameters. Therefore, the re-projection error can be a comprehensive evaluation indicator. As shown in Table 1, the average re-projection error of different objects has different values. The YOLO-based method has the smallest reprojection error 3.66 pixels, indicating the good performance of pose estimation.

The visualization results are shown in Figure.3. It can be seen from Figure.3 that the projected bounding box covers the objects even in complex environments.

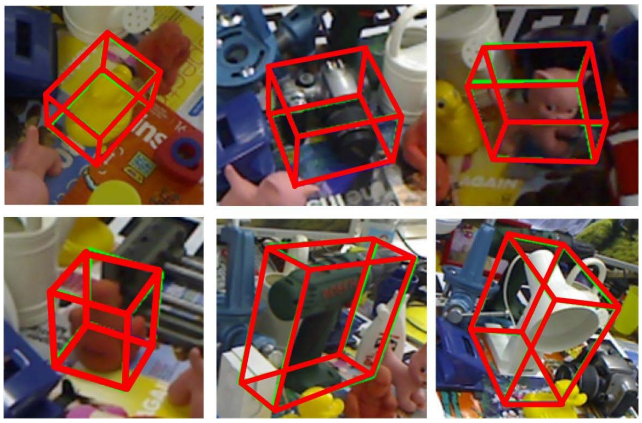
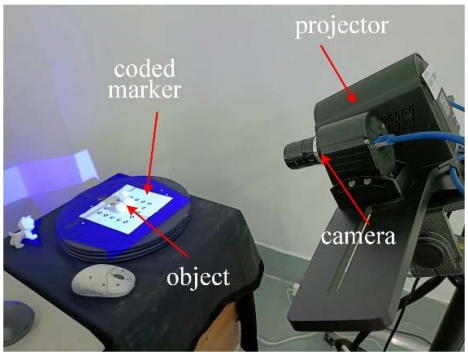
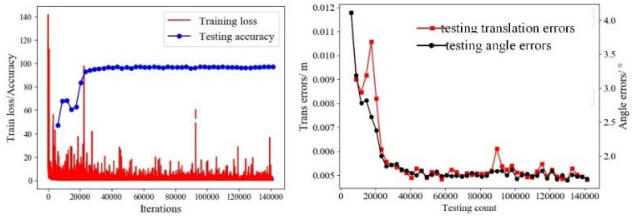


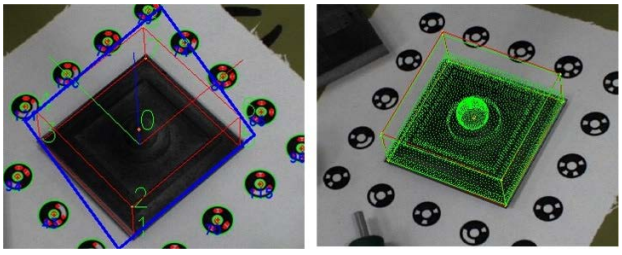
FIGURE 3. Example results on the LineMod dataset.



(a) setup.



(b) training loss and testing accuracy. (c) translation and rotation error.



(d) true pose determination. (e) pose estimation visualization.

FIGURE 4. Experimental results.

The computational efficiency is also compared. The results are shown in Table 2.

As indicated in Table 2, the YOLO-based network runs more than 5 times faster than the existing approaches. The refinement process is never needed. Moreover, the YOLO-based method can research real-time performance, which contributes to the high efficiency of the MLSM system.

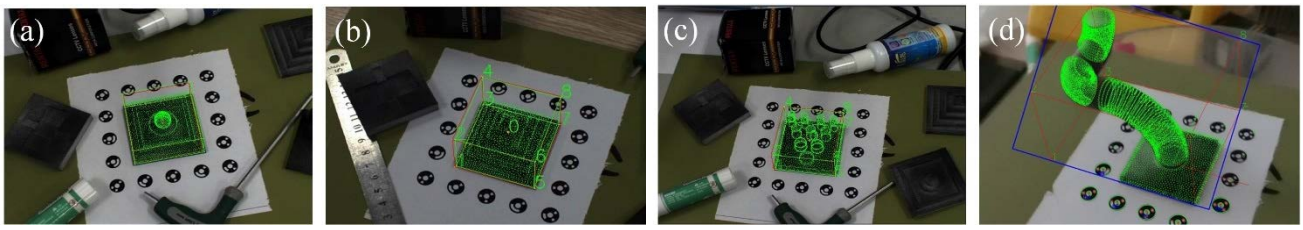


FIGURE 5. More case presentations of the object pose estimation. (a) sphere. (b) pyramid. (c) pillars. (d) elbow. Noting that the red lines are drawn according to the ground-truth 3-D bounding box, the green lines and points are drawn based on the estimated pose. The artifacts are from MMT, University of Nottingham [34].

TABLE 2. The comparison of runtime.

Method	Overall speed (fps)	Refinement time (ms/object)
Brachmann[31]	3	120
SSD[32]	4	23
BB8[33]	9	25
YOLO-Based	48	-

TABLE 3. Error computation of pose estimation.

Object	Sphere	Pyramid	Pillars	Elbow
Mean re-projecting error/pixel	2.526	0.846	1.753	3.216
Translation error/mm	3.541	1.897	3.019	5.187
Angle error/(°)	0.85	0.42	0.55	1.05

B. DATA ACQUISITION AND POSE ESTIMATION ANALYSIS OF MSLM SYSTEM

An experimental platform is built to verify the effectiveness of multi-view structured light measurement based on pose estimation. As shown in Figure.4 (a), the hardware includes a commercial projector (NEC np43 +) and an industrial camera (DMK 31BU03). The resolution of the camera is 1024×768 pixels. Figure.4 (b)-(e) shows the network training results of the prediction results.

As indicated in Figure.4 (a), the structured light measurement system consists of a single camera and an off-the-shelf optical projector. A 3-D printing part is considered the target measurement object. Figure.4 (b) shows the training loss and the testing accuracy, where the training loss is defined as the weighted distance offset between the true and estimated image points, and the testing accuracy is considered as the percentage of the number of points whose offset is less than 5 pixels to the total number of estimated points. The training loss drops rapidly and the testing accuracy keeps 100% with the iterations increasing. It can be confirmed from Figure.4 (b) that the network reaches a high precision for the object pose estimation. As can be seen from Figure.4 (c) that the total translation error of the three directions is less than 5 mm and the total rotation error is less than 1 degree. This satisfies the accuracy requirement of the initial data registration. Figure.4 (d) presents the marker detection and the 3-D and 2-D bounding box determination for true pose calculation from one perspective. Figure.4 (e) visualizes the projection model, the estimated 3-D bounding box, and the centroid of the measured object under the view of Figure.4 (e). It can be seen that the projection fits the object well considering the offset of the corners and centroid is less than 5 pixels. More pose estimation results for more shapes are shown in Figure.5.

Figure.5 indicates that, from different perspectives, the estimated pose keeps consistent with the true one. The artifacts, including the sphere, pyramid, pillar, and elbow, show the robustness and effectiveness of the network in multi-shape pose estimation. In Figure.5, each object needs network training for pose estimation, so the pose estimation method takes a lot of time in the early training stage compared with the marker-based registration. However, because it does not need to paste marker points, and there are no requirements for the placement pose, the pose estimation method is more flexible and convenient. The corresponding pose estimation error is shown in Table 3.

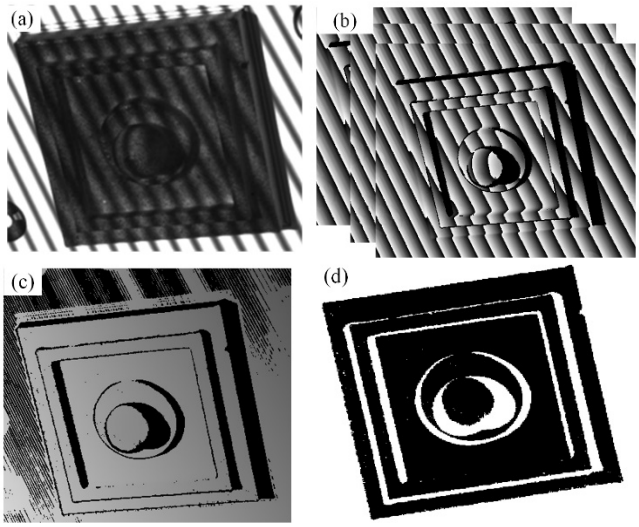


FIGURE 6. Structured light reconstruction based on the proposed system. (a) Projection image; (b) wrapping phase; (c) absolute phase; (d) 3D point cloud.

As can be seen from Figure.5 and Table 3 that the pixel errors, which are determined by the gap between the

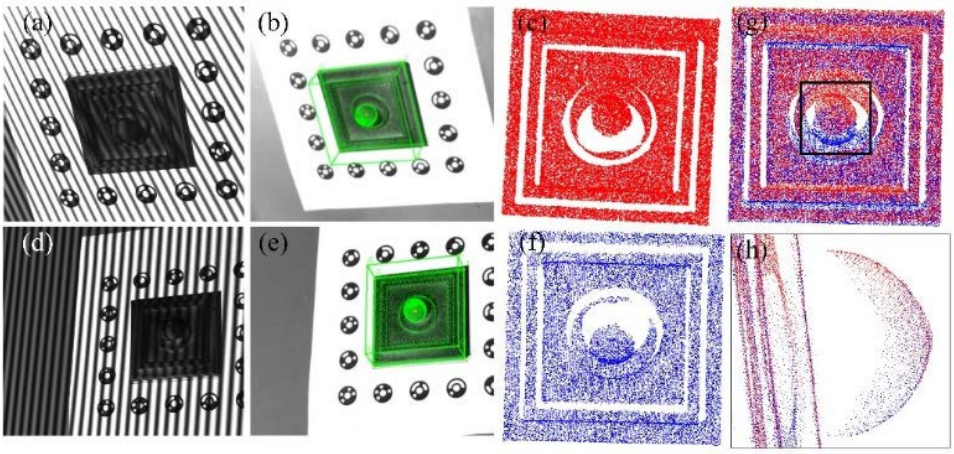


FIGURE 7. Point cloud alignment using estimated pose. (a) (d) Projection images in two views; (b) (e) pose estimation results; (c) (f) point clouds in two views; (g) data splicing result; (h) zoom-in view of box in (g).

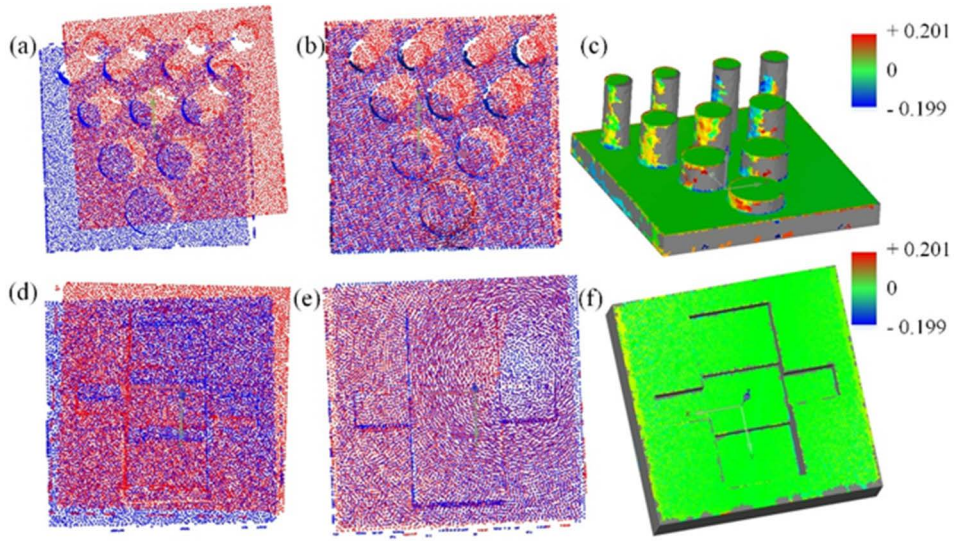


FIGURE 8. Data registration with the estimated pose. (a) (d)Two-view registration of pillars and recess, with deep learning-based pose estimation; (b) (e)global refinement using ICP algorithm based on rough rigid transformation; (c) (f) error distributions of the final registration of fused data in (b), (e) and CAD model, where the error is determined by the point-to-model distance.

estimated 8 corners and the true corners, vary from 0.846 pixels to 3.216 pixels, proving the high-precision of the network in 3D target detection. Furthermore, the errors of the estimated pose (translation and rotation) are determined through the Euclidean distance of the estimation and the ground truth. It can be seen that the translation errors range from 1.8 mm to 5.2 mm and the angle errors distribute around 0.5-1.0 degrees. Such results are acceptable in single-shot-pose estimation. The precise estimation contributes to the following data registration.

C. POINT CLOUD REGISTRATION IN MSLM

Figure. 6 shows the process of single-view reconstruction using the built structured light measurement system.

It can be seen in Figure. 6(c) that a part of the three-dimensional data is obtained from the current measurement view. Namely, the complete surface morphology of the sphere cannot be obtained from a single perspective, Further, three-dimensional measurement is carried out from another perspective. Meanwhile, the target poses in two perspectives are estimated from the collected images and are then utilized for data coarse alignment. The results are shown in Figure. 7.

Figure. 7 (a) and (d) show the structure light measurement in two views, respectively. Figure. 7 (c) and (f) are the corresponding 3D point clouds, respectively. Figure. 7 (b) and (e) show the pose estimation results of the two views. Figure. 7 (g)(h) show the results of point cloud alignment using the estimated pose, indicating the acceptable results and the effectiveness of the proposed strategy.

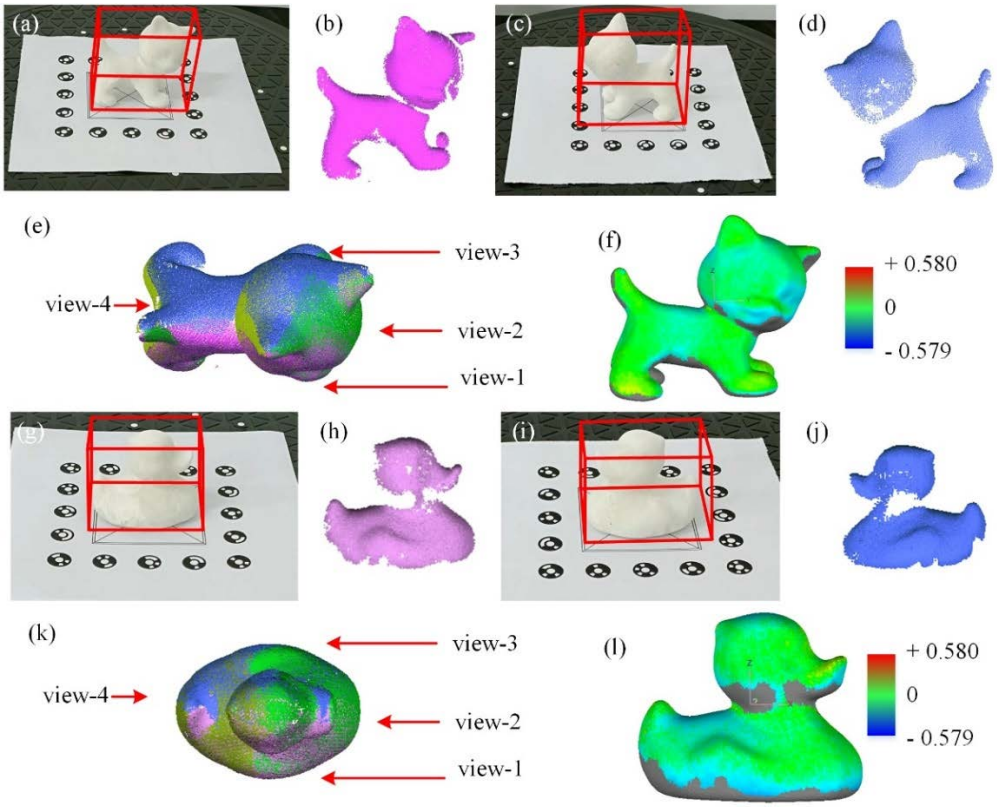


FIGURE 9. Experiments of multi-view measurement with the proposed strategy. (a)(g)are the 1st view pose estimation results, respectively. (b)(h)are the measured point clouds under the 1st view, respectively. (c) and (i) are the 3rd view pose estimation results, respectively. (d)(j)are the measured point clouds 3rd view, respectively. (e) (k) are the multi-view data registration with the estimated pose, respectively. (f)(l) are the error distributions of the final registration with ICP.

D. EVALUATION OF REGISTRATION ACCURACY

More objects are tested to validate the effectiveness and accuracy of the point cloud registration results. The refined point cloud is compared with the theoretical digital model to compute the overall measurement error. The results are shown in Figure. 8.

Figure. 8 shows the data registration with the estimated pose measurement and error evaluation results. Figure.8 (a) is a rough alignment result with the estimated pose. Figure.8 (b) shows the global refinement using the ICP algorithm. The point cloud in Figure.8 (b) is registered with the theoretical CAD model to evaluate the registration accuracy. Figure.8 (c) indicates the error distribution of the two view alignments. As shown in Figure.8 (c), using the estimated pose, the maximum and minimum registration errors are −0.184 mm and 0.203 mm, respectively. The average error is 0.063 mm and the standard deviation is 0.07 mm. Figure.8(d) is a rough alignment of the recess using the estimated pose. The fine registration and error distribution are shown in Figure.8 (e) and (f), respectively, indicating the effectiveness of multi-view data alignment using deep learning. Figure.8 (a) and Figure.8 (d) are the results of direct alignment using the estimated pose, which does not depend on the real value of feature points. Due to the deviation of the estimated pose under multiple views, the data deviation of the final point clouds is the comprehensive deviation of the pose estimation and the measurement. The control mark points are used to fuse the point clouds of the two views, and the average error results are shown in Table 4.

Table 4 compares the errors of data registration results using the proposed method and the marker-based method. It can be seen that the deviation of the average error is about 0.02 mm, and the error number obtained by the estimation method is also close to the marker-based stitching results, which verifies the comparability between the MSLM based on deep learning and the existing feature-based measurement method. Under our existing configuration, the average pose estimation time of a single image is 0.05 s, and the detection and matching optimization time of two view marker points is 0.08 s. Further, It takes 1.6 s for15 views of the data rough alignment. Comparatively, the average reduction time of a single image is 0.056 s when using the pose estimation method. Therefore, the advantages of the direct pose estimation method will be more obvious in more view measurements.

Figure.9 shows more experimental results with the proposed measurement strategy. The model are both from the public dataset LineMod [30]. Figure.9(a)(c)(g)(i) are the

TABLE 4. Error comparison of data fusion using markers and deep learning.

Object	Sphere	Pyramid	Pillars
Pose estimation/mm	0.065	0.058	0.067
Coded markers/mm	0.044	0.038	0.046
Error/mm	0.021	0.02	0.023

TABLE 5. Accuracy evaluation and comparison.

Method \ Object	Cat (mm)	Duck (mm)
Brachmann [31]	0.13	0.12
SSD [32]	0.15	0.14
BB8 [33]	0.1	0.13
YOLO-Based	0.07	0.08

pose estimation results, which can be compared with the marker-based pose estimation results to evaluate the accuracy. Figure.9(b)(d)(h)(j) are the single-view point cloud. Figure.9 (e)(k) are the data registration results with the estimated poses. It can be seen that the alignment results are fine even though the two views do not have a common area. Concretely, Figure.9(b) and (d) are the front and black view of a cat model. The point clouds have few common areas. Therefore, it is hard to register the point clouds with the common feature. However, as shown in Figure.9(e), the two-point cloud can be aligned in good condition, indicating the effectiveness of the proposed strategy. To verify the final measurement accuracy, we further compare the measured point cloud with the CAD model and draw the error distribution. As indicated in Figure.9(d)(l), the global error is low and the average error is 0.068 mm and 0.076 mm. The results show that the accuracy of the multi-view point clouds registration with the estimated pose is relatively high and can meet the measurement acquirement of the structured light measurement system.

We further carried out the accuracy evaluation with the existing pose estimation network and the public dataset LineMod [30]. The main idea of our single-shot pose estimation network is the combination of the key point extraction and the PnP algorithm. Brachmann, SSD, and BB8, which all pose estimation networks with a single image, are selected for accuracy comparison. The accuracy is also described with the standard deviation of the final registration error. The results are shown in Table 5.

It can be seen from Table 5, the standard deviation of the proposed strategy with the YOLO network is relatively lower than others. Combining Tables 1 and 2, the pose estimation accuracy and the processing time of the proposed strategy are also better than the others. The main advantage is we do not need marker points to create a common feature but also achieve complete data registration with a relatively high speed for the multi-view structured light measurement system.

IV. CONCLUSION

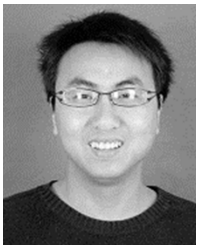
A multi-view structured light measurement method integrating deep learning pose estimation is proposed. The 6D pose of the measured object under the current measurement perspective is directly estimated by the pose estimation network. Compared with the traditional methods based on an auxiliary turntable, marker- and tracking-based measurement, the proposed method has good convenience and flexibility. Compared with the alignment method based on feature recognition, the proposed method has wider applicability. The average error of the proposed strategy is 0.02 mm, which is close to marker-based multi-view alignment. More importantly, this method can be applied to some special application scenarios, such as in-situ measurement of parts that are not allowed to paste superfluous objects. However, compared with the marker point stitching method, the proposed method is dependent on early training information. In addition, the proposed method needs to upgrade the pose estimation network for the target reconstruction that cannot contain all objects in a single view. Further study will focus on the multi-view measurement of large-scale objects using deep learning-based pose estimation.

REFERENCES

[1] A. G. Marrugo, F. Gao, and S. Zhang, "State-of-the-art active optical techniques for three-dimensional surface metrology: A review [invited]," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 37, no. 9, pp. B60–B77, Sep. 2020.
[2] S. Zhang, "High-speed 3D shape measurement with structured light methods: A review," *Opt. Lasers Eng.*, vol. 106, pp. 119–131, Jul. 2018.
[3] H. Cui, T. Jiang, X. Cheng, W. Tian, and W. Liao, "A general gamma nonlinearity compensation method for structured light measurement with off-the-shelf projector based on unique multi-step phase-shift technology," *J. Mod. Opt.*, vol. 66, no. 15, pp. 1579–1589, Sep. 2019.
[4] S. Barone, A. Paoli, and A. V. Razionale, "Shape measurement by a multi-view methodology based on the remote tracking of a 3D optical scanner," *Opt. Lasers Eng.*, vol. 50, no. 3, pp. 380–390, Mar. 2012.
[5] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Simultaneous reconstruction and calibration for multi-view structured light scanning," *J. Vis. Commun. Image Represent.*, vol. 39, pp. 120–131, Aug. 2016.
[6] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.
[7] J. Shi and Z. Sun, "Large-scale three-dimensional measurement based on LED marker tracking," *Vis. Comput.*, vol. 32, no. 2, pp. 179–190, Feb. 2016.
[8] J. Shi, Z. Sun, and S. Bai, "3D reconstruction framework via combining one 3D scanner and multiple stereo trackers," *Vis. Comput.*, vol. 34, no. 3, pp. 377–389, Mar. 2018.
[9] T. Jiang, X. Cheng, H. Cui, and X. Li, "Combined shape measurement based on locating and tracking of an optical scanner," *J. Instrum.*, vol. 14, no. 1, Jan. 2019, Art. no. P01006.
[10] T. Jiang, H. Cui, and X. Cheng, "Accurate calibration for large-scale tracking-based visual measurement system," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.
[11] J. Wang, B. Tao, Z. Gong, S. Yu, and Z. Yin, "A mobile robotic measurement system for large-scale complex components based on optical scanning and visual tracking," *Robot. Comput.-Integr. Manuf.*, vol. 67, Feb. 2021, Art. no. 102010.
[12] Z. Zhou, W. Liu, Q. Wu, Y. Wang, B. Yu, Y. Yue, and J. Zhang, "A combined measurement method for large-size aerospace components," *Sensors*, vol. 20, no. 17, p. 4843, Aug. 2020.
[13] J. Wang, B. Tao, Z. Gong, W. Yu, and Z. Yin, "A mobile robotic 3-D measurement method based on point clouds alignment for large-scale complex surfaces," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[14] S. Barone, A. Paoli, and A. V. Razionale, "Multiple alignments of range maps by active stereo imaging and global marker framing," *Opt. Lasers Eng.*, vol. 51, no. 2, pp. 116–127, Feb. 2013.
[15] W.-C. Chang and C.-H. Wu, "Candidate-based matching of 3-D point clouds with axially switching pose estimation," *Vis. Comput.*, vol. 36, no. 3, pp. 593–607, Mar. 2020.
[16] Z. Yao, Q. Zhao, X. Li, and Q. Bi, "Point cloud registration algorithm based on curvature feature similarity," *Measurement*, vol. 177, Jun. 2021, Art. no. 109274.
[17] W. Liu, Z. Lan, Y. Zhang, Z. Zhang, H. Zhao, F. Ye, and X. Li, "Global data registration technology based on dynamic coded points," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 2, pp. 394–405, Feb. 2018.
[18] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognit.*, vol. 47, no. 6, pp. 2280–2292, 2014.
[19] P. Psota, H. Tang, K. Pooladvand, C. Furlong, J. J. Rosowski, J. T. Cheng, and V. Lédl, "Multiple angle digital holography for the shape measurement of the unpainted tympanic membrane," *Opt. Exp.*, vol. 28, no. 17, pp. 24614–24628, 2020.
[20] L. Yu and B. Pan, "Single-camera stereo-digital image correlation with a four-mirror adapter: Optimized design and validation," *Opt. Lasers Eng.*, vol. 87, pp. 120–128, Dec. 2016.
[21] J. Xu, P. Wang, Y. Yao, S. Liu, and G. Zhang, "3D multi-directional sensor with pyramid mirror and structured light," *Opt. Lasers Eng.*, vol. 93, pp. 156–163, Jun. 2017.
[22] V. Villena-Martinez, S. Oprea, M. Saval-Calvo, J. Azorin-Lopez, A. Fuster-Guillo, and R. B. Fisher, "When deep learning meets data alignment: A review on deep registration networks (DRNs)," *Appl. Sci.*, vol. 10, no. 21, p. 7524, Oct. 2020.
[23] W.-C. Chang and V.-T. Pham, "3-D point cloud registration using convolutional neural networks," *Appl. Sci.*, vol. 9, no. 16, p. 3273, Aug. 2019.
[24] D. Jack, J. K. Pontes, S. Sridharan, C. Fookes, S. Shirazi, F. Maire, and A. Eriksson, "Learning free-form deformations for 3D object reconstruction," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 317–333.
[25] G. D. Pais, S. Ramalingam, V. M. Govindu, J. C. Nascimento, R. Chellappa, and P. Miraldo, "3DRegNet: A deep neural network for 3D point registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7193–7203.
[26] S. Zakharov, I. Shugurov, and S. Ilic, "DPOD: 6D pose object detector and refiner," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1941–1950.
[27] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6D object pose prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 292–301.
[28] L. Ding and C. Feng, "DeepMapping: Unsupervised map estimation from multiple point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8650–8659.
[29] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua, "GarNet: A two-stream network for fast and accurate 3D cloth draping," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8739–8748.
[30] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 548–562.
[31] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother, "Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3364–3372.
[32] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again," in *Proc. ICCV*, Oct. 2017, pp. 1521–1529.
[33] M. Rad and V. Lepetit, "BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth," in *Proc. ICCV*, Oct. 2017, pp. 3828–3836.
[34] J. Eastwood, D. Sims-Waterhouse, S. Piano, R. Weir, and R. Leach, "Autonomous close-range photogrammetry using machine learning," in *Proc. ISMTII*, 2019, pp. 1–6.

TAO JIANG received the bachelor's degree in mechanical engineering from the Xuzhou University of Technology, Xuzhou, China, in 2016, and the Ph.D. degree in mechanical engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2021. He is currently an Assistant Professor with Suqian University, Suqian, China. His current research interests include vision-based measurement and robotic vision.



KAITUO FANG received the bachelor's and master's degrees in mechanical engineering from Jiangnan University, Wuxi, China, in 2009 and 2012, respectively. He is currently a Lecturer with Suqian University, Suqian, China. His current research interests include electromechanical measurement and control technology.

HAIFANG ZHAO received the master's degree in mechanical engineering from the China University of Mining and Technology, Xuzhou, China, in 2012. Her current research interests include computer 3D modeling and graphic processing.

GUOBIN CHEN (Graduate Student Member, IEEE) was born in Hulun Buir, Inner Mongolia, China, in 1987. He received the B.S., M.S., and Ph.D. degrees from the North University of China, Taiyuan, China, in 2015. In 2015, he joined the Suqian College, Suqian, China, as a Lecturer. Since 2018, he has been working with the Peter Grünberg Research Center, Nanjing University of Posts and Telecommunications, Nanjing, China, as a Postdoctoral Researcher. His current research interests include magnetic field vectorial sensing and imaging from dc to high frequency.

YANFENG WANG was born in Liaocheng, China, in 1980. He received the M.S. degree in operations research and cybernetics and the Ph.D. degree in control theory and control engineering from Northeastern University, in 2007 and 2013, respectively. From 2013 to 2022, he was an Associate Professor at the School of Engineering, Huzhou University, Huzhou, Zhejiang, China. He is currently a Full Professor with the School of Mechanical and Electrical Engineering, Suqian University, Suqian, China. He was supported by the National Natural Science Funds of China and the Natural Science Funds of Zhejiang. He has authored three books and more than 30 articles. His main research interests include networked control systems, fault detection, and fault-tolerant control.